# STUDIES ON CLASSIFICATION PROBLEMS AND APPLICATION IN CONSUMER LENDING OPTIMIZATION SYSTEM

by

Mingxing Gong

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Financial Services Analytics

Spring 2022

**STUDIES ON CLASSIFICATION PROBLEMS AND APPLICATION IN CONSUMER LENDING OPTIMIZATION SYSTEM**

by

Mingxing Gong

Approved: _____
Bintong Chen, Ph.D.
Chair of Financial Services Analytics

Approved: _____
Bruce Weber, Ph.D.
Dean of Alfred Lerner College of Business & Economics

Approved: _____
Louis F. Rossi, Ph.D.
Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Dr. Bintong Chen, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Dr. Gonzalo Arce, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Dr. Xiao Fang, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Dr. Fengqin Zhao, Ph.D.
Member of dissertation committee

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## ABSTRACT

Classification is a fundamental problem in machine learning and has wide range of applications. In this dissertation, we introduce and study the likelihood and Bayesian probability transformations on feature variables of the binary classification problem. We systematically investigate their theoretical properties and associated benefits when applied in many machine learning topics, algorithms, and applications.

In the first part of the dissertation, we define the likelihood and Bayesian probability transformations and study their theoretical properties. The transformations are then used to improve confusion-matrix based classification performance measure characterization, ROC curve, and cost-sensitive analysis. In particular, we propose a unified framework for all existing classification performance measures and show that the transformations lead to dominant performance measurements and efficient threshold calculations, guaranteed concavity of ROC curve, and milder assumptions for cost-sensitive analysis.

The second part of the dissertation focuses on two major extensions of the binary classification problem. Extension 1 introduces an ambiguous region in the binary classification problem, which usually requires collecting additional information for a classification decision. We obtain optimal ambiguity thresholds under various performance measures and cost-sensitive environments. Extension 2 proposes a dynamic ensemble of two binary classifiers. We show that the ensemble dominates both component classifiers in terms of the ROC curves and also verify the dominance numerically and empirically by using Lending Club data.

The last part of the dissertation applies some of our proposed classification techniques in consumer lending. We develop a consumer lending optimization system by

taking risk and pricing sensitivity in a holistic view to determine customers' risk category and pricing. We numerically demonstrate the enhancement of the optimization system over the popular risk-based pricing currently used in lending industry.

# Chapter 1

# INTRODUCTION

## 1.1 Background

Machine learning refers to the capability of a model to improve its own performance automatically from experience and by the use of data without being explicitly programmed (Mitchell, 1997). It does so by utilizing a statistical model to make decisions and integrating the result of each new iteration into that model. The model, in essence, is programmed to improve through trial-and-error iterations. Machine learning models traditionally fall into three broad categories: supervised, unsupervised, and reinforcement learning. Supervised learning is characterized by using a labeled dataset to train algorithms to make predictions in contrast to unsupervised learning, where instances in the training dataset are unlabeled (Kotsiantis et al., 2007). Reinforcement learning is a subarea of machine learning to study how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward (Das et al., 2015). Majority of machine learning applications are supervised (Bhavsar and Ganatra, 2012; Singh et al., 2016). Machine learning models can be used to perform classification and regression based on the targeted outputs, where classification categorizes given instances into discrete classes as the output and regression predicts continuous values as the output (see Figure 1.1). In our dissertation, we mainly focus on supervised machine learning classification problems.

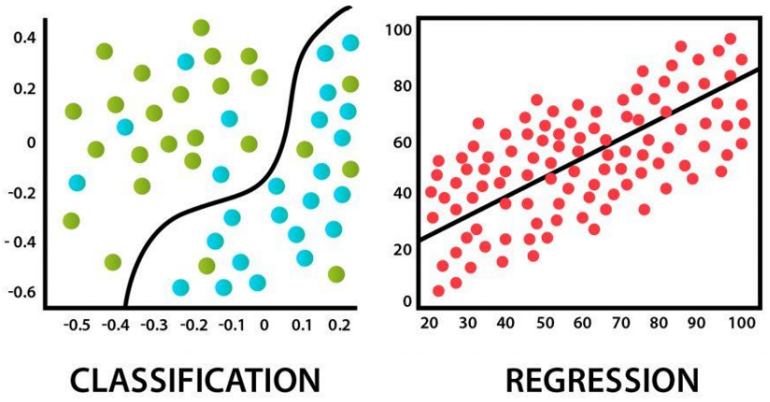


Figure 1.1: Classification vs. Regression [1]

Machine learning classification is a task to automatically categorize the data into a set of given classes on the basis of a training set of data, in which the outcome of the observations are known. Examples of machine learning classification problems include text categorization (e.g., sentiment analysis, email spam filter), credit risk forecasting (e.g., predict whether customers will default on a loan), fraud detection, marketing (e.g., predict if customers will response to a promotion offer), medicine and biology (e.g., predict whether a disease is Covid or not), natural language processing (e.g., intent and entity classification for natural language understanding), optical character recognition, etc. Domingos (2012) defines machine learning classifier as a system that inputs (typically) a vector of discrete and/or continuous *feature values* and outputs a single discrete value, the *class*. The machine learning classification model components can be illustrated by Figure 1.2, where the core component consists of a learning algorithm and a decision making model. The learning algorithm is responsible for extracting patterns from the input data and generating a continuous output, while the decision making model is used to choose proper threshold to transform the continuous output into discrete outcomes.

Recently, there are some advocates of the practice by incorporating the decision model into the learning algorithm to automate the decision process. A typical example

[1] Source: https://www.analytixlabs.co.in/blog/classification-in-machine-learning/

is to create cost-sensitive machine learning classifiers (Zhou and Liu, 2005; Masnadi-Shirazi et al., 2012; Xia et al., 2017; Nami and Shajari, 2018; Yu et al., 2019; Olowookere and Adewale, 2020). We believe that the black-box machine learning classifier with incorporated decision making model is subject to various drawbacks (Figure 1.2): 1. It adds more complexity and opacity to the machine learning model, which is well known for lack of transparency and interpretability due to the black-box nature of the decision process (De Laat, 2018; McGovern et al., 2019; Carvalho et al., 2019); 2. Decision criteria (e.g., cost matrix) needs to be pre-defined. When situation changes, the model has to be retrained, resulting in inefficiency and high computation cost; 3. It also leads to bias/unfairness issues when the decision process is embedded into the algorithm (Corbett-Davies and Goel, 2018; Mehrabi et al., 2021). Therefore, Zarsky (2013) underscores the importance of separation of the decision-making model from algorithm learning. With the model transparency and interpretability in mind, we embrace the separation of decision making model from the learning algorithm and conduct extensive research on each item. Part I and Part II of this dissertation cover our research on the improvement of learning algorithm, including a unified framework for all existing classification measures, ROC curve characterization and enhancement, extension of classification problem, ensemble learning, etc., while Part III focuses on the decision making model with application in consumer loan optimization system.

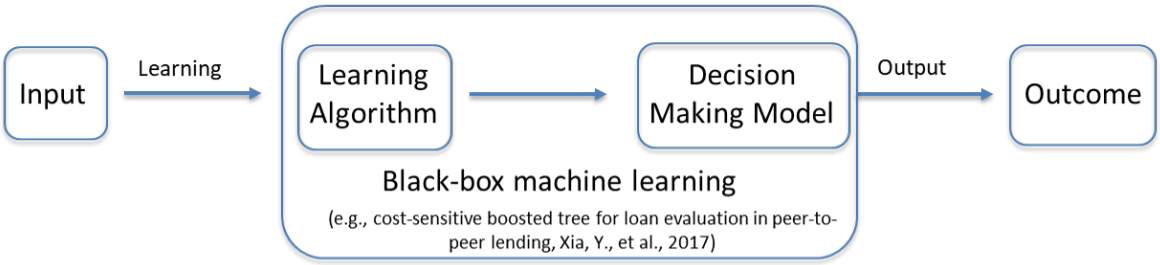


Figure 1.2: Machine Learning Classification Model Components

There are numerous studies on machine learning classifiers from theoretical perspectives. Majority of these studies focus on improving specific machine learning algorithms, such as logistic regression (Wright, 1995; Ranganathan et al., 2017; Sur and

Candès, 2019), decision tree (Safavian and Landgrebe, 1991; Myles et al., 2004; Song and Ying, 2015), support vector machine (Huang et al., 2018; Iranmehr et al., 2019; Cervantes et al., 2020), neural network classification (Zhou et al., 2016; Li et al., 2018; Yao et al., 2019), extreme gradient boosting tree (Chen et al., 2015; Chen and Guestrin, 2017; Duan et al., 2021), random forest (Breiman, 2001; Paul et al., 2018; Chai and Zhao, 2019), Naive Bayes classifier (Rish et al., 2001; Xu et al., 2017), k-Nearest Neighbors (KNN) (Guo et al., 2003; Zhang et al., 2017), etc. The subareas in improving machine learning algorithms are characterized by objective function (Mukhoti et al., 2020; Qiu et al., 2020), optimization methodology (Le et al., 2011; Bottou et al., 2018; Sun et al., 2019), the choice of hyperparameters (Bardenet et al., 2013; Feurer and Hutter, 2019; Probst et al., 2019), overfitting prevention (Hawkins, 2004; Srivastava et al., 2014; Ying, 2019), or initialization process (Feurer et al., 2015; Humbird et al., 2018), etc.

These studies are model specific, requiring a deep understanding of specific algorithms and their functioning. Another category of studies on machine learning classification problems with arising interest is model agnostic, which can be applied to any classification algorithms. Those model-agnostic studies mainly concentrate on the output of the machine learning model by treating the underlying classification algorithms as a black box, which requires no knowledge of the algorithms or their functioning. Examples of these research areas cover how to choose a proper cut-off value to transform a continuous output into discrete decisions (e.g., cost-sensitive learning (Domingos, 1999; Elkan, 2001; Wang and Chen, 2017)), ensemble learning by a combination of outputs derived from multiple machine learning classifiers (Dietterich et al., 2002; Polikar, 2012; Sagi and Rokach, 2018; Zhou, 2021), model evaluation (Kotsiantis et al., 2007; Sokolova and Lapalme, 2009; Chicco and Jurman, 2020), and ROC analysis (Fawcett, 2006; Powers, 2011; Majnik and Bosnić, 2013), etc. A common deficiency of these studies is to treat the machine learning outputs as true probability (Elkan, 2001) or simply assuming a 'well-calibrated' classifier (Chase Lipton et al., 2014) without

pointing out what is the 'well-calibrated' classifier or explicitly discussing the properties of those 'well-calibrated' classifiers. Motivated by the gap in the existing literature, we introduce and study the likelihood and Bayesian probability transformations on feature variables of the binary classification problem. We systematically investigate their theoretical properties and associated benefits when applied in many machine learning topics, algorithms, and applications.

## 1.2 Overview of Part I

From theoretical perspective coupled with empirical evidence, Part I proposes a density-based likelihood transformation of machine learning raw score, uncovers the mathematical properties of the transformation, and characterizes the intrinsic relationship between maximal classification measurements and the associated thresholds. Our contributions in Part I mainly lie in following aspects:

- In Chapter 3, we propose likelihood and Bayesian probability transformations and identify structural properties of the density function associated with the transformation, which plays a critical role in my dissertation.

- In Chapter 4, we present a unified framework for all classification performance measurements and conduct systematical studies of theoretical properties of confusion-matrix based measures and disclose the relationships between the optimal measures and decision-making threshold in the context of likelihood and Bayesian probability transformation. Our research sheds light on choosing appropriate measures in model development and evaluation, improving the efficiency of finding the optimal threshold for corresponding threshold-dependent measures, and pointing out directions about designing new classification performance measures.

- In Chapter 5, we address the following two major limitations of current studies on ROC curve through likelihood ratio transformation (or its monotone variants):

    1. Little attention has been paid to its concavity. In most studies, concavity of ROC was implicitly assumed. Some papers recognized the non-concave ROC

and proposed ad-hoc methods to correct the problem. Some researchers recommend not to use the classifiers if the ROC curve is not concave.

2. The ROC curve is created based on one dimensional score output. There lacks discussion to construct ROC curve for multidimensional scores.

We rigorously prove that likelihood and Bayesian transformation can always result in a concave ROC curve regardless of original ROC curve is concave or not. In addition, we reveal that these transformations lead to the most discriminate ROC curve, which is in line with Neyman-Pearson lemma. Furthermore, we extend ROC curve to multi-dimensions and systematically discuss the characterization of ROC curves.

- In Chapter 6, we enhanced cost-sensitive learning through threshold adjustment based on the likelihood transformation. Our approach does not rely on any shape assumption of Type I and Type II error functions.

### 1.3 Overview of Part II

Based on the extensive studies in Part I, we extend our theoretical study in binary classification with binary outcomes to binary classification with an ambiguous region and propose a dynamic ensemble learning method in Part II.

In Chapter 7, we identify that most binary classification models are based on binary decisions, where the outcomes are either positive or negative. In many real-world applications, binary classification with ambiguous region (BCAR) is more appropriate for decision making. We expanded the model performance measures in general binary classification to a binary classification with an ambiguous region. In addition, we explore the relationships between the optimal measures and thresholds & base rate in BCAR. Finally, we expand the threshold-based cost sensitive learning to BCAR.

In Chapter 8, we recognize that density-based ensemble methods have been under-researched due to the need of empirical estimation of density functions, which can be impractical in case of high dimension. We propose a likelihood-ratio-based

dynamic ensemble of two classifiers in binary classification to address the computation concerns. We prove that our proposed ensemble method is theoretically and empirically better than both single classifiers in the ensemble in terms of ROC/AUC.

### 1.4 Overview of Part III

A pivotal element in consumer lending optimization system is pricing. The importance of having the right pricing strategies in place has been highlighted by many practitioners in the financial industry. Traditional risk-based pricing is the dominate approach in industry, which are subject to multiple limitations. In Chapter 9, we propose a consumer lending optimization system by jointly taking into consideration of risk and pricing sensitivity to determine customers' risk category and pricing.

### 1.5 Summary of Dissertation Structure

The rest of my dissertation is structured as following: Chapter 2 introduces classification problem and performance measures as background and overview; Chapter 3 discusses likelihood and Bayesian transformation and identifies nice properties of density functions associated with the transformed scores. This chapter serves as the basis of my dissertation; Chapter 4-6 focuses on the improving of the learning algorithms, which include a unified framework for all classification (Chapter 4), ROC curve characterization and enhancement (Chapter 5), cost-sensitive learning improvement (Chapter 6); Chapter 7 expands traditional binary classification to binary classification with an ambiguous region motivated by many real-world problems. In this chapter, we extend our performance measures optimization and cost sensitive learning framework to the extended problem; Chapter 8 provides a density-based dynamic ensemble of two classifiers, which are theoretically and empirically proven to be superior over any single classifier in the ensemble pool; Chapter 9 proposes a consumer lending optimization system, which aims to identify the optimal decision boundaries to assign each applicant into corresponding category as well as the optimal pricing for each category.

# Part I

# CLASSIFICATION PERFORMANCE MEASURE CHARACTERIZATION AND ENHANCEMENT

# Chapter 2

# CLASSIFICATION PROBLEM AND PERFORMANCE MEASURES: BACKGROUND AND OVERVIEW

Consider a dataset consisting of $m$ i.i.d instances $\{(x_i, y_i)\}_{i=1}^m$ of a random variable pair $(X, Y)$, where $X \in \Omega \subseteq R^n$ represents the features associated with an instance, or a predictive score generated by certain machine learning model for the instance, and $Y \in \{0, 1\}$ represents the actual class category of the instance. The objective of binary classification is to find a proper rule, based on the $m$ given instances, to "accurately" predict the class variable $Y$ using the feature (or score) variable $X$. Let $\hat{Y} \in \{0, 1\}$ denote the predicted class category. In this dissertation, classes 'positive/negative' are interchangeably used with classes '1/0'.

## 2.1 Classification Performance Measures

To evaluate the performance of a classification algorithm, we compare its predicted classes v.s. the true classes. More specifically, we apply a classification algorithm to a data set with $P$ positives and $N$ negatives. Suppose the algorithm predicts $\hat{P}$ positives and $\hat{N}$ negatives, among which $TP$ and $FN$ are number of positive instances correctly and mistakenly predicted (classified), respectively, and $TN$ and $FP$ are number of negative instances correctly and mistakenly classified, respectively. $TP$, $FN$, $TN$, and $FP$ correspond to the four elements in a two-by-two confusion matrix (as illustrated in Table 9), where

$$P = TP + FN, \ N = FP + TN,$$

and

$$\hat{P} = TP + FP, \ \hat{N} = FN + TN.$$

| | | **Actual Class** | | |
|---|---|---|---|---|
| | | $Y = 1$ | $Y = 0$ | **total** |
| **Predicted Class** | $\hat{Y} = 1$ | True Positive ($TP$) | False Positive ($FP$) | $\hat{P}$ |
| | $\hat{Y} = 0$ | False Negative ($FN$) | True Negative ($TN$) | $\hat{N}$ |
| | **total** | P | N | |

Table 2.1: A Confusion Matrix for Binary Classification

Confusion matrix is the most commonly accepted measure in evaluating classification tasks (Susmaga, 2004; Fatourechi et al., 2008; Ferri et al., 2009; Sokolova and Lapalme, 2009; Powers, 2011; Hossin and Sulaiman, 2015; Gong, 2021). Many measures have been proposed based on confusion-matrix, since each measure only focuses one particular angle and it is very difficult to represent all the information of a confusion matrix based on one single measure (Cao et al., 2020; Gong, 2021). Out of those measures, accuracy and error rate are the fundamental single confusion-matrix measures.

### Accuracy and Error Rate Based on Actual Class

True Positive Rate (TPR) refers to the proportion of those who are classified as positive out of those who are actual positive. It is also known as sensitivity or recall. TPR is considered as prediction accuracy on the actual positive instances and False Negative Rate (FNR) measures the error rate on the predictions of those instances. On the other hand, True Negative Rate (TNR, or known as specificity) is considered as accuracy on the prediction of actual negative instances and False Positive Rate (FPR) measures the error rate on the prediction of actual negative instances.

$$TPR = \frac{TP}{P} \tag{2.1}$$

$$FNR = \frac{FN}{P} \tag{2.2}$$

$$TNR = \frac{TN}{N} \tag{2.3}$$

$$FPR = \frac{FP}{N} \tag{2.4}$$

Clearly, $TPR = 1 - FNR$ and $TNR = 1 - FPR$. In statistical hypothesis, FPR and FNR are called Type I error and Type II error respectively.

**Accuracy and Error Rate Based on Predicted Class**

Similarly, there also exists metrics on performance on the predicted positive instances, such as Positive Predictive Value (PPV, or known as precision) and False Discovery Rate (FDR). They measure the reliability of the positive predictions. Negative Predictive Value (NPV) and False Omission Rate (FOR) are performance metrics on the predicted negative instances. Similar to precision and FDR, they measure the reliability of the negative predictions.

$$PPV = \frac{TP}{\hat{P}} \tag{2.5}$$

$$FDR = \frac{FP}{\hat{P}} \tag{2.6}$$

$$NPV = \frac{TN}{\hat{N}} \tag{2.7}$$

$$FOR = \frac{FN}{\hat{N}} \tag{2.8}$$

**Accuracy and Error Rate on Overall Instances**

The overall ACCuracy (ACC) is the most commonly used performance metric (Dinga et al., 2019; Gong, 2021). The accuracy measures the proportion of instances that are correctly predicted and the ERRor rate (ERR) measures the proportion of instances

that are mis-classified.

$$ACC = \frac{TP + TN}{P + N} \tag{2.9}$$

$$ERR = \frac{FP + FN}{P + N} \tag{2.10}$$

Note that none of above-mentioned single measure alone is sufficient to represent the model performance (Gong, 2021). Although accuracy and error rate alike metrics are simple and intuitive, it has long been criticized (Japkowicz et al., 2000; Van Hulse et al., 2007; He and Garcia, 2009; Longadge and Dongre, 2013; Krawczyk, 2016), especially when dealing with imbalanced data where one class is very rare, but important. Take fraud detection as an example, where only 1% transactions are fraudulent. A model that simply claims all transactions are legitimate can lead to 99% accuracy, which is hard to surpass. The model, however, is useless, because no fraudulent transactions are captured by the model. Similarly, if TPR is used as the performance metric, a model that claims all transactions are fraudulent can result in 100% fraud capture rate. To address the issue, various balanced metrics are formed through a combination of those basic single metrics such as F-measure, Youden's index, balanced accuracy, etc.

**Receiver Operating Characteristic Curve**

Notice that all performance measures described so far depend on a fixed confusion-matrix. Since the predictions are usually made via a threshold based decision in most machine learning algorithms, these performance measures vary as the threshold changes. In addition, metrics such as NPV and precision, are dependent on the prior class distribution (i.e., ratio of number of positive and negative instances in the data). In contrast, TPR and FPR are independent of the prior class distribution, since they are calculated based on the actual positives and negatives, respectively. Therefore, in order to measure the overall model performance across all thresholds and ensure the measure is independent of prior distributions, Receiver Operating Characteristic

(ROC) curve, which represents a trade-off between TPR and FPR, is deemed as a useful tool to measure overall ranking and separation power in model evaluation (Powers, 2011). A typical traditional ROC curve is constructed based on two assumptions:

1. The predictive score $X$ generated by the classification algorithm is of a scalar and $\Omega = R$.

2. A threshold rule is used to make the class label prediction $\hat{Y}$: given a threshold $t$, the instance is classified as "positive" or $\hat{Y} = 1$ if $X \geq t$, otherwise, it is classified as "negative" or $\hat{Y} = 0$.

Various pairs of $(FPR, TPR)$ are created by varying the thresholds. An example of ROC curve is depicted in Figure 2.1. The $x-$axis is FPR, and the $y$-axis is TPR. The diagonal connecting (0,0) and (1,1) implies that the model does not provide any discriminatory power to separate the two classes. Classifiers with ROC curves far above the diagonal curve are desirable in terms of discriminatory power.

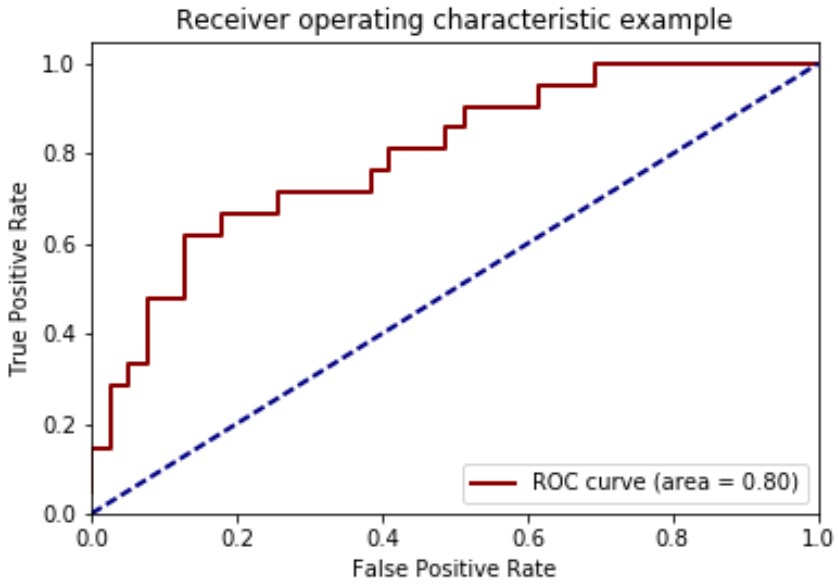


Figure 2.1: An Example of ROC Curve *Source: simulation (Pedregosa et al., 2011)*

### Cost-Sensitive Measure

In addition to confusion-matrix measures and ROC curve, recently cost-sensitive based measures have gained more and more popularity (Elkan, 2001; Bahnsen et al., 2013, 2014; George et al., 2016; Wang and Chen, 2017; Olowookere and Adewale, 2020). Most classifiers assume equal misclassification cost, which is not true in many real-world applications. For example, the cost of accepting a fraudulent transaction is much higher than rejecting a legitimate transaction. Cost-sensitive learning was proposed to address the unequal cost issue. The effectiveness of cost-sensitive learning relies strongly on the supplied cost matrix (Fernández et al., 2018). Table 2.2 is a typical cost matrix for binary classification, where $C(i,j), (i,j \in \{0,1\})$ is the cost of predicting class $i$ when the true class is $j$. Let $c_{ij}$ denote $C(i,j)$ for simplicity. It is reasonable to assume $c_{01} > c_{11}$ and $c_{10} > c_{00}$. Although most research uses the terminology of costs, the matrix in terms of profits generally is preferable (Elkan, 2001). Without specific instructions, the default matrix we use in this dissertation is profit matrix. In addition, typical cost-sensitive methods assume a fixed cost for each cell in the matrix. In other words, the cost depends on the class and is the same across instances. This type of cost-matrix is called 'class-dependent' matrix.

The class-dependent matrix is not practical in many real-world scenarios (Correa Bahnsen et al., 2015). In the case of credit card fraud detection, accepting a fraudulent transaction may result in an amount of loss varying by the transaction amount (Ngai et al., 2011). In churn analysis, which is to predict the customers' likelihood of abandoning a service provider, failing to differentiate a profitable or unprofitable churner has a significant different financial impact (Verbraken et al., 2012). In such cases, the elements in the cost-matrix vary by instances. This type of cost-matrix is known as instance-dependent matrix.

| | | Actual Class | |
|---|---|---|---|
| | | $Y = 1$ | $Y = 0$ |
| **Predicted Class** | $\hat{Y} = 1$ | $C(1,1) = c_{11}$ | $C(1,0) = c_{10}$ |
| | $\hat{Y} = 0$ | $C(0,1) = c_{01}$ | $C(0,0) = c_{00}$ |

Table 2.2: Binary Misclassification Cost Matrix

## 2.2 Related Work and Research Motivations

There are a myriad of research papers on performance measures (Fatourechi et al., 2008; Sokolova and Lapalme, 2009; Hossin and Sulaiman, 2015; Jiao and Du, 2016; Canbek et al., 2017; Ballabio et al., 2018). After reviewing those papers, we identify that those studies implicitly include two major assumptions:

1. The score $X$ is of a scalar and $\Omega = R$.

2. The score $X$ is a 'perfect' or 'well-calibrated' score or the score can result in a concave ROC.

The first assumption is often unnoticeable. Note that all above-mentioned measures are based on a threshold classification rule to make the class label prediction $\hat{Y}$. The threshold rule implies that $X$ is a scalar. In reality, the predictive score can be a multi-dimensional vector.

Second assumption is explicitly made in many literature studies. For example, Cohen and Goldszmidt (2004); Lipton et al. (2014); Böken (2021) discussed the properties of the confusion-matrix measures and studied how to choose the optimal threshold to maximize corresponding measures in confusion-matrix based on the assumption that the score is a 'perfect' or from a 'well-calibrated' classifier. Nevertheless, the authors fail to discuss how to transform a non-perfect score into a 'perfect' one. Another example is from ROC analysis, where there are numerous theoretical and empirical research papers. Majority of those studies implicitly assume concavity of ROC (Bradley, 1997; Fawcett, 2006; Majnik and Bosnić, 2013; Narkhede, 2018; Wang and Ni, 2019; Kannan and Vasanthi, 2019; Omar and Ivrissimtzis, 2019; Janssens and Martens, 2020; Gong,

2021). However, there is much evidence showing that non-concavity is unavoidable. Vardhan et al. (2020) recommended not to use the classifier if the ROC curve is not concave. The properties and characterization of ROC curve is not well understood. In addition, there is lack of study regarding how to transform a non-concave ROC into a concave ROC.

For cost-sensitive measures, Elkan (2001) proposed a direct cost-sensitive threshold adjustment method to transform an accuracy based threshold into a cost-sensitive decision. This method heavily relies on the estimation accuracy. Wang and Chen (2017) enhanced Elkan's method by minimizing the expected cost subject to the convex assumption of Type I and Type II error functions, which can result in concave ROC.

In summary, current research on performance measures are subject to the violation of the above-mentioned assumptions, since all these studies fail to discuss when $X$ is multi-dimensional or $X$ is not 'perfect'. Motivated by the identified gaps, we propose likelihood and Bayesian probability transformation to convert those non-perfect multi-dimensional scores and in-depth discuss their properties after transformation.

# Chapter 3

# LIKELIHOOD AND BAYESIAN TRANSFORMATIONS

Suppose the class label of the instance is a binary random variable, $Y \in \{0, 1\}$, such that $P(Y = 1) = \pi_1$ and $P(Y = 0) = \pi_0 = 1 - \pi_1$. $\pi_1$ is also called base rate. The classification algorithm generates a predictive score $X \in \Omega \subseteq R^n$ for the instance. The predictive score can be a multi-dimensional vector such as customer attributes or a single dimensional scalar such as a credit score. Assume $X$ is a continuous random variable following a probability density distribution $f(x)$ and cumulative distribution function $F(x)$. The effectiveness of the classification algorithm is determined by the conditional distributions of the predictive score $X$ for both classes: $f_i(x) \equiv f(x|Y = i)$, $i \in \{0, 1\}$. Denote $F_i(x)$ as the corresponding cumulative distribution functions of $f_i(x)$. Clearly, $f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$.

## 3.1 Likelihood and Bayesian Transformation

In this section, we study the properties of random variable transformations that convert the multi-dimensional predicted score $X$ to a scalar $S$. Consider a differential variable transformation $T$ that maps $X \in \Omega \subseteq R^n$ to $S \in \Omega_T \subseteq R$:

$$S = T(X),\ \Omega_T = \{s|s = T(x), x \in \Omega\}. \tag{3.1}$$

Let $g^T(\cdot)$ be the corresponding probability density function of random variable $S$. We have

$$g^T(s) = \sum_{k=1}^{n(s)} \left| \frac{d}{ds} T_k^{-1}(s) \right| f(T_k^{-1}(s)), \tag{3.2}$$

where $T_k^{-1}(s)$ are solutions of equation $T(x) = s$ and $n(s)$ is the total number of solutions. Similarly, the conditional density functions $g_0^T(s)$ and $g_1^T(s)$ of $S$, corresponding to negative and positive instances, respectively, can be expressed as follows:

$$g_0^T(s) = \sum_{k=1}^{n(s)} \left| \frac{d}{ds} T_k^{-1}(s) \right| f_0(T_k^{-1}(s)), \tag{3.3}$$

$$g_1^T(s) = \sum_{k=1}^{n(s)} \left| \frac{d}{ds} T_k^{-1}(s) \right| f_1(T_k^{-1}(s)). \tag{3.4}$$

Since $f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$, using (Eqs. 3.2 - 3.4), we have

$$g^T(s) = \pi_0 g_0^T(s) + \pi_1 g_1^T(s) \tag{3.5}$$

holds for all $s \in \Omega_T$ and any differential transformation $T(\cdot)$.

Now consider the likelihood transformation $L$ that maps $X \in \Omega \subseteq R^n$ to $S \in \Omega_L \subseteq R_+$:

$$S = L(x) = \frac{f_1(x)}{f_0(x)}, \ \Omega_L = \{s | s = L(x), x \in \Omega\}. \tag{3.6}$$

Since $L(x) \geq 0$ by definition, we have $\Omega_L \subseteq R_+$. Let $g^L(\cdot)$, $g_0^L(\cdot)$, and $g_1^L(\cdot)$ be the density and class-conditional density functions of the resulting random scalar $S$.

**Proposition 1.** *The following properties hold for $g^L(\cdot)$, $g_0^L(\cdot)$, and $g_1^L(\cdot)$ resulting from the likelihood transformation (Eq. 3.6):*

1. $g_0^L(s) = [1/(\pi_0 + \pi_1 s)] g^L(s), \quad g_1^L(s) = [s/(\pi_0 + \pi_1 s)] g^L(s), \quad$ *and*
2. $g_1^L(s)/g_0^L(s) = s$

*for all $s \in \Omega_L \subseteq R_+$.*

*Proof.* We first prove the result 2. Notice that

$$\frac{g_1^L(s)}{g_0^L(s)} = \frac{\sum_{k=1}^{n(s)} \left| \frac{d}{ds} L_k^{-1}(s) \right| f_1(L_k^{-1}(s))}{\sum_{k=1}^{n(s)} \left| \frac{d}{ds} L_k^{-1}(s) \right| f_0(L_k^{-1}(s))}.$$

For each $k$, since $f_1(L_k^{-1}(s))/f_0(L_k^{-1}(s)) = L(L_k^{-1}(s)) = s$, we have $g_1^L(s)/g_0^L(s) = s$. Result 2 then follows immediately. Combining result 2 and (Eq. 3.5) with $T(\cdot) = L(\cdot)$, we obtain result 1. □

We have two observations based on Proposition 1.

- The density function $g^L(\cdot)$ resulting from the likelihood transformation contains more information than the density function $f(\cdot)$ in the $x$ space. Indeed, together with the prior probability $\pi_0$ ($\pi_1 = 1 - \pi_0$), we can recover the class-conditional density functions $g_0^L(\cdot)$ and $g_1^L(\cdot)$. In other words, the likelihood transformed density function $g^L(\cdot)$, together with the prior probability of the instance, would be sufficient information to capture the performance of the classification algorithm.
- $S$ is a now scalar random variable and its likelihood ratio $g_1^L(s)/g_0^L(s) = s$ monotonically increases in $s$.

Let us consider another transformation $B$ based on popular Bayesian probability that maps $X \in \Omega \subseteq R^n$ to $S \in \Omega_B \subseteq [0, 1]$:

$$B(x) = P(Y = 1|x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_0 f_0(x)} \tag{3.7}$$

Clearly, $B(x) = K(L(x))$, where $K(s) = \pi_1/(\pi_1 + \pi_0/s)$ is a monotone increasing function. Similar to Proposition 1 for the likelihood transformation, we have the following results for the Bayesian transformation:

**Proposition 2.** *The following properties hold for $g^B(\cdot)$, $g_0^B(\cdot)$, and $g_1^B(\cdot)$ resulting from the Bayesian transformation (Eq. 3.7):*

1. $g_0^B(s) = (1 - s)g^B(s)/\pi_0, \quad g_1^B(s) = sg^B(s)/\pi_1, \quad$ *and*
2. $g_1^B(s)/g_0^B(s) = (\pi_0/\pi_1)[s/(1 - s)]$

*for all $s \in \Omega_B \subseteq [0, 1]$.*

### 3.2 Dominance Properties of Likelihood Transformation

Recall that the threshold rule is used to make the class label prediction $\hat{Y}$: given a threshold $t$, an instance is classified as "positive" or $\hat{Y} = 1$ if $X \geq t$, otherwise, it is classified as "negative" or $\hat{Y} = 0$. Let $R_L$ denote the set of $x$ such that the associated instances are classified as positive based on the likelihood ratio transformation, and $R_T$ based on another transformation:

$$\begin{aligned} R_L &= \{x : L(x) = \frac{f_1(x)}{f_0(x)} \geq t_L\}, \\ R_T &= \{x : T(x) \geq t_T\}. \end{aligned} \tag{3.8}$$

Denote their respective complementary sets as follows:

$$\begin{aligned} R_L^c &= \{x : L(x) = \frac{f_1(x)}{f_0(x)} < t_L\}, \\ R_T^c &= \{x : T(x) < t_T\}. \end{aligned} \tag{3.9}$$

We have the following additional properties for likelihood transformation:

**Proposition 3.** *For any thresholds $t_L \in [0, +\infty)$, $t_T \in (-\infty, +\infty)$ and scalars $\alpha, \beta \geq 0$ such that $\alpha + \beta \neq 0$, if*

$$\alpha \int_{R_L} f_0(x)dx + \beta \int_{R_L} f_1(x)dx = \alpha \int_{R_T} f_0(x)dx + \beta \int_{R_T} f_1(x)dx,$$

*then*

$$\int_{R_L} f_1(x)dx \geq \int_{R_T} f_1(x)dx \text{ and } \int_{R_L} f_0(x)dx \leq \int_{R_T} f_0(x)dx.$$

*Proof.* If $t_L = 0$, the proposition holds with $t_T = -\infty$. We assume $t_L > 0$ in the rest of the proof. Since

$$\alpha \int_{R_L} f_0(x)dx + \beta \int_{R_L} f_1(x)dx = \alpha \int_{R_T} f_0(x)dx + \beta \int_{R_T} f_1(x)dx,$$

we have

$$\alpha P(R_L|Y=0)+\beta P(R_L|Y=1)=\alpha P(R_T|Y=0)+\beta P(R_T|Y=1).$$

This implies

$$\alpha P(R_L\cap R_T^c|Y=0)+\beta P(R_L\cap R_T^c|Y=1)=\alpha P(R_T\cap R_L^c|Y=0)+\beta P(R_T\cap R_L^c|Y=1), \tag{3.10}$$

since

$$R_L=(R_L\cap R_T)\cup(R_L\cap R_T^c),\ R_T=(R_L\cap R_T)\cup(R_L^c\cap R_T). \tag{3.11}$$

By (Eq. 3.8), we have

$$\begin{aligned}
&\alpha P(R_L\cap R_T^c|Y=0)+\beta P(R_L\cap R_T^c|Y=1)\\
\geq\ &\alpha P(R_L\cap R_T^c|Y=0)+t_L\beta P(R_L\cap R_T^c|Y=0)\\
=\ &(\alpha+t_L\beta)P(R_L\cap R_T^c|Y=0).
\end{aligned}$$

By (Eq. 3.9), we have

$$\begin{aligned}
&\alpha P(R_T\cap R_L^c|Y=0)+\beta P(R_T\cap R_L^c|Y=1)\\
\leq\ &\alpha P(R_T\cap R_L^c|Y=0)+t_L\beta P(R_T\cap R_L^c|Y=0)\\
=\ &(\alpha+t_L\beta)P(R_T\cap R_L^c|Y=0).
\end{aligned}$$

Since $\alpha,\beta\geq 0$, $\alpha+\beta\neq 0$, and $t_L>0$, $\alpha+t_L\beta>0$. Using (Eq. 3.10), we get

$$P(R_L\cap R_T^c|Y=0)\leq P(R_T\cap R_L^c|Y=0).$$

According to (Eq. 3.11), we have $P(R_L|Y=0)\leq P(R_T|Y=0)$, which is equivalent to $\int_{R_L} f_0(x)dx\leq\int_{R_T} f_0(x)dx$.

Using the inequalities below:

$$\alpha P(R_L \cap R_T^c | Y=0) \leq \frac{\alpha}{t_L} P(R_L \cap R_T^c | Y=1) \qquad \text{because of (Eq. 3.8)}$$
$$\alpha P(R_T \cap R_L^c | Y=0) \geq \frac{\alpha}{t_L} P(R_T \cap R_L^c | Y=1), \qquad \text{because of (Eq. 3.9)}$$

we can prove $\int_{R_L} f_1(x)dx \geq \int_{R_T} f_1(x)dx$ with the same logic. □

When $\alpha = 0$ or $\beta = 0$, we have the following corollary:

**Corollary 1.** *For any thresholds $t_L \in [0, +\infty)$ and $t_T \in (-\infty, +\infty)$,*

1. *if $\int_{R_L} f_0(x)dx = \int_{R_T} f_0(x)dx$, then $\int_{R_L} f_1(x)dx \geq \int_{R_T} f_1(x)dx$;*
2. *if $\int_{R_L} f_1(x)dx = \int_{R_T} f_1(x)dx$, then $\int_{R_L} f_0(x)dx \leq \int_{R_T} f_0(x)dx$.*

The above propositions and corollary will be used later in the dissertation to show dominance properties of various classification measurements and dominance of ROC curve under the likelihood transformation.

### 3.3 Summary

Although likelihood ratio and Bayesian probability are well understood, very little attention has been paid to the properties of the density functions after the likelihood and Bayesian transformation. In this Chapter, we proposed three forms of transformations.

- Our universal transformation (Eq. 3.1) converts a multi-dimensional score into a scalar with domain in $R$.
- The likelihood transformation (Eq. 3.6) converts the score into a scalar with domain in $R_+$. More importantly, we identify that the density function $g^L(\cdot)$ resulting from the likelihood transformation is much more informative than the density function $f(\cdot)$ in the $x$ space.

- The Bayesian probability transformation (Eq. 3.7) coverts the score into a scalar with domain in $[0, 1]$. We also note the similar nice properties of the density function associated with the transformed score.

Our proposed transformation methods can aid to convert a non-perfect multi-dimensional score into a new score. Based on the transformed new score, the threshold classification rule can be applied, which results in Neyman-Pearson classification (Neyman and Pearson, 1933). It is worth mentioning that the monotonic properties as well as the additional properties in integration identified in this chapter for the two transformations play a critical role in the rest of the dissertation.

# Chapter 4

# A UNIFIED FRAMEWORK FOR CLASSIFICATION MEASURES

Performance measures play an essential role in machine learning classification model evaluation (Gong, 2021). There are a plethora of classification measures. Ferri et al. (2009) and Gong (2021) classified these performance measures into three broad categories:

- Metrics based on the deviation from the actual probability: mean absolute error, accuracy ratio, Hinge loss, mean squared error (or known as Brier score), LogLoss (or known as cross-entropy), etc.
- Metrics based on separation power and visualization tool: ROC and its variants, precision-recall curve, AUC, Gini, cost-curve, Kolmogorov-Smirnov, total operating characteristic (TOC), lift chart, etc.
- Metrics based on a confusion matrix: accuracy, balanced accuracy, positive predictive value (PPV), recall, $F_\beta$, Jaccard index, Youden index, Matthews Correlation Coefficient (MCC), etc.

These performance measures have been extensively studied and widely used in many machine learning classification applications as a tool for model evaluation (Sokolova and Lapalme, 2009; Jiao and Du, 2016; Canbek et al., 2017; Ballabio et al., 2018). The evaluation of machine learning classifiers generally involve two steps: the first step is to determine which measure to use to compare the model performance and the second step is to estimate the corresponding measures and draw conclusions about which model is better. There are numerous empirical and experimental studies for the first step to select appropriate measure. These studies mainly focus on the advantages

and disadvantages of the measures in various situations. As Ferri et al. (2009) pointed out, given a dataset, a learning method that achieves the best model based on a given measure, might not be the best method if a different measure is used. For instance, Ling et al. (2003b) revealed that Naive Bayes and decision tree perform similarly in accuracy. Nevertheless, Naive Bayes is significantly better than decision tree in AUC. Even more surprisingly, Rosset (2004a) showed that if using AUC for selecting models based on a validation dataset results in better accuracy in a test dataset than using accuracy for selecting the models. It is commonly believed that each measure has its own advantages and disadvantages and no measure dominates another. Huang and Ling (2005) defined the criteria (degree of consistency and degree of discriminancy) to compare two different measures for learning algorithms. There also exist papers focusing on the machine learning performance measures in the specific fields. For example, Diab and Sartawi (2017) discussed the performance measures in text classification; Strøm et al. (2018) researched machine learning performance metrics and diagnostic context in radiology; Suragala et al. (2021) compared performance metrics of data mining algorithms on medical data. Another category of research papers study the measure selection by empirically exploring the correlations among various measures. For example, Ferri et al. (2009) analyzed experimentally the behaviour of eighteen different performance measures in multiple scenarios, discovering clusters and relationships among measures.

For the second step of the model evaluation, there are very limited literature. Most of the available research discussed the criteria to draw conclusions about which model is better. For example, Santafe et al. (2015) reviewed the important aspects of the evaluation process in supervised classification and highlighted the statistical significance tests of differences of the measures. There lack discussions regarding how to estimate the potential optimal measurement for each given performance measure. The gap motivates us to perform a deep dive of the three categories of measures and discuss their analytical properties in decision making process.

Metrics based on the deviation from the actual probability are used to evaluate the probability estimation accuracy performance. They are generally built into the

objective function of machine learning algorithms and are optimized during model training (Wu and Liu, 2007; Chai and Draxler, 2014; Zhang and Sabuncu, 2018; Zhou et al., 2019). These measures are independent of thresholds and are calculated without the need to specify thresholds. Therefore, this category of measures is out of our discussion scope. For metrics based on discriminatory power, they measure the overall rank-ordering ability of probability output without requiring exact estimation of the probability and are independent of threshold. Their calculations, however, rely on threshold classification rule (see Chapter 2). While the rank-ordering metrics provide an overall picture of model performance across all thresholds, there remains a practical need to determine the specific threshold that should be used for individual instance decisions. Confusion-matrix based measures are often those measures which are used to determine the optimal decision threshold for ROC alike rank-ordering measures (Fluss et al., 2005; Leeflang et al., 2008; Habibzadeh et al., 2016; Unal, 2017). As introduced in Chapter 2, most frequently used confusion-matrix based performance measures (Ling et al., 2003a; Ferri et al., 2009; Sokolova and Lapalme, 2009; Powers, 2011; Jiao and Du, 2016; Chicco and Jurman, 2020; Gong, 2021) include accuracy, balanced accuracy, recall, precision, specificity, $F$ score, Youden index, etc. These measures are also often used in ROC analysis to determine an optimal decision threshold. According to Unal (2017), the most popular method for threshold selection in ROC analysis is the Youden index method. This method defines the optimal threshold as the point maximizing the Youden function, which is the difference between TPR and FPR over all possible thresholds. The maximum Youden index has been proven equivalent to the Maximum Kolmogorov-Smirnov difference (Krzanowski and Hand, 2009), which is a commonly used performance measure in credit score modeling (Fang and Chen, 2019). Another approach is called the point closest-to-(0,1) corner criteria in ROC space (called CC), which defines the optimal threshold as the point maximizing the Euclidean distance between ROC curve and (0,1) point (Perkins Neil and Schisterman Enrique, 2006). A more recent approach proposed by Liu (2012), known as Concordance probability method (CP), characterizes the optimal threshold as the point maximizing the product

of TPR and TNR. Majority of the research, however, stopped there without in-depth discussing on determining the exact threshold. Only very limited discussions exist regarding how to find the optimal threshold to maximize the corresponding confusion-matrix based measures. The very limited examples include: Cohen and Goldszmidt (2004); Böken (2021) mentioned that for a 'well-recalibrated' classifier, the optimal threshold to maximize accuracy is 1/2; Lipton et al. (2014) provided insights into maximizing $F1$ scores and concluded that the optimal threshold is half the optimal $F1$ score.

These individual studies on some measures are ad-hoc by simply assuming the classifiers are well-calibrated and there lacks a systematic framework to discuss all classification measures. The absence of such discussions is mainly because the properties of the density function $g^L(\cdot)$ resulting from the likelihood transformation are not well recognized (see Propositions 1 and 2) and most research (Fluss et al., 2005; Liu, 2012) implicitly assume exhaustive search to find the optimal thresholds, which is inefficient and computationally expensive. In this Chapter, we develop a *unified framework* to study threshold-dependent classification measures, which include all popularly used classification measures. We systematically study the properties of the performance measurement as a function of the corresponding threshold. More specifically, we show that a monotone transformation of the score variable preserves the maximum performance measure and the uni-modality of the measurement function, while the likelihood transformation leads to a dominant maximum performance measure. We also classify the performance measurement functions according to their shapes: monotone, uni-modal, and multi-modal. These properties allow us to design efficient algorithms to search for the optimal threshold. To the best of our knowledge, this is the first work to provide a generalized framework to systematically study the properties of confusion-matrix based measures.

This chapter is structured as follows: Section 4.1 briefly introduces all classification measures and their relationships with the two fundamental basic measures TPR and FPR; Section 4.2 proposes a unified class of performance measures, which is a

monotone function of TPR and FPR. The unified class is shown to be general enough to include all existing confusion-matrix based measures as special cases; Section 4.3 examines the invariant properties of the underlying confusion-matrix and show that a monotone transformation of the score variable preserves the uni-modality of measurement function and the maximum measurement value; Section 4.4 reveals that the likelihood transformation of the score variable leads to the dominance of the maximum measurement value; Section 4.5 formulates measurement threshold optimization problem and studies the shape properties of the measurement function. All existing performance measures are classified into three categories according to their corresponding measurement function shapes: monotone, uni-modal, and multi-modal; Section 4.6 further discusses the intrinsic relationship between the measures and base rate; Section 4.7 summarizes our results and highlights our major contributions.

## 4.1 Background

Recall the definition in Chapter 2: $X$ is a scalar random variable following a continuous probability density distribution and the following threshold prediction for the classification algorithm applies: predicting $\hat{Y} = 1$ if $x \geq t$ and $\hat{Y} = 0$ otherwise, where $t \in R$ is a threshold parameter to be determined. A confusion-matrix (Table 9) is characterized by four elements depending on the threshold $t$ and base rate $\pi_1$:

$$CM(f_0, f_1, t, \pi_1) = (TP(t, \pi_1), FP(t, \pi_1), TN(t, \pi_1), FN(t, \pi_1)). \tag{4.1}$$

The four elements in the confusion matrix corresponds to the joint probabilities $P(Y = i, \hat{Y} = j)$, $i, j \in \{0, 1\}$:

$$TP(t, \pi_1) = P(Y = 1, \hat{Y} = 1) = \pi_1 P(\hat{Y} = 1 | Y = 1) = \pi_1 \int_t^{+\infty} f_1(s) ds, \tag{4.2}$$

$$FP(t, \pi_1) = P(Y = 0, \hat{Y} = 1) = \pi_0 P(\hat{Y} = 1 | Y = 0) = \pi_0 \int_t^{+\infty} f_0(s) ds, \tag{4.3}$$

$$TN(t, \pi_1) = P(Y = 0, \hat{Y} = 0) = \pi_0 P(\hat{Y} = 0 | Y = 0) = \pi_0 \int_{-\infty}^{t} f_0(s) ds, \tag{4.4}$$

$$FN(t, \pi_1) = P(Y = 1, \hat{Y} = 0) = \pi_1 P(\hat{Y} = 0 | Y = 1) = \pi_1 \int_{-\infty}^{t} f_1(s) ds. \tag{4.5}$$

By definitions (Eqs. 2.1 & 2.4),

$$TPR(t) = \frac{TP(t)}{TP(t) + FN(t)} = \frac{TP(t)}{\pi_1} = \int_t^{+\infty} f_1(t) dt$$
$$FPR(t) = \frac{FP(t)}{FP(t) + TN(t)} = \frac{FP(t)}{\pi_0} = \int_t^{+\infty} f_0(t) dt.$$

Therefore, for a fixed $\pi_1$, the confusion-matrix is fully described by $TPR(t)$ and $FPR(t)$, both of which are monotonically decreasing in $t$. It follows that for a given $\pi_1$, any performance measure derived from a confusion-matrix should be a function of $TPR(t)$ and $FPR(t)$.

Recall the definitions of single measures Chapter 2: TPR (Eq. 2.1), FNR (Eq. 2.2), TNR (Eq. 2.3), FPR (Eq. 2.4), PPV (Eq. 2.5) and NPV (Eq. 2.7), as well as various balanced measures by combining the above single measures, such as accuracy (ACC), error rate (ERR), balanced accuracy (BA), $F_\beta$, Jaccard index (JI), Markedness (MK), and ROC threshold methods such as Youden index (J), product of TPR and TNR (or called CP method), Euclidean distance between ROC curve and the (0,1) point (or called CC), Matthews Correlation Coefficient (MCC). Let $a$ denote TPR and $b$ denote FPR. Table 4.1 summarizes these measure in their original definition and in terms of TPR and FPR.

| Name | Definition | Function in TPR/FPR |
|---|---|---|

| | | |
|---|---|---|
| $PPV$ | $\pi_1 TPR/(\pi_1 TPR + \pi_0 FPR)$ | $\pi_1 a/(\pi_1 a + \pi_0 b)$ |
| $NPV$ | $\pi_0 TNR/(\pi_0 TNR + \pi_1 FNR)$ | $\pi_0(1-b)/(\pi_0(1-b) + \pi_1(1-a))$ |
| $FDR$ | $\pi_0 FPR/(\pi_1 TPR + \pi_0 FPR)$ | $\pi_0 b/(\pi_1 a + \pi_0 b)$ |
| $FOR$ | $\pi_1 FNR/(\pi_0 TNR + \pi_1 FNR)$ | $\pi_1(1-a)/(\pi_0(1-b) + \pi_1(1-a))$ |
| $ACC$ | $\pi_1 TPR + \pi_0 TNR$ | $\pi_1 a + \pi_0(1-b)$ |
| $ERR$ | $\pi_1 FNR + \pi_0 FPR$ | $\pi_1(1-a) + \pi_0 b$ |
| $BA$ | $c * TPR + (1-c) * TNR$ | $c * a + (1-c) * (1-b)$ |
| $J$ | $TPR + TNR - 1$ | $a - b$ |
| $JI$ | $TP/(TP + FN + FP)$ | $\pi_1 a/(\pi_1 + \pi_0 b)$ |
| $F1$ | $2PPV * TPR/(PPV + TPR)$ | $2a/(\pi_1 a + \pi_1 + \pi_0 b)$ |
| $F_\beta$ | $(1+\beta^2)PPV * TPR/(\beta^2 * PPV + TPR)$ | $\pi_1 a(1+\beta^2)/(\pi_1 \beta^2 + \pi_1 a + \pi_0 b)$ |
| $MK$ | $PPV + NPV - 1$ | $\pi_1 a/(\pi_1 a + \pi_0 b) + \pi_0(1-b)/(1-\pi_0 b - \pi_1 a) - 1$ |
| $MCC$ | $\sqrt{PPV * TPR * TNR * NPV}$ $- \sqrt{FDR * FNR * FPR * FOR}$ | $(a-b)\sqrt{\pi_0 \pi_1/[(\pi_1 a + \pi_0 b)(1 - \pi_1 a - \pi_0 b)]}$ |
| $CC$ | $\sqrt{(1-TPR)^2 + FPR^2}$ | $\sqrt{(1-a)^2 + b^2}$ |
| $CP$ | $TPR * TNR$ | $a(1-b)$ |

Table 4.1: Confusion-Matrix Measures in Terms of TPR/FPR

We next briefly introduce and elaborate the popular measures in the above table. PPV, NPV, FDR and FOR are discussed in Chapter 2. Accuracy (error rate) is used to measure how well a binary classifier correctly identifies the true positives and true negatives (Eqs. 2.9 - 2.10). Balanced accuracy is defined as the arithmetic mean of true positive rate and true negative rate. A generalized balanced accuracy measure assigns different weight ($c \in [0, 1]$) to the accuracy on the positive/negative classes. Youden index (also known as J statistic or Bookmaker Informedness) is a single statistic to measure the model separation power. Youden index is strongly related to balanced accuracy. On the contrary, Markedness is the arithmetic mean of positive predictive value and negative predictive value which measures the reliability of the positive and negative predictions by the model.

The widely adopted $F1$ measure is the harmonic mean of TPR and PPV. Jaccard index, also known as threat score (TS) or critical success index (CSI), is used to measure the similarity between the set of actual positives and predicted positives. Clearly

$$F1 = \frac{2JI}{1 + JI},$$

which indicates that $F1$ is a monotonically increasing function of Jaccard index. Note that $F1$ weighs PPV and TPR equally in the formula. A generalized measure is given by $F_\beta$ which assigns $\beta$ times as much importance to TPR as PPV.

The measures described so far only consider part of information of a confusion-matrix. In contrast, the Matthews correlation coefficient takes into account all four elements of the confusion matrix. MCC ranges from $[-1, 1]$, with the higher, the better agreement between the actual and prediction.

Besides the above-mentioned measures, the following two methods are often used to determine the optimal thresholds in ROC analysis:

1. The closest-to-(1, 0) Corner Criteria (CC). In this criteria, the "optimal" threshold is defined as the point closest to the point (1, 0) on the ROC curve.

2. Concordance Probability Method (CP). The concordance probability method defines the optimal threshold as the point maximizing the product of true positive rate and true negative rate.

## 4.2 A Unified Class of Classification Measures

Based on the argument in the previous section, all confusion matrix-based measures are functions of TPR and FPR. This motivates us to propose a unified classification performance measure in the following form:

$$\widetilde{MM}(t) = MM(TPR(t), FPR(t)),\ t \in R, \tag{4.6}$$

with the mappings specified by $MM : [0,1]^2 \to R_+$ and $\widetilde{MM} : R \to R_+$. If TPR, FPR are derived from density functions $g_0^T(\cdot)$ and $g_1^T(\cdot)$, which are obtained from a $T$ transformation of predicted score $X$ (Eq. 3.1) described in Chapter 3, we denote the unified measure function as $\widetilde{MM}^T(t), t \in \Omega_T$. For simplicity, we often drop the superscript $T$, if the transformation is not the focus of our study, and use $\widetilde{MM}$ in the remaining section. If $T$ is the likelihood or the Bayesian transformation, the unified measure is denoted as $\widetilde{MM}^L(t), t \in R_+$, or $\widetilde{MM}^B(t), t \in [0,1]$, respectively. For example, $FPR^L(t)$ and $MCC^B(t)$ represent the resulting FPR and MCC measures when the likelihood and the Bayesian transformations are applied to $X$, respectively.

We now introduce a class of unified measures $\mathcal{M}$ by restricting $MM$ to be a monotone increasing function of $TPR$ and monotone decreasing function of $FPR$. Denote $MM_1'(a,b) = \partial MM(a,b)/\partial a$ and $MM_2'(a,b) = \partial MM(a,b)/\partial b$.

**Definition 1.** *$\mathcal{M}$ consists of all measurement functions $MM(TPR(t), FPR(t))$, $t \in R$, such that $MM(\cdot,\cdot)$ satisfies $MM_1'(a,b) \geq 0$, $MM_2'(a,b) \leq 0$, $\forall a, b \in [0,1]$.*

We show that all existing classification performance measures belong to this class.

**Proposition 4.** *TPR, TNR, PPV, NPV,-FOR, -FDR, -FNR, -FPR, ACC, -ERR, BA, J, CP, -CC, JI, $F_\beta$, MK, MCC $\in \mathcal{M}$.*

*Proof.* In light of Table 4.1, it is straightforward to verify that measures TPR, TNR, PPV, NPV,-FOR, -FDR, -FNR, -FPR, ACC, -ERR, BA, J, CP, -CC, JI, $F_\beta$ are monotonically increasing in TPR and monotonically decreasing in FPR. Since $MK = PPV + NPV - 1$ and $PPV, NPV \in \mathcal{M}$, we have $MK \in \mathcal{M}$. For MCC, it can be re-expressed as:

$$MCC = \sqrt{PPV * TPR * TNR * NPV} - \sqrt{FDR * FNR * FPR * FOR}.$$

Since the first part of $MCC$ is monotonically increasing in TPR and decreasing in FPR, while the second part is reversely true, we obtain the proof. □

### 4.3 Measure Invariance under Monotone Transformation

In this section, we show that a monotone increasing transformation on $X \in R$ will preserve, in certain sense, the corresponding confusion matrix $CM$ and all the measures $\widetilde{MM}$ based on the confusion matrix. We start by presenting the invariance property for the confusion matrix.

**Lemma 1.** *Suppose $X \in R$ and transformation $T(\cdot)$ in (Eq. 3.1) is a monotonically increasing function, then $CM(g_0^T, g_1^T, t_g, \pi_1) = CM(f_0, f_1, t_f, \pi_1)$, where $t_g = T(t_f)$.*

*Proof.* According to (Eq. 4.1),

$$CM(f_0, f_1, t_f, \pi_1) = (TP(t_f, \pi_1), FP(t_f, \pi_1), TN(t_f, \pi_1), FN(t_f, \pi_1))$$

and

$$CM(g_0^T, g_1^T, t_g, \pi_1) = (TP^T(t_g, \pi_1), FP^T(t_g, \pi_1), TN^T(t_g, \pi_1), FN^T(t_g, \pi_1)).$$

To prove the two confusion-matrices are equal, we need to show each element is equal. Notice that $s = T(x)$ by definition and $t_g = T(t_f)$ by assumption. Since $T(x)$ is a monotone increasing function, $x \geq t_f$ implies $T(x) \geq T(t_f)$, and vice versa. Therefore, we have

$$\begin{aligned} TP(t_f, \pi_1) &= \pi_1 P(x \geq t_f | Y = 1) \\ &= \pi_1 P(T(x) \geq T(t_f) | Y = 1) \\ &= \pi_1 P(s \geq t_g | Y = 1) \\ &= TP^T(t_g, \pi_1). \end{aligned}$$

The same logic can be followed to prove the equalities associated with the other three elements. □

According to Lemma 1, $TPR^T(t_g) = TPR(t_f)$ and $FPR^T(t_g) = FPR(t_f)$. In light of measurement definition (Eq. 4.6), we have the following proposition:

**Proposition 5.** *Suppose transformation $T(\cdot)$ in (Eq. 3.1) is a monotonically increasing function, then $\widetilde{MM}^T(t_g) = \widetilde{MM}(t_f)$, where $t_g = T(t_f)$.*

Using Proposition 5, we show next that a monotone increasing transformation $T$ preserves the maximal value of and the uni-modality of $\widetilde{MM}(t)$, if it holds before the transformation.

**Corollary 2.** *Suppose $X \in R$ and transformation $T(\cdot)$ in (Eq. 3.1) is a monotonically increasing function. Let*

$$t^T = \arg\max_{t \in \Omega_T} \widetilde{MM}^T(t) \text{ and } t^* = \arg\max_{t \in R} \widetilde{MM}(t).$$

*The following properties hold:*

1. $\widetilde{MM}^T(t^T) = \widetilde{MM}(t^*)$.
2. $\widetilde{MM}^T(t)$, $t \in \Omega_T$, *is a unimodal function if and only if* $\widetilde{MM}(t)$, $t \in R$, *is a unimodal function.*

*Proof.* For $t_g \in \Omega_T, t_f \in R$, note that $t_g = T(t_f)$ by definition and $t_{g^*} = T(t^*)$ by assumption. According to Proposition 5, we have $\widetilde{MM}^T(t_g) = \widetilde{MM}(t_f)$ and $\widetilde{MM}^T(t_{g^*}) = \widetilde{MM}(t^*)$. Because $\widetilde{MM}(t_f) \leq \widetilde{MM}(t^*)$, we have $\widetilde{MM}^T(t_g) \leq \widetilde{MM}^T(t_{g^*})$. Therefore, $t_{g^*} = t^T$ and $\widetilde{MM}^T(t^T) = \widetilde{MM}(t^*)$.

For second property, we can prove sufficiency condition and necessary condition can be proven by following the same logic. Because $\widetilde{MM}(t_f)$ is a unimodal function,

$$\text{if } t_{f1} < t_{f2} < t^*, \widetilde{MM}(t_{f1}) < \widetilde{MM}(t_{f2}) < \widetilde{MM}(t^*); \tag{4.7}$$

$$\text{if } t_{f1} > t_{f2} > t^*, \widetilde{MM}(t_{f1}) < \widetilde{MM}(t_{f2}) < \widetilde{MM}(t^*). \tag{4.8}$$

Note that $t_{g1} = T(t_{f1})$ and $t_{g2} = T(t_{f2})$ by definition. By Proposition 5, $\widetilde{MM}^T(t_{g1}) = \widetilde{MM}(t_{f1})$, $\widetilde{MM}^T(t_{g2}) = \widetilde{MM}(t_{f2})$. Since $T(x)$ is a monotone increasing function,

$t_{f1} < t_{f2} < t^*$ (or $t_{f1} > t_{f2} > t^*$) implies that $T(t_{f1}) < T(t_{f2}) < T(t^*)$ (or $T(t_{f1}) > T(t_{f2}) > T(t^*)$), and vice versa. Because of (Eqs. 4.7 - 4.8), we have

$$\text{if } t_{g1} < t_{g2} < t_{g^*}, \widetilde{MM}^T(t_{g1}) < \widetilde{MM}^T(t_{g2}) < \widetilde{MM}^T(t_{g^*});$$
$$\text{if } t_{g1} > t_{g2} > t_{g^*}, \widetilde{MM}^T(t_{g1}) < \widetilde{MM}^T(t_{g2}) < \widetilde{MM}^T(t_{g^*}).$$

As a result, $\widetilde{MM}^T(t)$ is a unimodal function and $t_{g^*} = t^T$, which leads to $\widetilde{MM}^T(t_{g^*}) = \widetilde{MM}^T(t^T) = \widetilde{MM}(t^*)$. □

Recall the likelihood transformation $L(\cdot)$ (Eq. 3.6) and Bayesian transformation $B(\cdot)$ (Eq. 3.7) in Chapter 3. Since $B(x) = K(L(x))$ and $K(s) = \pi_1/(\pi_1 + \pi_0/s)$ is a monotone increasing function, according to Proposition 5, we have the following relationship between the likelihood and Bayesian transformations:

**Corollary 3.** *Suppose $X \in \Omega \subseteq R^n$. Consider the likelihood transformation (Eq. 3.6) and the Bayesian transformation (Eq. 3.7). Let*

$$t^L = \arg\max_{t \in R_+} \widetilde{MM}^L(t) \text{ and } t^B = \arg\max_{t \in [0,1]} \widetilde{MM}^B(t).$$

*The following properties hold:*

1. $\widetilde{MM}^L(t^L) = \widetilde{MM}^B(t^B)$.
2. $\widetilde{MM}^L(t), t \in R_+$, *is a unimodal function if and only if* $\widetilde{MM}^B(t), t \in [0,1]$, *is a unimodal function.*

### 4.4 Measure Dominance under the Likelihood Transformation

Different transformation on predicted score $X$ may lead to different confusion-matrix, and therefore different performance measures. We show that the likelihood transformation results in the maximal performance measure values.

**Proposition 6.** *Let $T(\cdot) : R^n \to R$ be any transformation on $X \in R^n$ and $L(\cdot) : R^n \to R_+$ be the likelihood transformation (Eq. 3.6). Let*

$$t^L = \arg\max_{t \in R_+} \widetilde{MM}^L(t) \text{ and } t^T = \arg\max_{t \in R} \widetilde{MM}^T(t).$$

*Then $\widetilde{MM}^L(t^L) \geq \widetilde{MM}^T(t^T)$.*

*Proof.* Since $TPR^L(\cdot), TPR^T(\cdot) \in [0, 1]$, there exists $t$ such that $TPR^L(t) = TPR^T(t^T)$. According to Corollary 1, we have $FPR^L(t) \leq FPR^T(t^T)$. Since $MM$ is monotonically decreasing in $FPR$, we have

$$\begin{aligned}
\widetilde{MM}^T(t^T) &= MM(TPR^T(t^T), FPR^T(t^T)) \\
&= MM(TPR^L(t), FPR^T(t^T)) \\
&\leq MM(TPR^L(t), FPR^L(t)) \\
&\leq MM(TPR^L(t^L), FPR^L(t^L)) \\
&= \widetilde{MM}^L(t^L).
\end{aligned}$$

□

According to Propositions 4 and 6, the likelihood transformation results in the best values of all measures in $\mathcal{M}$. In other words, measures under the likelihood transformation are better representation of the maximal potential of a classification algorithm and should be better measures to compare performances of different algorithms.

### 4.5 Optimal Threshold under the Likelihood Transformation

Section 4.4 reveals that maximal measurements under the likelihood transformation dominate other transformations. We show in this section that measures under the likelihood transformation are more likely to show nice structural properties, such as

monotonicity or uni-modality with regard to the threshold decision. These properties allow us to search for the optimal threshold more efficiently.

### 4.5.1 Optimal Threshold for Measures in $\mathcal{M}$

Consider a performance measure $\widetilde{MM}(t) \in \mathcal{M}$, where $t \in \Omega \subseteq R$ and $\Omega = [t_{min}, t_{max}]$. We are interested in finding the optimal threshold $t^*$ to achieve the maximal measurement value:

$$\max_{t \in \Omega} \quad \widetilde{MM}(t).$$

For measures where smaller values are desired, we may simply reverse the sign of the objective function. In order to efficiently locate $t^*$, we investigate the shape of $\widetilde{MM}(t)$ for various classification measures by classifying them into three shape categories: monotone, unimodal, and multi-modal with respect to the threshold $t$.

We start by providing a condition for a measurement function to be monotone.

**Proposition 7.** *Suppose $\widetilde{MM} \in \mathcal{M}$. Then $\widetilde{MM}(t)$, $t \in \Omega$, is a monotone function if there exist functions $A(t)$ and $D(t)$ such that*

$$\frac{d\widetilde{MM}(t)}{dt} = A(t)D(t), \ A(t) \geq 0, \ \textit{and } D(t) \geq 0, \ \forall t \in \Omega. \tag{4.9}$$

*Proof.* Clearly, $d\widetilde{MM}(t)/dt \geq 0$ if $A(t) \geq 0$ and $D(t) \geq 0$. Therefore, $\widetilde{MM}(t)$, $t \in \Omega$, is a monotone function. □

Based on the above proposition, we define the following sub-measure class of $\mathcal{M}$ in which all measures are monotone.

**Definition 2.** *$\mathcal{M}_m$ is a sub-class of $\mathcal{M}$ that consists of measures $\widetilde{MM}(t)$ satisfying condition (Eq. 4.9).*

Since all measures in $\mathcal{M}_m$ are monotone, they attain their optimal thresholds at the boundaries $t_{min}$ or $t_{max}$. The next results show that the four simple measures (under any transformation $T$) belong to the monotone class.

**Corollary 4.** *-TPR, -FPR, TNR, FNR* $\in \mathcal{M}_m$.

*Proof.* Note that $d(-TPR(t))/dt = dFNR(t)/dt = f_1(t) > 0$ and $d(-FPR(t))/dt = dTNR(t)/dt = f_0(t) > 0$. The result follows immediately. □

We next provide conditions for a measure to be unimodal within the threshold domain.

**Proposition 8.** *Suppose* $\widetilde{MM} \in \mathcal{M}$. *Then* $\widetilde{MM}(t)$, $t \in \Omega$, *is a unimodal function if there exist functions* $A(t)$ *and* $D(t)$ *such that*

$$\frac{d\widetilde{MM}(t)}{dt} = A(t)D(t),\ A(t) \geq 0,\ \frac{dD(t)}{dt} \leq 0\ \forall t \in \Omega \text{ and } D(t_{min}) \geq 0,\ D(t_{max}) \leq 0. \tag{4.10}$$

*Proof.* Since $dD(t)/dt \leq 0, D(t_{min}) \geq 0,\ D(t_{max}) \leq 0$, $D(t)$ is monotonically decreasing and there exists a $t^*$ such that $D(t^*) = 0$. As a result, $\forall t < t^*$, we have $D(t) \geq 0$ and $\forall t > t^*$, we have $D(t) \leq 0$. Since $A(t) \geq 0$, we have $\forall t < t^*, d\widetilde{MM}(t)/dt \geq 0$ and $\forall t > t^*, d\widetilde{MM}(t)/dt \leq 0$. Therefore, $\widetilde{MM}(t)$ is a unimodal function of $t$. □

Based on the above proposition, we define the following sub-measure class of $\mathcal{M}$, in which all measures are unimodal functions of the threshold.

**Definition 3.** $\mathcal{M}_u$ *is a sub-class of* $\mathcal{M}$ *that consists of all measures* $\widetilde{MM}(t)$ *satisfying condition (Eq. 4.10).*

If $\widetilde{MM} \in \mathcal{M}_u$, then the optimal threshold can be located by setting $d\widetilde{MM}(t)/dt = 0$.

### 4.5.2 Monotone Measures under the Likelihood Transformation

In this subsection, we provide conditions for performance measures under the likelihood transformation (Eq. 3.6) to be monotone. As a result, we show additional single measures, which are not monotone in general, become monotone under the likelihood transformation.

Consider a performance measure under the likelihood transformation: $\widetilde{MM}^L(t)$, $t \in \Omega_L \subseteq R_+$. In view of Definition 2, we consider the following special form of $D(t)$ in (Eq. 4.11):

$$D^L(t) = Z_1(TPR^L(t) - tFPR^L(t)) + Z_2[t(1 - FPR^L(t)) - (1 - TPR^L(t))] + Z_3, \quad (4.11)$$

where $Z_1$, $Z_2$, and $Z_3$ are some fixed scalars.

**Proposition 9.** *If $Z_1 \geq 0$, $Z_2 \geq 0$, and $Z_3 \geq 0$, then $D^L(t) \geq 0$ for all $t \in \Omega_L$.*

*Proof.* For any $t \in \Omega_L$, because of Proposition 1, we have that

$$tFPR^L(t) = t \int_t^{+\infty} g_0^L(s)ds \leq \int_t^{+\infty} s g_0^L(s)ds = \int_t^{+\infty} g_1^L(s)ds = TPR^L(t),$$

where the second equality follows from Proposition 1. Therefore,

$$tFPR^L(t) - TPR(t) \leq 0. \quad (4.12)$$

Using the same approach, we can also show

$$t(1 - FPR^L(t)) - (1 - TPR^L(t)) \geq 0. \quad (4.13)$$

The proposition then follows from the assumption that $Z_1, Z_2, Z_3 \geq 0$. □

The above proposition allows us to define a subclass of monotone measures under the likelihood transformation.

**Definition 4.** *$\mathcal{C}_m$ is a subclass of $\mathcal{M}_m$, where the likelihood transformation is applied on $X$ and the corresponding $D(t)$ (Eq. 4.9) is given by (Eq. 4.11) with some parameters $Z_1, Z_2, Z_3 \geq 0$.*

We prove additional simple measures, such as PPV, -NPV, -FDR, FOR, are monotone under the likelihood transformation.

**Corollary 5.** $PPV^L, -NPV^L, -FDR^L, FOR^L \in \mathcal{C}_m$.

*Proof.* Following Definition 4, we rewrite the derivatives of the above four measures in the following form:

$$\frac{d\widetilde{MM}^L(t)}{dt} = A(t)D^L(t).$$

We summarize their corresponding functions $A(t)$ and parameters of $D^L(t)$ in Table 4.2:

| Measures | $A(t)$ | $D^L(t)$ | | |
|---|---|---|---|---|
| | | $Z_1$ | $Z_2$ | $Z_3$ |
| PPV | $\pi_0\pi_1 g_0^L(t)/(\pi_0 TPR^L(t) + \pi_1 FPR^L(t))^2$ | 1 | 0 | 0 |
| -NPV | $\pi_0\pi_1 g_0^L(t)/(\pi_0(1 - FPR^L(t)) + \pi_1(1 - TPR^L(t)))^2$ | 0 | 1 | 0 |
| -FDR | $\pi_0\pi_1 g_0^L(t)/(\pi_0 TPR^L(t) + \pi_1 FPR^L(t))^2$ | 1 | 0 | 0 |
| FOR | $\pi_0\pi_1 g_0^L(t)/(\pi_0(1 - FPR^L(t)) + \pi_1(1 - TPR^L(t)))^2$ | 0 | 1 | 0 |

Table 4.2: A Class of Monotone Measures

It is straightforward to verify that $A(t) \geq 0$ and $Z_1, Z_2, Z_3 \geq 0$, and thus these measures satisfy Definition 4. □

While simple measures, such as -TPR, -FPR, TNR, FNR, are monotone ($\in \mathcal{M}_m$) under *any* $T$ transformation, according to Proposition 4, the monotone property, however, is not true in general for measures PPV, -NPV, -FDR, FOR, if $T$ is not the likelihood transformation, or a monotone of it. To demonstrate this, we construct a counter example for the PPV measure. Consider class-conditional densities given in Figure 4.1. They are clearly not resulted from the likelihood or monotone of the likelihood transformation, since $g_1(t)/g_0(t)$ is not monotone in $t$. Figure 4.2 clearly shows that the resulting PPV is not monotone.

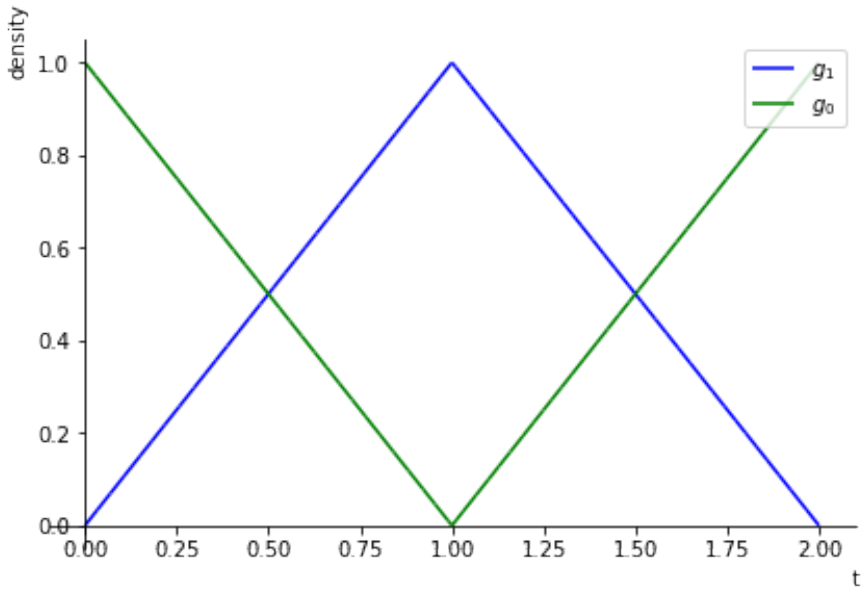


Figure 4.1: Class-conditional Density Distributions

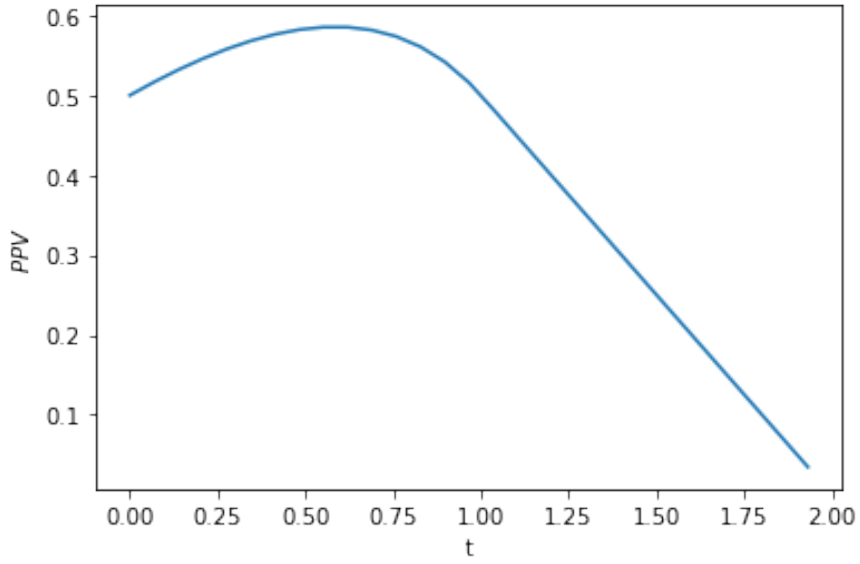


Figure 4.2: PPV is NOT Monotone

### 4.5.3 Unimodal Measures under the Likelihood Transformation

If a performance measure is a unimodal function of the threshold $t$, then we can efficiently locate the optimal threshold that leads to the best performance. Unfortunately, the unimodality property does not hold in general for all existing measures if

a general transformation $T$ is applied to the predicted score $X$. We show in this subsection that many popular balanced classification performance measures are unimodal under the likelihood transformation. Most results we obtained in this subsection are new to the classification measurement literature.

We start by showing two popular measures CC and CP, which are widely used in ROC threshold analysis, belong to $\mathcal{M}_u$ under the likelihood transformation.

**Corollary 6.** $-CC^L, CP^L \in \mathcal{M}_u$.

*Proof.* We first prove the result for measure $-CC^L$. By calculation, we have -$dCC^L(t)/dt = A(t)D^L(t)$, where

$$A(t) = \frac{g_0^L(t)}{\sqrt{(1-TPR^L(t))^2 + (FPR^L(t))^2}}, \text{ and } D^L(t) = -t + tTPR^L(t) + FPR^L(t).$$

Since $A(t) \geq 0$, $D^L(0) = 1$, $D^L(+\infty) < 0$, and

$$\frac{dD^L(t)}{dt} = -1 + TPR^L(t) - tg_1^L(t) - g_0^L(t) \leq 0 \quad \forall t \in \Omega_L,$$

by Definition 3, $-CC^L \in \mathcal{M}_u$.

For measure $CP^L$, By calculation, we have $dCP^L(t)/dt = A(t)D^L(t)$, where $A(t) = g_0^L(t)$ and $D^L(t) = -t + tFPR^L(t) + TPR^L(t)$. Since $A(t) \geq 0$, $D^L(0) = 1$, $D^L(+\infty) < 0$, and

$$\frac{dD^L(t)}{dt} = -1 + FPR^L(t) - tg_0^L(t) - g_1^L(t) \leq 0 \quad \forall t \in \Omega_L, \tag{4.14}$$

by Definition 3, $CP^L \in \mathcal{M}_u$. □

Next we provide conditions for $\widetilde{MM}^L$ to be unimodal when $D(t)$ in Definition 3 takes the special form $D^L(t)$ (Eq. 4.11).

**Lemma 2.** *If $Z_1 \geq 0, Z_2 < 0$, and $Z_3 \geq -Z_1$, then $dD^L(t)/dt \leq 0$ $\forall t \in R_+$ and $D^L(0) \geq 0, D^L(+\infty) \leq 0$ .*

*Proof.* In view of (Eq. 4.11), we have

$$\begin{aligned}\frac{dD^L(t)}{dt} &= Z_1(-g_1^L(t) + tg_0^L(t)) + Z_2(tg_0^L(t) - g_1^L(t)) - Z_1 FPR^L(t) + Z_2(1 - FPR^L(t)) \\ &= -Z_1 FPR^L(t) + Z_2(1 - FPR^L(t)) \\ &\leq 0,\end{aligned}$$

where the second equality follows from Proposition 1 and the inequality follows from the assumption $Z_1 \geq 0$ and $Z_2 < 0$. Since $Z_3 \geq -Z_1$, we have $D^L(0) = Z_1 + Z_3 \geq 0$. Finally, we have $D^L(+\infty) = \lim_{t\to\infty} Z_2(t-1) + Z_3 < 0$, since $Z_2 < 0$. □

Similar to Definition 4 for a subclass of monotone measures, based on Lemma 2, we are able to define a subclass of $\mathcal{M}_u$ when $D^L(t)$ meets the conditions specified in the lemma.

**Definition 5.** *$\mathcal{C}_u$ is a subclass of $\mathcal{M}_u$, where the likelihood transformation is applied on $X$ and the corresponding $D(t)$ (Eq. 4.9) is given by (Eq. 4.11) with some parameters $Z_1 \geq 0, Z_2 < 0$, and $Z_3 \geq -Z_1$.*

It turns out many commonly used measures that can be written as a linear factional functions of $TPR$ and $FPR$, such as $ACC^L, -ERR^L, BA^L, J^L, JI^L, F_\beta^L$, belong to this new class of unimodal measures $\mathcal{C}_u$ under the likelihood transformation. Consider the following measure in the fractional form:

$$\widetilde{MF}^L(t) = MF(TPR^L(t), FPR^L(t)) = \frac{\alpha_1 TPR^L(t) + \beta_1 FPR^L(t) + \gamma_1}{\alpha_2 TPR^L(t) + \beta_2 FPR^L(t) + \gamma_2}, \tag{4.15}$$

where $\alpha_2^2 + \beta_2^2 + \gamma_2^2 \neq 0$. The next result provides conditions for $\widetilde{MF}^L$ to be unimodal.

**Proposition 10.** *$\widetilde{MF}^L \in \mathcal{C}_u$ if its parameters satisfy the following conditions: $\alpha_1\gamma_2 - \alpha_2\gamma_1 > 0, \beta_1\gamma_2 - \beta_2\gamma_1 \leq 0$, and $\alpha_1\beta_2 - \alpha_2\beta_1 \geq max(\beta_1\gamma_2 - \beta_2\gamma_1, \alpha_2\gamma_1 - \alpha_1\gamma_2)$.*

*Proof.* Since $\mathcal{C}_u \subseteq \mathcal{M}$, we first show that $\widetilde{MF}^L \in \mathcal{M}$. Note that

$$\begin{aligned} MF_1^{'}(a,b) &= \frac{\alpha_1(\beta_2 b + \gamma_2) - \alpha_2(\beta_1 b + \gamma_1)}{(\alpha_2 a + \beta_2 b + \gamma_2)^2} \\ &\geq \frac{\min(\alpha_1\beta_2 + \alpha_1\gamma_2 - \alpha_2\beta_1 - \alpha_2\gamma_1, \alpha_1\gamma_2 - \alpha_2\gamma_1)}{(\alpha_2 a + \beta_2 b + \gamma_2)^2} \\ &\geq 0, \end{aligned}$$

where the first inequality holds since $b \in [0,1]$ and the numerator is a linear function of $b$. The second inequality follows from the assumptions that $\alpha_1\gamma_2 - \alpha_2\gamma_1 > 0$ and $\alpha_1\beta_2 - \alpha_2\beta_1 \geq \alpha_2\gamma_1 - \alpha_1\gamma_2$. Using similar logic, next we show that $MF_2^{'}(a,b) \leq 0$. Note that

$$\begin{aligned} MF_2^{'}(a,b) &= \frac{\beta_1(\alpha_2 a + \gamma_2) - \beta_2(\alpha_1 a + \gamma_1)}{(\alpha_2 a + \beta_2 b + \gamma_2)^2} \\ &\leq \frac{\max(\alpha_2\beta_1 + \beta_1\gamma_2 - \alpha_1\beta_2 - \beta_2\gamma_1, \beta_1\gamma_2 - \beta_2\gamma_1)}{(\alpha_2 a + \beta_2 b + \gamma_2)^2} \\ &\leq 0, \end{aligned}$$

where the first inequality holds since $a \in [0,1]$ and the numerator is a linear function of $a$. The second inequality follows from the assumptions that $\beta_1\gamma_2 - \beta_2\gamma_1 \leq 0$ and $\alpha_1\beta_2 - \alpha_2\beta_1 \geq \beta_1\gamma_2 - \beta_2\gamma_1$.

By Definition 1, $\widetilde{MF}^L \in \mathcal{M}$. Next, we prove $\widetilde{MF}^L \in \mathcal{C}_u$. By simple calculation, we have $d\widetilde{MF}^L(t)/dt = A(t)D^L(t)$, where

$$\begin{aligned} A(t) &= \frac{g_0^L(t)}{(\alpha_2 TPR^L(t) + \beta_2 FPR^L(t) + \gamma_2)^2} \\ D^L(t) &= -(\alpha_1\beta_2 - \alpha_2\beta_1)(tFPR^L(t) - TPR^L(t)) - t(\alpha_1\gamma_2 - \alpha_2\gamma_1) - (\beta_1\gamma_2 - \beta_2\gamma_1). \end{aligned}$$

Note that $D^L(t)$ can be re-written as:

$$D^L(t) = Z_1(TPR^L(t) - tFPR^L(t)) + Z_2[t(1 - FPR^L(t)) - (1 - TPR^L(t))] + Z_3,$$

where

$$Z_1 = (\alpha_1\beta_2 - \alpha_2\beta_1) + (\alpha_1\gamma_2 - \alpha_2\gamma_1)$$
$$Z_2 = -(\alpha_1\gamma_2 - \alpha_2\gamma_1)$$
$$Z_3 = -(\alpha_1\gamma_2 - \alpha_2\gamma_1) - (\beta_1\gamma_2 - \beta_2\gamma_1)$$

By assumptions, we have $Z_1 \geq 0, Z_2 < 0$ and $Z_3 \geq -Z_1$. By Definition 5, $\widetilde{MF}^L \in \mathcal{C}_u$ and the proposition is proven. □

**Corollary 7.** $ACC^L, -ERR^L, BA^L, J^L, JI^L, F_\beta^L \in \mathcal{C}_u$.

*Proof.* Note that these measures can be written in the form of (Eq. 4.15) with the parameters summarized in the following table, respectively:

| Measures | $\alpha_1$ | $\beta_1$ | $\gamma_1$ | $\alpha_2$ | $\beta_2$ | $\gamma_2$ |
|---|---|---|---|---|---|---|
| $ACC^L$ | $\pi_1$ | $-\pi_0$ | $\pi_0$ | 0 | 0 | 1 |
| $-ERR^L$ | $-\pi_1$ | $\pi_0$ | $\pi_1$ | 0 | 0 | 1 |
| $BA^L$ | c | -(1-c) | 1-c | 0 | 0 | 1 |
| $J^L$ | 1 | -1 | 0 | 0 | 0 | 1 |
| $F_\beta^L$ | $(1+\beta^2)\pi_1$ | 0 | 0 | $\pi_1$ | $\pi_0$ | $\beta^2\pi_1$ |
| $JI^L$ | $\pi_1$ | 0 | 0 | 0 | $\pi_0$ | 1 |

Table 4.3: Measures in Format (Eq. 4.15)

It is straightforward to verify that the parameters for each measure meet the condition of Proposition 10. □

Note that for measures CC, CP, ACC, -ERR, BA, J, $F_\beta$, JI, unimodality property holds under the likelihood transformation of predicted score $X$. The property, however, is not true in general for other $T$ transformations. Take accuracy for an example; Figure 4.3 represents the accuracy plot based on the class-conditional densities shown in Figure 4.1. Clearly, the ACC measure is not unimodal in this case.

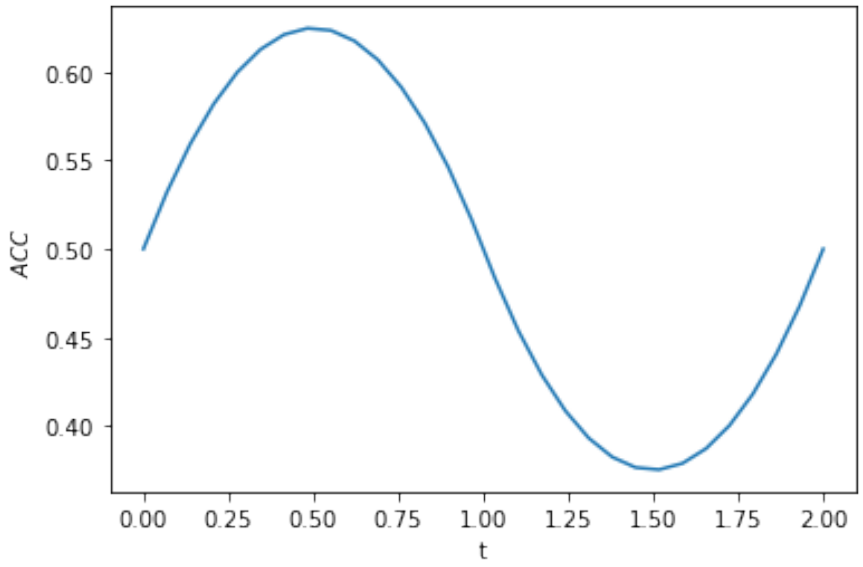


Figure 4.3: ACC is NOT Unimodal

Finally, we show how to calculate the optimal threshold for a unimodal measure under the likelihood transformation $\widetilde{MM}^L(t) \in \mathcal{C}_u$. Since $\widetilde{MM}^L(t)$ is unimodal, the optimal threshold can be obtained by setting its derivative to zero:

$$\begin{aligned}\frac{d\widetilde{MM}^L(t)}{dt} &= -MM_1^{'}(TPR^L(t), FPR^L(t))g_1^L(t) - MM_2^{'}(TPR^L(t), FPR^L(t))g_0^L(t) \\ &= g_0^L(t)(-tMM_1^{'}(TPR^L(t), FPR^L(t)) - MM_2^{'}(TPR^L(t), FPR^L(t))) = 0.\end{aligned}$$

It follows that the optimal threshold $t^L$ satisfies the following equation:

$$t^L = -\frac{MM_2^{'}(TPR^L(t^L), FPR^L(t^L))}{MM_1^{'}(TPR^L(t^L), FPR^L(t^L))}. \tag{4.16}$$

We summarize the optimal thresholds for all existing performance measures in Table 4.4.

### 4.5.4 Multi-Modal Measures

In subsections 4.5.2 - 4.5.3, we reveal that most currently used confusion-matrix measures are either monotone or unimodal functions of the threshold under the likelihood transformation. In this subsection, we show that two measures, Markedness and MCC, are exceptions even under the likelihood transformation. In fact, they are multi-modal in general w.r.t. the threshold, in which case an exhaustive search is needed to find the optimal threshold to achieve maximum value of the respective measurements.

Markedness (MK) is the arithmetic mean of positive predictive value and negative predictive value, which measures the reliability of the respective predictions by the model. It is defined as follows:

$$MK^L(t) = \frac{\pi_1 TPR^L(t)}{\pi_1 TPR^L(t) + \pi_0 FPR^L(t)} + \frac{\pi_0(1 - FPR^L(t))}{\pi_1(1 - TPR^L(t)) + \pi_0(1 - FPR^L(t))} - 1.$$

Clearly, MK ranges from -1 to +1, and the higher the value, the more reliable prediction. We now provide an example to show that MK is multi-modal under the likelihood transformation. Consider the class-conditional densities $g_1(t)$, $g_0(t)$ depicted in Figure 4.4 and $\pi_1 = 0.25$. It can be verified that ratio $g_1(t)/g_0(t)$ monotonically increases in $t$.

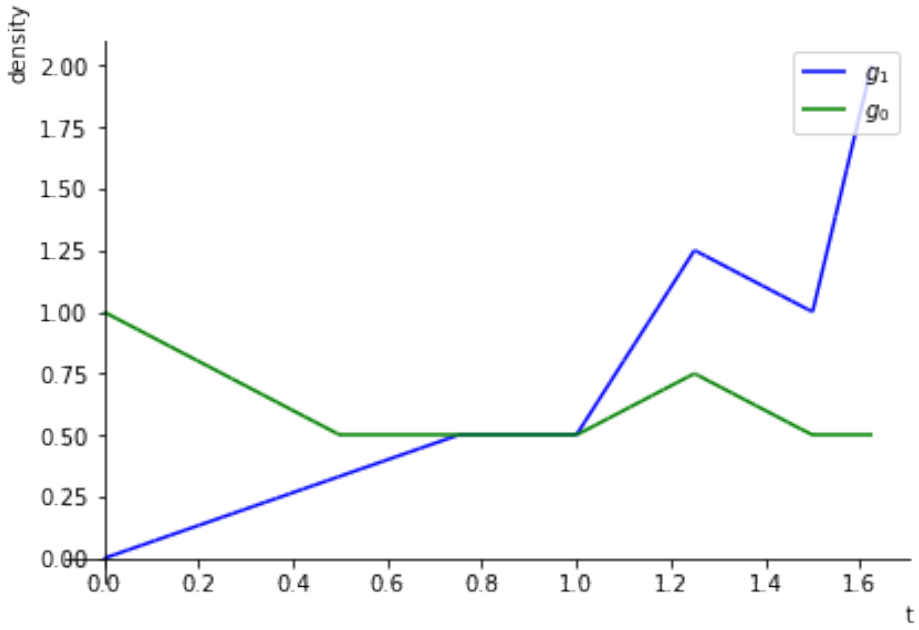


Figure 4.4: $g_0(t)$ and $g_1(t)$

We plotted the corresponding MK in Figure 4.5, which is clearly multi-modal in $t$. Since $g_1(t)/g_0(t)$ is monotone in $t$, by Corollary 2 in Section 4.3, the corresponding $MK^L$ is also multi-modal.

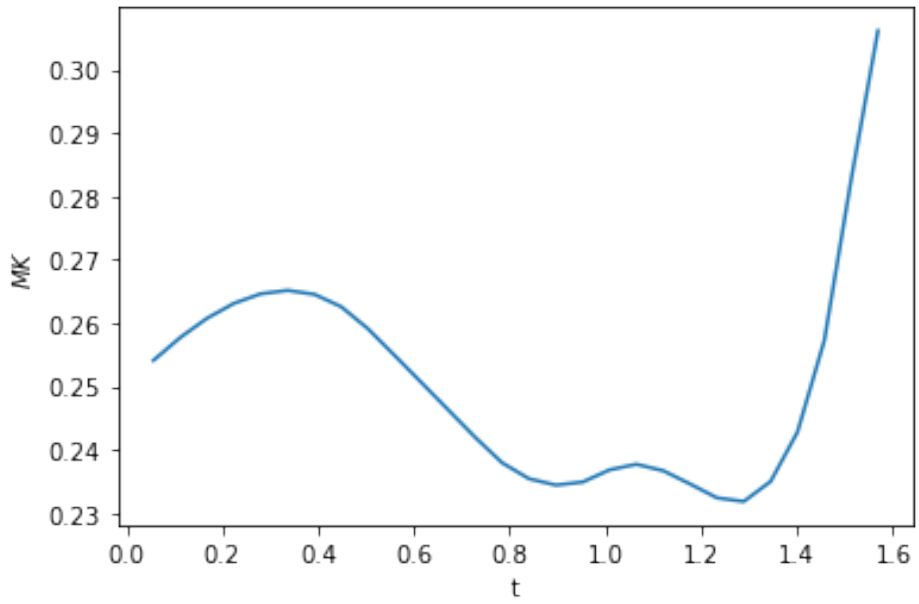


Figure 4.5: Markedness is Multi-Modal

Next we show that Matthews Correlation Coefficient (MCC) is in general multi-modal w.r.t. the threshold even after the likelihood transformation is applied to the predicted score $X$. After some algebraic manipulations (see Appendix A for details), we may rewrite MCC as follows:

$$MCC^L(t) = \sqrt{\frac{\pi_0\pi_1}{\omega(t)(1-\omega(t))}}(TPR^L(t) - FPR^L(t)),$$

where $\omega(t) = \pi_1 TPR^L(t) + \pi_0 FPR^L(t)$. Since the square operation preserves the modality properties of a non-negative function, we focus on the shape of $MCC^{L2}(t)$ instead. With additional extensive algebraic manipulations (in Appendix A), we obtain:

$$\frac{dMCC^{L2}(t)}{dt} = \frac{\pi_0\pi_1(TPR^L(t) - FPR^L(t))g_0^L(t)}{\omega(t)(1-\omega(t))} * \left[\frac{1 - TPR^L(t) - t(1 - FPR^L(t))}{\pi_1(1 - TPR^L(t)) + \pi_0(1 - FPR^L(t))} + \frac{TPR^L(t) - tFPR^L(t)}{\pi_1 TPR^L(t) + \pi_0 FPR^L(t)}\right]. \tag{4.17}$$

After investigating (Eq. 4.17), we construct the two class-conditional density functions $g_0(t)$ and $g_1(t)$ illustrated in Figure 4.6 (the Python code to construct the density functions is described in Appendix B).

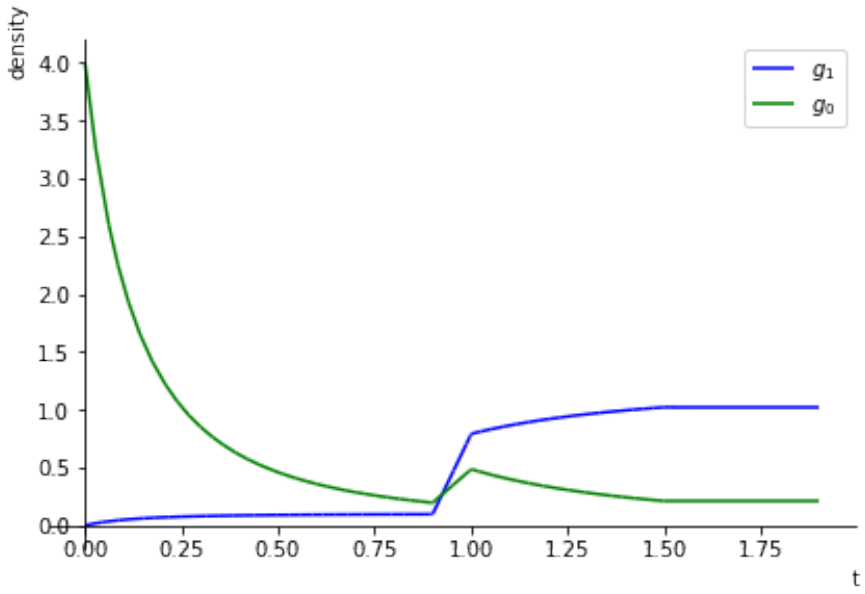


Figure 4.6: $g_0(t)$ and $g_1(t)$

Figure 4.7 shows that $g_1(t)/g_0(t)$ is a monotone increasing function of threshold $t$. We then calculate $MCC^{L2}(t)$ and plot it in Figure 4.8, which clearly indicates that $MCC$ is a multi-modal function. Since $g_1(t)/g_0(t)$ is a monotone function, by Corollary 2 discussed in Section 4.3, we show that $MCC^{L2}(t)$, and thus $MCC^{L}(t)$ is also in general a multi-modal function under the likelihood transformation.

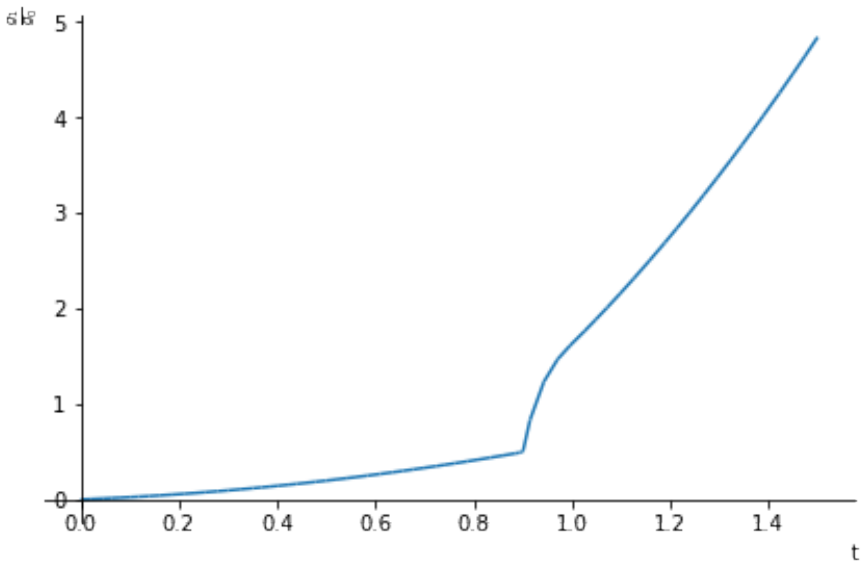


Figure 4.7: Likelihood Ratio $g_1(t)/g_0(t)$ is Monotone in $t$

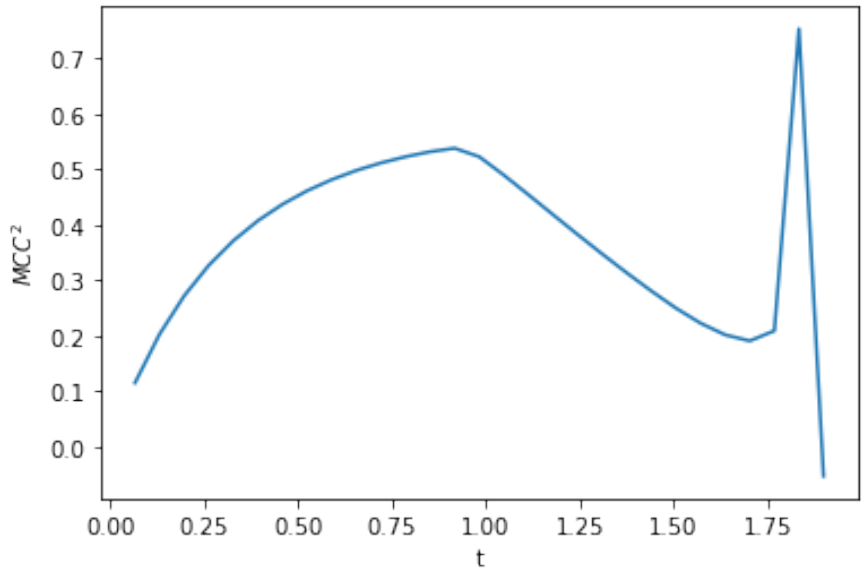


Figure 4.8: $MCC^2(t)$ vs $t$

## 4.6 Impact of Base Rate $\pi_1$

While measures such as $TPR(t)$, $FNR(t)$, $TNR(t)$, and $FPR(t)$ depend on only the threshold $t$, many other measures are also impacted by the base rate $\pi_1$, in addition to the threshold. For example, given a fixed threshold (t), $PPV(t, \pi_1)$

and $FOR(t, \pi_1)$ are monotonically increasing in base rate $\pi_1$, while $FDR(t, \pi_1)$ and $NPV(t, \pi_1)$ are monotonically decreasing in $\pi_1$. For $ACC(t, \pi_1)$ and $ERR(t, \pi_1)$, if $TPR(t) \geq TNR(t)$, $ACC(t, \pi_1)$ ($ERR(t, \pi_1)$) is monotonically increasing (decreasing) in base rate $\pi_1$; otherwise, $ACC(t, \pi_1)$ ($ERR(t, \pi_1)$) is monotonically decreasing (increasing) in $\pi_1$. For balanced measures, it is easy to verify that ACC and Youden index are independent of the base rate, while $F1$ measure is monotonically increasing in base rate as demonstrated below:

$$\frac{\partial F1(t, \pi_1)}{\partial \pi_1} = \frac{2 \int_t^{+\infty} f_1(x)dx \int_t^{+\infty} f_0(x)dx}{(2\pi_1 \int_t^{+\infty} f_1(x)dx + \pi_0 \int_t^{+\infty} f_0(x)dx + \pi_1 \int_{-\infty}^{t} f_1(x)dx)^2} > 0.$$

For completeness, we summarize the relationship between the measurements and the base rate for all existing measures in Table 4.4. We should be mindful when utilizing base-rate related measure to compare model performance across datasets with varying class priors. Unlike those performance measures (e.g., TPR, FPR) that are independent of the prior distributions, base-rate related measures may vary significantly on different data sets. For example, models with the same TPR and FPR may lead to varying $F1$ scores for different datasets. In this case, relying on $F1$ alone may draw a misleading conclusion (Gong, 2021).

## 4.7 Summary and Discussion

Confusion-matrix based performance measures have been widely adopted in machine learning classification model development, selection, and evaluation (Ferri et al., 2009). Little attention, however, has been paid to their theoretical properties. In this chapter, we propose a *unified framework* for all threshold-dependent classification measures with the following nice structural properties:

1. All measures are monotonically increasing in TPR and monotonically decreasing in FPR.

2. All measures are invariant under monotonic transformation. More specifically, the unimodality and maximum values of the measurements can be preserved.

3. The likelihood transformation results in the best optimal values of the measurements.

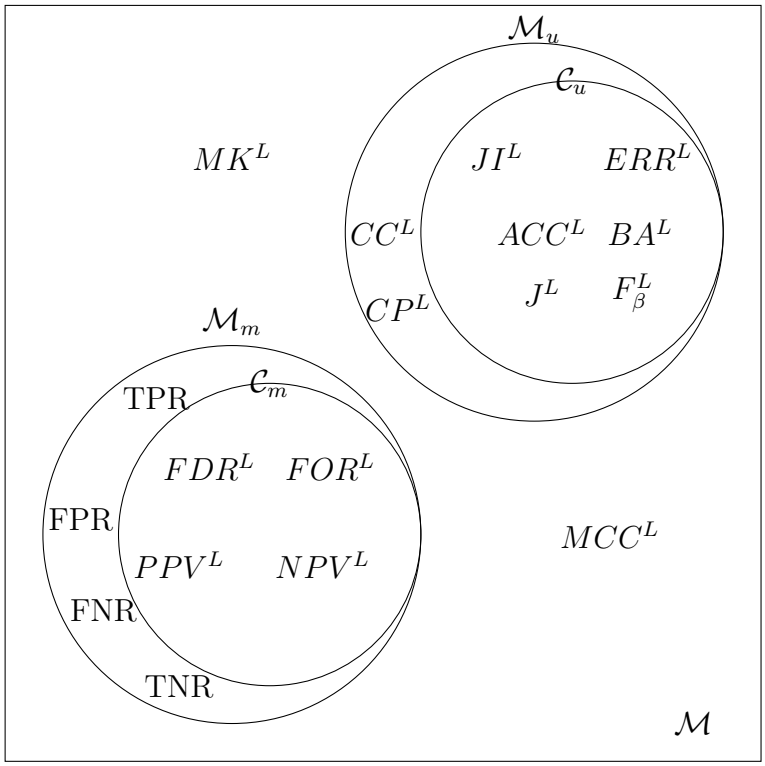


Figure 4.9: Unified Framework for Classification Measures

We further investigate the properties of the measures in the context of the likelihood transformation. We systematically reveal that classification measures composed in certain mathematical form achieve maximal values monotonically or unimodally. For those multi-modal measures, we provide examples. Our unified framework and resulting measure classifications are cleanly summarized in Figure 4.9. We also systematically disclose the relationships between the optimal values of various measurements and the corresponding decision-making thresholds. The detailed results are summarized in Table 4.4.

| $MM$ | Threshold (L) | Threshold (B) | t | $\pi_1$ |
|---|---|---|---|---|
| $TPR$ | 0 | 0 | $\searrow$ | – |
| $FNR$ | 0 | 0 | $\nearrow$ | – |

| | | | | |
|---|---|---|---|---|
| $TNR$ | $+\infty$ | 1 | $\nearrow$ | – |
| $FPR$ | $+\infty$ | 1 | $\searrow$ | – |
| $PPV^L$ | $+\infty$ | 1 | $\nearrow$ | $\nearrow$ |
| $FDR^L$ | $+\infty$ | 1 | $\searrow$ | $\searrow$ |
| $NPV^L$ | 0 | 0 | $\searrow$ | $\searrow$ |
| $FOR^L$ | 0 | 0 | $\nearrow$ | $\nearrow$ |
| $ACC^L$ | $\pi_0/\pi_1$ | 1/2 | U | $\nearrow, \searrow$ |
| $ERR^L$ | $\pi_0/\pi_1$ | 1/2 | U | $\searrow, \nearrow$ |
| $BA^L$ | $(1-c)/c$ | $(1-c)\pi_1/((1-c)\pi_1 + c\pi_0)$ | U | – |
| $F1^L$ | $\pi_0 JI(t_f^L)/\pi_1$ | $F1(t_f^B)/2$ | U | $\nearrow$ |
| $J^L$ | 1 | $\pi_1$ | U | – |
| $CC^L$ | $FPR^L(t_{cc}^L)/$ $(1-TPR^L(t_{cc}^L))$ | $FPR^B(t_{cc}^B)\pi_1/$ $[FPR^B(t_{cc}^B)\pi_1 + (1-TPR^B(t_{cc}^B))\pi_0]$ | U | – |
| $CP^L$ | $TPR^L(t_{cp}^L)/$ $(1-FPR^L(t_{cp}^L))$ | $TPR^B(t_{cp}^B)\pi_1/$ $[TPR^B(t_{cp}^B)\pi_1 + (1-FPR^B(t_{cp}^B))\pi_0]$ | U | – |

Table 4.4: Summary of Binary Confusion-Matrix Measure w.r.t. Thresholds and Base Rate. Note: L and B represent likelihood and Bayesian transformations, respectively. $\nearrow$ ($\searrow$) denotes monotonically increasing (decreasing). U means unimodal function. The monotone relationship for $ACC, ERR$ w.r.t. $\pi_1$ depends on the condition $TPR(t) > TNR(t)$.

In addition, we note that some measures are base rate related and others are not. Those base-rate related measures are monotone when certain conditions are met, which implies that when dealing with imbalanced data, where $\pi_0 \gg \pi_1$, the optimal decision to maximize accuracy is to predict all instances to be negatives ($t_{acc}^L = +\infty$). The implication is aligned with the criticism of using accuracy to measure model performance in imbalanced data, where the classification will be biased towards majority class. Our study and results shed lights on the following aspects in model development and evaluation:

- Our research provides theoretical support for choosing appropriate measures in

model development and evaluation, especially when dealing with imbalanced data;

- Our research presents a directional guidance on determining the threshold for optimizing corresponding measures, which can help significantly improve the computation efficiency;
- By investigating the properties of existing performance measures, we provide directions to define additional performance measures to optimize.

# Chapter 5

# ROC CURVE CHARACTERIZATION AND ENHANCEMENT

Receiver Operating Characteristic (ROC) curve was firstly used in World War II to discern between signals and noises by the military radar operators. Since then, ROC curve has been extended to various fields such as psychophysics (Green et al., 1966; Markowitz and Swets, 1967; Nevin, 1969), medicine (Lusted, 1971; Beck and Shultz, 1986; Zweig and Campbell, 1993), radiology (Goodenough et al., 1974; Hanley and McNeil, 1982; Van Erkel and Peter, 1998), epidemiology (Erdreich and Lee, 1981; Hanley and McNeil, 1982; Lijmer et al., 1996), biometrics (DeLong et al., 1988; Wayman, 1996; Obuchowski, 1997), etc.

More recently, ROC curve is increasingly used as a visualization tool to measure the discrimination ability of a machine learning model in classification problems (Bradley, 1997; Fawcett, 2006; Majnik and Bosnić, 2013; Narkhede, 2018; Wang and Ni, 2019; Kannan and Vasanthi, 2019; Omar and Ivrissimtzis, 2019; Janssens and Martens, 2020; Gong, 2021). It plays a very important role in machine learning model development, selection, and evaluation (Rosset, 2004b; Alemayehu and Zou, 2012; Majnik and Bosnić, 2013; Gong, 2021). The ROC curve represents a trade-off between true positive rate (TPR) and false positive rate (FPR). ROC demonstrates several advantages over confusion-matrix based metrics (Metz, 1978; Majnik and Bosnić, 2013; Gong, 2021): 1. ROC does not depend on threshold selection (Janssens and Martens, 2020); 2. The change of data prior distribution will not impact the ROC curve, since TPR and FNR are independent of base rate; 3. ROC analysis does not necessarily require exact probability, and what it matters is to separate positive instances from negative instances.

ROC and its variants have been extensively studied with numerous published papers in machine learning area,. For example, Drummond and Holte (2000) explicitly represented the expected cost of a classifier in cost curve as an alternative to ROC representation, which maintains many of the advantages of ROC and can visually tell the range of costs and class prior distribution where a particular classifier is the best. Flach (2004) utilized the ROC convex hull method to select among categorical classifiers and determines the optimal decision threshold for a ranking classifier, as well as regions where one classifier dominates another. Fawcett (2006) provided a detailed discussion of ROC space and pointed out that one attractive property of ROC curve is that they are independent of the prior class distribution. Swamidass et al. (2010) developed the general concentrated ROC (CROC) framework to address the early retrieval problem, where only the very top of the ranked list of predictions is of any interest. Robin et al. (2011) designed a tool called 'pROC' to facilitate ROC curve analysis and apply proper statistical tests for their comparison. Davis and Goadrich (2006); Su et al. (2015) found out that ROC curve is closely related to or can be transformed into precision-recall curve, lift chart, Kolmogorov–Smirnov (K-S) alike curves, etc. Fieldsend and Everson (2005); Landgrebe and Duin (2006); Deng et al. (2006); Hua and Tian (2020) extended ROC to multi-class by designing a three-dimensional measure called volume under ROC space, or via one vs. one (OVO) & one vs. all (OVA) methodologies.

All the above-mentioned studies implicitly assume concavity of the ROC curve by default. In many cases, non-concavity ROC is inevitable and often ignored. There are very few papers discussing non-concave ROC. Hilden (1991) pointed out the problem in the medical literature that "some authors do seem to overlook the concavity problem". Pesce et al. (2010) argued that direct use of a decision model "with a non-concave ROC curve must be considered irrational and unethical when applied to medical decisions". Vardhan et al. (2020) recommended not to use the classifier if the ROC curve is not concave. These research studies focus on empirical studies but fail to discuss how concavity can be addressed and how to transform the non-concave ROC

into concave ROC without re-training the model from a theoretical perspective. In addition, most papers treat ROC as a tool to measure model performance on analyzing empirical data. The properties and characterization of ROC curve is not well understood. Recently Medlock and Oppenheim (2019) proved that Neyman-Pearson ROC curve with the likelihood ratio transformation is concave. However, the paper lacks a systematical discussion of the characterization of ROC curves and fails to disclose the properties of the transformed ROC. Our study aims to bridge the gap by systematically discussing the properties and characterization of ROC curves, addressing the non-concave issue, and identifying nice properties of the transformed ROC, which lay a solid foundation for our following research topics.

This chapter is structured as follows: Section 5.1 briefly reviews ROC curve in the context of class conditional densities; Section 5.2 highlights the cases where ROC curve is non-concave and discloses the sufficiency and necessity conditions for a ROC curve to be concave; Section 5.3 is our main contribution, which discusses the properties and characterization of the ROC curve, proposes a systematical transformation to enhance ROC curve, highlights the nice properties of the transformed scores, and performs simulation studies to convert a non-concave ROC into a concave ROC. In addition, we explore the dominance properties of ROC curve and other visualization tools; Section 5.4 discusses and summarizes our key findings and contributions.

## 5.1 Introduction of ROC Curve

According to Table 4.1, recall the definitions of TPR and FPR:

$$TPR(t) = P(\hat{Y} = 1|Y = 1) = P(X \geq t|Y = 1) = \int_t^{+\infty} f_1(x)dx = 1 - F_1(t) \quad (5.1)$$

$$FPR(t) = P(\hat{Y} = 1|Y = 0) = P(X \geq t|Y = 0) = \int_t^{+\infty} f_0(x)dx = 1 - F_0(t) \quad (5.2)$$

ROC curve is defined as a parametric curve on the threshold $t$, where the vertical axis is the true positive rate $TPR(t)$ and the horizontal axis is the false positive rate $FPR(t)$:

$$ROC(f_0, f_1) = \{(FPR(t), TPR(t)), t \in (-\infty, +\infty)\} \tag{5.3}$$

A typical ROC curve is shown in Figure 5.1. By definitions of $TPR(t)$ and $FPR(t)$ (Table 4.1), ROC curve depends on only the conditional score density functions $f_0(x)$ and $f_1(x)$. Indeed, the slope of the ROC curve at threshold $t$ is given by

$$\frac{d\ TPR(t)}{d\ FPR(t)} = \frac{TPR'(t)}{FPR'(t)} = \frac{f_1(t)}{f_0(t)}, \tag{5.4}$$

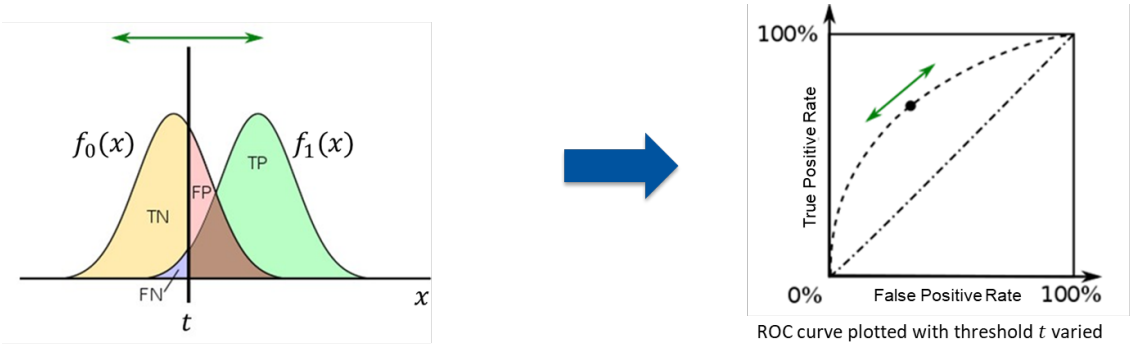


Figure 5.1: ROC Curve Based on Density Distributions with Varying $t$

In case the two conditional density functions are identical, i.e. $f_1(x) = f_0(x)\ \forall x \in R$, the ROC degenerates to a diagonal line (also known as 'chance line'), which implies that the classification algorithm can not statistically separate the class label of an instance. In general, a good classification algorithm should have its ROC located above the diagonal (Gong, 2021).

To measure the performance of a classifier and aid the selection of classifiers, a single summary metric AUC (area under the curve), is derived from the associated ROC. Intuitively, an AUC close to 0.5 indicates random guessing, while AUC=1 represents perfect prediction. It has been shown (Marzban, 2004), however, these intuition is precise only if the associated ROC is concave. In particular, for a concave ROC, its AUC is shown (Fawcett, 2006) to be equal to the probability that the classifier

will rank a randomly chosen positive instance higher than a randomly chosen negative instance:

$$AUC = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathbb{1}(x_1, x_0) f_1(x_1) f_0(x_0) dx_1 \; dx_0 = P(X_1 > X_0),$$

where the indicator function $\mathbb{1}(x_1, x_0) = 1$ if $x_1 > x_0$, and 0 otherwise. Random variable $X_1 = \{X|Y = 1\}$ is the conditional predictive score for a positive instance and $X_0 = \{X|Y = 0\}$ is the conditional predictive score for a negative instance. Since all these results rely critically on the concavity of the ROC curve, we discuss the associated conditions in the next section.

## 5.2 Concavity of ROC Curve

The ROC curve constructed based on (Eq. 5.3) can be concave as shown in Figure 5.1, where the predictive score distributions of the two classes have the same variance but different means. It can also be non-concave as demonstrated by four scenarios in Figures 5.2, where the two distributions have different variances or one of the distributions is bimodal. These scenarios are commonly reported (Parodi et al., 2008; Silva-Fortes et al., 2012; Martinez-Camblor et al., 2017; Vardhan et al., 2020) in medical research papers and carry useful scientific insights. For example, Martinez-Camblor et al. (2017) pointed out that it may not be true that the higher (lower) scores generated by a machine learning model are always associated with a higher probability of a positive instance. There also exist real situations where both lower and higher scores are associated with a higher probability of positive instances (Scenario 2 in Figure 5.2). The associated non-concave ROC curves, however, lead to all AUCs close to 0.5, which renders usefulness of these classification models according to the mostly used model selection metric (AUC).

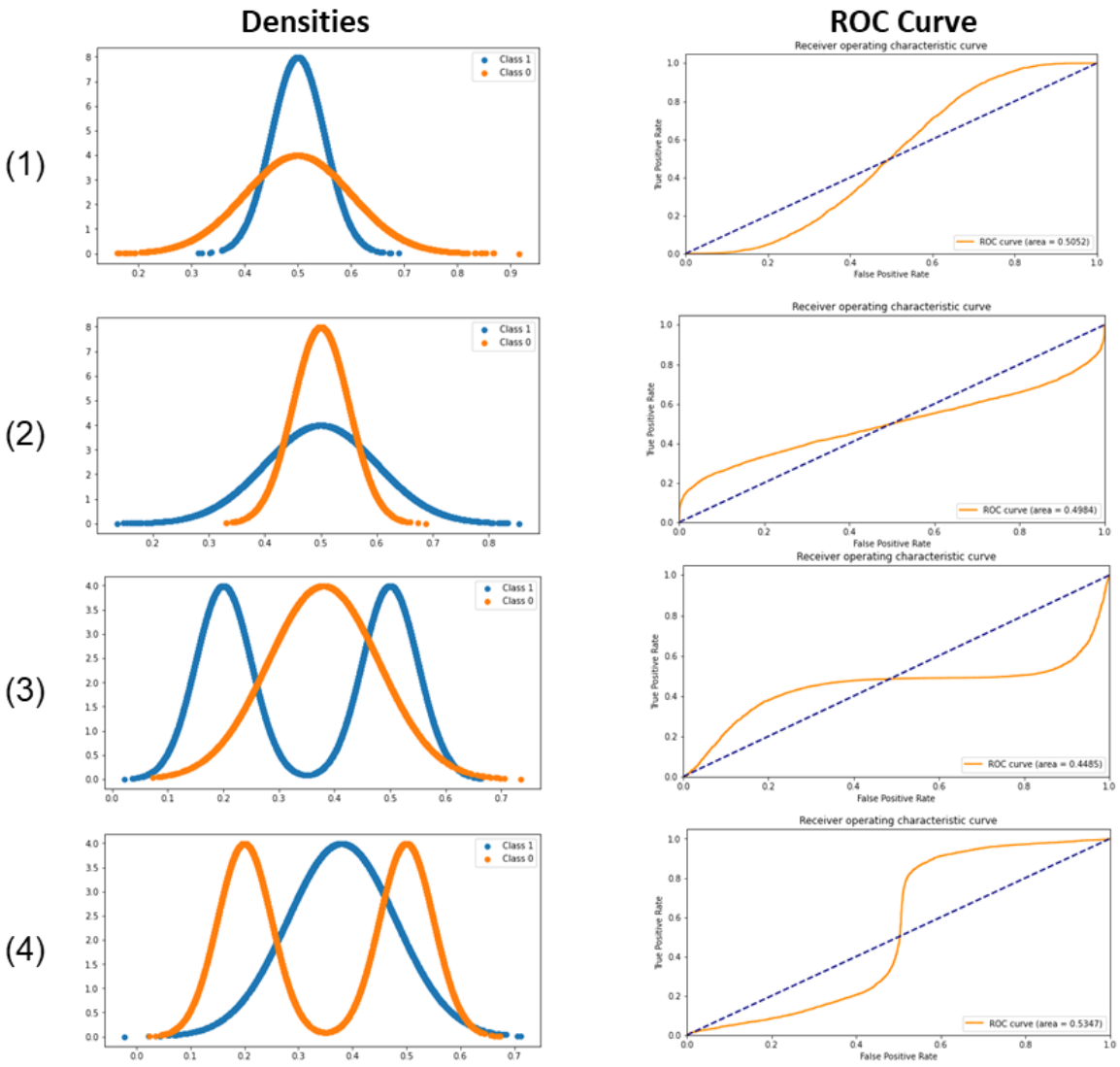


Figure 5.2: Scenarios of Density Distributions and Their Associated Non-Concave ROCs

We will now explore the condition that guarantees the concavity of the ROC curve and develop a process to construct a concave ROC in the next subsection. It turns out that the concavity of ROC curve is closely related to the likelihood ratio between the two conditional score distributions $f_0(x)$ and $f_1(x)$ defined as in (Eq. 3.6).

**Lemma 3.** *Suppose score $x \in \Omega \subseteq R$ is a scalar. Then $ROC(f_0, f_1)$ is concave if and only if $L(x)$ monotonically increases in $x$.*

*Proof.* Sufficiency: if $L(x)$ monotonically increases in $x$, then $ROC(f_0, f_1)$ is concave. Since $FPR(x)$ is monotonically decreasing in $x$ (according to Lemma 4) and $L(x)$ monotonically increases in $x$, we have that $L(x)$ is monotonically decreasing with regards to $FPR$, which means the slope of the ROC curve (Eq. 5.4) monotonically decreases with regards to $FPR$ (the $x$-axis of ROC curve). Therefore, the ROC curve is concave. The necessity condition can be proven by similar reverse logic. □

With this result, one can easily verify if a classification algorithm has a concave ROC curve by simply plotting the likelihood ratio $L(x)$. Figure 5.3 below demonstrates Lemma 3 for the four scenarios in Figure 5.2. Clearly, none of the likelihood functions are monotonically increasing in $x$.

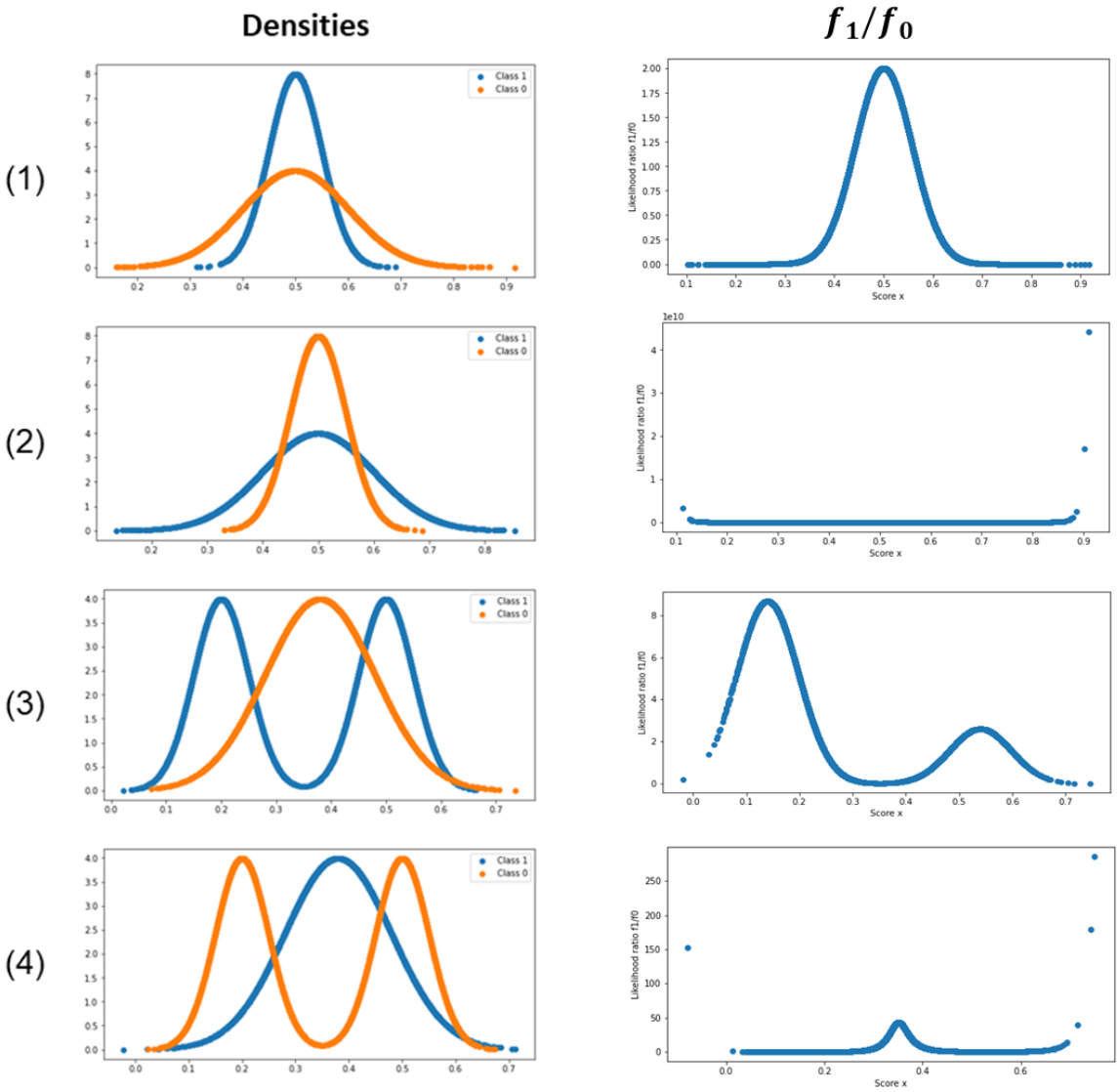


Figure 5.3: Scenarios of Density Distributions and Their Likelihood Ratio $f_1/f_0$

Various ad-hoc empirical methodologies are proposed to fix the non-concave ROC issues. Instead of using one varying threshold, Martinez-Camblor et al. (2017) proposed an empirical ROC to address scenario 2 by taking into account all pairs of possible thresholds, one for the lower scores and another for the higher scores. Those modifications, however, are ad-hoc in nature, since they are applicable for only specific situations. Furthermore, extant ROC literature (Medlock and Oppenheim, 2019) requires the predicted score $X$ to be a scalar, which can be compared and ordered.

In most classification applications, however, $X$ is usually a multi-dimensional vector, such as customer features. As a result, the ROC curve cannot be readily constructed using (Eqs. 5.1 and 5.2).

## 5.3 Properties of ROC Curve Based on the Likelihood Transformation

In this section, we explore the properties of ROC and its variants based on the likelihood transformation.

### 5.3.1 ROC curve under the Likelihood Transformation

Suppose score $X$ is a scalar random variable. We first prove the invariant properties of a *ROC* curve with regards to a monotone increasing transformation of $X$.

**Lemma 4.** *Suppose $X$ is a scalar random variable and transformation $T(\cdot)$ in (Eq. 3.1) is a monotonically increasing function, then $ROC(g_0^T, g_1^T) = ROC(f_0, f_1)$.*

*Proof.* Let $(FPR^T(t_g), TPR^T(t_g))$ be a point on $ROC(g_0^T, g_1^T)$, $(FPR(t_f), TPR(t_f))$ be a point on $ROC(f_0, f_1)$. Given

$$s = T(x) \qquad t_g = T(t_f),$$

since $T(x)$ is a monotone increasing function with regards to $x$, $x \geq t_f$ implies $T(x) \geq T(t_f)$, and vice versa. Therefore,

$$FPR(t_f) = P(x \geq t_f | Y = 0) = P(T(x) \geq T(t_f) | Y = 0) = P(s \geq t_g | Y = 0) = FPR^T(t_g)$$

and

$$TPR(t) = P(x \geq t_f | Y = 1) = P(T(x) \geq T(t_f) | Y = 1) = P(s \geq t_g | Y = 1) = TPR^T(t_g),$$

and hence $ROC(g_0^T, g_1^T)$ is identical to $ROC(f_0, f_1)$. □

As a result of this lemma, the ROC curve generated from the likelihood transformation $ROC(g_0^L, g_1^L)$ is identical to the ROC curve based on the Bayesian transformation $ROC(g_0^B, g_1^B)$ and both of them are concave.

**Proposition 11.** *$ROC(g_0^L, g_1^L)$ and $ROC(g_0^B, g_1^B)$ are concave and $ROC(g_0^L, g_1^L) = ROC(g_0^B, g_1^B)$.*

The proposition can be clearly proven by Proposition 1 and Lemma 3 & 4. Based on discussion in Subsection 3.1, different variable transformations may lead to different ROC curves. We are interested in finding out which transformation will lead to the most discriminate ROC curve. To this end, we first introduce the concept of dominance for a pair of ROC curves. Roughly speaking, ROC curve A dominates ROC curve B if curve B is beneath curve A. The formal definition is given below.

**Definition 6.** *Let $f_0(\cdot)$, $f_1(\cdot)$ and $g_0(\cdot)$, $g_1(\cdot)$ be the respective class-conditional density functions associated with two classification algorithms. Then $ROC(g_0, g_1)$ dominates $ROC(f_0, f_1)$, if for any threshold $t_g, t_f \in \Omega$ such that $FPR_g(t_g) = FPR_f(t_f)$, $TPR_g(t_g) \geq TPR_f(t_f)$.*

**Theorem 1.** *Let $L(\cdot)$ be the likelihood transformation (3.6) and $T(\cdot)$ be any transformation on $X$. Then $ROC(g_0^L, g_1^L)$ dominates $ROC(g_0^T, g_1^T)$.*

*Proof.* With the set notation in (Eqs. 3.8 and 3.9), TPR and FPR can be re-expressed as $P(R|Y = 1)$ and $P(R|Y = 0)$, respectively, where $R$ represents the set of $x$ where the associated instances are classified as positive ($R_L$ or $R_T$). Therefore, according to Corollary 1, for an thresholds $t_L$ and $t_T$ such that $FPR^L(t_L) = FPR^T(t_T)$, we have $TPR^L(t_L) \geq TPR^T(t_T)$. Therefore, $ROC(g_0^L, g_1^L)$ dominates $ROC(g_0^T, g_1^T)$ by definition. □

### 5.3.2 Other Classification Measurement Curves under the Likelihood Transformation

Precision-recall curve (PR) is another popular visualization tool for model evaluation (Davis and Goadrich, 2006; Boyd et al., 2013; Ozenne et al., 2015), where the

$x$-axis is TPR (recall) and the $y$-axis is PPV (precision). We can extend Corollary 1 to show that precision-recall curve generated from the likelihood transformation ($PR(g_0^L, g_1^L)$) dominates ones generated from other transformation ($PR(g_0^T, g_1^T)$).

**Corollary 8.** *$PR(g_0^L, g_1^L)$ dominates $PR(g_0^T, g_1^T)$.*

*Proof.* According to Proposition 4, PPV and NPV are monotonically increasing in TPR and decreasing in FPR. Based on Corollary 1, for any thresholds $t_L \in [0, +\infty)$ and $t_T \in (-\infty, +\infty)$ such that $TPR^T(t_T) = TPR^L(t_L)$, we have $PPV^L(t_L) \geq PPV^T(t_T)$ and $NPV^L(t_L) \geq NPV^T(t_T)$. □

Another extension of ROC curve is Total Operating Characteristic (TOC) curve, which is considered to reveal much more information than ROC (Pontius Jr and Si, 2014; Halligan et al., 2015). The $x$-axis of TOC is $TP + FP$ and $y$-axis is $TP$. For each threshold, ROC only tells TPR and FPR, while TOC reveals the total information in the confusion-matrix for each threshold (Pontius and Parmentier, 2014). By leveraging Proposition 3, we can prove that TOC curve generated from the likelihood transformation ($TOC(g_0^L, g_1^L)$) dominates ones generated from any other transformation ($TOC(g_0^T, g_1^T)$).

**Corollary 9.** *$TOC(g_0^L, g_1^L)$ dominates $TOC(g_0^T, g_1^T)$.*

*Proof.* According to Proposition 3, when $\alpha = \pi_0$ and $\beta = \pi_1$, if

$$\pi_0 \int_{R_L} f_0(x)dx + \pi_1 \int_{R_L} f_1(x)dx = \pi_0 \int_{R_T} f_0(x)dx + \pi_1 \int_{R_T} f_1(x)dx,$$

then $\int_{R_L} f_1(x)dx \geq \int_{R_T} f_1(x)dx$. Here $\int_R f_0(x)dx$ and $\int_R f_1(x)dx$ represents $FPR$ and $TPR$ (see Table 4.1), respectively, where $R$ represents the set of $x$ where the associated instances are classified as positive ($R_L$ or $R_T$). Proposition 3 implies when $FP^L + TP^L = FP^T + TP^T$, we have $TP^L \geq TP^T$. Therefore, $TOC(g_0^L, g_1^L)$ dominates $TOC(g_0^T, g_1^T)$. □

### 5.3.3 Simulation Study

Figure 5.2 shows that the corresponding ROC curves for some density distributions are non-concave, while Figure 5.3 indicates that their likelihood ratio $f_1/f_0$ is not monotone. As we discussed, the likelihood transformation (Eq. 1) results in concave ROC regardless of the concavity of original ROC and the likelihood transformed ROC dominates other ROC associated with other transformation. In this subsection, we performed a simulation study to transform the non-concave ROCs in Figures 5.2 into concave ROCs and provide empirical support for our theoretical studies.

Two samples of size 10k are generated from normal distributions for four scenarios: (1) Class 1: N(0.5, 0.05); Class 0: N(0.5, 0.1) (2) Class 1: N(0.5, 0.1); Class 0: N(0.5, 0.05) (3) Class 1: 0.5*N(0.2,0.05)+0.5*N(0.5, 0.05); Class 0: N(0.38, 0.1) (4) Class 1: N(0.38, 0.1); Class 0: 0.5*N(0.2,0.05)+0.5*N(0.5, 0.05). Density plots $f_1$ and $f_0$ are obtained based on kernel density estimation from the two samples. Apply $f_1$ and $f_0$ to each instance to estimate the density and calculate $s$ according to (Eq. 3.6). The original ROC is denoted as $ROC(f_0, f_1)$ and newly transformed ROC is denoted as $ROC(g_0^L, g_1^L)$. Both $ROC(f_0, f_1)$ and $ROC(g_0^L, g_1^L)$ are generated by following the same procedure through varying one threshold. TP is counted as the number of positive instances with score $x$ ($s$) higher than the threshold, while FP is counted as the number of negative instances with score $x$ ($s$) higher than the threshold. TPR and FPR can be calculated for each threshold.

As illustrated in Figure 5.4, all $ROC(g_0^L, g_1^L)$ curves are concave and dominate original ROC after the likelihood transformation with significant improvement in terms of AUC. For scenarios (1) and (2), where density distributions of the two classes have the same means but different variances, the AUC is improved from 0.5 to 0.7. For scenarios (3) and (4) with bimodal class distributions, the AUC is improved from 0.46 (0.54) to 0.83. Traditionally, those original models are considered as poor models and model re-train is needed. Nevertheless, our study shows that those original models with poor performance in terms of AUC actually demonstrates excellent discrimination power after the likelihood transformation and no re-train is necessary. The simulation

study fully supports our theoretical investigation.

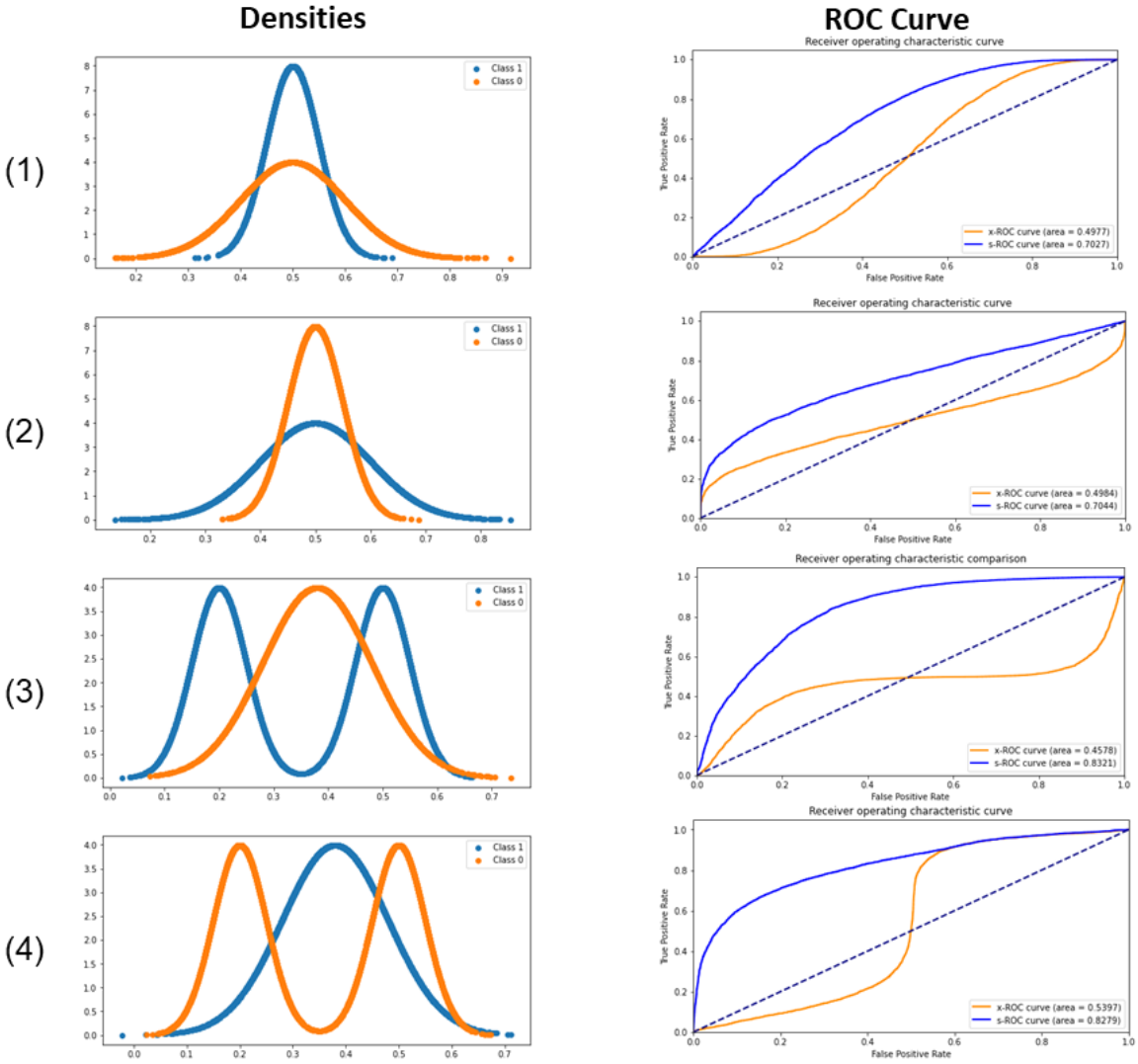


Figure 5.4: Enhanced ROC Based on Likelihood Ratio Transformation

## 5.4 Summary

ROC curve is a commonly used tool to measure the model performance, which has been extensively studied empirically and theoretically. However, those studies are subject to two major limitations:

1. Little attention has been paid to its concavity. In most studies, concavity of ROC was implicitly assumed. Some papers recognized the non-concave ROC

and proposed ad-hoc methods to remediate it. Some researchers recommend not to use the classifiers if the ROC curve is not concave.

2. The ROC curve is created based on one dimensional score output. There is lack of discussion to construct ROC curve for multidimensional scores.

In this chapter, we identify the necessity and sufficiency conditions for a ROC curve to be concave. In addition, we reveal that the likelihood (or monotone of likelihood) transformations will lead to the most discriminate ROC curve, which is in line with Neyman-Pearson lemma (Neyman and Pearson, 1933). Additionally, we discover that two alternative curves (TOC, PR) resulting from likelihood transformation dominates the ones from other transformations.

# Chapter 6

# ENHANCEMENT OF COST-SENSITIVE LEARNING

## 6.1 Introduction

Most classification methods aim to maximize estimation accuracy, assuming all the instances of mis-classification are equally costly. However, this assumption is not true in many real-world applications: the phenomenon of unequal cost is often observed in case of data imbalance, where one class (minority class) has significantly fewer instances than the other (majority class) (Sun et al., 2007). For example, in credit card fraud detection problem, approving a fraud transaction (minority class) costs much more than rejecting a legitimate transaction (majority class); Consider a cancer diagnosing situation: missing a positive cancer diagnose is a much more serious mistake than mis-diagnosing a healthy patient with cancer and following-up with more medical tests (Sammut and Webb, 2011); In bank loan underwriting, it is much more costly to accept a loan from a bad customer who will never repay than to reject a loan from a good customer. Under these situations, traditional non-cost-sensitive classifier cannot provide satisfactory results (Sun et al., 2009; Bahnsen et al., 2013). Developing a cost-sensitive classifier is necessary with the objective to minimize the overall cost.

Various methods have been proposed to deal with unequal cost issues in last two decades. In general, these methods fall into three categories: data adjustment, algorithm adjustment, and threshold adjustment. The data adjustment is mainly achieved by re-sampling, including over-sampling (replicating instances from the minority), down-sampling (removing instances from the majority), generating synthetic data (e.g., synthetic minority oversampling technique Chawla et al. (2002)). Elkan (2001) proposed a method to re-balance the data prior distribution to achieve cost-sensitive

classification decisions. The data adjustment can also be realized by relabeling. For example, Domingos (1999) proposed *Meta cost*, where the training examples are relabeled based on their optimal cost-sensitive classes. Drummond et al. (2003) and Mease et al. (2007) pointed out that data adjustment introduces problematic consequences that can potentially impact model performance. For example, for down-sampling, important information regarding majority class may be missed due to the removal of instances from the majority class. For over-sampling, overfitting issue has been observed in (Mease et al., 2007). In addition, modifying the dataset may alter the true distribution of classes and change the representation of the reality. In many cases, the improvement on model performance is also debatable (Poolsawad et al., 2014; More, 2016; Khushi et al., 2021). Furthermore, data adjustment method is only applicable to problems, where cost matrix is fixed. The prior class distribution cannot be altered based on instance-dependent cost (refer to Chapter 2 for instance-dependent cost matrix), which causes issues for dealing with such problems (Wang and Chen, 2017).

For algorithm adjustment, there are also a lot of paper to individually make each classification learner cost-sensitive, which aims to incorporate the cost information into the algorithms. For example, Wan et al. (1999) designed a cost-sensitive neural network by modifying the standard back-propagation algorithm; Ting (2002) proposed an instance-weighting method to introduce cost-sensitive decision trees; Fan et al. (1999) and Sun et al. (2007) studied how to reduce mis-classification cost in boosting algorithm; Bahnsen et al. (2014) developed an instance-dependent cost-sensitive logistic regression for credit scoring; Iranmehr et al. (2019) modified support vector machine's standard hinge loss function with cost information to demonstrate reliable generalization performance. The drawback of algorithm adjustment method is obvious: it requires an in-depth understanding of each learning algorithm and precise identification of root causes for its failure in learning from unequal cost data. It cannot be generalized to any classifier, since individually making each classifier cost-sensitive is burdensome and inefficient. In particular, many machine learning models are considered as black-box (Castelvecchi, 2016), making it very difficult to be cost-sensitive.

Different from data adjustment and algorithm adjustment, which are considered as pre-processing method, threshold adjustment belongs to post-processing, which aims to convert non cost-sensitive output into cost-sensitive decisions (Wang and Chen, 2017). Threshold adjustment is a fundamental method to transform an accuracy based threshold into a cost-sensitive threshold. For a binary classification, Elkan (2001) calculated the cost-sensitive threshold based on the break-even of classifying the instance into either class. The accuracy of probability estimate is critical to the threshold adjustment method. Zadrozny and Elkan (2001) proposed decision tree and naive Bayesian learning methods to enhance the accuracy of probability estimation. Sheng and Ling (2006) developed an empirical thresholding method which does not require accurate estimation of probabilities, but the ranking order of the probabilities. To incorporate the accuracy of probability estimation, Wang and Chen (2017) proposed a cost-sensitive threshold adjustment framework by minimizing the overall expected prediction cost based on the assumption that the Type I and Type II error functions are convex.

In this chapter, we improve Wang and Chen's work further and propose a simple and effective threshold-adjustment method to address the limitations in current cost-sensitive methods mentioned above. Our method is based on the likelihood transformation introduced in Chapter 3 and it does not require knowledge of the learning algorithms and can apply to any classifiers. More importantly, unlike Wang and Chen's method, our proposed method does not rely on convex assumptions of Type I and Type II error functions. The rest of this Chapter is structured as follows: Section 6.2 introduces the break-even method for cost-sensitive threshold adjustment; Section 6.3 reviews threshold adjustment method proposed by Wang and Chen via minimizing overall expected cost; Section 6.5 obtains and proves the optimal cost-sensitive threshold; Section 6.5 summarizes our work and highlights our major improvement over Wang and Chen's methods.

## 6.2 Direct Cost-Sensitive Decision-Making Method

Recall the definition of cost-matrix in Chapter 2. Let $C(i,j), (i,j \in \{0,1\})$ be the cost of predicting an instance as class $i$ when the true class is $j$. Let $c_{ij}$ denote $C(i,j)$ for simplicity. Consider an instance with score $x$. Let $P_j(x)$ be the true probability that the instance belongs to class $j$. Then the expected cost of predicting the instance as class $i$ is

$$H_i(x) = P_1(x)c_{i1} + P_0(x)c_{i0}, \ i = 1, 0. \tag{6.1}$$

Elkan (2001) proposed a direct cost-sensitive threshold (DCS) method, which calculates the break-even probability by setting $H_1(x) = H_0(x)$ and using the fact $P_1(x)+P_0(x) = 1$.

$$P_1^* = \frac{c_{10} - c_{00}}{c_{10} - c_{00} + c_{01} - c_{11}}. \tag{6.2}$$

In theory, for a given instance with score $x$, we should compare $P_1(x)$ with the threshold $P_1^*$ to classify its status. In reality, however, since $P_1(x)$ is generally unknown, the author suggests to use $x$ as an approximation of $P_1(x)$ to make the classification decision: if $x \geq P_1^*$, the instance will be classified as positive and negative, otherwise.

## 6.3 Minimizing Overall Expected Cost

By design, the above-mentioned direct cost-sensitive threshold (DCS) method works well only if all instance scores $x$ are accuracy estimates of probability $P_1(x)$. This usually requires carefully calibrations and can be difficult to achieve. Wang and Chen (2017) proposed another framework to minimize the expected cost (we call it expected cost method - ECM). It is designed to compare the instance score $x$ directly with a threshold $t$. For a given threshold $t$, let $H_0^{ecm}$ be the expected cost for those instances classified as negative and $H_1^{ecm}$ be the expected cost for those instances classified as

positive. We have

$$
\begin{aligned}
H_0^{ecm}(t) &= P(Y=0|x<t)c_{00} + P(Y=1|x<t)c_{01} \\
&= \frac{P(Y=0)P(x<t|Y=0)}{P(x<t)}c_{00} + \frac{P(Y=1)P(x<t|Y=1)}{P(x<t)}c_{01} \\
H_1^{ecm}(t) &= P(Y=0|x\geq t)c_{10} + P(Y=1|x\geq t)c_{11} \\
&= \frac{P(Y=0)P(x\geq t|Y=0)}{P(x\geq t)}c_{10} + \frac{P(Y=1)P(x\geq t|Y=1)}{P(x\geq t)}c_{11}
\end{aligned}
$$

To minimize the expected overall cost, the optimization problem can be formulated as:

$$
\begin{aligned}
\min_t \; H^{ecm}(t) \quad &= P(x<t) * H_0^{ecm}(t) + P(x\geq t) * H_1^{ecm}(t) \\
&= \pi_0 F_0(t)c_{00} + \pi_1 F_1(t)c_{01} + \pi_0(1-F_0(t))c_{10} + \pi_1(1-F_1(t))c_{11} \quad (6.3)
\end{aligned}
$$

where $F_0(t) = P(x < t|Y = 0)$ and $F_1(t) = P(x < t|Y = 1)$. Notice that $1 - F_0(t)$ and $F_1(t)$ are called Type I error and Type II error, respectively, in statistics literature. Wang and Chen (2017) proved that if $F_0(t)$ is concave and $F_1(t)$ is convex, the objective function $H^{ecm}(t)$ is convex and any local minimum of the optimization problem (6.3) is also a global minimum of the problem. The assumptions imply that $f_0(t)$ is monotonically decreasing and $f_1(t)$ is monotonically increasing, which are hard to satisfy in practice.

Wang and Chen (2017) showed that the optimal solution $t^*$ of (6.3) is the unique solution of the following equation:

$$
B(t^*_{ecm}) = \frac{c_{10} - c_{00}}{c_{10} - c_{00} + c_{01} - c_{11}}, \tag{6.4}
$$

where $B(t) = \pi_1 f_1(t)/(\pi_1 f_1(t) + \pi_0 f_0(t))$. The convexity assumptions made on $F_0(x)$ and $F_1(x)$ guarantee that function $B$ is monotone and has a unique inverse for all $t \in (0, 1)$.

### 6.4 Proposed Cost-Sensitive Enhancement

Notice that the objective function $H^{ecm}(t)$ may not be convex if the convexity assumptions on $F_0(x)$ and $F_1(x)$ are not met. In that case, it will be difficult to find the optimal threshold $t^*_{ecm}$. We apply the likelihood ratio transformation on the instance score $x$ and obtain the transformed score $s$ (Eq. 3.6) and threshold $t \in [0, +\infty)$. As in Chapter 3, let $g_i^L(\cdot)$, $i = 0, 1$, be the resulting conditional probability density functions of $s$ and $G_i^L(\cdot)$, $i = 0, 1$ be the corresponding cumulative probability functions. Similar to the derivations used in Wang and Chen (2017), we obtain the following optimization problem, for the transformed variable $s$, to minimize the expected overall cost:

$$\min_t \quad H^L(t) = \pi_0 G_0^L(t) c_{00} + \pi_1 G_1^L(t) c_{01} + \pi_0 (1 - G_0^L(t)) c_{10} + \pi_1 (1 - G_1^L(t)) c_{11} \tag{6.5}$$

**Proposition 12.** *For any given instance, the optimal threshold $t^L$ of (Eq. 6.5) is given by*

$$t^L = \frac{\pi_0 (c_{00} - c_{10})}{\pi_1 (c_{11} - c_{01})}.$$

*Proof.* We have

$$\frac{dH^L(t)}{dt} = \pi_0 g_0^L(t)(c_{00} - c_{10}) - \pi_1 g_1^L(t)(c_{11} - c_{01}). \tag{6.6}$$

Setting $H^L(t)$'s derivative to zero with respect to $t$ results in:

$$\frac{g_1^L(t^L)}{g_0^L(t^L)} = \frac{\pi_0 (c_{00} - c_{10})}{\pi_1 (c_{11} - c_{01})}$$

According to Proposition 1, we have $g_1^L(t^L)/g_0^L(t^L) = t^L$. In view of (Eq. 6.6), it is easy to show that $dH^L(t)/dt < 0$ when $t < t^L$ and $dH^L(t)/dt > 0$, otherwise. Therefore, $H^L(t)$ is a unimodal function and $t^L$ is the optimal threshold. □

With the above result, an instance is classified as positive (1) if the transformed score $s \geq t^L$, and negative, otherwise.

**Corollary 10.** *For any given instance, the optimal threshold $t^B$ for a Bayesian probability transformed score (Eq. 3.7) is given by*

$$t^B = \frac{c_{10} - c_{00}}{c_{10} - c_{00} + c_{01} - c_{11}}.$$

### 6.5 Summary

Threshold adjustment is a generalized cost-sensitive method to make non-cost-sensitive output into cost-sensitive decisions. The fundamental challenge for traditional direct cost-sensitive decision-making threshold adjustment method is the estimation accuracy. Wang and Chen (2017) recognized the challenge and proposed an enhanced algorithm by minimizing the expected cost. Their method relies on the assumption that Type I and Type II error functions are convex. Compared with Wang and Chen's method, our approach, which is built on likelihood or Bayesian probability transformation, demonstrates at least two advantages: 1. Our approach is more generally applicable since it does not rely on the convex assumption of Type I and Type II error functions, which is a very strong assumption. We are able to prove that the expected cost function in our optimization problem is unimodal for any general conditional density functions of the instance score; 2. We provide an explicit formula for the optimal threshold, while Wang and Chen's method requires inverse function $B$.

# Part II

# CLASSIFICATION PROBLEM EXTENSION AND SOLUTION

# Chapter 7

# BINARY CLASSIFICATION WITH AMBIGUITY

Majority binary classification models are based on binary decisions, which means the predicted outcomes are either positive or negative, e.g., Ferri et al. (2009); Powers (2011); Majnik and Bosnić (2013); Bahnsen et al. (2014); Awoyemi et al. (2017). However, in many real world applications, an ambiguous option is necessary for decision-making due to lack of information to produce a specific classification. For example, credit card companies develop behavioural credit scoring to evaluate credit risk of existing customers. The credit lines for those customers with scores at high (or low) end are proactively increased (or decreased), while those with scores at the middle remain unchanged or need to be further reviewed with additional information; and in fraud detection, transactions with high (low) risk scores are automatically declined (or approved), while a portion of transactions are required to be routed for manual review. Similar and related problems have been recognized in some literature (Herbei and Wegkamp, 2006; Yao, 2009; Yuan and Wegkamp, 2010; Lei, 2014; Campagner et al., 2019; Liu and Lin, 2019). Especially, Lei (2014) developed a framework of classification by extending the Neyman-Pearson lemma, which aims to minimize the ambiguous region, while achieving desired performance in the decided region. Current related research mainly focuses on modifying the model algorithm with adjusted objective functions to minimize the ambiguous region and little attention has been paid on the performance measures.

We notice that a better model in binary decision scenario may not perform better when an ambiguous region exists. For example, as illustrated in Table 7.1, the highlighted grey part is the ambiguous region and other part is decided region (DR). Assume the threshold is 0.5 (if score $\geq 0.5$, then predict positive, otherwise negative)

and the ambiguous size is 0.4. In such case, model A is deemed better than model B in terms of overall accuracy (80% vs. 70%). However, when only accuracy in decided region is considered, model B demonstrates much better accuracy than model A (100% vs. 83%).

| Score A | Score B | Actual Label | Model A Prediction | Model B Prediction |
|---|---|---|---|---|
| 0.98 | 0.99 | + | + | + |
| 0.91 | 0.71 | + | + | + |
| 0.85 | 0.31 | - | + | - |
| 0.55 | 0.49 | + | + | - |
| 0.51 | 0.48 | + | + | - |
| 0.48 | 0.47 | + | - | - |
| 0.47 | 0.46 | - | - | - |
| 0.31 | 0.21 | - | - | - |
| 0.21 | 0.11 | - | - | - |
| 0.05 | 0.12 | - | - | - |
| | | Overall Accuracy | 80% | 70% |
| | | DR Accuracy | 83% | 100% |

Table 7.1: Model Accuracy Comparison When Ambiguous Region is Considered

Conceptually, those threshold dependent confusion-matrix based measures (see Section 9) used in binary classification with binary decisions (BCBD) can be extended to binary classification with ambiguous region (BCAR). In the context of likelihood and Bayesian probability transformations, Chapter 4 provided in-depth systematic discussions regarding the relationship between the confusion-matrix measures and optimal threshold & base rate in BCBD and Chapter 6 enhanced cost-sensitive learning by proposing a simple and intuitive threshold-adjustment method. In this chapter, we extend such theoretical investigations to BCAR, where the confusion-matrix optimal thresholds and cost-sensitive thresholds are impacted by the ambiguous region.

The rest of this chapter is structured as following: Section 7.1 formulates the classification problem with ambiguity region. We aim to disclose the relationship between the optimal measure and decision-making thresholds; Section 7.2 constructs comprehensive measures in BCBR and discloses the intrinsic relationship between

the optimal measures and decision-making thresholds; Section 7.3 extends our cost-sensitive learning framework in BCBD to BCAR and identifies the optimal threshold pair to minimize the expected cost; Section 7.4 summarizes our results and highlights our contributions.

## 7.1 Problem Formulation

BCAR can be illustrated by Figure 7.1: an instance with a score less than $t_0$ is classified as 0 (or negative) and an instance with a score greater than $t_1$ is classified as 1 (or positive). Instances with score between $t_0$ and $t_1$ are ambiguous. The problem is to identify the pair of thresholds that lead to certain optimal measure. When $t_0 = t_1$, the problem is reduced to BCBD.


Figure 7.1: Illustration of Classification with Ambiguity with Likelihood Transformation

We first apply the likelihood transformation to convert score $X \in R^n$ to score $S \in R^+$ (Eq. 3.6). Recall the definitions developed in Chapter 3: $g^L(\cdot)$, $g_0^L(\cdot)$, and $g_1^L(\cdot)$ denote the density and class-conditional density functions of the resulting random scalar $S$. Let $G^L(\cdot)$, $G_0^L(\cdot)$, and $G_1^L(\cdot)$ be the corresponding cumulative probability distribution functions. The BCAR threshold optimization problem identifies the boundaries of the ambiguity region of a fixed size to maximize certain specified performance measure. It can be formulated as:

$$\begin{aligned} \max_{t_0,t_1} \quad & PM(t_0,t_1) \\ \text{s.t.} \quad & G^L(t_1) - G^L(t_0) = \alpha \\ & 0 \le t_0 \le t_1, \end{aligned}$$

where $PM$ represents certain performance measure to be optimized, which is a function of two thresholds $t_0$ and $t_1$. $\alpha$ is the size of the ambiguous region, which is assumed to be given.

## 7.2 Measures Based on Confusion-Matrix in BCAR

Table 7.2 represents the confusion-matrix for BCAR with three rows (for predicted classes) and two columns (for actual classes). Different from traditional $2 \times 2$ confusion-matrix, the BCAR confusion-matrix contains an additional row for ambiguous status.

| | | **Actual Class** | |
|---|---|---|---|
| | | $Y = 1$ | $Y = 0$ |
| **Predicted Class** | $\hat{Y} = 1$ | $TP$ | $FP$ |
| | $\hat{Y} = u$ | $UP$ | $UN$ |
| | $\hat{Y} = 0$ | $FN$ | $TN$ |

Table 7.2: Confusion-Matrix for BCAR

The six elements in the BCAR confusion-matrix can be described as the joint probabilities $P(Y = i, \hat{Y} = j)$, where $i \in \{0, 1\}, j \in \{0, u, 1\}$, which are functions of threshold(s) and base rate ($\pi_1$):

$$TP(t_0, t_1, \pi_1) = P(Y = 1, \hat{Y} = 1) = \pi_1 P(\hat{Y} = 1|Y = 1) = \pi_1 \int_{t_1}^{+\infty} g_1^L(s)ds,$$

$$FP(t_0, t_1, \pi_1) = P(Y = 0, \hat{Y} = 1) = \pi_0 P(\hat{Y} = 1|Y = 0) = \pi_0 \int_{t_1}^{+\infty} g_0^L(s)ds,$$

$$UP(t_0, t_1, \pi_1) = P(Y - 1, \hat{Y} = u) = \pi_1 P(\hat{Y} = u|Y = 1) = \pi_1 \int_{t_0}^{t_1} g_1^L(s)ds,$$

$$UN(t_0, t_1, \pi_1) = P(Y - 0, \hat{Y} = u) = \pi_0 P(\hat{Y} = u|Y = 0) = \pi_0 \int_{t_0}^{t_1} g_0^L(s)ds,$$

$$TN(t_0, t_1, \pi_1) = P(Y = 0, \hat{Y} = 0) = \pi_0 P(\hat{Y} = 0|Y = 0) = \pi_0 \int_{0}^{t_0} g_0^L(s)ds,$$

$$FN(t_0, t_1, \pi_1) = P(Y = 1, \hat{Y} = 0) = \pi_1 P(\hat{Y} = 0|Y = 1) = \pi_1 \int_{0}^{t_0} g_1^L(s)ds.$$

From these expressions, we observe that $UP$ and $UN$ are impacted by both thresholds, while all other matrix elements depend on only one threshold. Like confusion-matrix measures defined in Chapter 4, various measures can be constructed based on the six elements. We construct a list of fundamental accuracy and error alike measures, which are summarized in Table 7.3.

| Measures | Formulas |
|---|---|
| Overall Performance with Ambiguity Treated as Misclassification | |
| $OTPR(t_1)$ | $TP(t_1, \pi_1)/\pi_1$ |
| $OTNR(t_0)$ | $TN(t_0, \pi_1)/\pi_0$ |
| $OFNR(t_0)$ | $FN(t_0, \pi_1)/\pi_1$ |
| $OFPR(t_1)$ | $FP(t_1, \pi_1)/\pi_0$ |
| $OACC(t_0, t_1, \pi_1)$ | $TP(t_1, \pi_1) + TN(t_0, \pi_1)$ |
| Performance based on the Decided Region Only | |
| $TPR(t_0, t_1, \pi_1)$ | $TP(t_1, \pi_1)/(FN(t_0, \pi_1) + TP(t_1, \pi_1))$ |
| $TNR(t_0, t_1, \pi_1)$ | $TN(t_0, \pi_1)/(TN(t_0, \pi_1) + FP(t_1, \pi_1))$ |
| $PPV(t_1, \pi_1)$ | $TP(t_1, \pi_1)/(TP(t_1, \pi_1) + FP(t_1, \pi_1))$ |
| $NPV(t_0, \pi_1)$ | $TN(t_0, \pi_1)/(TN(t_0, \pi_1) + FN(t_0, \pi_1))$ |
| $FNR(t_0, t_1, \pi_1)$ | $FN(t_0, \pi_1)/(FN(t_0, \pi_1) + TP(t_1, \pi_1))$ |
| $FPR(t_0, t_1, \pi_1)$ | $FP(t_1, \pi_1)/(FP(t_1, \pi_1) + TN(t_0, \pi_1))$ |
| $FDR(t_1, \pi_1)$ | $FP(t_1, \pi_1)/(TP(t_1, \pi_1) + FP(t_1, \pi_1))$ |
| $FOR(t_0, \pi_1)$ | $FN(t_0, \pi_1)/(TN(t_0, \pi_1) + FN(t_0, \pi_1))$ |
| $ACC(t_0, t_1, \pi_1)$ | $(TP(t_1, \pi_1) + TN(t_0, \pi_1))/(1 - \alpha)$ |

Table 7.3: BCAR Confusion-Matrix Based Single Performance Measures

These measures are grouped into two categories. The first category is to measure the model performance in overall region by treating the ambiguous region as a misclassification. We observe that:

- Among the overall metrics, $OACC$ is related to both thresholds and base rate, while other overall single metrics are only related to one specific threshold.

- Given a fixed ambiguous region size $\alpha$, the pair of decision-making thresholds that maximize $OACC$ will result in maximum $ACC$ as well.

- $OTPR$ and $OFPR$ are monotonically decreasing in $t_1$, while $OTNR$ and $OFNR$ are monotonically increasing in $t_0$.

The second category only focuses on the performance in the decided region by skipping the ambiguous region. For the measures in the second category, we have the following observations:

- $PPV$ and $FDR$ are only related to threshold $t_1$ and base rate. As discussed in Proposition 4, $PPV$ is monotonically increasing with threshold $t_1$, while $FDR$ is monotonically decreasing in $t_1$.

- $NPV$ and $FOR$ are only related to threshold $t_0$. As discussed in Proposition 4, $NPV$ is monotonically decreasing in threshold $t_0$, while $FOR$ is monotonically increasing in $t_0$.

- $TPR$, $TNR$, $FNR$ and $FPR$ are related to both thresholds as well as base rate. Given a fixed ambiguous region size $\alpha$, $TPR$ and $FPR$ are monotonically decreasing with threshold $t_0$ (or $t_1$), while $TNR$ and $FNR$ are monotonically increasing with threshold $t_0$ (or $t_1$).

The above-mentioned accuracy and error alike metrics are fundamental single measures, which are subject to limitations (refer to detailed discussions in Chapter 2 & 4). A combination of them can help address some of the limitations. Below we consider some combined metrics such as accuracy, Jaccard index and $F1$ in more details, since their values do not vary monotonically with the corresponding thresholds.

**Accuracy**

Given $\alpha$ and base rate $\pi_1$, according to Table 7.3, the problem can be formulated as below:

$$\begin{aligned} \max_{t_0,t_1} \quad & ACC(t_0,t_1) = \pi_0 \int_0^{t_0} g_0^L(s)ds + \pi_1 \int_{t_1}^{+\infty} g_1^L(s)ds \\ \text{s.t.} \quad & G^L(t_1) - G^L(t_0) = \alpha \\ & 0 \le t_0 \le t_1 \end{aligned}$$

**Proposition 13.** *The optimal decision-making thresholds* $(t^L_{0(acc)}, t^L_{1(acc)})$ *to maximize the accuracy measure satisfies* $t^L_{0(acc)} t^L_{1(acc)} = \pi_0^2/\pi_1^2$.

*Proof.* Note that

$$t_1 = (G^L)^{-1}(\alpha + G^L(t_0)) \tag{7.1}$$

$$\frac{dt_1}{dt_0} = \frac{g^L(t_0)}{g^L((G^L)^{-1}(\alpha + G^L(t_0)))} = \frac{\pi_0 g_0^L(t_0) + \pi_1 g_1^L(t_0)}{\pi_0 g_0^L(t_1) + \pi_1 g_1^L(t_1)}. \tag{7.2}$$

Since

$$\frac{\partial(ACC)}{\partial t_0} = \pi_0 g_0^L(t_0) - \pi_1 g_1^L(t_1)\frac{dt_1}{dt_0},$$

and because of 7.2, we have

$$\frac{\partial(ACC)}{\partial t_0} = \frac{\pi_0^2 g_0^L(t_0) g_0^L(t_1) - \pi_1^2 g_1^L(t_0) g_1^L(t1)}{\pi_0 g_0^L(t_1) + \pi_1 g_1^L(t_1)}.$$

Setting $\partial(ACC)/\partial t_0$ to zero results in

$$\frac{g_1^L(t^L_{0(acc)}) g_1^L(t^L_{1(acc)})}{g_0^L(t^L_{0(acc)}) g_0^L(t^L_{1(acc)})} = \frac{\pi_0^2}{\pi_1^2}.$$

Since $g_1^L(t_{1(acc)}^L)/g_0^L(t_{1(acc)}^L) = t_{1(acc)}^L$ and $g_1^L(t_{0(acc)}^L)/g_0^L(t_{0(acc)}^L) = t_{0(acc)}^L$ according to Proposition 1, we have $t_{0(acc)}^L t_{0(acc)}^L = \pi_0^2/\pi_1^2$.

Besides, when $t_0 t_1 \leq \pi_0^2/\pi_1^2$, $\partial(ACC)/\partial t_0 \geq 0$; Else, $\partial(ACC)/\partial t_0 < 0$. Therefore, $(t_{0(acc)}^L, t_{1(acc)}^L)$ are the optimal thresholds. □

**Corollary 11.** *The optimal decision-making thresholds $(t_{0(acc)}^B, t_{1(acc)}^B)$ to maximize the accuracy for the Bayesian probability transformed score (Eq. 3.7) satisfies $t_{0(acc)}^B + t_{1(acc)}^B = 1$.*

*Proof.* The corollary can be proven by following the same procedure as Proposition 13 by applying Proposition 2. □

Clearly, with the above results, we can calculate the two optimal thresholds by solving respective joint equations with the constraint: $G^L(t_1) - G^L(t_0) = \alpha$.

### $F1$ in Decided Region

For $F1$ in decided region, the $F1$ is defined as harmonic mean of $TPR(t_0, t_1)$ and $PPV(t_1)$. Recall the definition of Jaccard Index (JI) and $F1$ in Chapter 4:

$$JI_{DR}(t_0, t_1) = \frac{TP(t_1)}{TP(t_1) + FP(t_1) + FN(t_0)}$$

and

$$F1_{DR}(t_0, t_1) = \frac{2TP(t_1)}{2TP(t_1) + FP(t_1) + FN(t_0)}.$$

Clearly, $F1_{DR} = 2JI_{DR}/(1 + JI_{DR})$, which indicates that $F1$ is a monotonically increasing function of Jaccard index. Therefore, the threshold that maximizes $F1$ will result in optimal Jaccard index as well. Given $\alpha$, we aim to maximize $F1$ in the decided region. Notice that

$$\begin{aligned} F1_{DR}(t_0, t_1) &= \frac{2\pi_1 \int_{t_1}^{+\infty} g_1^L(s)ds}{2\pi_1 \int_{t_1}^{+\infty} g_1^L(s)ds + \pi_0 \int_{t_1}^{+\infty} g_0^L(s)ds + \pi_1 \int_0^{t_0} g_1^L(s)ds} \\ &= \frac{2\pi_1(1 - G_1^L(t_1))}{1 + \pi_1 - 2\pi_1 G_1^L(t_1) - \pi_0 G_0^L(t_1) + \pi_1 G_1^L(t_0)}. \end{aligned}$$

The problem can be formulated as:

$$\begin{aligned} \max_{t_0,t_1} \quad & F1_{DR}(t_0,t_1) = \frac{2\pi_1(1-G_1^L(t_1))}{1+\pi_1-2\pi_1 G_1^L(t_1)-\pi_0 G_0^L(t_1)+\pi_1 G_1^L(t_0)} \\ \text{s.t.} \quad & G^L(t_1)-G^L(t_0)=\alpha \\ & 0 \le t_0 \le t_1 \end{aligned}$$

**Proposition 14.** *Let $(t^L_{0(DRF)}, t^L_{1(DRF)})$ denote the optimal decision-making thresholds to maximize the $F1$ in the decided region and $JI_{DR}(t^L_{0(DRF)}, t^L_{1(DRF)})$ is the Jaccard index achieved by the optimal thresholds. The following properties hold:*

1. $(\pi_0\pi_1 t^L_{1(DRF)} + \pi_1^2 t^L_{0(DRF)} t^L_{1(DRF)})/(\pi_0^2 + \pi_0\pi_1 t^L_{1(DRF)}) = JI_{DR}(t^L_{0(DRF)}, t^L_{1(DRF)})$,
2. $t^L_{0(DRF)} \le \pi_0 JI_{DR}(t^L_{0(DRF)}, t^L_{1(DRF)})/\pi_1 \le t^L_{1(DRF)}$.

*Proof.* We have

$$\begin{aligned} \frac{\partial F1_{DR}(t_0,t_1)}{\partial t_0} = & \frac{-2\pi_1 g_1^L(t_1)\frac{dt_1}{dt_0}[\pi_0 - \pi_0 G_0^L(t_1) + \pi_1 G_1^L(t_0)]}{[1+\pi_1-2\pi_1 G_1^L(t_1)-\pi_0 G_0^L(t_1)+\pi_1 G_1^L(t_0)]^2} \\ & - \frac{2\pi_1(1-G_1^L(t_1))[-\pi_0 g_0^L(t_1)\frac{dt_1}{dt_0} + \pi_1 g_1^L(t_0)]}{[1+\pi_1-2\pi_1 G_1^L(t_1)-\pi_0 G_0^L(t_1)+\pi_1 G_1^L(t_0)]^2}. \end{aligned}$$

In view of (Eq. 7.2), we have

$$\pi_0 g_0^L(t_1)\frac{dt_1}{dt_0} - \pi_1 g_1^L(t_0) = \frac{\pi_0^2 g_0^L(t_0) g_0^L(t_1) - \pi_1^2 g_1^L(t_0) g_1^L(t_1)}{\pi_0 g_0^L(t_1) + \pi_1 g_1^L(t_1)} \tag{7.3}$$

and

$$g_1^L(t_1)\frac{dt_1}{dt_0} = \frac{\pi_0 g_0^L(t_0) g_1^L(t_1) + \pi_1 g_1^L(t_0) g_1^L(t_1)}{\pi_0 g_0^L(t_1) + \pi_1 g_1^L(t_1)}. \tag{7.4}$$

According to Proposition 1, (Eqs. 7.3-7.4) can be combined and simplified as:

$$\frac{g_1^L(t_1)\frac{dt_1}{dt_0}}{\pi_0 g_0^L(t_1)\frac{dt_1}{dt_0} - \pi_1 g_1^L(t_0)} = \frac{\pi_0 t_1 + \pi_1 t_0 t_1}{\pi_0^2 - t_0 t_1 \pi_1^2}.$$

Setting the partial derivative $\partial F1_{DR}(t_0, t_1)/\partial t_0$ to zero results in:

$$\frac{1 - G_1^L(t^L_{1(DRF)})}{\pi_0 - \pi_0 G_0^L(t^L_{1(DRF)}) + \pi_1 G_1^L(t^L_{0(DRF)})} = \frac{\pi_0 t^L_{1(DRF)} + \pi_1 t^L_{0(DRF)} t^L_{1(DRF)}}{\pi_0^2 - t^L_{0(DRF)} t^L_{1(DRF)} \pi_1^2} \tag{7.5}$$

After rearranging (Eq. 7.5), we have

$$\frac{\pi_1(1 - G_1^L(t^L_{1(DRF)}))}{\pi_1(1 - G_1^L(t^L_{1(DRF)})) + \pi_0 - \pi_0 G_0^L(t^L_{1(DRF)}) + \pi_1 G_1^L(t^L_{0(DRF)})} = \frac{\pi_0 \pi_1 t^L_{1(DRF)} + \pi_1^2 t^L_{0(DRF)} t^L_{1(DRF)}}{\pi_0^2 + \pi_0 \pi_1 t^L_{1(DRF)}}$$

The left side of the equation is $JI_{DR}(t^L_{0(DRF)}, t^L_{1(DRF)})$. Therefore,

$$\frac{\pi_0 \pi_1 t^L_{1(DRF)} + \pi_1^2 t^L_{0(DRF)} t^L_{1(DRF)}}{\pi_0^2 + \pi_0 \pi_1 t^L_{1(DRF)}} = JI_{DR}(t^L_{0(DRF)}, t^L_{1(DRF)}).$$

Given $t_0 \leq t_1$, we get

$$\frac{\pi_0 \pi_1 t^L_{0(DRF)} + \pi_1^2 t^L_{0(DRF)} t^L_{0(DRF)}}{\pi_0^2 + \pi_0 \pi_1 t^L_{0(DRF)}} \leq JI_{DR}(t^L_{0(DRF)}, t^L_{1(DRF)}) \leq \frac{\pi_0 \pi_1 t^L_{1(DRF)} + \pi_1^2 t^L_{1(DRF)} t^L_{1(DRF)}}{\pi_0^2 + \pi_0 \pi_1 t^L_{1(DRF)}}$$

Then we obtain

$$t^L_{0(DRF)} \leq \frac{\pi_0}{\pi_1} JI_{DR}(t^L_{0(DRF)}, t^L_{1(DRF)}) \leq t^L_{1(DRF)}.$$

We can also prove that $F1_{DR}(t_0, t_1)$ is a unimodal function (see Appendix C) and thus $(t^L_{0(DRF)}, t^L_{1(DRF)})$ are the optimal thresholds. □

**Corollary 12.** *Let $(t^B_{0(DRF)}, t^B_{1(DRF)})$ denote the optimal decision-making thresholds for a Bayesian probability transformed score to maximize the $F1$ in the decided region. The following properties hold:*

1. $F1_{DR}(t^B_{0(DRF)}, t^B_{1(DRF)}) = 2t^B_{1(DRF)}/(1 - t^B_{0(DRF)} + t^B_{1(DRF)})$
2. $t^B_{0(DRF)} \leq F1_{DR}(t^B_{0(DRF)}, t^B_{1(DRF)})/2 \leq t^B_{1(DRF)}$.

*Proof.* According to Proposition 2, (Eqs. 7.3-7.4) can be combined and simplified as:

$$\frac{g_1^B(t_1)\frac{dt_1}{dt_0}}{\pi_0 g_0^B(t_1)\frac{dt_1}{dt_0} - \pi_1 g_1^B(t_0)} = \frac{t_1}{\pi_1(1-t_0-t_1)}.$$

Setting the partial derivative to zero results in:

$$\frac{1 - G_1^B(t^B_{1(DRF)})}{\pi_0 - \pi_0 G_0^B(t^B_{1(DRF)}) + \pi_1 G_1^B(t^B_{0(DRF)})} = \frac{t^B_{1(DRF)}}{\pi_1(1 - t^B_{0(DRF)} - t^B_{1(DRF)})} \tag{7.6}$$

After rearranging (Eq. 7.6), we have

$$\frac{2\pi_1(1 - G_1^B(t^B_{1(DRF)}))}{2\pi_1(1 - G_1^B(t^B_{1(DRF)})) + \pi_0 - \pi_0 G_0^B(t^B_{1(DRF)}) + \pi_1 G_1^B(t^B_{0(DRF)})} = \frac{2t^B_{1(DRF)}}{1 - t^B_{0(DRF)} + t^B_{1(DRF)}}$$

The left side of the equation is $F1_{DR}(t^B_{0(DRF)}, t^B_{1(DRF)})$. Therefore, we have

$$F1_{DR}(t^B_{0(DRF)}, t^B_{1(DRF)}) = \frac{2t^B_{1(DRF)}}{1 - t^B_{0(DRF)} + t^B_{1(DRF)}}.$$

Given $t_0 \leq t_1$, we have

$$t^B_{0(DRF)} \leq \frac{F1_{DR}(t^B_{0(DRF)}, t^B_{1(DRF)})}{2} \leq t^B_{1(DRF)}.$$

We can also prove that $F1_{DR}(t_0, t_1)$ is a unimodal function (see Appendix C) and thus $(t^B_{0(DRF)}, t^B_{1(DRF)})$ are the optimal thresholds. □

### 7.3 Cost-Sensitive Learning in BCAR

In previous discussions in this chapter so far, we assume equal decision costs to optimize the confusion-matrix measures based on a fixed ambiguous region. However, this equal cost assumption is not valid in many real-world applications as we discussed in Chapter 6. Besides, it is difficult to pre-define the size of the ambiguous region. Given the cost matrix of binary classification with ambiguous option, it is important

to apply cost-sensitive analysis to optimize the ambiguous region in order to achieve optimal decision to align with a profit-oriented goal.

Table 7.4 is a customized cost matrix with ambiguous region, where $c_{ij}$ ($i \in \{0,1\}, j \in \{0,1\}$) is the cost of predicting class $i$ when the true class is $j$ and $c_{uj}$ is the cost of decision to assign the instances into ambiguous category when the true class is $j$. $c_{01}$ and $c_{10}$ represent the misclassification costs associated with false negatives and false positives.

| | | **Actual Class** | |
|---|---|---|---|
| | | $Y = 1$ | $Y = 0$ |
| **Predicted Class** | $\hat{Y} = 1$ | $c_{11}$ | $c_{10}$ |
| | $\hat{Y} = u$ | $c_{u1}$ | $c_{u0}$ |
| | $\hat{Y} = 0$ | $c_{01}$ | $c_{00}$ |

Table 7.4: Cost Matrix for BCAR

We make the following reasonableness assumptions regarding the cost matrix: it costs more to assign an instance into ambiguous category than to a correct category but less than to assign to an incorrect category. That is

$$c_{00} < c_{u0} < c_{10} \text{ and } c_{11} < c_{u1} < c_{01}. \tag{7.7}$$

Given an instance with score $s$, let $H_0(t_0, t_1)$ be the expected cost for predicting negative, $H_u(t_0, t_1)$ be the expected cost for predicting ambiguous, and $H_1(t_0, t_1)$ be the

expected cost for predicting positive. We have

$$
\begin{aligned}
H_0(t_0,t_1) &= P(Y=0|s<t_0)c_{00} + P(Y=1|s<t_0)c_{01} \\
&= \frac{P(Y=0)P(s<t_0|Y=0)}{P(s<t_0)}c_{00} + \frac{P(Y=1)P(s<t_0|Y=1)}{P(s<t_0)}c_{01}, \\
H_u(t_0,t_1) &= P(Y=0|t_0\le s<t_1)c_{u0} + P(Y=1|t_0\le s<t_1)c_{u1} \\
&= \frac{P(Y=0)P(t_0\le s<t_1|Y=0)}{P(t_0\le s<t_1)}c_{u0} + \frac{P(Y=1)P(t_0\le s<t_1|Y=1)}{P(t_0\le s<t_1)}c_{u1}, \\
H_1(t_0,t_1) &= P(Y=0|s\ge t_1)c_{00} + P(Y=1|s\ge t_1)c_{01} \\
&= \frac{P(Y=0)P(s\ge t_1|Y=0)}{P(s\ge t_1)}c_{10} + \frac{P(Y=1)P(s\ge t_1|Y=1)}{P(s\ge t_1)}c_{11}.
\end{aligned}
$$

It follows that the overall expected cost is given by:

$$
\begin{aligned}
H(t_0,t_1) &= P(s<t_0)H_0(t_0,t_1) + P(t_0\le s<t_1)H_u(t_0,t_1) + P(s\ge t_1)H_1(t_0,t_1) \\
&= \pi_0 G_0^L(t_0)c_{00} + \pi_1 G_1^L(t_0)c_{01} \\
&\quad + \pi_0(G_0^L(t_1) - G_0^L(t_0))c_{u0} + \pi_1(G_1^L(t_1) - G_1^L(t_0))c_{u1} \\
&\quad + \pi_0(1-G_0^L(t_1))c_{10} + \pi_1(1-G_1^L(t_1))c_{11}.
\end{aligned}
$$

We are interested in finding the optimal thresholds $(t_0,t_1)$ to minimize the expected overall cost:

$$
\min_{t_0,t_1} \quad H(t_0,t_1) \text{ s.t. } \quad 0\le t_0\le t_1. \tag{7.8}
$$

**Proposition 15.** *For any given instance, the optimal threshold pair $(t_0^L, t_1^L)$ of optimization problem 7.8 is given by*

$$
(t_0^L,t_1^L) = \begin{cases} \left(\frac{\pi_0(c_{00}-c_{u0})}{\pi_1(c_{u1}-c_{01})}, \frac{\pi_0(c_{10}-c_{u0})}{\pi_1(c_{u1}-c_{11})}\right), & \text{if } \frac{c_{00}-c_{u0}}{c_{u1}-c_{01}} \le \frac{c_{10}-c_{u0}}{c_{u1}-c_{11}} \\ \left(\frac{\pi_0(c_{10}-c_{u0})}{\pi_1(c_{u1}-c_{11})}, \frac{\pi_0(c_{10}-c_{u0})}{\pi_1(c_{u1}-c_{11})}\right), & \text{otherwise.} \end{cases} \tag{7.9}
$$

*Proof.* By setting the following partial derivatives to zero,

$$\frac{\partial H(t_0,t_1)}{\partial t_0} = \pi_0 g_0^L(t_0)c_{00} + \pi_1 g_1^L(t_0)c_{01} - \pi_0 g_0^L(t_0)c_{u0} - \pi_1 g_1^L(t_0)c_{u1} \tag{7.10}$$

$$\frac{\partial H(t_0,t_1)}{\partial t_1} = \pi_0 g_0^L(t_1)c_{u0} + \pi_1 g_1^L(t_1)c_{u1} - \pi_0 g_0^L(t_1)c_{10} - \pi_1 g_1^L(t_1)c_{11}, \tag{7.11}$$

we obtain

$$\frac{g_1^L(t_0^L)}{g_0^L(t_0^L)} = \frac{\pi_0(c_{00}-c_{u0})}{\pi_1(c_{u1}-c_{01})}$$
$$\frac{g_1^L(t_1^L)}{g_0^L(t_1^L)} = \frac{\pi_0(c_{10}-c_{u0})}{\pi_1(c_{u1}-c_{11})}.$$

According to Proposition 1, we have

$$t_0^L = \frac{\pi_0(c_{00}-c_{u0})}{\pi_1(c_{u1}-c_{01})}$$
$$t_1^L = \frac{\pi_0(c_{10}-c_{u0})}{\pi_1(c_{u1}-c_{11})}.$$

(Eqs. 7.10 and 7.11) can be re-expressed as:

$$\frac{\partial H(t_0,t_1)}{\partial t_0} = g_0^L(t_0)(\pi_0 c_{00} + \pi_1 t_0 c_{01} - \pi_0 c_{u0} - \pi_1 t_0 c_{u1}) \tag{7.12}$$

$$\frac{\partial H(t_0,t_1)}{\partial t_1} = g_0^L(t_1)(\pi_0 c_{u0} + \pi_1 t_1 c_{u1} - \pi_0 c_{10} - \pi_1 t_1 c_{11}) \tag{7.13}$$

Using the assumptions about the costs (7.7), we can prove that when $t_0 \leq t_0^L$, (Eq. 7.12)$\leq 0$; else, (Eq. 7.12)$> 0$. And when $t_1 \leq t_1^L$, (Eq. 7.13)$\leq 0$; else, (Eq. 7.13)$> 0$. Therefore, $H(t_0,t_1)$ is a unimodal function of $t_0$ and $t_1$.

In addition, assumptions (7.7) guarantee that $0 \leq t_0^L$ and $0 \leq t_1^L$. To ensure that $t_0^L \leq t_1^L$, the following condition must hold:

$$(c_{00}-c_{u0})(c_{u1}-c_{11}) \geq (c_{10}-c_{u0})(c_{u1}-c_{01}).$$

If the above condition does not hold, since $H(t_0,t_1)$ is monotonically decreasing when

$t_0 \leq t_0^L$, the optimal threshold pair of optimization problem (7.8) becomes $(t_1^L,\ t_1^L)$, which implies that the ambiguous region shrinks to $\emptyset$. □

**Corollary 13.** *For any given instance, the optimal threshold pair $(t_0^B, t_1^B)$ of optimization problem (7.8) under the Bayesian transformation is given by*

$$(t_0^B, t_1^B) = \begin{cases} \left(\frac{c_{00}-c_{u0}}{c_{00}-c_{u0}+c_{u1}-c_{01}}, \frac{c_{10}-c_{u0}}{c_{10}-c_{u0}+c_{u1}-c_{11}}\right), & \text{if } \frac{c_{00}-c_{u0}}{c_{u1}-c_{01}} \leq \frac{c_{10}-c_{u0}}{c_{u1}-c_{11}} \\ \left(\frac{c_{10}-c_{u0}}{c_{10}-c_{u0}+c_{u1}-c_{11}}, \frac{c_{10}-c_{u0}}{c_{10}-c_{u0}+c_{u1}-c_{11}}\right), & \text{otherwise.} \end{cases} \tag{7.14}$$

*Proof.* The corollary can be proven by following the same procedure as Proposition 15 and by applying Proposition 2. □

## 7.4 Summary

Most binary classification models are based on binary decisions, which means the outcomes are either positive or negative. In many real-world applications, binary classification with ambiguous region (BCAR) is more suitable for decision making. We expanded the model performance measures in general binary classification to a binary classification with ambiguous region. In addition, We explored the relationships between the optimal measures and thresholds & base rate in BCAR. Finally, we expanded the threshold-based cost sensitive learning to BCAR.

# Chapter 8

# DYNAMIC ENSEMBLE OF TWO CLASSIFIERS

Ensemble learning is an integration of individually trained classifiers, seeking to achieve more accurate predictive performance than any of the single classifiers in the ensemble. Currently, ensemble learning is regarded as the state-of-the art approach for solving a plethora of machine learning challenges such as imbalanced data (Sun et al., 2009; Fernández et al., 2018; Yu et al., 2018; Tanha et al., 2020), overfitting issue (Ying, 2019; Papouskova and Hajek, 2019; Yu et al., 2020), curse of dimensionality (Sagi and Rokach, 2018; Mejri and Mejri, 2020; Wu et al., 2020; Lim, 2020). The main idea behind ensemble methodology is from the wisdom of crowds, which states that aggregation of information from a group will lead to better decision than any individual, if certain conditions are satisfied. Surowiecki (2005) pointed out that not all groups are wise and in order to form a wise group, several key requirements are necessary, including diversity of opinions, independence, decentralization, and aggregation mechanism. To meet those requirements, generating the base classifiers plays a crucial role in ensemble learning. There are five main strategies to generate a diverse pool of classifiers (Cruz et al., 2018):

- Different initializations: Different initializations may result in different classifiers with varying predictive accuracy. For example, the initialization is a key step in neural network model training and different initialization methods may converge differently.

- Different hyperparameters: The base classifiers are constructed with different combinations of the hyper-parameters, which results in different model architecture. For instance, the combination of neural network classifiers are trained with distinct number of layers, number of neurons and activation functions.

- Different classification algorithms: This method refers to the combination of different classification algorithms. For example, a new model can be formed by combining logistic regression, decision tree, K-Nearest Neighbor, and Support Vector Machine.

- Different training sets: Each base classifier is trained using a subset of the training set. Typical examples include bagging, boosting and clustering-based classifiers.

- Different feature space: This methodology is often used in situations where the data can be represented in different feature spaces.

Common ensembles include bootstrap aggregating (also known as bagging), boosting, and stacking (or call it meta-learner): Bootstrap aggregating is a method to generate multiple classifiers based on bootstrap samplings, which are random sampling with replacement. The multiple classifiers are combined via average for regression models or majority voting for classification models. A well-established bootstrap aggregation ensemble method is random forest (Breiman, 1996a); Boosting means sequentially building a series of weak learners (e.g., decision tree) to combine into a strong learner. Each weak learner aims to correct the errors of its predecessor. Most widely used boosting methods are AdaBoost (Rojas et al., 2009), XGBoost (Chen et al., 2015), LightGBM (Ke et al., 2017), CatBoost (Dorogush et al., 2018), etc.; Stacking involves creating two-level models. The first level is to build different types of base models and the second level is to train a learning algorithm to combine the predictions of those base models. For example, Li and Chen (2020) built a logistic regression to combine the outputs from random forest and LightGBM. Given bootstrap and boosting methods are well-established with many well-known variants (Breiman, 1996a; Rojas et al., 2009;

Chen and Guestrin, 2017; Ke et al., 2017; Dorogush et al., 2018) and the research on stacking is more ad-hoc in nature (Breiman, 1996b), our study in this chapter mainly focuses on stacking.

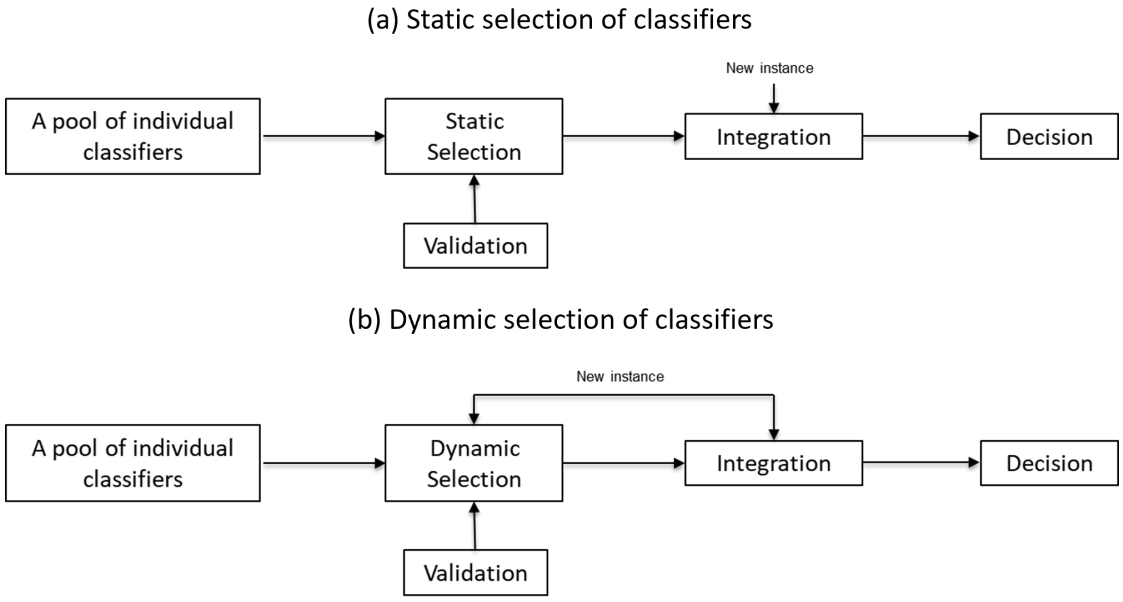


Figure 8.1: Static vs. Dynamic Classifiers Selection

The combination strategies in the ensembles can be conceptually grouped into two categories: static or dynamic. In static combination approaches, the ensemble of classifiers is selected based on the training or validation data (Figure 8.1a). Commonly used static classifier combination methods are majority voting (Breiman, 1996a), stacked regressions (Breiman, 1996b), feature-weighted linear stacking (Sill et al., 2009), Bayesian model average (Hoeting et al., 1999), etc. In the dynamic combination approaches, the combination is based on each new instance to be classified (Figure 8.1b). The rationale behind dynamic combination is that not every classifier in the pool is an expert in classifying all unknown samples; rather, each base classifier is an expert in a different local region of the feature space (Zhu et al., 2004). There are a myriad of research on dynamic combination methods based on various dynamic combination criteria. Majority of these research focuses either on how to define the local regions such as K-Nearest Neighbors (Ko et al., 2008; Wang and Mao, 2020; Mahfouz

et al., 2021), clustering techniques (Kuncheva, 2000; Dong et al., 2020; Casa et al., 2020), etc., or on how to define criteria to measure the level of competence of base classifiers, such as local accuracy estimates and ranking (Cruz et al., 2017). Those research studies are ad-hoc and lack solid systematical theoretical support. Furthermore, empirically those dynamic ensemble classifiers show no better performance than static ensemble (Ko et al., 2008; Dong et al., 2020).

We also observe that very few papers discuss the density-based Bayesian dynamic selection due to the impractical estimation of density function for high-dimensional data (Wang and Scott, 2019). To make the computation efficient, Naive Bayes based selection has been discussed relying on conditional independence of each classifier (Li and Hao, 2009; Bang and Wu, 2016). However, the performance of Naive Bayes is questionable due to easy violation of conditional independence assumption (Rennie et al., 2003). It is commonly believed that the computation efficiency and estimation accuracy of the density function will be significantly enhanced when the ensemble size is reduced. A recent study (Bonab and Can, 2019) advocates that there is an optimal number of component classifiers in the ensemble that gives the highest accuracy. The study indicates that the optimal number of component classifiers equal to the number of classes, which provided support for the research of the ensemble of two classifiers for binary classification. Therefore, we propose a density-based likelihood dynamic ensemble of two classifiers, which can be theoretically and empirically proven to be superior over any of single classifiers in the ensemble in terms of ROC, a key performance measure in evaluating machine learning classification.

This chapter is structured as follows: Section 8.1 introduces our proposed dynamic ensemble of two classifiers and theoretically proves that the ensemble classifiers perform better than any single classifier in the ensemble in terms of ROC. In addition, this section outlines the estimation procedure with pseudo code; Section 8.2 reviews the kernel-based density estimation method, which aids the estimation of density functions; Section 8.3 provides several simulated examples to demonstrate the effectiveness of our proposed methods; Section 8.4 tests our ensemble method on real-world data,

which shows satisfactory results; Section 8.5 summarizes our findings.

## 8.1 A Dynamic Ensemble of Two Classifiers

Let us consider an ensemble classifier with two models in binary classification, where a combination of two classification algorithms are applied to predict the class label of an instance. The two classification algorithms generate predictive score $X_1 \in R$ and $X_2 \in R$ for the instance, respectively. Assume $X_j (j \in \{1, 2\})$ is a continuous random variable following a probability density distribution $f^j(x_j)$ and cumulative distribution function $F^j(x_j)$. The effectiveness of the single classification algorithm is determined by the conditional distributions of the predictive score $X_j$ for both classes: $f_i^j(x_j) \equiv f^j(x_j|Y = i), i \in \{0, 1\}$. Clearly, $f^j(x_j) = \pi_0 f_0^j(x_j) + \pi_1 f_1^j(x_j)$. Let $S_j$ denote the corresponding likelihood ratio transformation (Eq. 3.6) of $X_j$ and $g_i^{Lj}(s_j)$ be the class-conditional density functions. According to (Eq. 5.3), the ROC curves associated with $X_j$ and $S_j$ are defined as $ROC(f_0^j, f_1^j)$ and $ROC(g_0^{Lj}, g_1^{Lj})$, respectively.

Let $f_i(x_1, x_2)$ $i \in \{0, 1\}$ represent the joint conditional density probability. The marginal probability density function can be obtained by integrating the joint conditional probability distribution over $X_j$. That is $f_i^1(x_1) = \int_{-\infty}^{+\infty} f_i(x_1, x_2)dx_2$ and $f_i^2(x_2) = \int_{-\infty}^{+\infty} f_i(x_1, x_2)dx_1$.

According to Bayes' theorem, the probability of being class 1 given $x_1$ and $x_2$ is:

$$P(Y = 1|x_1, x_2) = \frac{\pi_1 f_1(x_1, x_2)}{\pi_1 f_1(x_1, x_2) + \pi_0 f_0(x_1, x_2)}$$

which is monotonically increasing function of $f_1(x_1, x_2)/f_0(x_1, x_2)$. Define a random variable transformation

$$S_{12} = f_1(X_1, X_2)/f_0(X_1, X_2)$$

and let $g_i^L(s_{12})$ be the corresponding class-conditional density function. The ROC curve associated with $S_{12}$ is expressed as $ROC(g_0^L, g_1^L)$.

Note $s_{12}$ leverages the joint density probability information, which is believed to be more informative than a single model. For the ensemble system, a threshold rule is used to make the class label prediction $\hat{Y}$: given a threshold $t_{12}$, the instance is classified as "positive" or $\hat{Y} = 1$ if $S_{12} \geq t_{12}$, otherwise, it is classified as "negative" or $\hat{Y} = 0$.

We prove that the ROC associated with the ensemble score $S_{12}$ dominates the ROC associated with any single model score.

**Theorem 2.** *$ROC(g_0^L, g_1^L)$ dominates both $ROC(g_0^{Lj}, g_1^{Lj})$ and $ROC(f_0^j, f_1^j)$.*

*Proof.* According to Theorem 1, $ROC(g_0^{Lj}, g_1^{Lj})$ dominates $ROC(f_0^j, f_1^j)$. We only need to prove $ROC(g_0^L, g_1^L)$ dominates $ROC(g_0^{Lj}, g_1^{Lj})$. We prove that $ROC(g_0^L, g_1^L)$ dominates $ROC(g_0^{L1}, g_1^{L1})$ and the same procedure can be followed to prove $ROC(g_0^L, g_1^L)$ dominates $ROC(g_0^{L2}, g_1^{L2})$ . Given two scalars $t_{12}$ and $t_1$ that $FPR^L(t_{12}) = FPR^{L1}(t_1)$, the sub regions of the $x_1x_2$-space ($R_L(t_{12})$) where the instances are assigned as positive by $s_{12}$ are given by

$$R_L(t_{12}) = \{(x_1, x_2) | f_1(x_1, x_2) \geq t_{12} f_0(x_1, x_2)\}. \tag{8.1}$$

Based on the marginal conditional density function transformation, the sub regions in $x_1x_2$-space ($(R_{L1}(t_1))$) where instances are assigned as positive by $s_1$ are given by

$$\begin{aligned} R_{L1}(t_1) &= \{x_1 | f_1^1(x_1) \geq t_1 f_0^1(x_1)\} \\ &= \{(x_1, x_2) | \int_{-\infty}^{+\infty} f_1(x_1, x_2) dx_2 \geq t_1 \int_{-\infty}^{+\infty} f_0(x_1, x_2) dx_2\}. \end{aligned} \tag{8.2}$$

Therefore, the false positive rates are

$$\begin{aligned} FPR^L(t_{12}) &= P(R_L(t_{12}) | Y = 0) \\ &= \int \int_{\{(x_1,x_2) | f_1(x_1,x_2) \geq t_{12} f_0(x_1,x_2)\}} f_0(x_1, x_2) dx_1 dx_2 \end{aligned}$$

and

$$FPR^{L1}(t_1) = P(R_{L1}(t_1)|Y=0)$$
$$= \int\int_{\{(x_1,x_2)|\int_{-\infty}^{+\infty} f_1(x_1,x_2)dx_2 \geq t_1 \int_{-\infty}^{+\infty} f_0(x_1,x_2)dx_2\}} f_0(x_1,x_2)dx_1dx_2.$$

Note that

$$R_L(t_{12}) = (R_L(t_{12}) \cap R_{L1}(t_1)) \cup (R_L(t_{12}) \cap R^c_{L1}(t_1)),$$
$$R_{L1}(t_1) = (R_{L1}(t_1) \cap R_L(t_{12})) \cup (R_{L1}(t_1) \cap R^c_L(t_{12})),$$

where the superscript $c$ indicates the complement of the set. Since $FPR^L(t_{12}) = FPR^{L1}(t_1)$, we have

$$P(R_L(t_{12}) \cap R^c_{L1}(t_1)|Y=0) = P(R_{L1}(t_1) \cap R^c_L(t_{12})|Y=0). \tag{8.3}$$

For true positive rates, we need to prove $TPR^L(t_{12}) \geq TPR^{L1}(t_1)$, which is equivalent to

$$P(R_L(t_{12}) \cap R^c_{L1}(t_1)|Y=1) \geq P(R_{L1}(t_1) \cap R^c_L(t_{12})|Y=1).$$

Indeed, we have

$$
\begin{aligned}
& P(R_L(t_{12}) \cap R^c_{L1}(t_1)|Y=1) \\
&= \int\int_{R_L(t_{12})\cap R^c_{L1}(t_1)} f_1(x_1,x_2)dx_1dx_2 \\
&\geq \int\int_{R_L(t_{12})\cap R^c_{L1}(t_1)} t_{12}f_0(x_1,x_2)dx_1dx_2 && \text{(because of (Eq. 8.1))} \\
&= t_{12}P(R_L(t_{12}) \cap R^c_{L1}(t_1)|Y=0) \\
&= t_{12}P(R_{L1}(t_1) \cap R^c_L(t_{12})|Y=0) && \text{(because of (Eq. 8.3))} \\
&= t_{12}\int\int_{R_{L1}(t_1)\cap R^c_L(t_{12})} f_0(x_1,x_2)dx_1dx_2 \\
&\geq \int\int_{R_{L1}(t_1)\cap R^c_L(t_{12})} f_1(x_1,x_2)dx_1dx_2 && \text{(reversal logic of (Eq. 8.1))} \\
&= P(R_{L1}(t_1) \cap R^c_L(t_{12})|Y=1)
\end{aligned}
$$

and this completes the proof. □

Below is the procedure for our density-based ensemble algorithm.

**Algorithm 1** Density-Based Ensemble Procedure

**Input:**

$\zeta$ is the training data,
$Y$ is true label of the training data,
$\zeta_i$ are the instances with $Y = i, i \in \{0, 1\}$ in the training data,
$\mathcal{L}_1, \mathcal{L}_2$ are the classification algorithms,
$KD$ is the joint kernel density estimation algorithm.

1: Let $E_1$, $E_2$ be machine learning models learned by applying $\mathcal{L}_1, \mathcal{L}_2$ to $\zeta$
2: $X_{1i} = E_1(\zeta_i) \& X_{2i} = E_2(\zeta_i), i \in \{0, 1\}$
3: Let $\mathcal{KD}_i$ be the kernel density estimation engine learned by applying $KD$ to $(X_{1i}, X_{2i}), i \in \{0, 1\}$
4:
5: **for** a testing instance $\iota$ **do**
6: $x_1 = E_1(\iota) \& x_2 = E_2(\iota)$
7: $f_i(x_1, x_2) = \mathcal{KD}_i(x_1, x_2), i \in \{0, 1\}$
8: Calculate the ensemble score $s_{12} = f_1(x_1, x_2)/f_0(x_1, x_2)$
9: **end for**

## 8.2 Kernel-based Density Estimation

Let ($\chi_1$, $\chi_2$,...,$\chi_n$) be instances randomly sampled from a univariate distribution with an unknown density $f$ at any given point $\chi$. The kernel density estimator is defined as:

$$\hat{f}_h(\chi) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{K}_h(\chi - \chi_i) = \frac{1}{nh}\sum_{i=1}^{n}\mathcal{K}(\frac{\chi - \chi_i}{h}), \tag{8.4}$$

where $\mathcal{K}$ is the kernel function and $h > 0$ is a smoothing parameter known as the bandwidth. A kernel with subscript $h$ is called the scaled kernel and defined as $\mathcal{K}_h(\chi) = \frac{1}{h}\mathcal{K}(\frac{\chi}{h})$. The selection of bandwidth $h$ plays a crucial role in the estimation accuracy. This area is well studied with numerous published papers (Sheather and Jones, 1991; Terrell and Scott, 1992; Kim and Scott, 2012). Commonly used kernel functions including uniform, Epanechnikov, triangular, biweight, triweight, Gaussian, and others. Given the density estimation is not the purpose of our study, we mainly use Gaussian kernel with default settings in the Python Statsmodels package.

## 8.3 Simulations

For model $j$, $j \in \{1, 2\}$, let $X_{j1}$ be a random variable for positive instance scores and $X_{j0}$ be a random variable for negative instance scores. Assume $[X_{10}\ X_{20}] \sim N(\mu_0, Q_0)$ and $[X_{11}\ X_{21}] \sim N(\mu_1, Q_1)$. We randomly sample $n = 40000$ instances, where the number of positive instances $n_1 = 10000$ and the number of negative instances $n_0 = 30000$. The data is split into training/testing by 80/20 ratio. Training data is used for density estimation and testing is purely for validating.

Here we list two cases to illustrate effectiveness of our proposed ensemble strategy. Consider case 1 (Figure 8.2): $\mu_0 = \begin{bmatrix}6 & 5\end{bmatrix}$ $\mu_1 = \begin{bmatrix}6 & 8\end{bmatrix}$ $Q_0 = \begin{bmatrix}3 & 1.25 \\ 1.25 & 3\end{bmatrix}$ $Q_1 = \begin{bmatrix}2 & 1.99 \\ 1.99 & 2\end{bmatrix}$ and case 2 (Figure 8.3): $\mu_0 = \begin{bmatrix}4 & 6\end{bmatrix}$ $\mu_1 = \begin{bmatrix}6 & 8\end{bmatrix}$ $Q_0 = \begin{bmatrix}3 & 2.77 \\ 2.77 & 3\end{bmatrix}$ $Q_1 = \begin{bmatrix}2 & 1.94 \\ 1.94 & 2\end{bmatrix}$. The left part of the figures is the probability density distribution by each class and model. For the right part, in addition to ensemble score ROC, ROCs associated with each individual model score ($x$) as well as the transformed score ($s$) (refer to Section 5.3) were plotted as well.

In case 1, model 1 provides little separation power over random guessing, although ROC with transformed score $s_1$ slightly improves (AUC 0.55 vs 0.50). According to convex hull theory (Majnik and Bosnić, 2013), model 2 dominates model 1 and the sub-optimal model 1 should not be selected. However, our proposed ensemble shows better performance than any single model in the ensemble and our results indicate that even the model with poor performance can still provide additional value to achieve a better model.

In case 2, all individual scores as well as their transformed scores exhibit similar performance and the transformation does not improve the ROC given the original ROCs are concave (see Section 5.3 for details). Our proposed ensemble demonstrates significant improvement over any single model (AUC improvement from 0.81 to 0.94).

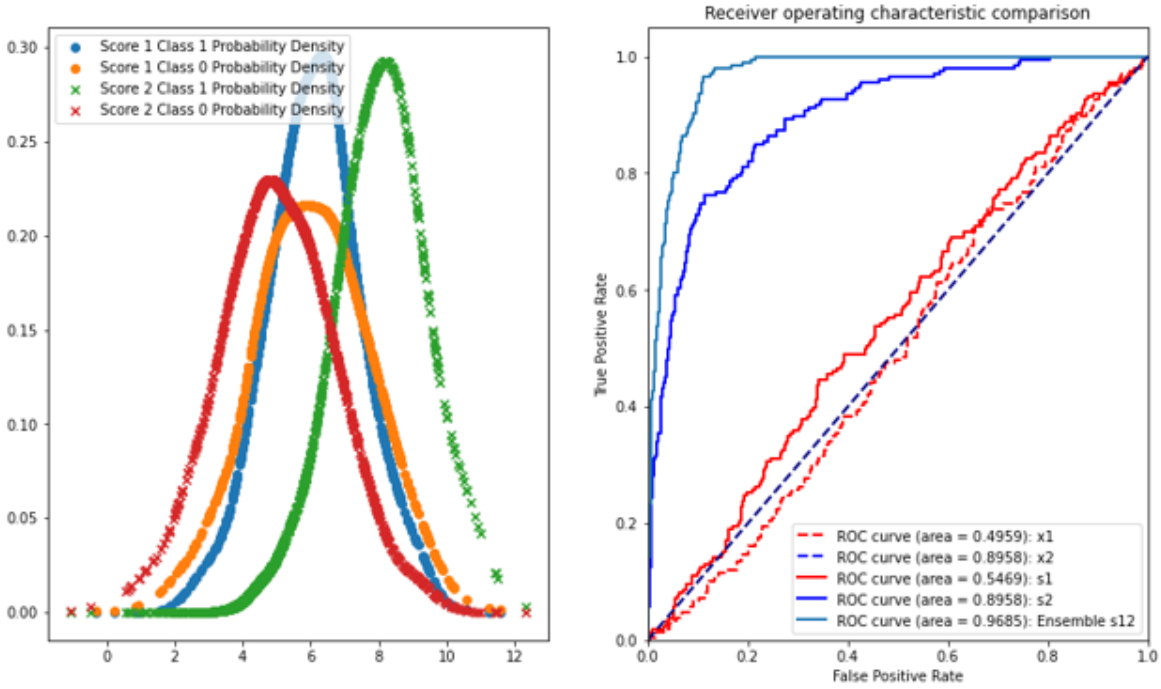


Figure 8.2: Simulated Case 1

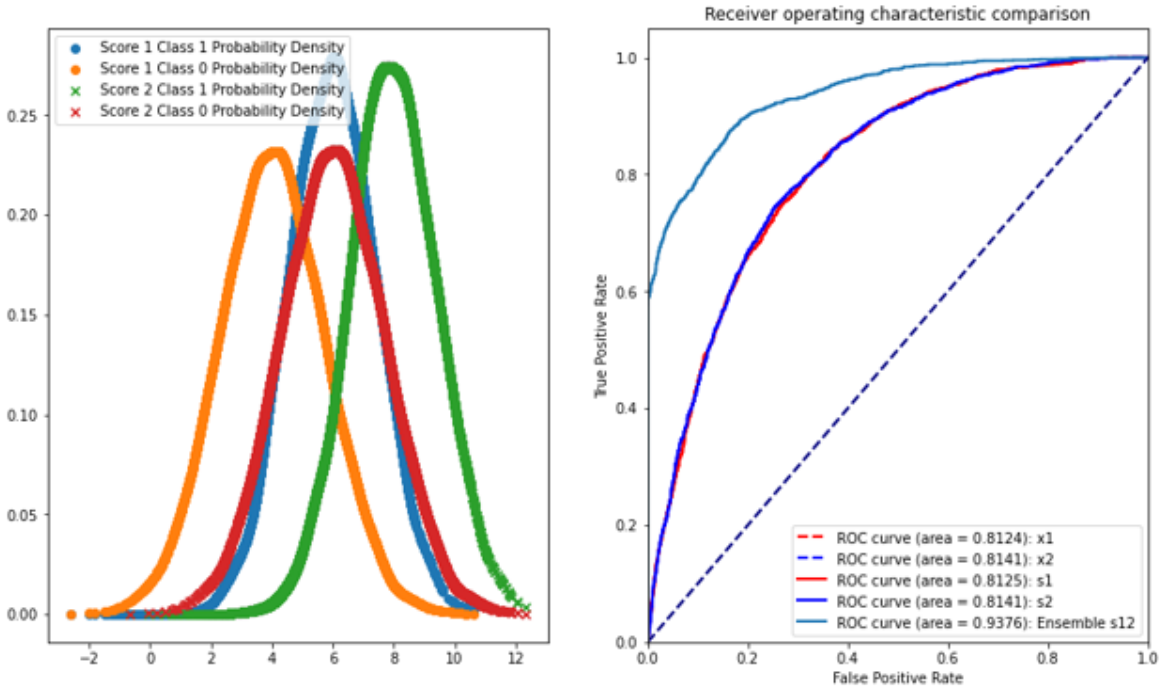


Figure 8.3: Simulated Case 2

## 8.4 Empirical Study with Real-world Data

In this section, we apply our ensemble strategy to real-world data to test its applicability. The data we use is from Lending Club, which is a fintech company to provide personal loans. Their loan data information is publicly available. The data can also be downloaded from Kaggle (Lending Club). Its 2016Q1 loan data was used with 66,362 observations and 70 variables. The models were built to predict the likelihood of loan default. The data was randomly split into training/testing by 80/20 ratio. The training data was for model development and density estimation and testing data was purely for testing. Multiple machine learning models were developed, including Logistic Regression (LR), Decision Tree (DT), and Neural Network (NN). Density estimation was based on Kernel Density Estimator in Python Statsmodels package. Our testing results are documented in Table 8.1. AUC ($s_1$) stands for the AUC for the likelihood ratio transformed score for the first model in column 1, while AUC ($s_2$) is for the second one. Table 8.1 shows that our proposed method demonstrates higher AUC than any

single model in the ensemble.

| Model 1 + Model 2 | AUC($s_1$) | AUC($s_2$) | Proposed AUC($s_{12}$) |
|---|---|---|---|
| LR + DT | 0.6611 | 0.6645 | 0.6803 |
| LR + NN | 0.6611 | 0.6486 | 0.6658 |
| DT + NN | 0.6645 | 0.6486 | 0.6791 |

Table 8.1: Performance Comparison between Proposed Ensemble and Single Model in the Pool

## 8.5 Summary

Density-based ensemble methods have not been paid sufficient attention due to numerical difficulties in estimating high-dimensional probability density. To address the computational concerns, we proposed a likelihood-ratio-based dynamic ensemble of two classifiers for binary classification. We demonstrate theoretically and empirically that our method performs better than any single classifier in the ensemble in terms of ROC/AUC.

# Part III

# APPLICATIONS

# Chapter 9

# CONSUMER LENDING OPTIMIZATION SYSTEM

As of Aug 2021, the Federal Reserve Bank estimated the value of consumer loans at all commercial banks to be $1.594 trillion in the United States. According to IMARC Group's latest report, the consumer loan market is expected to grow at a compound annual growth rate of around 5% in next five years. As the market size is huge and continues to grow, a consumer lending optimization system will result in significant positive financial impact. A pivotal element in consumer lending optimization system is pricing. The importance of having the right pricing strategies in place has been highlighted by many practitioners in the financial industry (PricewaterhouseCoopers, 2012; Goyal et al., 2016; Experian, 2018). For example, PricewaterhouseCoopers (2012) points out that "As companies realize the inefficiency and unreliability of the traditional loan pricing strategies, we believe a trend toward a pricing optimization" in loan pricing; Goyal et al. (2016) from Boston Consulting Group (BCG) estimates that a structured pricing setting at least increases revenue between 7% and 10%; Experian (2018) states that loan pricing optimization can help lenders significantly improve profitability and efficiency while achieving a superior customer experience. However, the pricing optimization in consumer lending draws very little attention in academic area and there are only very few papers discussing this problem. See the limited examples in (Edelberg, 2006; Phillips, 2013; Terblanche and De La Rey, 2014; Chun and Lejeune, 2020).

Risk-based pricing is the dominant approach in current lending industry (Edelberg, 2006; Chun and Lejeune, 2020), where the price (or interest rate) is dependent on the customer's credit risk plus a markup profit. The rationale behind risk-based pricing is simple. In order to compensate the potential loss of the principals, a lender

should charge a higher interest rate for the customer the lender deems a higher risk borrower. Therefore, accurately estimating the risk is of the crucial importance in risk-based pricing. Most of the literature, however, on the topic of credit pricing optimization fails to consider the accuracy of the risk estimation in their framework by simply assuming 'true risk' or accepting the probability output from machine learning classification models as the true probability of default (Phillips, 2013). In a consumer lending optimization system, the accurate estimation of the credit risk should be considered coupled with pricing.

In addition, traditional risk-based pricing does not consider the price sensitivity and adverse selection, which means different customers may response differently to the offered pricing. The effect of price sensitivity and adverse selection in consumer lending pricing is recognized in (Phillips et al., 2011; Phillips, 2013; Terblanche and De La Rey, 2014). Phillips et al. (2011) reveals that lower-risk customers, on the average, demonstrate more sensitive to pricing than higher-risk customers. As price is increased, lower-risk customers tend to seek out alternatives more quickly than those with higher risk – and vice versa. Since the price sensitivity and adverse selection significantly impact the customers' response to the offered pricing, they should be taken into account as part of the consumer lending optimization system.

Theoretically, an ideal consumer lending optimization system can offer a customized pricing to different customers based on their risk assessment and price sensitivity. Nevertheless, thanks to operational, regulatory and practical concerns, most lenders have a relatively small number of risk categories (Phillips, 2013). For example, lending club divides the customers into A-G seven categories with the A class assigned the lowest interest rate and G class the highest interest rate. Instead of finding the optimal interest rate for each customer by treating the pricing as a continuous variable, a more practical problem in reality is to properly assign the customers to the corresponding risk category and identify optimal pricing for each category to optimize overall profits. Motivated by the limitations of the risk-based pricing and the real-world problem in lending industry, we propose a consumer lending optimization

system by taking risk and pricing sensitivity in a holistic view to determine customers' risk category and pricing.

This chapter is structured as follows: Section 9.1 gives a brief introduction of a typical consumer lending decision process and risk-based pricing in the lending industry. Section 9.2 introduces a cost matrix in consumer lending setting. Section 9.3 formulates the problem, breaks the problem into sub-problems and informatively resolves the problem in the cost-sensitive framework by considering both the risk and customers' response. Section 9.4 designs a simulation experiment and empirically proves that our method can improve the profitability by a significant amount over the traditional risk-based pricing method. Section 9.5 concludes this chapter with discussion of the limitations of our research and potential future research.

## 9.1 Background

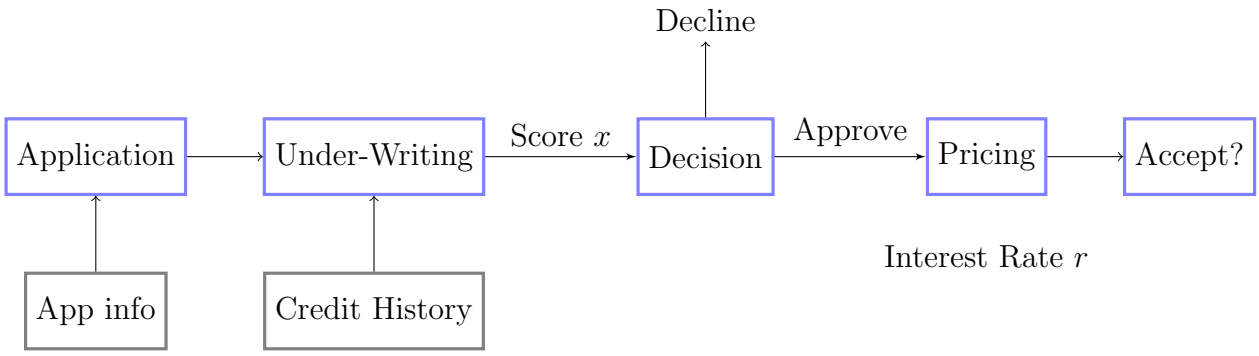


Figure 9.1: Typical Consumer Lending Decision Process, Where $x$ is Estimated Default Probability Derived from Machine Learning Algorithms Based on Applicants' Information

A consumer lending strategy consists of two key decisions: credit approval / denial (or underwriting decision) and pricing. These two decisions are determined in a sequential process as illustrated in Figure 9.1. Lenders collect the information about the customers' profile & their desired loan amount and estimate the risk associated with the application. Based on estimated risk, under-writing credit decisions are made, and

then pricing is offered for approved customers. Customers will decide whether to accept the offer or not. The process details are elaborated as below:

1. Application: A prospective borrower submits an application for a loan. Generally, the lenders require customers to disclose information regarding their personal identity (such as social security number, address, and birth date), financial situation (e.g., annual income, monthly housing payment, employment) and desired loan attributes including loan amount and terms (how long the customers can pay-off the loan). Some lenders may also require loan purpose to assess risk and pricing.

2. Underwriting: The key task in this stage is to quantify the credit risk for the application. In addition to the information customers provided, lenders may seek internal information (if the customers have relationship with the lender), credit history from bureau (e.g., Experian, Equifax, TransUnion), and/or other alternative information from third parties. Based on those accumulated information, the lenders assign each applicant a score generated from machine learning models which indicates the likelihood of default as denoted by $x$ in Figure 9.1. Different lenders have different approaches or data sources to evaluate customers' risk so that the same customer deemed high risky by one lender might be considered low risky by another lender. Developing sophisticated machine learning models to accurately reflect the applicants true default probability plays a very important role in the consumer lending decision. The machine learning approaches used in credit risk modeling in consumer lending have been widely explored. For details see (Thomas, 2000).

3. Credit Decision: The lenders rely on the risk score to decide whether to approve or reject the applicants. There is no specific guideline to choose approximate score cut-off to make the approval / denial decision, which may not align with business objective in profit-oriented organization. Nowadays, most lenders rely on the decision engines like TSYS, TRIAD, etc. to automate the decisions.

Customers can receive the decisions instantly. A small percentage of applicants may be routed for manual review and the approval / denial decisions are made by credit analysts.

4. Pricing: For those approvals, lenders charge different interest rate ($r$) corresponding to the risk levels in terms of the default probability. Risk-based pricing is the most popular approach to pricing consumer lending in United States (Edelberg, 2006). The logic for risk-based approach is simple that a higher interest rate is charged to compensate the cost due to the likelihood of default. Under risk-based pricing framework, the applicants are categorized into different risk categories based on the risk assessment. The interest rate offered in each category is given by $r_i = r_c + m + l_i$, where $r_i$ is the rate offered to the customers in category $i$, $r_c$ is the lender's cost of fund, $z$ is a standard fixed margin and $l_i$ the premium that varies by risk categories to at least compensate the expected loss with the borrowers in that category. For example, given three categories (A, B, C), the corresponding interest rates are (13.99%, 17.99%, 25.99%). Assume $r_c = 2\%$ and $z = 3\%$. When $x <= 8.99\%$, the applicant is assigned to A category, and when $8.99\% < x <= 12.99\%$, then B category, and when $12.99\% < x <= 20.99\%$, then C category. A customers with probability of default greater than 20.99% would be rejected.

5. Accept?: Customers decide whether to accept the lenders' offer or not. The consumer lending is a very competitive market. Most lenders accept online applications that make the customers easier to shop around. Moreover, customers' response to the pricing is ignored in the risk-based pricing, which significantly impacts the lenders' profit. Borrower preferences for lower prices must be recognized in developing a theory for the determination of optimal offers for consumer lending. Phillips (2013) points out adverse selection is another issue in loan decision process since lenders only hold partial information of the borrowers and

those risky borrowers who anticipate pitfalls are more likely to accept the high priced loans.

In summary, the current lending strategy is subject to the following limitations:

- Predictive accuracy of risk estimation plays a pivotal role in current strategy. However, lenders generally accept the probability output from models as the risk estimation and the accuracy of risk estimation is not considered holistically with pricing.

- Customers in different pricing categories demonstrate different profitability. Risk-based pricing is purely based on the estimated risk, which is not aligned with maximizing-profit goal.

- The current consumer lending strategy is a sequential decision process and fails to take the customers' acceptance rate and adverse selection into the optimization decision.

In the following sections, we aim to address those limitations by proposing a comprehensive consumer lending optimization system based on cost-sensitive study.

## 9.2 Cost Matrix in Consumer Lending

In our problem setting, the cost matrix is a profit matrix and we regard default loans as 'positive' instances and non-default loans as 'negative' instances. The acceptance of negative instances results in interest revenue for the lenders, while the approval of positive instances poses the lenders the risk of losing all principal amount. The cost matrix in consumer lending can be represented as Table 9.1, where $c_{i0}$ is the profit of assigning a negative instance to category $i$ and $c_1$ is the cost of accepting a positive instance. According to Wang et al. (2020), the elements of the cost matrix can be formulated as $c_{i0} = r_i - r_c$ and $c_1 = -(LGD + r_c)$, where $r_i$ is the interest rate in category $i$ ($i \in \{1, 2, 3..., n\}$), $r_c$ is the cost of fund and $LGD$ is loss given default. For any category $i < j$, we assume $r_i < r_j$. Both $r_c$ and $LGD$ are considered as constants

in the setting. For those rejected applications, the cash-flow for both negative and positive instances are zero.

| | actual negative | actual positive |
|---|---|---|
| Assigned to Class 1 | $c_{10}$ | $c_1$ |
| Assigned to Class 2 | $c_{20}$ | $c_1$ |
| Assigned to Class 3 | $c_{30}$ | $c_1$ |
| ... | ... | ... |
| Reject (D) | $c_{D0} = 0$ | $c_{D1} = 0$ |

Table 9.1: Cost Matrix for Consumer Lending System

## 9.3 Problem Formulation

The current lending system fails to consider pricing dependent risk, which means the pricing also impacts the customers' ability to pay. The pricing can be incorporated into the customers' risk denoted as (Eq. 9.1):

$$s = P(Y = 1|x, r_i) = \frac{f_{1i}(x)\pi_{1i}}{\pi_{1i}f_{1i}(x) + \pi_{0i}f_{0i}(x)}, \tag{9.1}$$

where $f_{1i}(x) = f(x|Y = 1, r_i)$, $f_{0i}(x) = f(x|Y = 0, r_i)$, $\pi_{1i} = P(Y = 1|r_i)$, and $\pi_{0i} = P(Y = 0|r_i)$. The density function for $s$ is denoted as $g(s)$. Ideally, conditional density functions $f_{0i}\&f_{1i}$ and $\pi_{0i}\&\pi_{1i}$ can be obtained from randomly charging customers different pricing. The random experiment generally is very expensive. Our future research will study how to construct $s$ based on limited/partial data information. To simplify the problem, our discussion in this Chapter focuses on $s$ space by assuming price-independent risk.

To address the drawback that customers' take rate and adverse selection are ignored, we consider a take rate function $TR^A(s, r)$ with adverse impact into account. The function is dependent on risk ($s$) and interest rate ($r$). We have two key assumptions for $TR^A(s, r)$:

1. Risky customers are more willing to accept the offer (adverse selection), which means $\partial TR^A(s, r)/\partial s > 0$.

2. Customers prefer lower interest rate, which means $TR^A(s, r_i) > TR^A(s, r_j)$ for $r_i < r_j$.

Our proposed consumer lending optimization system aims to assign applicants into proper risk category and optimize the pricing in that category as well. Given $n$ risk categories and assuming lenders charge the same interest rate for the same category, the problem can formulated as:

$$\begin{aligned}
\max_{r_1,r_2,...,r_n,s_1,s_2,...,s_n} \quad & R(r_1...,r_n,s_1...,s_n) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} [(1-s)c_{i0} + s * c_1] TR^A(s, r_i) g(s) ds \\
\text{s.t.} \quad & s_0 \leq s_1 \leq s_2 \leq ... \leq s_n \leq 1 \\
& 0 \leq r_1 \leq r_2 \leq ... \leq r_{n-1} \leq r_n \\
& s_0 = 0,
\end{aligned}$$

where $s_{i-1}$ and $s_i$ are the category boundary for interest rate category $r_i$, which means that an applicant with score $s$ between $s_{i-1}$ and $s_i$ will be charged interest rate $r_i$. Let $(n+1)th$ category be the reject category. We propose a divide-and-conquer method to solve this optimization problem step by step:

- Step 1: Given risk categories and fixed interest rate for each category, optimize the category boundary (ST1 - Boundary Problem).
- Step 2: Given category boundary, find the optimal interest rate for each category (ST2 - Pricing Problem).
- Step 3: Find the optimal category boundary and interest rate based on the iterations of Step 1 & 2.

### 9.3.1 ST1 - Boundary Problem

Our first problem can be described as: given $n$ rating categories and interest rate for each category (denote as $r_1$, $r_2$, $r_3 \ldots$, $r_n$), our aim is to assign an applicant

into proper category to maximize the expected profit of the lender. The problem can be formulated as:

$$\begin{aligned} \max_{s_1,s_2,...,s_n} \quad & R(s_1...,s_n) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} [(1-s)c_{i0} + s * c_1] TR^A(s, r_i) g(s) ds \\ \text{s.t.} \quad & s_0 \leq s_1 \leq s_2 \leq ... \leq s_n \leq 1, \\ & s_0 = 0. \end{aligned} \tag{9.2}$$

Firstly, we propose a solution to the optimization problem based on pairwise comparison of any two categories. Then, we identify situations where category boundaries are ordered and the category boundary can be given by the comparison of two adjacent categories via un-constrained optimization.

**Solution based on Pairwise Comparison**

With the basis of the assumption that risk is independent of pricing, the expected profit for an instance being in category $i$ is:

$$R_i(s) = [(1-s)c_{i0} + sc_1] TR^A(s, r_i).$$

Let $s_{ij}$ be the break-even score which leads to $R_i(s) = R_j(s)$. We prove that under certain conditions, $s_{ij}$ exists and is unique.

**Lemma 5.** *Let* $\Phi_s = \{s | \min(R_i(s), R_j(s)) > 0\}$. *Suppose* $TR^A(s, r_i)/TR^A(s, r_j)$ *is a monotonically decreasing function in* $s$ *when* $r_i < r_j$. *If* $R_i(0) > R_j(0)$, *there is a unique solution* $s_{ij}$ *for equation* $R_i(s) = R_j(s), (s \in \Phi_s)$ *and for* $s < s_{ij}$, *we have* $R_i(s) > R_j(s)$ *and* $R_i(s) < R_j(s)$, *otherwise.*

*Proof.* Given $s^*$ such that

$$s^* = \frac{c_{i0}}{c_{i0} - c_1} = \frac{r_i - r_c}{r_i + LGD},$$

we have $R_i(s^*) = 0$. Since $R_j(s^*) = [(1 - s^*)c_{j0} + s^*c_1]TR^A(s^*, r_j)$, we can show $R_j(s^*) > 0$ (because $(1 - s^*)c_{j0} + s^*c_1 > 0$). Given $R_i(0) > R_j(0)$ and $R_i(s)$ is a continuous function, it is guaranteed at least one solution for $R_i(s) = R_j(s)$. Note that

$$R_i(s) = R_i(s) \longrightarrow \frac{(1-s)c_{i0} + s * c_1}{(1-s)c_{j0} + s * c_1} \frac{TR^A(s, r_i)}{TR^A(s, r_j)} = 1$$

It is easy to prove that $[(1-s)c_{i0}+s*c_1]/[(1-s)c_{j0}+s*c_1]$ is a monotonically decreasing function of $s$. Given $TR^A(s, r_i)/TR^A(s, r_j)$ is also a monotonically decreasing function of $s$, $R_i(s)/R_i(s)$ is a monotonically decreasing function of $s$. Therefore, there is unique solution to $R_i(s) = R_j(s)$. In addition, for $s < s_{ij}$, we have $R_i(s) > R_j(s)$ and otherwise, $R_i(s) < R_j(s)$ for $s > s_{ij}$. □

It is worth noting that the assumption that $TR^A(s, r_i)/TR^A(s, r_j)$ is a monotonically decreasing function of $s$ when $i < j$ requires that

$$\frac{TR^{A'}(s, r_i)}{TR^{A'}(s, r_j)} < \frac{TR^A(s, r_i)}{TR^A(s, r_j)}.$$

Since $TR^A(s, r_i)/TR^A(s, r_j) > 1$, $TR^{A'}(s, r_i)/TR^{A'}(s, r_j) < 1$ can satisfy the assumption, which means the adverse selection effect is more obvious in higher interest rate region than low interest rate region. The assumption is valid in line with business intuition.

To further illustrate the Lemma, we performed a simulation study (see results in Figure 9.2) based on a sigmoid take rate function

$$TR^A(s, r) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 s + \beta_2 r)}},$$

where $\beta_1 > 0$ and $\beta_2 < 0$ that satisfies the two assumptions of $TR^A(s, r)$. The simulation results show that any pair of $R_i(s), R_j(s)$ intersects and only intersects once when $s \in \Phi_s$. It also demonstrate that when $s < s_{ij}$, the expected profit in category $i$ $(R_i(s))$ dominates the one in category $j$ $(R_j(s))$.

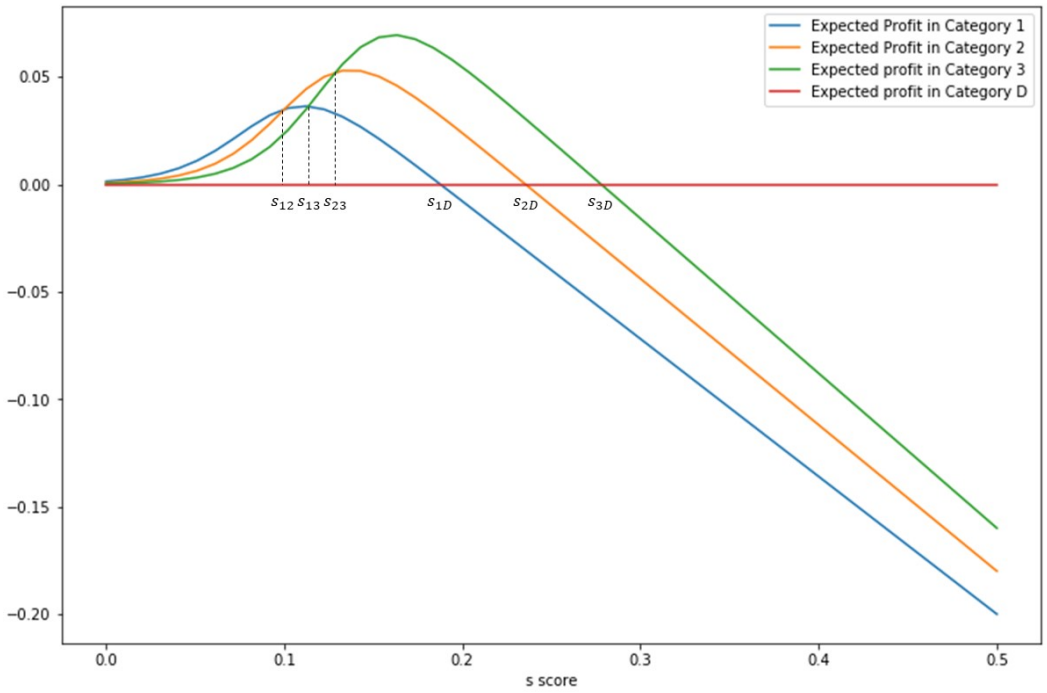


Figure 9.2: Simulation Results Based on Sigmoid Take Rate Function (1)

According to Lemma 5, each pair of categories has a break-even score, with a total of $n(n-1)/2$ scores. These break-even scores are represented as:

$$s_1^{BE} = \{s_{12}, s_{13}, s_{14}..., s_{1i}..., s_{1n}\}$$
$$s_2^{BE} = \{s_{23}, s_{24}..., s_{2i}..., s_{2n}\}$$
$$.....$$
$$s_i^{BE} = \{s_{i(i+1)}..., s_{in}\}$$
$$.....$$
$$s_{n-1}^{BE} = \{s_{(n-1)n}\}$$

Suppose $s_{ik}, s_{ij} \in s_i^{BE}, i < k < j$. We reveal that if $s_{ik} > s_{ij}$, categories between $i$ and $j$ are inferior and should be out of consideration since we can either assign the instance to category $i$ or $j$ with higher profit than any categories between $i$ and $j$.

**Proposition 16.** *Suppose* $i < k < j$ *and* $R_i(0) > R_k(0) > R_j(0)$. *If* $s_{ik} > s_{ij}$, *then* $\max(R_i(s), R_j(s)) > R_k(s) \forall s \in \Phi_s = \{s | \min(R_i(s), R_j(s)) \geq 0\}$.

*Proof.* According to Lemma 5, we have $R_i(s) > R_k(s)$ for $s < s_{ik}$.

- Firstly, we show $s_{kj} < s_{ik}$. In light of Lemma 5, we have $R_j(s) > R_i(s) > R_k(s)$ for $s_{ij} < s < s_{ik}$. Given $R_k(0) > R_j(0)$, $R_k(s)$ and $R_j(s)$ intersect before $s_{ik}$. Therefore, $s_{kj} < s_{ik}$.

- Given $s_{kj} < s_{ik}$, we have $R_i(s) > R_k(s)$ for $s < s_{kj}$ and $R_j(s) > R_k(s)$ for $s > s_{kj}$. Therefore, $\max(R_i(s), R_j(s)) > R_k(s)$.

Hence, the proposition is proven. □

Figure 9.3 indicates that when $s_{13} < s_{12}$, we have $\max(R_1(s), R_3(s)) > R_2(s)$ for $s \in \Phi_s$, which means category 2 is inferior. The simulation results support the Proposition.

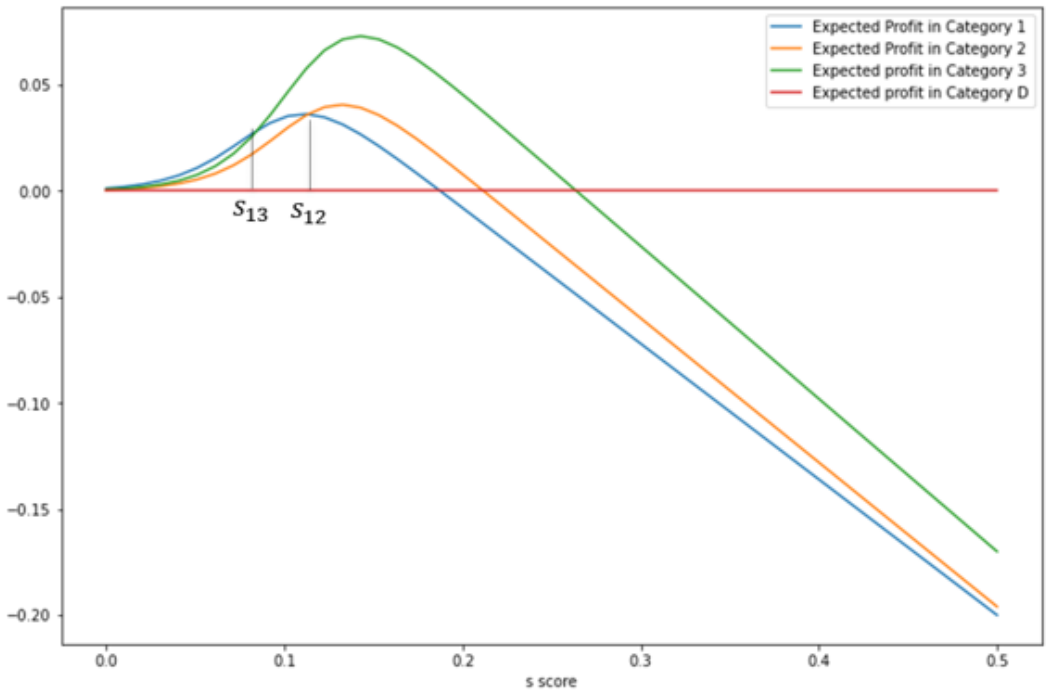


Figure 9.3: Simulation Results Based on Sigmoid Take Rate Function (2)

Next we show the optimal category boundaries from pairwise comparison are ordered.

**Proposition 17.** *Suppose* $s_i^* = \min\{s_{i(i+1)}, s_{i(i+2)}, ..., s_{in}\}$ *and* $s_{ij} = s_i^*$ *for some* $j \in \{i+1, i+2, ..., n\}$ *and* $s_j^* = \min(s_{j(j+1)}, s_{j(j+2)}, ..., s_{jn})$*, we have* $s_i^* < s_j^*$.

*Proof.* We prove the proposition by contradiction. Suppose $k \in \{j+1, j+2, ..., n\}$ such that $s_{jk} = s_j^*$. We have $s_{ik} > s_{ij}$ by definition. If $s_i^* \geq s_j^*$, then $s_{ik} > s_{ij} \geq s_{jk}$. According to Lemma 5, if $s_{ij} < s < s_{ik}$, then $R_j(s) > R_i(s) > R_k(s)$.

On the other hand, because $s_{ij} \geq s_{jk}$, if $s > s_{ij}$, we have $s > s_{jk}$, which leads to $R_j(s) < R_k(s)$ in light of Lemma 5.

Therefore, it is contradictory and the proposition is proven. □

According to Propositions 16 and 17, we show that the pairwise comparison results in ordered category boundaries, which are solution of the objective function (Eq. 9.2).

### Solution based on Un-constrained Optimization

In previous section, we prove that the category boundary can be achieved from pairwise comparison. Let $s_i^*$ denote the category boundary. In this subsection, we disclose the conditions that results in $s_0^* < s_1^* < s_2^* < s_3^* < ..., < s_n^*$ and $s_0^* = 0$.

Assume take rate is independent of risk (no adverse selection impact) and let $TR(r)$ denote the take rate function. The objective function can be expressed as:

$$R(s_1, s_2..., s_n) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} [(1-s)c_{i0} + s * c_1] TR(r_i) g(s) ds. \tag{9.3}$$

**Proposition 18.** *If take rate function* $TR(r)$ *is a log-concave function and* $rTR(r)$ *is a monotone decreasing function, then the optimal boundary for (Eq. 9.3) satisfies that* $s_i^* < s_{i+1}^*, i \in \{0, 1, ..., n-1\}$.

*Proof.* We have

$$\begin{aligned}\frac{\partial R(s_1, s_2..., s_n)}{\partial s_i} &= [(1-s_i)c_{i0} + s_i * c_1]TR(r_i)g(s_i) - [(1-s_i)c_{(i+1)0} + s_i * c_1]TR(r_{i+1})g(s_i) \\ &= D_s(s_i)g(s_i),\end{aligned}$$

where $D_s(s_i) = [(1-s_i)c_{i0} + s_i * c_1]TR(r_i) - [(1-s_i)c_{(i+1)0} + s_i * c_1]TR(r_{i+1})$. It is easy to show that $D_s(s_i)$ is monotonically deceasing when $rTR(r)$ is a monotone decreasing function, which means $R(s_1, s_2..., s_n)$ is a unimodal function.

Setting $\partial R(s_1, s_2..., s_n)/\partial s_i$ to zero leads to

$$\frac{TR(r_i)}{TR(r_{i+1})} = \frac{(1-s_i^*)c_{(i+1)0} + s_i^* c_1}{(1-s_i^*)c_{i0} + s_i^* c_1}.$$

Re-arranging the equation results in:

$$s_i^* = \frac{TR(r_{i+1})r_{i+1} - TR(r_i)r_i + r_c(TR(r_i) - TR(r_{i+1}))}{TR(r_{i+1})r_{i+1} - TR(r_i)r_i + LGD(TR(r_{i+1}) - TR(r_i))} \tag{9.4}$$

Let $M_i = TR(r_{i+1})r_{i+1} - TR(r_i)r_i$ and $N_i = TR(r_{i+1}) - T(r_i)$. In order to prove $s_i^* < s_{i+1}^*$, we need

$$\frac{M_i - r_c N_i}{M_i + LGDN_i} < \frac{M_{i+1} - r_c N_{i+1}}{M_{i+1} + LGDN_{i+1}}. \tag{9.5}$$

Since $rTR(r)$ is a monotone decreasing function,

$$M_i = r_{i+1}TR(r_{i+1}) - r_i TR(r_i) < 0.$$

If (Eq. 9.5) holds, we need

$$\frac{M_i}{N_i} < \frac{M_{i+1}}{N_{i+1}}.$$

Recall

$$\frac{M_i}{N_i} = \frac{TR(r_{i+1})r_{i+1} - TR(r_i)r_i}{TR(r_{i+1}) - TR(r_i)}.$$

Since $r_{i+1} = r_i + \Delta r$ and, if $\Delta r \to 0$, then $TR(r_{i+1}) \approx TR(r_i) + \Delta r TR'(r_i)$, we have

$$\frac{M_i}{N_i} = \frac{TR(r_i) + TR'(r_i)r_i}{TR'(r_i)} + \Theta,$$

where $\Theta \to 0$. Define $\psi(r) = (TR(r) + TR'(r)r)/TR'(r)$. We prove that if $TR(r)$ is log-concave, $\psi(r)$ is monotone increasing function.

$$\begin{aligned}(\psi(r))' &= \frac{(2TR'(r) + TR''(r)r)TR'(r) - TR''(r)(TR(r) + TR'(r)r)}{(TR'(r))^2} \\ &= \frac{2(TR'(r))^2 - TR''(r)TR(r)}{(TR'(r))^2}\end{aligned}$$

We prove that if

$$(TR'(r))^2 - TR''(r)TR(r) > 0,$$

then $\psi(r)$ is a monotonically increasing function, which means $M_i/N_i < M_{i+1}/N_{i+1}$. So, if $rTR(r)$ is monotone decreasing and $TR(r_i)$ is a log-concave function (because of $(TR'(r_i))^2 - TR''(r_i)TR(r_i) > 0$), we prove that $s_i^* < s_{i+1}^*$. □

### 9.3.2 ST2 - Pricing Problem

Our second problem aims to find optimal pricing for each category given the category boundary. Assume no adverse select impact, which means that take rate is independent of risk. The problem can be formulated as:

$$\begin{aligned}\max_{r_1, r_2, ..., r_n} \quad & R(r_1, r_2, ..., r_n) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} [(1-s)c_{i0}(r_i) + s * c_1] TR(r_i) g(s) ds \qquad (9.6) \\ \text{s.t.} \quad & 0 \le r_1 \le r_2 \le ... \le r_n\end{aligned}$$

**Proposition 19.** *If $TR(r)$ is concave, then objective function $R(\cdot)$ (Eq. 9.6) is concave and the optimal interest rate $r_i^*$ satisfies the following condition:*

$$[c_{i0} TR^{'}(r_i^*) + TR(r_i^*)] \int_{s_{i-1}}^{s_i} g(s) ds = [TR(r_i^*) + TR^{'}(r_i^*)(c_{i0} - c_1)] \int_{s_{i-1}}^{s_i} s g(s) ds.$$

*Proof.* In order to achieve the maximum value of the objective function, calculate the first and second order derivative:

$$\frac{\partial R(r_1, r_2, ..., r_n)}{\partial r_i} = TR^{'}(r_i) \int_{s_{i-1}}^{s_i} [(1-s)c_{i0}(r_i) + s * c_1]g(s)ds + TR(r_i) \int_{s_{i-1}}^{s_i} (1-s)g(s)ds$$
$$\frac{\partial^2 R(r_1, r_2, ..., r_n)}{\partial r_i^2} = TR^{''}(r_i) \int_{s_{i-1}}^{s_i} [(1-s)c_{i0}(r_i) + s * c_1]g(s)ds + 2TR^{'}(r_i) \int_{s_{i-1}}^{s_i} (1-s)g(s)ds$$

Since $TR'(r_i) < 0$ and $(1-s)c_{i0}(r_i) + s * c_1 > 0$, if $TR^{''}(r_i) < 0$, then $R(\cdot)$ is a concave function. And to optimize $R(\cdot)$, it requires:

$$[c_{i0}TR^{'}(r_i^*) + TR(r_i^*)] \int_{s_{i-1}}^{s_i} g(s)ds = [TR(r_i^*) + TR^{'}(r_i^*)(c_{i0} - c_1)] \int_{s_{i-1}}^{s_i} sg(s)ds \quad (9.7)$$

Rearranging the equation, we have

$$\frac{\int_{s_{i-1}}^{s_i} sg(s)ds}{\int_{s_{i-1}}^{s_i} g(s)ds} = \frac{c_{i0}TR^{'}(r_i^*) + TR(r_i^*)}{T(r_i^*) + TR^{'}(r_i^*)(c_{i0} - c_1)} = \phi(r_i^*)$$

$\phi(\cdot)$ is a monotone increasing function of $r$

Since

$$s_{i-1} \leq \phi(r_i^*) = \frac{\int_{s_{i-1}}^{s_i} sg(s)ds}{\int_{s_{i-1}}^{s_i} g(s)ds} \leq s_i, \quad (9.8)$$

$$s_i \leq \phi(r_{i+1}^*) = \frac{\int_{s_i}^{s_{i+1}} sg(s)ds}{\int_{s_i}^{s_{i+1}} g(s)ds} \leq s_{i+1}, \quad (9.9)$$

we have $\phi(r_i^*) \leq \phi(r_{i+1}^*)$. Since $\phi(r)$ is monotone increasing function, $r_i^* \leq r_{i+1}^*$. □

### 9.3.3 ST3 - Iterations of ST1 and ST2

With fixed interest rate in each category, our first step is to find an ordered category boundary for each category, which is given by (Eq. 9.10). An instance with score $s$ will be assigned to category $i$, if $s_{i-1} \leq s \leq s_i$. We call this step as optimal cut-off method. Given the category boundary, optimal interest rates in each category

need to meet the condition of (Eq. 9.11). The iteration of step 1 and step 2 result in the consumer lending optimization system.

$$s_i = \frac{M_i - r_c N_i}{M_i + LGDN_i} \tag{9.10}$$

$$[c_{i0} TR^{'}(r_i) + TR(r_i)] \int_{s_{i-1}}^{s_i} g(s)ds = [TR(r_i) + TR^{'}(r_i)(c_{i0} - c_1)] \int_{s_{i-1}}^{s_i} sg(s)ds, \tag{9.11}$$

where $M_i = T(r_{i+1})r_{i+1} - T(r_i)r_i$ and $N_i = T(r_{i+1}) - T(r_i)$.

For (Eq. 9.11), we have applied approximation methods, where $T'(r_i)$ is directly calculated from the take rate function, $\int_{s_{i-1}}^{s_i} sg(s)ds$ is the average $s$ score between $s_{i-1}$ and $s_i$, and $\int_{s_{i-1}}^{s_i} g(s)ds$ is approximate density given by trapezoidal rule. The detailed proposed procedure is described as follows:

**Algorithm 2** Proposed Procedure

1: Score engines $E$ learned from the training data after applying machine learning algorithms
2: $E$ score the training data to get $X$
3: Estimate the density functions of $X$ as well as the conditional density function for instances in class $i \in \{0, 1\}$: $f_i(X)$
4: **for** a testing instance $z$ **do**
5: Apply score engines $E$ to score the testing instance to get score $x$
6: Apply $f_0(X)$ and $f_1(X)$ to the instance to estimate score $s = \pi_1 f_1(x)/(\pi_1 f_1(x) + \pi_0 f_0(x))$
7: **end for**
8: **for** risk category $i$ **do**
9: Calculate $m_i = T(r_{i+1})r_{i+1} - T(r_i)r_i$ and $n_i = T(r_{i+1}) - T(r_i)$
10: According (Eq. 9.10), calculate the decision boundary for category $i$ $s_i$
11: **end for**
12: The maximum boundary is determined by (max rate-$r_c$)/(max rate+LGD)
13: Given the boundary determined by Step 9-13, solve (Eq. 9.11) to achieve the corresponding optimal interest rate.
14: Iterate the process (Steps 9-14) until the optimal interest rate is not changed.

## 9.4 Simulations

### Data Preparation

We use Python make_classification() function to create a synthetic binary classification problem with sample size (100k) and total 20 input features with 15 informative features, 2 redundant features and 3 noisy features. make_classification() is a popular python package, which is used to generate a random n-class classification problem. The detailed introduction of this package can be referred to (Guyon, 2003). Standard data pre-processing was applied to the data and a logistic regression was built to output a probability random variable $X$. The true probability of an instance is defined as $X + \epsilon$, where $\epsilon \sim N(0, 0.05)$. Assume 7 categories with given interest rates $[0.10, 0.22, 0.28, 0.30, 0.35, 0.40, 0.45]$ and the maximum legitimate interest rate is 0.5. Suppose take rate function

$$T(r) = -r^4 - \frac{14.2}{8}r + 1,$$

which is a concave function. The take or not take decision is based on the take rate function, which is only related to interest rate. Let $r_c = 0.03$ and $LGD = 0.5$.

### Risk-based Method

According to the risk-based method, the fixed margin plays a important role. We vary the fixed margin from 0.02 to 0.14 and the profit ranges from \$1,082 to \$1,586. Approval rate is defined as number of approvals divided by all applicants and enrolled rate measures the ratio of customers who accept the offers among the approvals. Results in Table 9.2 shows that high approval rate and enrollment rate don't result in highest profitability. Margin 0.10 brings the highest profitability with approval rate 59.1%, but only 26.5% approvals accept the offer.

| Margin | Approved % | Enrolled % | $ Profit |
|---|---|---|---|
| 0.02 | 69.7% | 43.2% | $1,082 |
| 0.04 | 69.1% | 42.2% | $1,145 |
| 0.06 | 66.8% | 35.0% | $1,552 |
| 0.08 | 63.2% | 29.5% | $1,564 |
| 0.10 | 59.1% | 26.5% | $1,586 |
| 0.12 | 57.2% | 24.6% | $1,537 |
| 0.14 | 53.8% | 20.4% | $1,450 |

Table 9.2: Risk-Based Method Results Varying by Margin

**Results Summary and Discussion**

Table 9.3 summarizes the simulation results based on risk-based method, optimal cut-off, and optimization system. We compare several metrics across these three methods: approval rate, acceptance rate, enrolled rate, average interest rate charged for enrolled, total profit, and profit per enrolled. Table 9.3 shows that the optimal cut-off method by assigning the customer into a proper category without changing the interest rate adds 7% more profit compared with traditional risk-based method, while optimization system can more than double the profit.

It is worth noting that the optimization system approves the most customers (74.4% approval rate) and 78.3% of the approvals accept the offer. The final enrollment rate of optimization system is more than double than risk-based method since our proposed method systematically considers customers' response rate, which explains the significant improvement of the profit. In addition, we note that the average interest rates across three method are similar.

| | Risk-based method | Optimal cut-off | Optimization system |
|---|---|---|---|
| Approved % | 59.1% | 73.3% | 74.4% |
| Accepted % | 44.9% | 40.1% | 78.3% |
| Enrolled % | 26.5% | 29.4% | 58.3% |
| Average rate % | 30.5% | 32.9% | 34.7% |
| $ Profit | $1,586 | $1,708 | $3,539 |
| $ Profit per enrolled | $0.199 | $0.194 | $0.202 |

Table 9.3: Simulated Results

## 9.5 Summary and Discussion

Current industry practice to pricing consumer lending in United States is risk-based pricing, which is purely based on the estimated risk and not aligned with maximizing profit goal. Besides, the current consumer lending strategy is a sequential decision process and fails to take the customers' acceptance rate and adverse selection into the optimization decision. To address the limitations of current consumer lending strategy, we proposed a consumer lending optimization system, which can significantly improve the efficiency and profitability of current lending system.

We divide the problems into several sub-problems and solve the problems based on assumptions. In problem 1, we show pairwise comparison can solve the problem with consideration of adverse selection impact (riskier customers have higher acceptance rate) with the fewest assumptions. We also identify conditions which will result in ordered boundary by assuming no-adverse impact and the acceptance rate function is a log-concave function. For problem 2, in addition to non-adverse-selection impact assumption, we further tighten the assumption of a concave acceptance function instead of log-concave function.

It is worth noting that all our discussion is based on $s$ space and assume price-independent risk, which means the pricing does not impact the risk. In our future research, a potential area is to construct $s$ based on the information in $x$ space and take the price-dependent risk into account.

# BIBLIOGRAPHY


Alemayehu, D. and Zou, K. H. (2012). Applications of roc analysis in medical research: recent developments and future directions. *Academic radiology*, 19(12):1457–1464.

Awoyemi, J. O., Adetunmbi, A. O., and Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pages 1–9. IEEE.

Bahnsen, A. C., Aouada, D., and Ottersten, B. (2014). Example-dependent cost-sensitive logistic regression for credit scoring. In *2014 13th International conference on machine learning and applications*, pages 263–269. IEEE.

Bahnsen, A. C., Stojanovic, A., Aouada, D., and Ottersten, B. (2013). Cost sensitive credit card fraud detection using bayes minimum risk. In *2013 12th international conference on machine learning and applications*, volume 1, pages 333–338. IEEE.

Ballabio, D., Grisoni, F., and Todeschini, R. (2018). Multivariate comparison of classification performance measures. *Chemometrics and Intelligent Laboratory Systems*, 174:33–44.

Bang, S.-J. and Wu, W. (2016). Naive bayes ensemble: A new approach to classifying unlabeled multi-class asthma subjects. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 460–465. IEEE.

Bardenet, R., Brendel, M., Kégl, B., and Sebag, M. (2013). Collaborative hyperparameter tuning. In *International conference on machine learning*, pages 199–207. PMLR.

Beck, J. R. and Shultz, E. K. (1986). The use of relative operating characteristic (roc) curves in test performance evaluation. *Archives of pathology & laboratory medicine*, 110(1):13–20.

Bhavsar, H. and Ganatra, A. (2012). A comparative study of training algorithms for supervised machine learning. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(4):2231–2307.

Böken, B. (2021). On the appropriateness of platt scaling in classifier calibration. *Information Systems*, 95:101641.

Bonab, H. and Can, F. (2019). Less is more: A comprehensive framework for the number of components of ensemble classifiers. *IEEE Transactions on neural networks and learning systems*, 30(9):2735–2745.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311.

Boyd, K., Eng, K. H., and Page, C. D. (2013). Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 451–466. Springer.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Breiman, L. (1996a). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (1996b). Stacked regressions. *Machine learning*, 24(1):49–64.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Campagner, A., Cabitza, F., and Ciucci, D. (2019). Three–way classification: Ambiguity and abstention in machine learning. In *International Joint Conference on Rough Sets*, pages 280–294. Springer.

Canbek, G., Sagiroglu, S., Temizel, T. T., and Baykal, N. (2017). Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 821–826. IEEE.

Cao, C., Chicco, D., and Hoffman, M. M. (2020). The mcc-f1 curve: a performance evaluation technique for binary classification. *arXiv preprint arXiv:2006.11278*.

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832.

Casa, A., Scrucca, L., and Menardi, G. (2020). Better than the best? answers via model ensemble in density-based clustering. *Advances in Data Analysis and Classification*, pages 1–25.

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, 538(7623):20.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., and Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215.

Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae). *Geoscientific Model Development Discussions*, 7(1):1525–1534.

Chai, Z. and Zhao, C. (2019). Enhanced random forest with concurrent analysis of static and dynamic nodes for industrial fault classification. *IEEE Transactions on Industrial Informatics*, 16(1):54–66.

Chase Lipton, Z., Elkan, C., and Narayanaswamy, B. (2014). Thresholding classifiers to maximize f1 score. *arXiv e-prints*, pages arXiv–1402.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, T. and Guestrin, C. (2017). Xgboost: Reliable large-scale tree boosting system. *Availiable online: http://learningsys. org/papers/LearningSys_2015_paper_32. pdf (accessed on 9 October 2019)*.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Chun, S. Y. and Lejeune, M. A. (2020). Risk-based loan pricing: Portfolio optimization approach with marginal risk contribution. *Management Science*, 66(8):3735–3753.

Cohen, I. and Goldszmidt, M. (2004). Properties and benefits of calibrated classifiers. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 125–136. Springer.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Correa Bahnsen, A., Aouada, D., and Ottersten, B. (2015). Ensemble of example-dependent cost-sensitive decision trees. *arXiv e-prints*, pages arXiv–1505.

Cruz, R. M., Sabourin, R., and Cavalcanti, G. D. (2017). Meta-des. oracle: Meta-learning and feature selection for dynamic ensemble selection. *Information fusion*, 38:84–103.

Cruz, R. M., Sabourin, R., and Cavalcanti, G. D. (2018). Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41:195–216.

Das, S., Dey, A., Pal, A., and Roy, N. (2015). Applications of artificial intelligence in machine learning: review and prospect. *International Journal of Computer Applications*, 115(9).

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

De Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & technology*, 31(4):525–541.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Deng, K., Bourke, C., Scott, S., and Vinodchandran, N. (2006). New algorithms for optimizing multi-class classifiers via roc surfaces. In *Proceedings of the ICML workshop on ROC analysis in machine learning*, pages 17–24.

Diab, S. and Sartawi, B. (2017). Classification of questions and learning outcome statements (los) into blooms taxonomy (bt) by similarity measurements towards extracting of learning outcome from learning material. *arXiv preprint arXiv:1706.03191*.

Dietterich, T. G. et al. (2002). Ensemble learning. *The handbook of brain theory and neural networks*, 2:110–125.

Dinga, R., Penninx, B. W., Veltman, D. J., Schmaal, L., and Marquand, A. F. (2019). Beyond accuracy: measures for assessing machine learning models, pitfalls and guidelines. *BioRxiv*, page 743138.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 155–164. ACM.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14(2):241–258.

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Drummond, C. and Holte, R. C. (2000). Explicitly representing expected cost: An alternative to roc representation. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 198–207.

Drummond, C., Holte, R. C., et al. (2003). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer.

Duan, J., Asteris, P. G., Nguyen, H., Bui, X.-N., and Moayedi, H. (2021). A novel artificial intelligence technique to predict compressive strength of recycled aggregate concrete using ica-xgboost model. *Engineering with Computers*, 37(4):3329–3346.

Edelberg, W. (2006). Risk-based pricing of interest rates for consumer loans. *Journal of monetary Economics*, 53(8):2283–2298.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd.

Erdreich, L. S. and Lee, E. T. (1981). Use of relative operating characteristic analysis in epidemiology: a method for dealing with subjective judgement. *American journal of epidemiology*, 114(5):649–662.

Experian (2018). Price optimization in retail consumer lending. https://www.experian.com/decision-analytics/loan-pricing-software.

Fan, W., Stolfo, S. J., Zhang, J., and Chan, P. K. (1999). Adacost: misclassification cost-sensitive boosting. In *Icml*, volume 99, pages 97–105.

Fang, F. and Chen, Y. (2019). A new approach for credit scoring by directly maximizing the kolmogorov–smirnov statistic. *Computational Statistics & Data Analysis*, 133:180–194.

Fatourechi, M., Ward, R. K., Mason, S. G., Huggins, J., Schloegl, A., and Birch, G. E. (2008). Comparison of evaluation metrics in classification applications with imbalanced datasets. In *2008 seventh international conference on machine learning and applications*, pages 777–782. IEEE.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from imbalanced data sets*, volume 10. Springer.

Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38.

Feurer, M. and Hutter, F. (2019). Hyperparameter optimization. In *Automated Machine Learning*, pages 3–33. Springer, Cham.

Feurer, M., Springenberg, J., and Hutter, F. (2015). Initializing bayesian hyperparameter optimization via meta-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Fieldsend, J. E. and Everson, R. M. (2005). Formulation and comparison of multi-class roc surfaces.

Flach, P. (2004). The many faces of roc analysis in machine learning. *ICML Tutorial*.

Fluss, R., Faraggi, D., and Reiser, B. (2005). Estimation of the youden index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(4):458–472.

George, N. I., Lu, T.-P., and Chang, C.-W. (2016). Cost-sensitive performance metric for comparing multiple ordinal classifiers. *Artificial intelligence research*, 5(1):135.

Gong, M. (2021). A novel performance measure for machine learning classification. *International Journal of Managing Information Technology*.

Goodenough, D. J., Rossmann, K., and Lusted, L. B. (1974). Radiographic applications of receiver operating characteristic (roc) curves. *Radiology*, 110(1):89–95.

Goyal, D., Karthikeyan, S., Kulkarni, V., Noguera, V., and Wachters, I. (2016). "moneyball" in commercial lending: From art to science in pricing. https://www.bcg.com/publications/2016/financial-institutions-moneyball-pricing-commercial-loans.

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics*, volume 1. Wiley New York.

Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). Knn model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 986–996. Springer.

Guyon, I. (2003). Design of experiments of the nips 2003 variable selection benchmark. In *NIPS 2003 workshop on feature extraction and feature selection*, volume 253.

Habibzadeh, F., Habibzadeh, P., and Yadollahie, M. (2016). On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia medica*, 26(3):297–307.

Halligan, S., Altman, D. G., and Mallett, S. (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European radiology*, 25(4):932–939.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Herbei, R. and Wegkamp, M. H. (2006). Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 709–721.

Hilden, J. (1991). The area under the roc curve and its competitors. *Medical Decision Making*, 11(2):95–101.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical science*, 14(4):382–417.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Hua, J. and Tian, L. (2020). A comprehensive and comparative review of optimal cut-points selection methods for diseases with multiple ordinal stages. *Journal of Biopharmaceutical Statistics*, 30(1):46–68.

Huang, J. and Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., and Xu, W. (2018). Applications of support vector machine (svm) learning in cancer genomics. *Cancer Genomics-Proteomics*, 15(1):41–51.

Humbird, K. D., Peterson, J. L., and McClarren, R. G. (2018). Deep neural network initialization with decision trees. *IEEE transactions on neural networks and learning systems*, 30(5):1286–1295.

Iranmehr, A., Masnadi-Shirazi, H., and Vasconcelos, N. (2019). Cost-sensitive support vector machines. *Neurocomputing*, 343:50–64.

Janssens, A. C. J. and Martens, F. K. (2020). Reflection on modern methods: revisiting the area under the roc curve. *International Journal of Epidemiology*.

Japkowicz, N. et al. (2000). Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68, pages 10–15. Menlo Park, CA.

Jiao, Y. and Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology*, 4(4):320–330.

Kannan, R. and Vasanthi, V. (2019). Machine learning algorithms with roc curve for predicting and diagnosing the heart disease. In *Soft Computing and Medical Bioinformatics*, pages 63–72. Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154.

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., and Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, 9:109960–109975.

Kim, J. and Scott, C. D. (2012). Robust kernel density estimation. *The Journal of Machine Learning Research*, 13(1):2529–2565.

Ko, A. H., Sabourin, R., and Britto Jr, A. S. (2008). From dynamic classifier selection to dynamic ensemble selection. *Pattern recognition*, 41(5):1718–1731.

Kotsiantis, S. B., Zaharakis, I., Pintelas, P., et al. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC curves for continuous data*. Crc Press.

Kuncheva, L. I. (2000). Clustering-and-selection model for classifier combination. In *KES'2000. Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies. Proceedings (Cat. No. 00TH8516)*, volume 1, pages 185–188. IEEE.

Landgrebe, T. and Duin, R. (2006). A simplified extension of the area under the roc to the multiclass domain. In *Seventeenth annual symposium of the pattern recognition association of South Africa*, pages 241–245.

Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., and Ng, A. Y. (2011). On optimization methods for deep learning. In *ICML*.

Leeflang, M. M., Moons, K. G., Reitsma, J. B., and Zwinderman, A. H. (2008). Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clinical chemistry*, 54(4):729–737.

Lei, J. (2014). Classification with confidence. *Biometrika*, 101(4):755–769.

Li, H., Parikh, N. A., and He, L. (2018). A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes. *Frontiers in neuroscience*, 12:491.

Li, K. and Hao, L. (2009). Naïve bayes ensemble learning based on oracle selection. In *2009 Chinese Control and Decision Conference*, pages 665–670. IEEE.

Li, Y. and Chen, W. (2020). A comparative performance assessment of ensemble learning for credit scoring. *Mathematics*, 8(10):1756.

Lijmer, J. G., Hunink, M. G., van den Dungen, J. J., Loonstra, J., and Smit, A. J. (1996). Roc analysis of noninvasive tests for peripheral arterial disease. *Ultrasound in medicine & biology*, 22(4):391–398.

Lim, J. S. (2020). *Ensemble learning of high dimension datasets*. PhD thesis, The University of Waikato.

Ling, C. X., Huang, J., and Zhang, H. (2003a). Auc: a better measure than accuracy in comparing learning algorithms. In *Conference of the canadian society for computational studies of intelligence*, pages 329–341. Springer.

Ling, C. X., Huang, J., and Zhang, H. (2003b). Auc: a statistically consistent and more discriminating measure than accuracy. In *IJCAI*, volume 3, pages 519–524.

Lipton, Z. C., Elkan, C., and Narayanaswamy, B. (2014). Thresholding classifiers to maximize f1 score. *arXiv preprint ArXiv:1402.1892*, 14.

Liu, X. (2012). Classification accuracy and cut point selection. *Statistics in medicine*, 31(23):2676–2686.

Liu, Y. and Lin, L. (2019). Classification with minimum ambiguity under distribution heterogeneity. *Journal of Statistical Computation and Simulation*, 89(12):2239–2260.

Longadge, R. and Dongre, S. (2013). Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*.

Lusted, L. B. (1971). Signal detectability and medical decision-making. *Science*, 171(3977):1217–1219.

Mahfouz, M. A., Shoukry, A., and Ismail, M. A. (2021). Eknn: Ensemble classifier incorporating connectivity and density into knn with application to cancer diagnosis. *Artificial Intelligence in Medicine*, 111:101985.

Majnik, M. and Bosnić, Z. (2013). Roc analysis of classifiers in machine learning: A survey. *Intelligent data analysis*, 17(3):531–558.

Markowitz, J. and Swets, J. A. (1967). Factors affecting the slope of empirical roc curves: Comparison of binary and rating responses. *Perception & Psychophysics*, 2(3):91–100.

Martinez-Camblor, P., Corral, N., Rey, C., Pascual, J., and Cernuda-Morollón, E. (2017). Receiver operating characteristic curve generalization for non-monotone relationships. *Statistical methods in medical research*, 26(1):113–123.

Marzban, C. (2004). The roc curve and the area under it as performance measures. *Weather and Forecasting*, 19(6):1106–1114.

Masnadi-Shirazi, H., Vasconcelos, N., and Iranmehr, A. (2012). Cost-sensitive support vector machines. *arXiv preprint arXiv:1212.0975*.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, 100(11):2175–2199.

Mease, D., Wyner, A. J., and Buja, A. (2007). Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research*, 8(Mar):409–439.

Medlock, C. and Oppenheim, A. (2019). Optimal roc curves from score variable threshold tests. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5327–5330. IEEE.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Mejri, M. and Mejri, A. (2020). Randomforestmlp: an ensemble-based multi-layer perceptron against curse of dimensionality. *arXiv preprint arXiv:2011.01188*.

Metz, C. E. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier.

Mitchell, T. (1997). Machine learning.

More, A. (2016). Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. H., and Dokania, P. K. (2020). Calibrating deep neural networks using focal loss. *arXiv preprint arXiv:2002.09437*.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., and Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6):275–285.

Nami, S. and Shajari, M. (2018). Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors. *Expert Systems with Applications*, 110:381–392.

Narkhede, S. (2018). Understanding auc-roc curve. *Towards Data Science*, 26.

Nevin, J. A. (1969). Signal detection theory and operant behavior: A review of david m. green and john a. swets' signal detection theory and psychophysics1. *Journal of the Experimental Analysis of Behavior*, 12(3):475.

Neyman, J. and Pearson, E. S. (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, 50(3):559–569.

Obuchowski, N. A. (1997). Nonparametric analysis of clustered roc curve data. *Biometrics*, pages 567–578.

Olowookere, T. A. and Adewale, O. S. (2020). A framework for detecting credit card fraud with cost-sensitive meta-learning ensemble approach. *Scientific African*, 8:e00464.

Omar, L. and Ivrissimtzis, I. (2019). Using theoretical roc curves for analysing machine learning binary classifiers. *Pattern Recognition Letters*, 128:447–451.

Ozenne, B., Subtil, F., and Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology*, 68(8):855–859.

Papouskova, M. and Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision support systems*, 118:33–45.

Parodi, S., Pistoia, V., and Muselli, M. (2008). Not proper roc curves as new tool for the analysis of differentially expressed genes in microarray experiments. *BMC bioinformatics*, 9(1):1–30.

Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., and Kundu, S. (2018). Improved random forest for classification. *IEEE Transactions on Image Processing*, 27(8):4012–4024.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perkins Neil, J. and Schisterman Enrique, F. (2006). The inconsistency of "optimal" cut-points using two roc based criteria. *Am J Epidemiol*, 163:670–675.

Pesce, L. L., Metz, C. E., and Berbaum, K. S. (2010). On the convexity of roc curves estimated from radiological test results. *Academic radiology*, 17(8):960–968.

Phillips, R. (2013). Optimizing prices for consumer credit. *Journal of Revenue and Pricing Management*, 12(4):360–377.

Phillips, R., Raffard, R., and Solutions, N. (2011). Price-driven adverse selection in consumer lending. Technical report, Working Paper. Columbia University.

Polikar, R. (2012). Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer.

Pontius, R. G. and Parmentier, B. (2014). Recommendations for using the relative operating characteristic (roc). *Landscape Ecology*, 29(3):367–382.

Pontius Jr, R. G. and Si, K. (2014). The total operating characteristic to measure diagnostic ability for multiple thresholds. *International Journal of Geographical Information Science*, 28(3):570–583.

Poolsawad, N., Kambhampati, C., and Cleland, J. (2014). Balancing class for performance of classification with a clinical dataset. In *proceedings of the World Congress on Engineering*, volume 1, pages 1–6.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

PricewaterhouseCoopers (2012). Enhancing origination revenue through price optimization modeling. https://www.pwc.com/us/en/industries/banking-capital-markets/consumer-finance/library/loan-origination-price-optimization-modeling.html.

Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301.

Qiu, B., Huang, Z., Liu, X., Meng, X., You, Y., Liu, G., Yang, K., Maier, A., Ren, Q., and Lu, Y. (2020). Noise reduction in optical coherence tomography images using a deep neural network with perceptually-sensitive loss function. *Biomedical optics express*, 11(2):817–830.

Ranganathan, P., Pramesh, C., and Aggarwal, R. (2017). Common pitfalls in statistical analysis: logistic regression. *Perspectives in clinical research*, 8(3):148.

Rennie, J. D., Shih, L., Teevan, J., and Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623.

Rish, I. et al. (2001). An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):1–8.

Rojas, R. et al. (2009). Adaboost and the super bowl of classifiers a tutorial introduction to adaptive boosting. *Freie University, Berlin, Tech. Rep.*

Rosset, S. (2004a). Model selection via the auc. In *Proceedings of the twenty-first international conference on Machine learning*, page 89. ACM.

Rosset, S. (2004b). Model selection via the auc. In *Proceedings of the twenty-first international conference on Machine learning*, page 89.

Safavian, S. R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674.

Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

Sammut, C. and Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

Santafe, G., Inza, I., and Lozano, J. A. (2015). Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44(4):467–508.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690.

Sheng, V. S. and Ling, C. X. (2006). Thresholding for making classifiers cost-sensitive. In *AAAI*, volume 6, pages 476–481.

Sill, J., Takács, G., Mackey, L., and Lin, D. (2009). Feature-weighted linear stacking. *arXiv preprint arXiv:0911.0460*.

Silva-Fortes, C., Turkman, M. A. A., and Sousa, L. (2012). Arrow plot: a new graphical tool for selecting up and down regulated genes and genes differentially expressed on sample subgroups. *BMC bioinformatics*, 13(1):1–15.

Singh, A., Thakur, N., and Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Song, Y.-Y. and Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Strøm, H., Albury, S., and Sørensen, L. T. (2018). Machine learning performance metrics and diagnostic context in radiology. In *2018 11th CMI International Conference: Prospects and Challenges Towards Developing a Digital Economy within the EU*, pages 56–61. IEEE.

Su, W., Yuan, Y., and Zhu, M. (2015). A relationship between the average precision and the area under the roc curve. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 349–352.

Sun, S., Cao, Z., Zhu, H., and Zhao, J. (2019). A survey of optimization methods from a machine learning perspective. *IEEE transactions on cybernetics*, 50(8):3668–3681.

Sun, Y., Kamel, M. S., Wong, A. K., and Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378.

Sun, Y., Wong, A. K., and Kamel, M. S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Suragala, A., Venkateswarlu, P., and Raju, M. C. (2021). A comparative study of performance metrics of data mining algorithms on medical data. In *ICCCE 2020*, pages 1549–1556. Springer.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Susmaga, R. (2004). Confusion matrix visualization. In *Intelligent Information Processing and Web Mining*, pages 107–116. Springer.

Swamidass, S. J., Azencott, C.-A., Daily, K., and Baldi, P. (2010). A croc stronger than roc: measuring, visualizing and optimizing early retrieval. *Bioinformatics*, 26(10):1348–1356.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., and Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, 7(1):1–47.

Terblanche, S. and De La Rey, T. (2014). Credit price optimisation within retail banking. *ORiON*, 30(2):85–102.

Terrell, G. R. and Scott, D. W. (1992). Variable kernel density estimation. *The Annals of Statistics*, pages 1236–1265.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16(2):149–172.

Ting, K. M. (2002). An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering*, 14(3):659–665.

Unal, I. (2017). Defining an optimal cut-point value in roc analysis: an alternative approach. *Computational and mathematical methods in medicine*, 2017.

Van Erkel, A. R. and Peter, M. (1998). Receiver operating characteristic (roc) analysis: basic principles and applications in radiology. *European Journal of radiology*, 27(2):88–94.

Van Hulse, J., Khoshgoftaar, T. M., and Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, pages 935–942. ACM.

Vardhan, R. V., Balaswamy, S., and Sameera, G. (2020). A straight forward approach in understanding the improperness of roc curve. *FOUNDED 1998*, page 59.

Verbraken, T., Verbeke, W., and Baesens, B. (2012). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE transactions on knowledge and data engineering*, 25(5):961–973.

Wan, C., Wang, L., and Ting, K. M. (1999). Introducing cost-sensitive neural networks. In *Proceedings of the second international conference on information, communications and signal processing*, pages 1–4. Citeseer.

Wang, B. and Mao, Z. (2020). A dynamic ensemble outlier detection model based on an adaptive k-nearest neighbor rule. *Information Fusion*, 63:30–40.

Wang, D. and Chen, B. (2017). Cost-sensitive learning algorithm. In *Working paper*. University of Delaware.

Wang, H., Kou, G., and Peng, Y. (2020). Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending. *Journal of the Operational Research Society*, pages 1–12.

Wang, Y. and Ni, X. S. (2019). Predicting class-imbalanced business risk using resampling, regularization, and model emsembling algorithms. *International Journal of Managing Information Technology (IJMIT) Vol*, 11.

Wang, Z. and Scott, D. W. (2019). Nonparametric density estimation for high-dimensional data—algorithms and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(4):e1461.

Wayman, J. L. (1996). Technical testing and evaluation of biometric identification devices. In *Biometrics*, pages 345–368. Springer.

Wright, R. E. (1995). Logistic regression.

Wu, Q., Lin, Y., Zhu, T., and Zhang, Y. (2020). Hiboost: A hubness-aware ensemble learning algorithm for high-dimensional imbalanced data classification. *Journal of Intelligent & Fuzzy Systems*, 39(1):133–144.

Wu, Y. and Liu, Y. (2007). Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983.

Xia, Y., Liu, C., and Liu, N. (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications*, 24:30–49.

Xu, S., Li, Y., and Wang, Z. (2017). Bayesian multinomial naïve bayes classifier to text classification. In *Advanced multimedia and ubiquitous engineering*, pages 347–352. Springer.

Yao, H., Zhu, D.-l., Jiang, B., and Yu, P. (2019). Negative log likelihood ratio loss for deep neural network classification. In *Proceedings of the Future Technologies Conference*, pages 276–282. Springer.

Yao, Y. (2009). Three-way decision: an interpretation of rules in rough set theory. In *International Conference on Rough Sets and Knowledge Technology*, pages 642–649. Springer.

Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of Physics: Conference Series*, volume 1168, page 022022. IOP Publishing.

Yu, L., Zhou, R., Tang, L., and Chen, R. (2018). A dbn-based resampling svm ensemble learning paradigm for credit classification with imbalanced data. *Applied Soft Computing*, 69:192–202.

Yu, S., Li, X., Zhang, X., and Wang, H. (2019). The ocs-svm: An objective-cost-sensitive svm with sample-based misclassification cost invariance. *IEEE Access*, 7:118931–118942.

Yu, X., Zhang, Z., Wu, L., Pang, W., Chen, H., Yu, Z., and Li, B. (2020). Deep ensemble learning for human action recognition in still images. *Complexity*, 2020.

Yuan, M. and Wegkamp, M. (2010). Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(1).

Zadrozny, B. and Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 204–213. ACM.

Zarsky, T. Z. (2013). Transparent predictions. *U. Ill. L. Rev.*, page 1503.

Zhang, S., Li, X., Zong, M., Zhu, X., and Cheng, D. (2017). Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(3):1–19.

Zhang, Z. and Sabuncu, M. R. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

Zhou, Y., Wang, X., Zhang, M., Zhu, J., Zheng, R., and Wu, Q. (2019). Mpce: a maximum probability based cross entropy loss function for neural network classification. *IEEE Access*, 7:146331–146341.

Zhou, Z.-H. (2021). Ensemble learning. In *Machine Learning*, pages 181–210. Springer.

Zhou, Z.-H. and Liu, X.-Y. (2005). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1):63–77.

Zhu, X., Wu, X., and Yang, Y. (2004). Dynamic classifier selection for effective mining from noisy data streams. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 305–312. IEEE.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577.

# APPENDIX A

# RE-EXPRESS MATTHEWS CORRELATION COEFFICIENT

Recall the definitions of PPV, NPV, FDR, FOR in Chapter 2. Let $\omega = \pi_1 TPR + \pi_0 FPR$. MCC can be re-expressed as:

$$\begin{aligned}
MCC &= \sqrt{PPV * TPR * TNR * NPV} - \sqrt{FDR * FNR * FPR * FOR} \\
&= \sqrt{\frac{\pi_1}{\omega} TPR * TPR * TNR * \frac{\pi_0}{1-\omega} TNR} \\
&- \sqrt{(1-TNR) * \frac{\pi_1}{\omega} * (1-TPR) * (1-TNR) * \frac{\pi_0}{1-\omega} * (1-TNR)} \\
&= \sqrt{\frac{\pi_0 \pi_1}{\omega(1-\omega)}} [TPR * TNR - (1-TNR) * (1-TPR)] \\
&= \sqrt{\frac{\pi_0 \pi_1}{\omega(1-\omega)}} (TPR + TNR - 1) \\
&= \sqrt{\frac{\pi_0 \pi_1}{\omega(1-\omega)}} J
\end{aligned} \tag{A.1}$$

Since $PPV = \pi_1 TPR/\omega$ and $NPV = \pi_0 TNR/(1-\omega)$, we have $TPR = \omega PPV/\pi_1$ and $TNR = (1-\omega)NPV/\pi_0$. MCC can also be re-expressed as:

$$\begin{aligned}
MCC &= \sqrt{PPV * \frac{\omega}{\pi_1} PPV * \frac{1-\omega}{\pi_0} NPV * NPV} \\
&- \sqrt{(1-PPV) * \frac{\omega}{\pi_1} (1-PPV) \frac{1-\omega}{\pi_0} (1-NPV) * (1-NPV)} \\
&= \sqrt{\frac{\omega(1-\omega)}{\pi_1 \pi_0}} (NPV + PPV - 1) \\
&= \sqrt{\frac{\omega(1-\omega)}{\pi_1 \pi_0}} MK
\end{aligned} \tag{A.2}$$

In light of (Eqs. A.1 - A.2), we have $MCC^2 = MK * J$. The relationships between MCC and MK/J are also recognized in (Chicco and Jurman, 2020).

According to (Eq.A.1), we have

$$MCC^2 = \frac{\pi_0\pi_1(a-b)^2}{\omega(1-\omega)}.$$

Note that

$$\begin{aligned}\frac{dMCC^2}{dt} &= -\pi_0\pi_1 g_1^L(t)\frac{2(a-b)\omega(1-\omega) - (a-b)^2\pi_1(1-2\omega)}{[\omega(1-\omega)]^2} \\ &- \pi_0\pi_1 g_0^L(t)\frac{-2(a-b)\omega(1-\omega) - (a-b)^2(1-2\omega)\pi_0}{[\omega(1-\omega)]^2}\end{aligned}$$

Given

$$2\omega(1-\omega) - (a-b)\pi_1(1-2\omega) = (1-b)(\pi_1 a + \pi_0 b) + b[\pi_1(1-a) + \pi_0(1-b)]$$

and

$$2\omega(1-\omega) + (a-b)\pi_0(1-2\omega) = (1-a)(\pi_0 b + \pi_1 a) + a[\pi_0(1-b) + \pi_1(1-a)],$$

$d(MCC^2)/dt$ can be re-expressed as

$$\frac{dMCC^2}{dt} = \frac{\pi_0\pi_1(a-b)g_0^L(t)}{\omega(1-\omega)}[\frac{1-a-t(1-b)}{\pi_1(1-a)+\pi_0(1-b)} + \frac{a-tb}{\pi_1 a + \pi_0 b}]. \tag{A.3}$$

## APPENDIX B

## CODE TO IDENTIFY CASE WHERE MCC IS MULTI-MODAL

```
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure
import numpy as np
from sympy import *

t=symbols('t')
q1=0.1*t;
q2=(87*t/90)-78/90;
q3=(1/10)*t+78/100;
p1=4*t+1;
p2=5*t-1;
a1=p1*(1-q1)/(p1-q1);b1=(1-q1)/(p1-q1)
a2=p2*(1-q2)/(p2-q2);b2=(1-q2)/(p2-q2)
a3=p2*(1-q2)/(p2-q3);b3=(1-q3)/(p2-q3)
g1_1=-diff(a1,t);g0_1=-diff(b1,t)
g1_2=-diff(a2,t);g0_2=-diff(b2,t)
g1_3=-diff(a3,t);g0_3=-diff(b3,t)
###Ensure continous density function
g1_4=((g1_2.subs(t,1)-g1_1.subs(t,0.90))/0.10)*t\
+g1_2.subs(t,1)-((g1_2.subs(t,1)-g1_1.subs(t,0.90))/0.10)
g0_4=((g0_2.subs(t,1)-g0_1.subs(t,0.90))/0.10)*t\
+g0_2.subs(t,1)-((g0_2.subs(t,1)-g0_1.subs(t,0.90))/0.10)
```

```
g1=Piecewise((g1_1, (t<0.90) & (t>=0)),\
(g1_4,(t>=0.90) & (t<1)),(g1_2,(1<=t) & (t<=1.50)),
(g1_2.subs(t,1.5),(t>1.5)&(t<=1.9)))
g0=Piecewise((g0_1, (t<0.90) & (t>=0)),\
(g0_4,(t>=0.90) & (t<1)),(g0_2,(1<=t) & (t<=1.50)),
(g0_2.subs(t,1.5),(t>1.5)&(t<=1.9)))
g10=g1/g0


x1=plot(g1,g0, (t,0,1.9),xlabel='t',ylabel='density',\
show=False, legend=True)
x1[0].line_color='b'
x1[1].line_color='g'
x1[0].label='g1'
x1[1].label='g0'
x1.show()


plot(g10,(t,0,1.5),xlabel='t',ylabel='$\\frac{g_1}{g_0}$')

z=symbols('z')


a=1-integrate(nsimplify(g1),(t,0,z))
b=1-integrate(nsimplify(g0),(t,0,z))

p=a/b; q=(1-a)/(1-b);

d=2*p*q+p+q-t*(p+q+2)
```

```
dlist=[];plist=[];mcc2=[]
for t1 in np.linspace(0,1.9,num=30):
    dlist.append(d.subs({z:t1,t:t1}))
    plist.append(p.subs(z,t1))
    mcc=(a.subs({z:t1,t:t1})-b.subs({z:t1,t:t1}))**2/\
((a.subs({z:t1,t:t1})+b.subs({z:t1,t:t1}))*\
(2-a.subs({z:t1,t:t1})-b.subs({z:t1,t:t1})))
    mcc2.append(mcc)

t1=list(np.linspace(0,1.9,num=30))
plt.plot(t1,dlist)
plt.xlabel('t')
plt.ylabel('$D_{pq}$')
plt.plot(t1,plist)

plt.plot(t1,mcc2)
plt.xlabel('t')
plt.ylabel('$MCC^2$')
```

# APPENDIX C

# UNIMODALITY PROOF IN CHAPTER 7

**Prove $F1_{DR}(t_0, t_1)$ is a Unimodal Function in Proposition 14**

*Proof.* We have

$$\begin{aligned}\frac{\partial F1_{DR}(t_0,t_1)}{\partial t_0} = &\frac{-2\pi_1 g_1^L(t_1)\frac{dt_1}{dt_0}[\pi_0 - \pi_0 G_0^L(t_1) + \pi_1 G_1^L(t_0)]}{[1+\pi_1 - 2\pi_1 G_1^L(t_1) - \pi_0 G_0^L(t_1) + \pi_1 G_1^L(t_0)]^2} \\ &- \frac{2\pi_1(1-G_1^L(t_1))[-\pi_0 g_0^L(t_1)\frac{dt_1}{dt_0} + \pi_1 g_1^L(t_0)]}{[1+\pi_1 - 2\pi_1 G_1^L(t_1) - \pi_0 G_0^L(t_1) + \pi_1 G_1^L(t_0)]^2}. \end{aligned} \tag{C.1}$$

As proven in Proposition 14,

$$\frac{g_1^L(t_1)\frac{dt_1}{dt_0}}{\pi_0 g_0^L(t_1)\frac{dt_1}{dt_0} - \pi_1 g_1^L(t_0)} = \frac{\pi_0 t_1 + \pi_1 t_0 t_1}{\pi_0^2 - t_0 t_1 \pi_1^2}. \tag{C.2}$$

Therefore, $\partial F1_{DR}(t_0, t_1)/\partial t_0$ can be re-expressed as:

$$\frac{\partial F1_{DR}(t_0,t_1)}{\partial t_0} = A(t)D(t),$$

where

$$A(t) = \frac{-2\pi_1[\pi_0 g_0^L(t_1)\frac{dt_1}{dt_0} - \pi_1 g_1^L(t_0)][\pi_0 - \pi_0 G_0^L(t_1) + \pi_1 G_1^L(t_0)]}{[1+\pi_1 - 2\pi_1 G_1^L(t_1) - \pi_0 G_0^L(t_1) + \pi_1 G_1^L(t_0)]^2},$$

and

$$D(t) = \frac{\pi_0 t_1 + \pi_1 t_0 t_1}{\pi_0^2 - t_0 t_1 \pi_1^2} - \frac{1 - G_1^L(t_1)}{\pi_0 - \pi_0 G_0^L(t_1) + \pi_1 G_1^L(t_0)}.$$

In light of (Eq. C.2), when $\pi_0^2 - t_0 t_1 \pi_1^2 < 0$, we have $\pi_0 g_0^L(t_1)\frac{dt_1}{dt_0} - \pi_1 g_1^L(t_0) < 0$ and $A(t) > 0$ & $D(t) < 0$. Therefore, when $t_0 t_1 > \pi_0^2/\pi_1^2$, $\partial F1_{DR}(t_0, t_1)/\partial t_0 < 0$.

When $t_0 t_1 < \pi_0^2/\pi_1^2$, we have $A(t) < 0$. Clearly, the first part of $D(t)$ is monotonically increasing. We can also prove that $\pi_0 - \pi_0 G_0^L(t_1) + \pi_1 G_1^L(t_0)$ is monotonically

increasing, since $\pi_0 g_0^L(t_1)\frac{dt_1}{dt_0} - \pi_1 g_1^L(t_0) > 0$ by $t_0 t_1 < \pi_0^2/\pi_1^2$. As a result, the second part of $D(t)$ is monotonically decreasing. Therefore, $D(t)$ is monotonically increasing. According to Definition 3, $F1_{DR}(t_0, t_1)$ is a unimodal function. We can apply the same logic to prove that the function is unimodal under Bayesian transformation. □