ABSTRACT

ZHANG, LIXIA. Classification and Variable Selection by Maximizing the Area under the Curve. (Under the direction of Dr. Howard Bondell.)

The receiver-operating characteristic (ROC) curve is considered as a gold standard to visualize and evaluate classification performance. The area under the curve is the most commonly used measure. The ROC curve and its area also have been applied in constructing classifiers and selecting the optimal subset of variables. When the class label distribution is highly unbalanced, in which the majority of the objects belong to one group, the precision recall (PR) curve is a desirable alternative to the ROC curve. In the first part of this thesis, we construct the optimal linear classifier by maximizing the area under the PR curve, where the area is estimated non-parametrically. We demonstrate that the ranking produced by this classifier outperforms the one produced by maximizing the area under ROC curve in terms of false discovery rate, when one group is dominant. In the second part of this thesis, we propose an ROC-based forward selection algorithm for variable selection to construct a linear classifier. The empirical area under the ROC curve is used as the objective function, and the selection sequence is based on the score statistics. We apply the sigmoid function as an approximation to the non-differentiable empirical expression, which is equivalent to the two-sample Mann-Whitney statistics. From the theory of two-sample U-statistics, we obtain the asymptotic distribution for the score statistic and obtain the associated p-value. Simulation studies show the superiority of the proposed algorithm over the traditional logistic regression forward selection in terms of variable selection solution path, especially when the covariates are highly correlated. The approach also extends naturally for both numerical and nominal attributes.

Classification and Variable Selection by Maximizing the Area under the Curve

by
Lixia Zhang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2016

APPROVED BY:

Dr. Lexin Li	Dr. Hua Zhou
Dr. Eric Laber	Dr. Howard Bondell Chair of Advisory Committee

DEDICATION

To my beloved family.

BIOGRAPHY

Lixia Zhang was born in Shanghai, China. She graduated with a bachelor's degree in mathematics from Shanghai University in June 2009. In August, she came to North Carolina State University to earn her master in statistics. To purse advanced degree, she joined the Ph.D. program in statistics in Spring 2011. Her primary research interest includes classification, variable selection, the receiver operating characteristic curve, and the precision-recall curve. During her Ph.D. study, she also served as a student intern United Therapeutics for four and half years. She is expected to graduate in the summer of 2016.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Howard Bondell for his supervision, thoughtful advice, kindness, patience, and his continued support and encouragement throughout my Ph.D. study. His guidance towards being a professional is an invaluable asset which I will carry in the future. My appreciation also goes to Dr.s Lexin Li, Hua Zhou and Eric Laber for serving on my committee and valuable comments on my research work. I thank Dr. John Stone as well for being the graduate representative in my committee.

Many thanks to my mentors at United Therapeutics, Dr. Yi Zhou and Jody Cleveland, when I worked as a statistician intern from 2010 to 2015. They offered me great opportunities to gain industrial experience. I also thank all of my co-workers and colleagues at United Therapeutics for being so supportive of me. Besides, I thank Dr. John Iacoviello at L.E.K. consulting company for advices on the future career path, when I served as a intern from 2015 to 2016.

Last but not the least, I would like to express my gratitude to my parents and my husband for their endless love. I would never make it without their motivation and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Binary Classification Problems

Binary classification is one of the most widely discussed topics, which have received a great amount of attention in various research areas, including medical diagnosis, credit scoring, and anomaly detection. Classification is to construct a rule that maps each object to a predicted class based on a set of training objects whose memberships are known. Consider a binary classification problem with outcome $Y \in \{0, 1\}$, where $Y = 1$ denotes the positive group and $Y = 0$ is the negative group. For each object, a p-dimensional covariate vector is measured and is denoted as $X \in \mathbb{R}^p$. Without loss of generality, assume that the classification rule is given by:

$$\hat{Y} = \mathbb{I}(g(X) \geq c), \tag{1.1}$$

for some function $g(\cdot)$ and threshold c, where $\mathbb{I}$ is the indicator function. The score, $g(X)$, generates an ordering among objects, while the threshold c creates a classification decision from this ordering. Thus, in general, there are two types of classifiers. One is a discrete classifier that gives a hard-labeled outcome. The other is called a scoring classifier that generates an ordered object ranking by score values. The object ranking indicates that the one with higher score value has higher chance to be assigned to the positive group. In many instances, researchers are more

interested in the ranking score rather than the hard-labeled outcome, since a ranking not only provides the potential label assignment but also conveys information about the ordering among objects. If the function $g(X)$ takes a linear combination of covariates X, $\beta^T X$, it indicates a linear classifier. Furthermore, suppose there are N objects involved, among which n subjects are from the positive group and the remaining $m = N - n$ ones belong to the negative group. The proportion of positives among the whole population is introduced to describe the class label distribution, which is defined as $\pi = \Pr(Y = 1)$ in population and can be estimated as $\hat{\pi} = n/N$ under a random sampling. The data is balanced when the number of objects in each group is approximately equal, e.g. $\pi = 0.5$; on the other hand, the data is unbalanced when the majority of the objects belong to one group, e.g. $\pi = 0.1$ or 0.9.

Table 1.1: Confusion Matrix

		Actual Class	
		p	**n**
Predicted Class	**P**	**T**rue **P**ositives	**F**alse **P**ositives
	N	**F**alse **N**egatives	**T**rue **N**egatives

Given an object and a threshold c in the classification rule (3.1), there are four possible classification outcomes:

- True positive: a correctly classified positive object;
- False positive: a misclassified negative object;
- True negative: a correctly labeled negative object;
- False negative: a incorrectly classified negative object.

The performance of a classification decision based on a set of objects can be summerized

by a confusion matrix as in Table 1.1. It is a 2 by 2 table, and each cell counts the number of objects having the corresponding outcome. As the threshold c in the classification rule is varied, a series of confusion matrices are generated.

Both the Receiver Operating Characteristic (ROC) curve and the Precision-Recall (PR) curve are widely used techniques to visualize a series of confusion matrices and evaluate the classification performance of a scoring classifier. Since the superiority of a classifier can be determined by the ROC or PR curve, we intend to implement these two curves to construct classifiers. Here are the introductions about the ROC curve and the PR curve.

1.2 Receiver Operating Characteristics Curve

The ROC curve was first developed for assessing the accuracy of radar equipments in battlefields during World War II, and was soon introduced to psychology for perceptual detection of stimuli study (Egan (1975), Swets (1973)). Swets and Pickett (1982) brought the theory and the practice of the ROC curve to the biomedical realm, and then the ROC technique was widely treated as a comprehensive guide to make medical decision. Over the years, the ROC curve also increasingly gained attention for model evaluation and classifier selection in many applications of machine learning (Bradley (1997), Spackman, Kent A. (1989)). In summary, the ROC curve is useful to: (1) evaluate the discriminatory ability of a ranking score to correctly assign class labels; (2) find an optimal cut-off point to best classify the two-group objects; (3) compare the classification performances among two or more classifiers; and (4) select variables to construct a optimal classifier that maximizes diagnosis accuracy.

The ROC curve represents the tradeoff between false positive rate (FPR = 1- specificity) vs. true positive rate (TPR = sensitivity). Since the metrics in ROC space are defined based on the confusion matrix, both of them are functions of the threshold c and given as FPR(c) and TPR(c). In order to simplify the notations, the threshold c is omitted in the metric expressions.

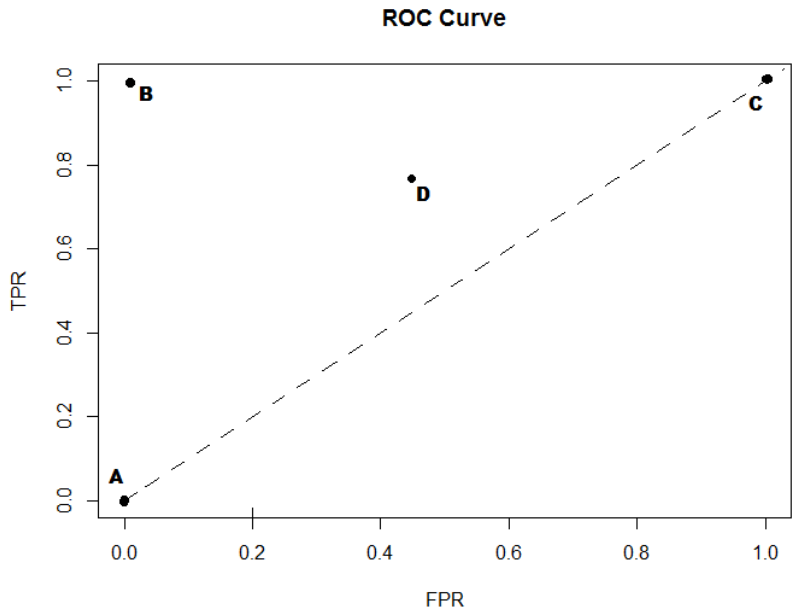


Figure 1.1: ROC Space

Then the metrics are written by:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \quad \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{1.2}$$

Since the denominators for the ROC metrics in (3.4) are fixed, the value for either FPR or TPR increases as the threshold c gets larger. Thus, TPR is a monotonically increasing function of FPR.

In ROC space in Figure 1.1, three important points need to be introduced. The origin point A (0, 0) is the curve starting point, which represents the strategy of never assigning an object to the positive group. On the other hand, the upper right corner point C (1, 1) is the curve endpoint and represents the strategy of unconditionally predicting objects as positives. If the curve reaches the point B (0,1), then it indicates a perfect separation, where all positives are positioned ahead of negatives. Besides, the diagonal line in ROC space represents the strategy of a random guess on class label assignment. Any point, e.g. the point D, that appears in the

upper left triangle leads to a better classification outcome than a random guess does.

In addition to TPR and FPR, the area under the ROC curve (AUCROC) is also one of the metrics that summarize the whole curve as well as the discriminatory accuracy of a scoring classifier. The AUCROC is defined by:

$$\text{AUCROC} = \int_0^1 \text{TPR} \quad d\text{FPR}. \tag{1.3}$$

This integral area is a portion of a unit square, so its value is always between 0 and 1. Without any assumptions on the data, one intuitive empirical approach to estimate the integral area in Equation (3.2) is interpolating points using a step function with jumps at the data values. The rationale is that any value between two adjacent data values doesn't change the label assignments. Given a decreasing ranking, by treating the score value of each object as a threshold, we obtain N rectangles in ROC space and the estimation of AUCROC is denoted as

$$\hat{\text{AUCROC}}_{\text{Rec}} = \sum_{i=1}^{N} (\text{FPR}_i - \text{FPR}_{i-1}) \times \text{TPR}_i, \tag{1.4}$$

where FPR_i and TPR_i are the corresponding rate values when the i^{th} object in the decreasing ranking is treated as the threshold in the classification rule (3.1), and $\text{FPR}_0 = 0$. This rectangle estimation $\hat{\text{AUCROC}}_{\text{Rec}}$ is also equivalent to the one by adopting the empirical cumulative density functions for positive and negative groups respectively. Besides, Bamber (1975) gave an important statistical property to AUCROC by linking the integral area with a probability expression as:

$$\text{AUCROC} = \Pr(g(\tilde{X}_1) > g(\tilde{X}_0)), \tag{1.5}$$

where $\tilde{X}_1$ denotes a randomly picked positive object and $\tilde{X}_0$ is a randomly picked negative object. Replacing the probability in (1.5) by its empirical version, we obtain the estimated AUCROC value as equivalent to the Mann-Whitney statistics for two-sample problem (Hanley

and McNeil (1982)), which is expressed as:

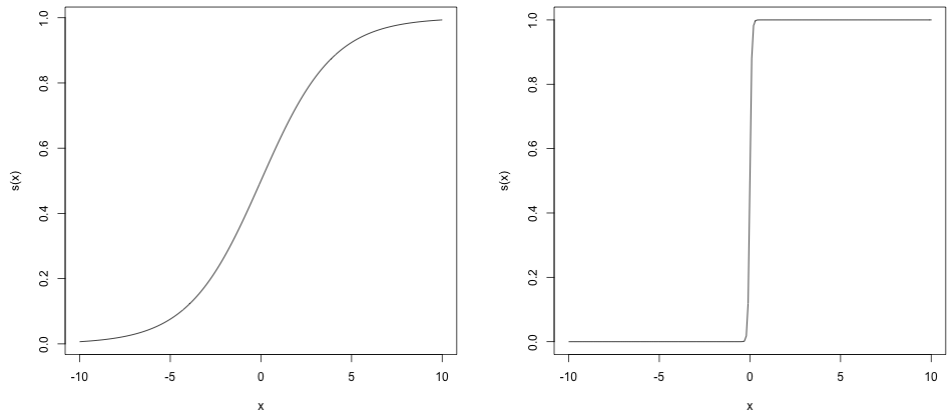
$$\widehat{\text{AUCROC}}_{\text{MW}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{I}(g(X_{1i}) - g(X_{0j}) > 0), \tag{1.6}$$

where X_{1i} is the covariate vector for the i^{th} object in the positive group, and X_{0j} is the covariate vector for the j^{th} object in the negative group. Although the empirical estimated $\widehat{\text{AUCROC}}_{\text{MW}}$ in Equation (2.11) has nice interpretation and is easy to compute, it is not a continuous or differentiable function. Thus, it becomes a challenging task to obtain the optimizer using $\widehat{\text{AUCROC}}_{\text{MW}}$ as the target function. In order to make $\widehat{\text{AUCROC}}_{\text{MW}}$ differentiable, the sigmoid function, $s(x) = 1/(1 + \exp(-x/\sigma_N))$, is adopted to smooth the indicator function in Equation (2.11), and then the approximated AUCROC is given by:

$$\widehat{\text{AUCROC}}_{\text{MWS}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{1+\exp\big(-\big(g(X_{1i}) - g(X_{0j})\big)/\sigma_N\big)}, \tag{1.7}$$

where σ_N is a scale parameter. The scale parameter σ_N determines the magnitude of the sigmoid function's approximation to the indicator function. Figure 3.1 shows the plots of the sigmoid function under different values of σ_N. The plot in Figure 3.1(a) displays a somewhat linear fashion under $\sigma_N = 2$, thus it is a poor approximation to the indicator function. On the other hand, Figure 3.1(b) gives a perfect approximation under $\sigma_N = 0.05$. Therefore, in order to make the sigmoid function approximate the indicator function excellently, we need to obtain a small σ_N value, which leads to large value of $\mid x/\sigma_N \mid$.

Recall that the empirical AUCROC estimations in Equations (1.4), (2.11) and (1.7) only depend on the ranking score produced by the function $g(X)$. There are two remarks about the functionality of the scoring classifiers. First, any constant adding to the score value does not alter the object ranking or the empirical AUCROC estimations. So it is fair to assume that no intercept is included in the scoring classifier $g(X)$. Second, any multiplier to the current score value does not alter the object ranking either. In other words, $g(X)$ is only identified up to a scalar-multiple. In order to make the classifier identifiable, one variable should be chosen as the



(a) $\sigma_N = 2$ (b) $\sigma_N = 0.05$

Figure 1.2: The approximation of the sigmoid function, $s(x) = 1/(1 + \exp(-x/\sigma_N))$, to the indicator function for different values of the scale parameter σ_N.

anchor and its coefficient estimate should be fixed with norm of 1.

In literature, in addition to the aforementioned empirical AUCROC estimations, there are also other non-parametric estimation methods proposed for AUCROC estimation, e.g. Faraggi and Reiser (2002) obtained a non-parametric estimation of AUCROC by adopting Gaussian kernels to estimate probability density functions for positive and negative groups respectively. Besides, the underlying binormal distribution is the most commonly considered assumption for parametric AUCROC estimation, where the covariate vectors for both positive and negative follow normal distributions. Hanley (1998) discussed the robustness of binormal latent distribution for ROC curve fit; Gu and Ghosal (2009) obtained AUCROC under binormal assumption by Bayesian approach using a rank likelihood; and Gonçalves et al. (2014) gave a broad overview on the Bayesian approach to estimate AUCROC both semi-parametrically and non-parametrically.

For a long time, AUCROC has been used as a standard tool to evaluate diagnosis accuracy and compare classification performance among algorithms. Nowadays, many researches have been done to construct an optimal classifier by maximizing AUCROC directly. Pepe et al.

(2003) adopted the ROC curve to rank diagnostic performance of genes in microarray experiments; Ma and Huang (2005) implemented the threshold gradient decent algorithm to obtain the coefficients for selected biomarkers by maximizing AUC and then applied cross-validation approach to determine the optimal classifier; and Graf and Bauer (2009) investigated the relationship between false discovery rate threshold for multiple tests control and the optimal AUC value for future independent population prediction so as to determine the number of variables to be chosen.

1.3 Precision-Recall Curve

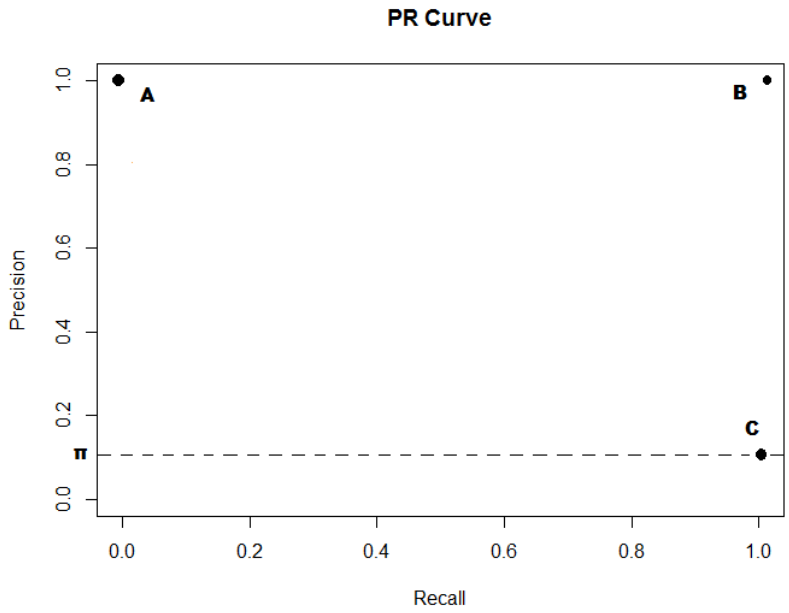


Figure 1.3: PR space

The PR curve is considered as a desirable alternative to the ROC curve when the class label distribution is highly unbalanced. It was originally introduced to evaluate information retrieval system by judging both efficiency and effectiveness of the whole retrieval process,

where the non-relevant queries are the majority (Raghavan et al. (1989a)). Recently, it gained great attention in the machine learning as well (Peltonen and Kaski (2011)). The PR curve describes the tradeoff between recall and precision as the threshold c in the classification rule (3.1) is varied. As functions of the threshold c, the two rate metrics in PR space are defined as:

$$\begin{aligned} \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} = 1-\text{FDR}. \end{aligned} \tag{1.8}$$

Since the denominator of precision changes as c varies, there is no certain patten displayed between recall and precision.

In PR space, the three important points shown in Figure 1.3 have the same interpretation as those in ROC space, however they have totally different locations. The PR curve starts from the upper left corner A (0,1), where no objects are misclassified, and ends at the point C (1,π). The endpoint of the PR curve is not a fixed point as the one in ROC space. It depends on one data's proportion of positives. More skewed the data is, lower the curve endpoint is located. Besides, the upper right corner B (1,1) is the point that represents the perfect separation.

Furthermore, one important measure to evaluate the classification performance in PR space, which is the area under the PR curve (AUCPR), is given by:

$$\text{AUCPR} = \int_0^1 \text{Precision} \quad d\text{Recall}. \tag{1.9}$$

Similar to AUCROC, this integral area is also ranged from 0 to 1. Besides, the empirical estimated $\hat{\text{AUCPR}}_{\text{Rec}}$ can be obtained by using a step function to interpolate points in PR space. Given a decreasing ranking, it is written as:

$$\hat{\text{AUCPR}}_{\text{Rec}} = \sum_{i=1}^{N} (\text{Recall}_i - \text{Recall}_{i-1}) \times \text{Precision}_i, \tag{1.10}$$

where $\text{Recall}_0 = 0$. In literature, $\hat{\text{AUCPR}}_{\text{Rec}}$ estimation is also called as average precision

(AP) estimation. It is known that the value of recall alters only when a true positive object is classified correctly and the difference between any two adjacent recall values is identical. So we can calculate $\hat{\text{AUCPR}}_{\text{Rec}}$ using the precision value when one positive object is treated as the threshold. Thus, $\hat{\text{AUCPR}}_{\text{Rec}}$ in Equation (1.10) can be simplified as

$$\hat{\text{AUCPR}}_{\text{Rec}} = \frac{1}{n}\sum_{j=1}^{n} \text{Precision}(X_{1j}), \tag{1.11}$$

where $\text{Precision}(X_{1j})$ is the precision value when the j^{th} positive object is classified as positive among a decreasing object ranking. In addition to $\hat{\text{AUCPR}}_{\text{Rec}}$ estimator, Boyd et al. (2013) discussed other AUCPR estimators as well, such as trapezoidal estimator, interpolation estimator, etc. Similar to AUCROC, the underlying binormal distribution is also a widely used assumption to calculate AUCPR (Brodersen et al. (2010)). Besides, Keilwagen et al. (2014b) derived and discussed AUCPR for binary classification with soft labels.

1.4 Relationship between the ROC Curve and the PR Curve

The metrics in both PR space and ROC space are defined based on the confusion matrices, so these two curves share many common features. Davis and Goadrich (2006) showed that for any dataset with fixed numbers of positive and negative objects, points can be one-to-one translated between the ROC and PR spaces. However, the two curves also have a lot of differences. The classifier that maximizes AUCROC doesn't always optimize AUCPR, and vice versa (Davis and Goadrich (2006)).

Comparing the metrics' definitions, we notice that recall is the same as TPR. The difference between the metrics of ROC and PR is the denominator of FPR versus that of FDR. We see that FPR considers the total number of true negative objects; while FDR instead takes the number of predicted positive objects into consideration. This difference is analogous to the difference between Type I error and FDR in that the ROC curve relates to Type I error, while the PR curve measures FDR. Furthermore, this difference results in the different magnitude of

sensitivity towards the skewness of class label distribution of the two curves. The ROC curve is insensitive to the class label distribution, while the PR curve is extremely sensitive. For the ROC curve, either TPR or FPR focuses only on the portion of objects that are labeled correctly within each group. Thus, in the situation where the negative objects are the majority, the ROC curve would not be able to detect the large number of incorrectly labeled negatives, because such great number results in only a small portion compared to the totally number of negative objects. For the PR curve, on the other hand, precision is able to capture and reflect such situation. Knowing the positive is the minority, the slightly changed numerator over the dramatically increased denominator would definitely alert that too many negatives are misclassified.

1.5 Two sample U-Statistics

The empirical estimated AUCROC in Equation (1.7) is a two-sample U-statistics. So does its first derivative with respect to the coefficient, if $g(X)$ takes a linear functionality as $\beta^T X$. In the third chapter, we took advantage of the property of the two-sample U-statistics to conduct forward selection, when treating $\widehat{\text{AUCROC}}_{\text{MWS}}$ as the target function. Thus, we give a general and brief introduction about U-statistics in this chapter.

U-statistics, a distribution-free unbiased estimator, was first introduced by Hoeffding (1948), and had asymptotic normal distribution in large samples. Hoeffding (1961) developed H-decomposition to further discuss its asymptotical property. In addition to Wassily Hoeffding's theoretical contribution to U-statistics, Sproule (1974) developed a Kolmogorov-like inequality to examine the converge properties; Ahmad (1981) discussed the rates of convergence in central limit theory; and Lenth (1983) investigated interesting characteristics of U-statistics by linking the jackknife procedure with the existed properties. Meanwhile, U-statistics has great impact and wide applications on non-parametric inference. Fu (2012) applied one-sample U-statistics to assess the asymptotic distribution of a non-parametric estimator for the multivariate functional data analysis; Xu (2007) adopted one-sample U-statistics in econometrics to investigate the asymp-

totic results for some inequality and poverty measures. Moreover, for two-sample U-statistics, Dehling and Fried (2012) derived the asymptotical distribution of the two-sample empirical U-quantiles in the case of correlated data; Dehling et al. (2013) studied a robust test for detecting change-points of a stochastic process based on the two-sample Wilcoxion test statistics. For problem of the variable selection, Song and Ma (2010) implemented U-statistics in penalized variable selection using LASSO and ridge.

Suppose we have $\tilde{X}_{11}, \ldots, \tilde{X}_{1n}$, $\tilde{X}_{01}, \ldots, \tilde{X}_{0m}$ independently and identically distributed (i.i.d.) with cumulative distribution functions F_1 and F_0, then a kernel ψ is a function defined with $k + l$ arguments

$$\psi(X_{11}, \ldots, X_{1k}; X_{01}, \ldots, X_{0l})$$

which is symmetric in each set of arguments. Then the corresponding two-sample U-statistics based on ψ has the form

$$U_{n,m} = \binom{n}{k}^{-1} \binom{m}{l}^{-1} \sum_{(n,k)} \sum_{(m,l)} \psi(X_{11}, \ldots, X_{1k}; X_{01}, \ldots, X_{0l}), \tag{1.12}$$

where (n, k) represents all possible unordered subsets of k elements chosen without replacement from the set $\{X_{11}, X_{12}, \ldots, X_{1n}\}$, and the same interpretation for (m, l) in the set $\{X_{01}, X_{02}, \ldots, X_{0m}\}$. Moreover, it can be shown that $U_{n,m}$ is an unbiased estimator for

$$\theta = \int \cdots \int \psi(u_1, \ldots, u_k; v_1, \ldots, v_l) \prod_{i=1}^{k} dF_1(u_i) \prod_{j=1}^{l} dF_0(v_j).$$

Based on Hoeffding (1961)'s decomposition, U-statistics can be decomposed into a sum of uncorrelated components. We take two-sample U-statistics with $k = l = 1$ as an example, and write it in general as

$$\psi(X_{1i}, X_{0j}) = \theta + h_{1,0}(X_{1i}) + h_{0,1}(X_{0j}) + g(X_{1i}, X_{0j}), \tag{1.13}$$

where

$$h_{1,0}(X_{1i}) = \psi_{1,0}(X_{1i}) - \theta$$
$$h_{0,1}(X_{0j}) = \psi_{0,1}(X_{0j}) - \theta$$
$$g(X_{1i}, X_{0j}) = \psi(X_{1i}, X_{0j}) - h_{1,0}(X_{1i}) - h_{0,1}(X_{0j}) - \theta.$$

Since both $\tilde{X_{1i}}$ and $\tilde{X_{0j}}$ are independent random variables with the same distribution respectively, it is obvious to find that $h_{1,0}(\tilde{X}_{1i})$ and $h_{0,1}(\tilde{X}_{0j})$ are also independently distributed with

$$E(h_{1,0}(\tilde{X}_{1i})) = 0,\ Var(h_{1,0}(\tilde{X}_{1i})) = \varsigma_{1,0}^2;$$
$$E(h_{0,1}(\tilde{X}_{0j})) = 0,\ Var(h_{0,1}(\tilde{X}_{0j})) = \varsigma_{0,1}^2.$$

Therefore, the two-sample U-statistics in (1.12) with $k = l = 1$ can be written as

$$U = \theta + H_{1,0} + H_{0,1} + R, \tag{1.14}$$

where $H_{1,0} = n^{-1}\sum_{i=1}^{n} h_{1,0}(\tilde{X}_{1i})$, $H_{0,1} = m^{-1}\sum_{j=1}^{m} h_{1,0}(\tilde{X}_{0j})$, and R is the remaining element whose variance has a lower order infinitesimal of sample size. Thus, it is easy to proof that, by large number theory, $n^{\frac{1}{2}}H_{1,0}$ and $m^{\frac{1}{2}}H_{0,1}$ both converge to normal distributions with mean zero and variances $\varsigma_{1,0}^2$ and $\varsigma_{0,1}^2$ respectively. Moreover, let $N = n + m$ and $n/N \to p_1 \in (0, 1)$ as n and $m \to \infty$, then we have

$$\sqrt{N}(U - \theta) = \sqrt{N}\big(H_{1,0} + H_{0,1} + R\big)$$
$$= \sqrt{\frac{N}{n}}\sqrt{n}\frac{\sum_{i=1}^{n} h_{(1,0)}(X_{1i})}{n} + \sqrt{\frac{N}{m}}\sqrt{m}\frac{\sum_{j=1}^{m} h_{(0,1)}(X_{0j})}{m} + o_p(1).$$

Since $H_{1,0}$ and $H_{0,1}$ are independent, it follows that $N^{\frac{1}{2}}(U - \theta)$ converges in distribution to a normal random variable with zero mean and variance $p_1^{-1}\varsigma_{1,0}^2 + (1 - p_1)^{-1}\varsigma_{0,1}^2$. Therefore, the asymptotic distribution for a single two-sample U-statistics in (1.14) is written by

$$\sqrt{N}(U^{(q)} - \theta^{(q)}) \xrightarrow{D} N(0, p_1^{-1}\varsigma_{1,0}^{(q)2} + (1 - p_1)^{-1}\varsigma_{0,1}^{(q)2}). \tag{1.15}$$

The above asymptotic distribution result for a single two-sample U-statistics can be gener-

alized to the case of a U-statistics vector $\mathbf{U} = (U^1, \ldots, U^{(p)})^T$, where the components in the vector $\mathbf{U}$ are defined based on the same two independent sample $X_{11}, \ldots, X_{1n}$ and $X_{01}, \ldots, X_{0m}$. Then the joint limiting distribution of $\mathbf{U}$ is a multivariate normal distribution. If F_1 and F_0 are continuous and both of their second moments exist, then as $n \to \infty$ and $m \to \infty$, by applying the central limit theorem, the joint distribution of

$$\left[\sqrt{N}(U^{(1)} - \theta^{(1)}), \sqrt{N}(U^{(2)} - \theta^{(2)}), \ldots, \sqrt{N}(U^{(p)} - \theta^{(p)})\right]$$

converges to a multivariate normal distribution with zero vector mean and the limiting variance-covariance matrix $\mathbf{\Sigma}_{lim}$. Furthermore, the asymptotic variance-covariance matrix can be obtained as $\mathbf{\Sigma}_{lim} = (\sigma_{lim}^{2((s),(t)}$), where

$$\sigma_{lim}^{2((s),(t)} = \frac{1}{p_1}\varsigma_{1,0}^{2((s),(t))} + \frac{1}{1-p_1}\varsigma_{0,1}^{2((s),(t))},$$

with $n/N \to p_1 \in (0, 1)$.

1.6 Dissertation Outline

In this dissertation, we investigate two patterns of datasets and develop the best approaches to obtain the optimal linear classifier using AUC respectively. In Chapter 2, we focus on the unbalanced data, where the majority of objects belong to the negative group. The gold standard tool AUCROC fails to select the optimal linear classifier in terms of the object ranking due to its insensitivity to the class label distribution. We take the advantage of PR curve's control on FDR, and construct the optimal linear classifier, which positions the majority of positive objects ahead of negative ones. Since the accuracy of AUC estimation is not the primary interest, we adopt the rectangle estimation for both AUCROC and AUCPR. The superiority of PR curve over ROC curve on the unbalanced data is shown in simulation studies.

In Chapter 3, we developed a ROC-based variable selection algorithm to obtain the optimal subset of variables. The algorithm treats AUCROC as the target function to be maximized, and implements forward selection to obtain the variable selection solution path. Inspired by the

likelihood-based hypothesis statistics, we adopt score statistics to be the selection criterion as well as the stopping rule. In order to make AUCROC differentiable, the empirical estimated $\widehat{\text{AUCROC}}_{\text{MWS}}$ is used. We further show the score function derived from $\widehat{\text{AUCROC}}_{\text{MWS}}$ is a two-sample U-Statistics, and follows an asymptotic normal distribution. Simulation studies in different scenarios, e.g. magnitude of correlation between covariates, are conducted to compare the proposed AUCROC forward selection algorithm with the regular linear logistic regression using forward selection by the whole solution path as well the relation between p-value and the selected solution.

Chapter 2

Classification for Highly Unbalanced Data by Maximizing the Area Under the Precision-Recall Curve

2.1 Introduction

Binary classification problems have received a great amount of attention in various research areas, including medical diagnosis, credit scoring, and anomaly detection. Consider a binary classification problem with class label $Y \in \{0, 1\}$ indicating negative and positive group assignment, and predictor variable vector $X \in \mathbb{R}^p$. If the number of objects in each group is approximately equal, then we say the data is balanced; on the other hand, if the majority of the objects belong to one group, the data is unbalanced. Without loss of generality, assume that the classification rule is given by

$$\hat{Y} = \mathbb{I}(g(X) \geq c) \tag{2.1}$$

for some function $g(\cdot)$ and a threshold c. The score, $g(X)$, gives the rank order of the objects, while the threshold c creates a hard classification outcome from this ordering. If the function

$g(X)$ takes a linear combination of the predictor variable vector X, which is $\boldsymbol{\beta}^T X$, it indicates a linear classifier. In many instances, researchers are more interested in the ranking score rather than the hard-labeled outcome, since a ranking not only provides the potential label assignment but also conveys information about the ordering among objects.

For a fixed threshold c, the performance of a classifier can be assessed by a confusion matrix as in Table 1.1. A confusion matrix is composed of four components:

- True positives (TP): the number of positive objects correctly classified as positive;
- False positives (FP): the number of negative objects incorrectly classified as positive;
- True negatives (TN): the number of correctly labeled negative objects;
- False negatives (FN): the number of misclassified positive objects.

An overall picture of a classifier can then be viewed as a series of confusion matrices as the threshold varies.

The receiver-operating characteristic (ROC) curve is a widely used approach and often considered as a gold standard to visualize a series of confusion matrices. The plot of the ROC curve represents the trade-off between true positive rate (TPR) and false positive rate (FPR), where TPR measures the fraction of positive objects that are correctly labeled and FPR measures the fraction of negative ones that are misclassified as positive. The area under the ROC curve (AUCROC), given in (3.2), is the most common scalar measurement that summarizes the curve, as well as the performance of a score ranking.

$$\text{AUCROC} = \int_0^1 \text{TPR} \quad d\text{FPR} \tag{2.2}$$

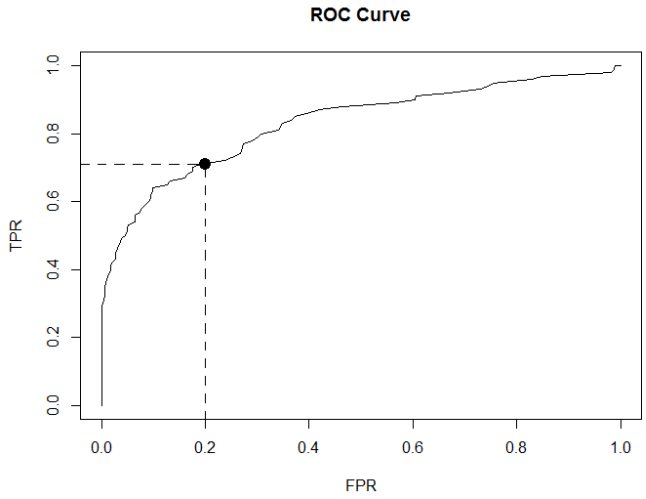
Larger AUCROC leads to a better overall classification performance. For a long time, the AUCROC is widely used to evaluate classification performance. Recently many researchers maximize AUCROC directly to obtain optimal classifiers. For example, Pepe et al. (2006) used

AUCROC as the objective function for combining biomarkers for classification; Ma and Huang (2005) implemented an AUCROC-based Threshold Gradient Descent Regularization algorithm to select important biomarkers for disease classification; Marrocco et al. (2008) constructed a nonparametric linear classifier that relied on the Wilcoxon–Mann–Whitney statistic; Castro and Braga (2008) proposed an algorithm to maximize AUCROC based on gradient descent; and Graf and Bauer (2009) selected a best classification model that optimizes AUCROC based on FDR-thresholding. However, despite its popularity, the ROC curve has some drawbacks that make it a less ideal tool to evaluate and construct classifiers under unbalanced data.

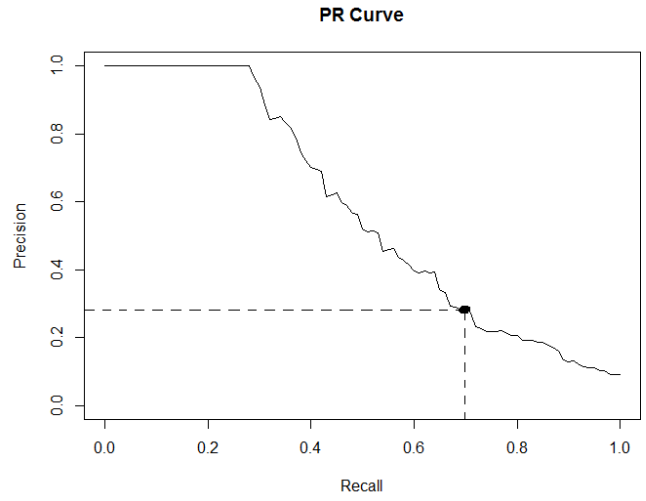
A disadvantage of the ROC curve is its ignorance to the proportion of the positive objects relative to the negative objects. In the situation where negative objects are the majority, we can observe two facts from ROC space, (1) a small change in FPR value indicates a large increase in the actual number of misclassified negative objects; (2) the majority of the AUCROC corresponds to the region of high FPR. Therefore, under unbalanced situations, false discovery rate (FDR) can be very high in the region that dominates the curve.

As an example to demonstrate this phenomenon, consider a dataset composed of 100 positive and 1000 negative objects. Figure 2.1(a) is an example of a ROC curve for a classification algorithm for this unbalanced dataset. The point $(0.2, 0.7)$ shown in Figure 2.1(a) represents that 270 out of 1100 objects are classified as positive. However, among the 270 classified positives, only 70 objects are correctly predicted and the remaining 200 are not. At the point $(0.2, 0.7)$, where FPR is still relatively low, i.e. 0.2, the FDR reaches $\frac{200}{270}$, which means nearly three-fourths of the classified positives are incorrect. Furthermore, we can see that the FDR would increase as FPR gets larger, in the region where the slope of the ROC curve becomes flat. We can see that the majority of the AUCROC is added in the region which is not of interest. Therefore, under unbalanced cases, the ignorance to class label distribution makes AUCROC a poor measure of classification performance, and the linear classifier obtained by maximizing AUCROC becomes a less desirable one.

In order to overcome the deficiencies of AUCROC when the class label distribution is skewed,



(a) The ROC curve



(b) The PR curve

Figure 2.1: The ROC curve and PR curve for an unbalanced data. The unbalanced example dataset is composed of 100 positive and 1000 negative objects. The point $(0.2, 0.7)$ shown in ROC space corresponds to the point $(0.7, 0.26)$ on PR space. The point represents that 270 out of 1100 objects are classified as positive. Among 270 classified positives, only 70 observations are correct and 200 observations are incorrect.

the partial area under the ROC curve (pAUCROC) in (2.3) has been proposed. The partial area is considered over a pre-specified range of FPR rather than the whole range.

$$\text{pAUCROC}(s) = \int_0^s \text{TPR} \quad d\text{FPR}, \quad \text{for some} \;\; 0 < s < 1 \tag{2.3}$$

For the example in Figure 2.1(a), we can constrain the range of FPR to be $(0, 0.2)$ and obtain the optimal classifier by maximizing pAUCROC in order to meet low FDR demand. McClish (1989) introduced pAUCROC under the binormal assumption to solve the problem when the entire range is not of interest and proposed a test statistics to compare two partial areas; Yousef, Waleed A. (2013) assessed the performance of a classifier by pAUCROC, and discussed its nonparametric estimator; Ma et al. (2013) discussed pAUCROC and standardized pAUCROC to evaluate diagnostic performance. In the context of obtaining optimal classifiers, Wang and Chang (2011), Hsu and Hsueh (2013b) and Ricamato and Tortorella (2011) proposed algorithms to select the best linear classifier by maximizing pAUCROC when low FPR is prerequisite; and Komori and Eguchi (2010) developed a new pAUCROC-focused method based on a boosting technique to construct nonlinear classifiers. However, in general, the cutoff, s, is an arbitrary choice, and thus pAUCROC has not gained in popularity.

The precision recall (PR) curve, has become an increasingly competitive and useful alternative to the ROC curve. It plots the recall, equivalent to TPR, against the precision, or positive predictive rate. The area under the PR curve (AUCPR), in (2.4), is also considered as a comprehensive performance measure and has attracted many recent investigations (e.g. Oliphant et al. (2010b)).

$$\text{AUCPR} = \int_0^1 \text{Precision} \quad d\text{Recall} \tag{2.4}$$

Due to the sensitivity of precision to the class label distribution, AUCPR controls the number of misclassified negatives, and consequently meets low FPR and FDR demands. Most importantly, the AUCPR does not require a chosen bound on the region under consideration as does the partial area, so it is threshold independent. Figure 2.1(b) is the PR curve generated from the

aforementioned algorithm for the unbalanced example data. The point $(0.2, 0.7)$ in ROC space in Figure 2.1(a) corresponds to the point $(0.7, 0.26)$ in PR space shown in Figure 2.1(b). It shows that the small partial area in ROC space corresponds to the majority of the area in PR space. It further indicates that even when the whole range of recall is taken, the FDR is relatively low in the dominated area of AUCPR. Therefore, AUCPR is an important tool for classification problems under unbalanced data.

Davis and Goadrich (2006) showed that for any dataset with fixed numbers of positive and negative objects, points can be one-to-one translated between the ROC and PR spaces. So the ROC curve and the PR curve share many common features, but they also differ a great deal. In this paper, we point out that the classifiers constructed by maximizing the two areas could be very different; and demonstrate that under unbalanced situations, the classifier obtained by maximizing AUCPR outperforms the one obtained by maximizing AUCROC in terms of FDR.

In section 2, we give brief introductions to the ROC curve and the PR curve, and demonstrate their different preferences on score rankings by example. We introduce the empirical area under the curve (AUC) estimation applied in this paper for the two curves in section 3. In section 4, the implementation of maximizing AUC to obtain linear classifiers is described in details. In section 5 we summarize the comparisons between the optimal linear classifiers obtained by the two curves from a simulation study. Conclusion is summarized in section 6.

2.2 The ROC and PR Curves

Suppose there are N objects involved in the classification problem, among which n subjects are from the positive group and the remaining $m = N - n$ ones belong to the negative group. The proportion of positives in the population is given by $\Pr(Y = 1)$ and denoted as π.

2.2.1 Metrics in Two Curves

Given a threshold c, the classification outcome can be represented in a confusion matrix as in Table 1.1. Then a single point in either ROC space or PR space can be constructed from a

given confusion matrix. The ROC curve is composed of TPR versus FPR; while the PR curve plots precision against recall. Equations (3.4) and (2.6) give the definitions for the ROC and PR metrics:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \quad \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{Recall}; \tag{2.5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = 1 - \text{FDR}. \tag{2.6}$$

Notice that recall is the same as TPR. The difference between the metrics of ROC and PR is the denominator of FPR versus that of FDR. We see that FPR considers the total number of true negative objects; while FDR instead takes the number of correctly predicted positive objects into consideration. This difference is analogous to the difference between Type I error and FDR in that the ROC curve relates to Type I error, while the PR curve measures FDR.

In either ROC or PR space, one classification decision determined by a given threshold is displayed as a single point. A ROC curve or a PR curve is obtained as the threshold varies. Then AUC is a common metric that measures the overall classification performance.

2.2.2 Receiver Operating Characteristic Curve

The ROC curve was first developed in the early 1950's as a signal detection technique (Swets (1986)). One of its first application was to assess the accuracy of radar equipment in World War II. Later the ROC curve gained attention in radiology imaging diagnosis (Lusted (1960), Metz (1986)), as well as solving medical decision problems (Metz (1978), Lusted (1978)). Recently it has been widely used for model evaluation and selection in many applications of machine learning (Spackman, Kent A. (1989), Bradley (1997)). Its wide applications not only owing to its comprehensive presentation of tradeoff between FPR and TPR, but also a summarized number measuring classification performance, which is AUCROC. In addition to the defined integral expression of AUCROC in Equation (3.2), Bamber (1975) gave an important statistical

property by linking the integral area with a probability expression as:

$$\text{AUCROC} = P(g(\mathbf{X}_{\text{pos}}) > g(\mathbf{X}_{\text{neg}})), \tag{2.7}$$

where $\mathbf{X}_{\text{pos}}$ denotes a randomly picked object from the positive group and $\mathbf{X}_{\text{neg}}$ is a randomly picked negative object. The estimation for AUCROC has been widely discussed in both parametric and non-parametric approaches. For parametric approaches, based on the underlying binormal distribution assumption, AUCROC can be represented using the standard normal distribution. Hanley (1988) discussed the robustness of binormal latent distribution for ROC curve fit; Tosteson and Begg (1988) applied ordinal regression model to estimate the ROC curve parameters and its AUC for categorical rating data; Gu and Ghosal (2009) obtained AUCROC under the binormal assumption via a Bayesian approach. For AUCROC estimation, replacing the probability in (2.7) by its empirical version, we obtain the AUCROC value as equivalent to the Mann-Whitney U-statistics (Hanley and McNeil (1982)), given by

$$\hat{\text{AUCROC}}(\boldsymbol{\beta}) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{I}(g(X_{\text{pos},i}) - g(X_{\text{neg},j}) > 0), \tag{2.8}$$

where $X_{\text{pos},i}$ indicates the i^{th} object in the positive group and $X_{\text{neg},j}$ refers to the j^{th} object in the negative group.

2.2.3 Precision Recall Curve

As an alternative to the ROC curve, the PR curve was originally introduced to evaluate information retrieval system by judging both efficiency and effectiveness of the whole retrieval process (Raghavan et al. (1989b)), and have been gaining popularity for classification performance evaluation under unbalanced data. AUCPR, as a counterpart of AUCROC, can be interpreted as average precision. Similar to AUCROC, an underlying binormal distribution is one widely used parametric assumption to calculate AUCPR (Brodersen et al. (2010)). For nonparametric approaches, different from the ROC curve, Davis and Goadrich (2006) pointed

that linear interpolation is not appropriate in PR space when two points are far apart due to its varied denominator of precision, and then proposed an expression for the intermediate PR point. Clémençon and Vayatis (2009) introduced a nonparametric approach to assess PR curve, and examined its related theoretical and practical issues; Boyd et al. (2013) discussed several different approaches to compute empirical AUCPR point estimators, e.g. trapezoidal estimator, interpolation estimators, and obtained confidence intervals by performing bootstrap procedure; Keilwagen et al. (2014a) pointed out AUCPR is highly informative about the classifier performance, and discussed AUCPR for weighted and un-weighted data.

2.2.4 A Simple Example

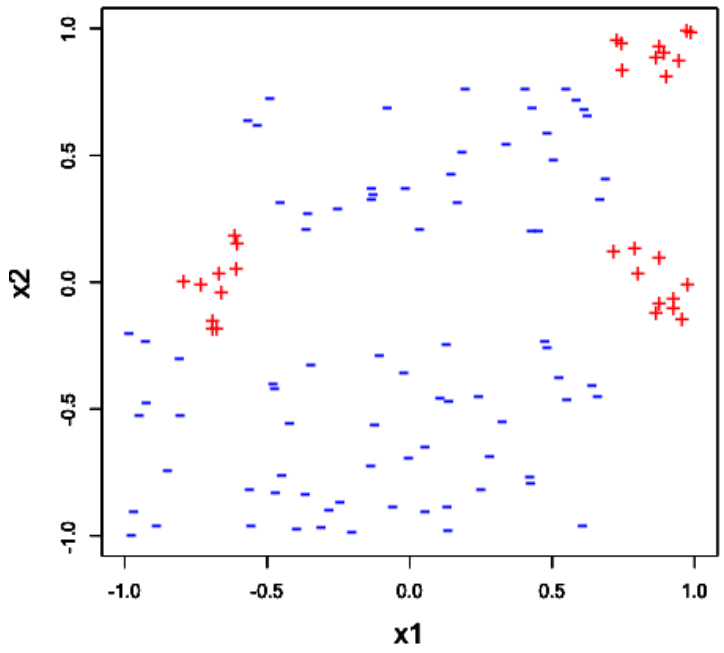


Figure 2.2: Scatter plot for the contrived example data. The bivariate X data is composed of 30 positives and 80 negatives. Red plus symbol refers to the object in the positive group; blue minus symbol indicates the object in the negative group.

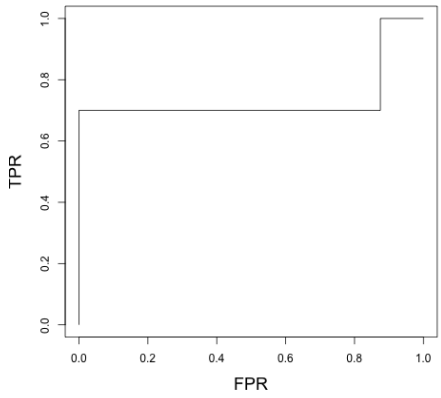
To further investigate the different behaviors for score rankings under the ROC curve and the PR curve, we will give a simple contrived example. Consider a bivariate X, where the data is composed of 30 positives and 80 negatives. Figure 2.2 displays the scatter plot for this example data. We investigate two candidate linear classifiers. The first linear classifier is the classifier according to the variable x_1, while the second one would be according to x_2. These two classifiers generate the ranking S_1 and S_2 correspondingly.

$$S_1 : \overbrace{1...1}^{20 \text{ times}} \quad \overbrace{0...0}^{70 \text{ times}} \quad \overbrace{1...1}^{10 \text{ times}} \quad \overbrace{0...0}^{10 \text{ times}}$$

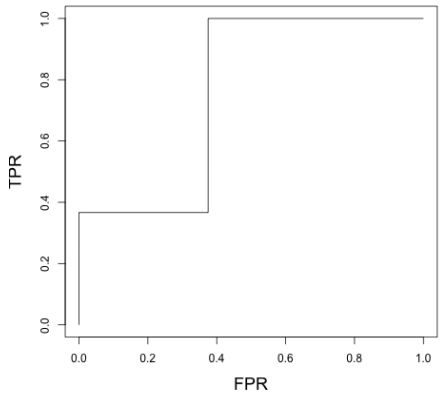
$$S_2 : \overbrace{1...1}^{10 \text{ times}} \quad \overbrace{0...0}^{30 \text{ times}} \quad \overbrace{1...1}^{20 \text{ times}} \quad \overbrace{0...0}^{50 \text{ times}}$$

Comparing the two rankings above, we see that, S_1 places 20 positive objects at the top of the ordering, so we are able to have 20 correctly classified positives without any mistakes. Meanwhile we would need to consider sacrificing the remaining 10 positives positioned at the tail, in favor of zero-misclassified 20 positives. In ranking S_2, there are only 10 positive objects listed at the top. Thus, if more than 10 positive objects are desired, we will be required to include 30 misclassified negative objects. Therefore, under the situation where the positive group is the minority, S_1 is still able to keep most of the positives at the top of a score ranking, while S_2 would result in many false positive objects if we want to keep a significant number of positive label correctly.

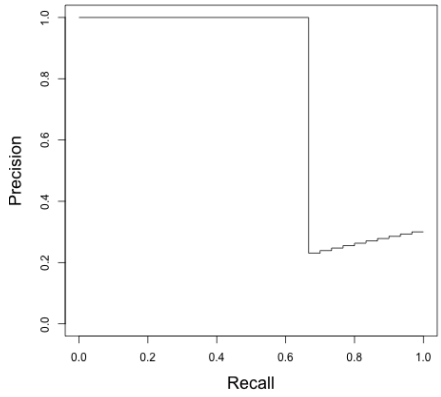
Observing the corresponding ROC and PR curves and AUC values for the ranking S_1 and S_2 in Figure 2.3, we are able to tell the different preferences these two curves have. In ROC space, the top 20 positive objects in the ranking S_1 lead to relative higher TPR than the ranking S_2 does when FPR is low. However, when we look at FPR above 0.4, then S_2 has a large jump. This region that is not likely of interest adds the most area. Because of this, the ROC curve prefers S_2 with an area of 0.75 over S_1 with an area of 0.71. On the other hand, the PR curve shows the opposite selection. The PR curve generated by the ranking S_1 stays at value 1 of precision longer than the one by S_2, so the overall AUCPR of S_1 is higher with area value 0.76. Therefore, in this particular example, the ROC curve prefers the linear classifier according to



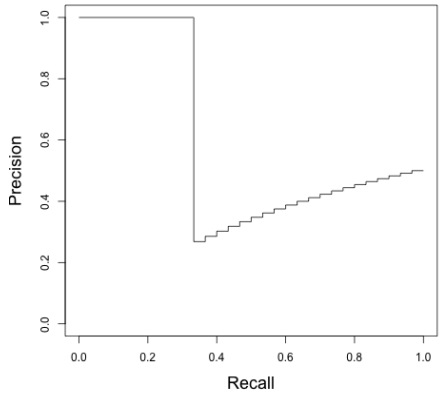
(a) the ROC curve for S_1 with AUCROC = 0.71



(b) the ROC curve for S_2 with AUCROC = 0.75



(c) the PR curve for S_1 with AUCPR = 0.76



(d) the PR curve for S_2 with AUCPR = 0.6

Figure 2.3: ROC curves and PR curves for the scoring S_1 and S_2

the x_2 direction, while the PR curve prefers the one towards the x_1 direction. Moreover, we can conclude that, in general, the linear classifier that positions more positive objects at the top would outperform in PR space. This seems like a more desirable outcome in typical cases.

2.3 AUC Calculation

The primary interest in this paper is to construct an optimal linear classifier by maximizing AUC, which takes the form of $\beta^T X$ in the classification rule (3.1). Thus, in this section, we compute the metrics for both two curves based on linear classifiers, and specify the corresponding AUC calculations.

Recall that either AUCROC or AUCPR is determined by a score ranking among objects. Since our focus is on constructing the optimal classifiers by maximizing the AUC rather than the accuracy of the AUC values, the method of interpolation that we use does not effect our results. Therefore, we implement the intuitive empirical approach resulting in a step function with jumps at the data values. By treating the score value of each object as a threshold in the classification rule (3.1), we are able to obtain n possible classification outcomes based on the data, which consequently generate n rectangles in either of the two spaces. Then given the ordered ranking, we can estimate AUCROC and AUCPR as the summation of n rectangle areas, and written by

$$\hat{\text{AUCROC}}(\beta) = \sum_{i=1}^{n} (\text{FPR}(\beta)_i - \text{FPR}(\beta)_{i-1}) \times \text{TPR}(\beta)_i, \tag{2.9}$$

$$\hat{\text{AUCPR}}(\beta) = \sum_{i=1}^{n} (\text{recall}(\beta)_i - \text{recall}(\beta)_{i-1}) \times \text{precision}(\beta)_i, \tag{2.10}$$

where the index i denotes the i^{th} object among the decreasing ranking, and the data-based metric rates are calculated as

- $\text{TPR}(\boldsymbol{\beta})_i = \text{recall}(\boldsymbol{\beta})_i = \dfrac{\sum_{j=1}^{n} \mathbb{I}(\boldsymbol{\beta}^T X_j \geq \boldsymbol{\beta}^T X_i | Y_j = 1)}{\sum_{j=1}^{n} \mathbb{I}(Y_j = 1)}$,

- $\text{FPR}(\boldsymbol{\beta})_i = \dfrac{\sum_{j=1}^{n} \mathbb{I}(\boldsymbol{\beta}^T X \geq \boldsymbol{\beta}^T X_i | Y = 0)}{\sum_{j=1}^{n} \mathbb{I}(Y_j = 0)}$,
- $\text{precision}(\boldsymbol{\beta})_i = \dfrac{\sum_{j=1}^{n} \mathbb{I}(Y_j = 1 | \boldsymbol{\beta}^T X_j \geq \boldsymbol{\beta}^T X_i)}{\sum_{j=1}^{n} \mathbb{I}(\boldsymbol{\beta}^T X_j \geq \boldsymbol{\beta}^T X_i)}$,
- $\text{FPR}(\boldsymbol{\beta})_0 = \text{recall}(\boldsymbol{\beta})_0 = 0$ under the assumption that the plot starts from the origin $(0, 0)$.

As an alternative to the expression for the AUCROC in (2.9), we can also estimate AUCROC empirically using the expression in (2.8). Given a linear classifier $\boldsymbol{\beta}^T X$, we have

$$\hat{\text{AUCROC}}(\boldsymbol{\beta}) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{I}(\boldsymbol{\beta}^T X_{\text{pos},i} - \boldsymbol{\beta}^T X_{\text{neg},j} > 0). \tag{2.11}$$

2.4 Optimization

In this section, we concentrate on describing the approach to obtain optimal linear classifiers on both ROC and PR spaces, and illustrating the optimization implementation in details.

Either of the AUCROC estimation in (2.9) or the AUCPR estimation in (2.10) is a function of β, so the target function is to find the optimal β that maximizes AUC and can be written, in general, as

$$\hat{\beta} = \text{argmax}\{\text{AUC}(\beta)\}. \tag{2.12}$$

First note that the objective function, which results from either (2.10) or (2.11), only depends on the ranking of $\beta^T X_i$, and changing β will not change the ranking on a small region. Therefore, the target function (2.12) is a non-linear and non-differentiable function with respect to β. To optimize, we use Nelder and Mead (1965) 'simplex' direct search method.

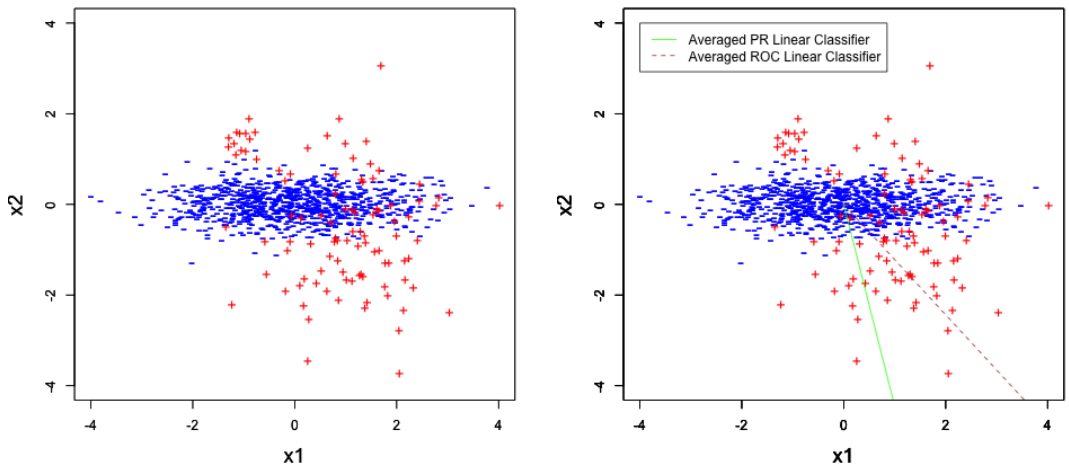
In addition to the non-linear and non-differentiable property of the target function, there are two additional issues involved. First is the identifiability of β, and second is the choice of initial starting point in the optimization.

Constructing a linear classifier is to find a vector direction that produces a score ranking maximizing AUC. Since only the ordering matters, an intercept is not included in β. Addition-

ally, β is only identified up to a scalar-multiple. In order to make β identifiable, we choose one variable as an anchor and fix its coefficient estimate to have norm of 1. The predictor variable that has the maximal marginal AUC is selected as the anchor, and then its corresponding estimate is set to be 1 or -1, depending on the direction of its marginal maximizer. The coefficient estimation from a linear logistic regression model is applied as the starting value for β, rescaling so that the anchor has magnitude of 1.

2.5 Simulation Study

We now conduct a simulation to support the conclusions that (1) the linear classifier by optimizing the PR curve is different from the one by optimizing the ROC curve; and (2) the optimal PR linear classifier outperforms the optimal ROC classifier in terms of FDR.



(a) (b)

Figure 2.4: (a) Simulated data pattern; (b) The averaged vectors over 100 datasets for PR maximizer vs. ROC maximizer.

We simulate data in two dimensions, and Figure 2.4(a) shows the data pattern to be inves-

tigated, where x_1 and x_2 represent the covariate values for each observation. The data pattern is composed of 1000 points with 100 positives, so $\pi = 0.1$. The data points belonging to the positive group denoted in red (+) in Figure 2.4(a) are generated from a mixture of normal distributions,

$$\mathbf{X}|\text{Y=1} \sim \text{p*N}\left(\begin{pmatrix} 1 \\ -0.8 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1.2 \end{pmatrix}\right) + \text{(1-p)*N}\left(\begin{pmatrix} -1 \\ 1.5 \end{pmatrix}, \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix}\right),$$

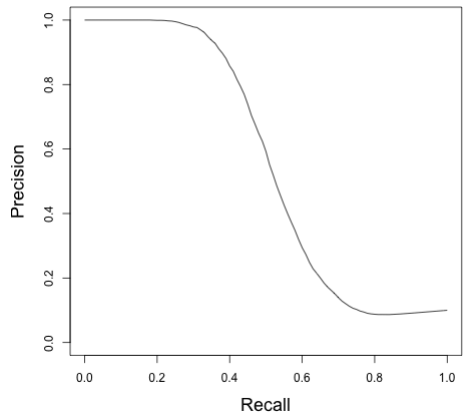
where N(μ,Σ) denotes a normal distribution with mean μ and covariance matrix Σ ,and p is the probability from the first normal distribution, which is set to be 0.85 in the simulation. The data points in the negative group denoted in blue (−) are generated from a normal distribution as

$$\mathbf{X}|\text{Y=0} \sim \text{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.2 & 0 \\ 0 & 0.35 \end{pmatrix}\right).$$

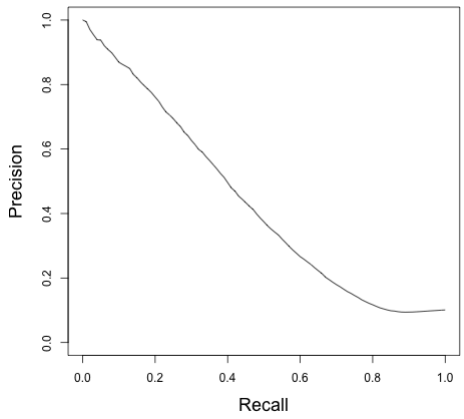
2.5.1 Different Preference in Optimal Linear Classifier

We generate 100 datasets using the data structure described above, and construct the optimal PR linear classifiers and ROC linear classifiers by maximizing the area under the two curves correspondingly for each dataset. Figure 2.4(b) displays two averaged vectors over the simulated 100 datasets for the PR maximizer in green solid line and the ROC maximizer in dashed brown line, where the two vectors point towards different directions. Figure 2.5 present the averaged PR and ROC curves for the optimal PR and ROC classifiers correspondingly. Specifically, Figure 2.5(a) shows the averaged PR curve that is obtained via the optimal PR linear classifiers. Then the Figures from 2.5(b) to 2.5(d) in order are: the PR curve by the optimal PR linear classifier; the PR curve by the optimal ROC linear classifier; the ROC curve by the optimal PR linear classifier; and he ROC curve by the optimal ROC linear classifier.

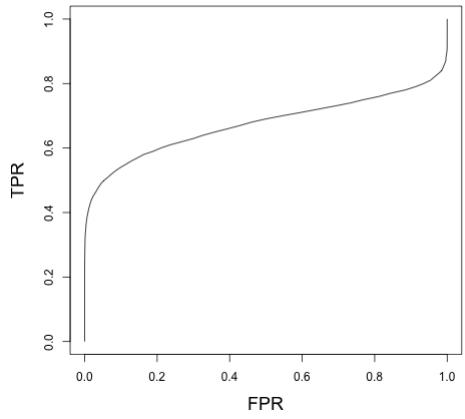
Comparing the top two figures, we see that the averaged AUCPR value at the top-left corner is higher than the averaged AUCPR at the top-right corner. It indicates that the linear classifier



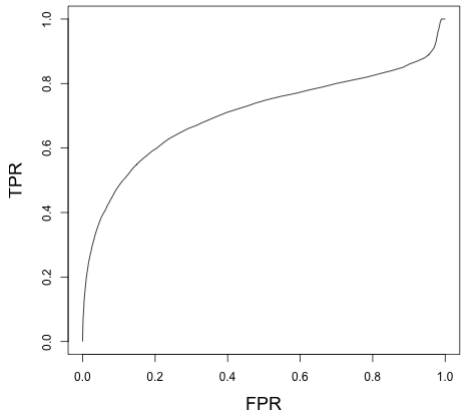
(a) The PR curve by the PR classifiers
Averaged area = 0.554



(b) The PR curve by the ROC classifiers
Averaged area = 0.429



(c) The ROC curve by the PR classifiers
Averaged area = 0.669



(d) The ROC cure by the ROC classifiers
Averaged area = 0.701

Figure 2.5: The averaged PR curves and ROC curves for the optimal linear PR classifiers and ROC classifiers over 100 datasets, and the averaged AUC values. The subfigures from (a) to (d) in order are: the PR curve produced by optimal PR classifiers; the PR curve produced by optimal ROC classifiers; the ROC curve produced by optimal PR classifiers; and the ROC curve produced by optimal ROC classifiers.

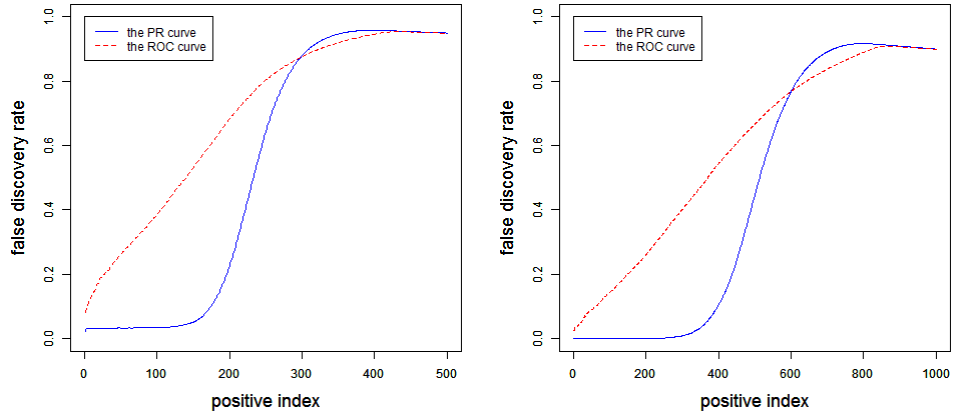
that optimizes AUCROC doesn't optimize AUCPR. The similar conclusion can be obtained for ROC space and the optimal ROC classifier from the bottom two figures. In addition, observing Figure 2.5(a) and 2.5(b), we find that there is a flat platform where precision stays at value 1 in PR space and a vertical line where TPR stays at value 0 in ROC space. It means the optimal PR linear classifier intends to position positives at the early part of the ranking, and the corresponding region likely adds the most area in PR space. However, the horizontal and vertical flat part at the end of two curves correspondingly also indicate that the optimal PR linear classifier meanwhile scarifies some positives by leaving at the tail of the ranking. On the other hand, the optimal ROC linear classifier positions positive and negative objects in a more balanced way, and try to avoid positives left behind, thus no obvious flat parts can be observed at either the early or tail part in Figure 2.5(c) and 2.5(d).

2.5.2 Comparison of Optimal Linear Classifiers by FDR

As discussed earlier, FDR is a common criterion to measure the degree of misclassification, particularly under the unbalanced case, and then can be used to determine the superiority of one classifier over another. Thus, the FDR values are compared between the optimal ROC and PR linear classifiers averaging over 100 simulated datasets. Since we are interested to examine the impact of the data's skewness, from a highly skewed situation to a balanced case, the datasets with proportion of positives in $\{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$ are investigated.

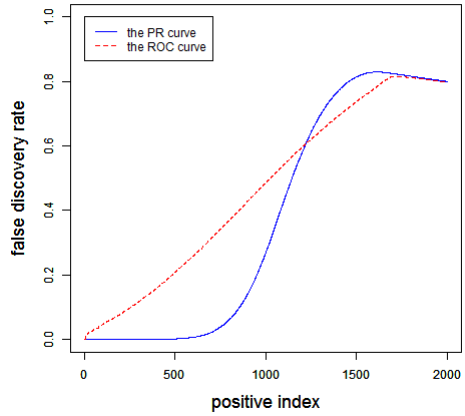
Figure 2.6 displays the FDR comparisons between the optimal PR and the ROC linear classifiers for each of the 6 positive proportions. The linear classifiers are trained on a dataset with 1000 observations and evaluated on a testing dataset with 10000 observations. Each graph plots the averaged FDR over 100 datasets as a function of the number of total positives correctly identified. Thus, the range for x-axis varies for each graph depending on the total number of positives in the dataset. In each graph, two FDR lines are generated, where the FDR line by the ROC classifier is in red dashed line and the one by the PR classifier is the blue solid line.

From the results, we see a large difference between the two approaches. The FDR by the

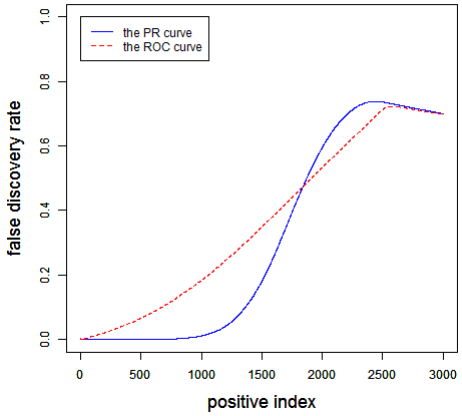


(a) $\pi = 0.05$

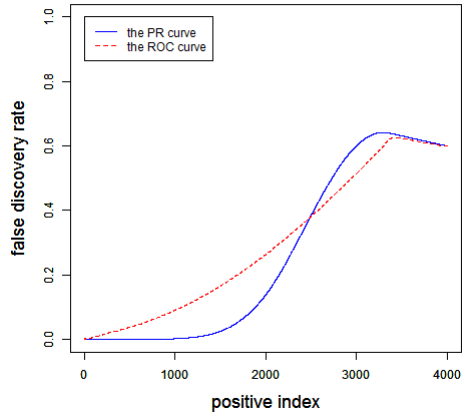
(b) $\pi = 0.1$



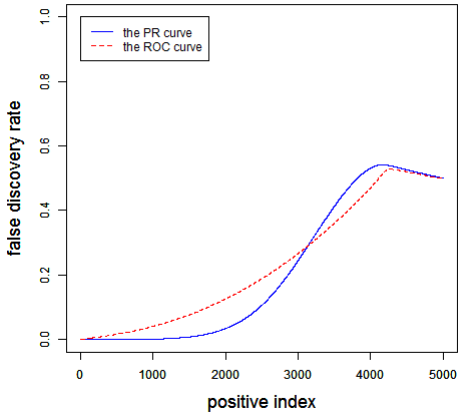
(c) $\pi = 0.2$



(d) $\pi = 0.3$

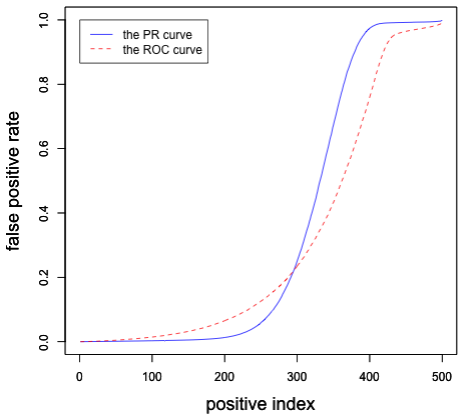


(e) $\pi = 0.4$

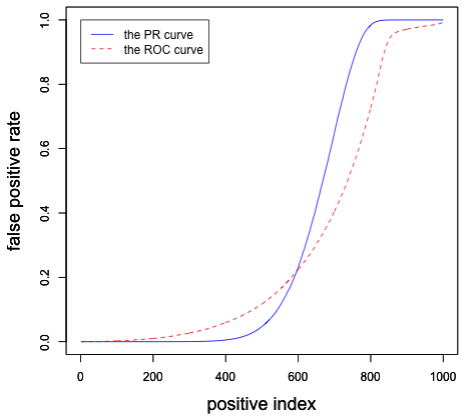


(f) $\pi = 0.5$

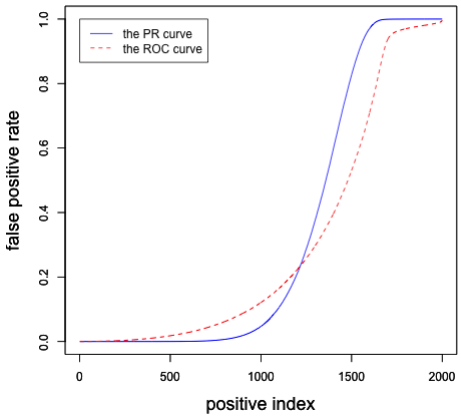
Figure 2.6: False discovery rate comparison between the optimal linear classifiers by the PR curve and the ROC curve on data with different proportions of positives π.



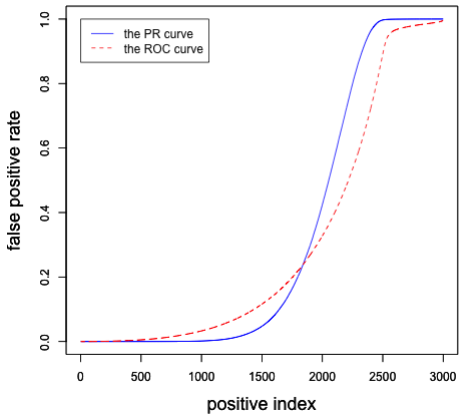
(a) $\pi = 0.05$



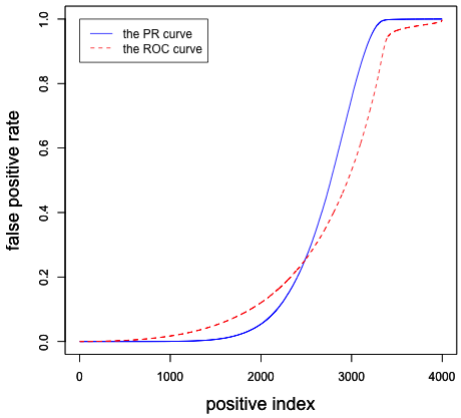
(b) $\pi = 0.1$



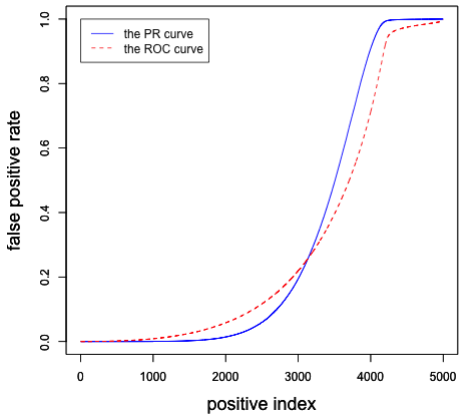
(c) $\pi = 0.2$



(d) $\pi = 0.3$



(e) $\pi = 0.4$



(f) $\pi = 0.5$

Figure 2.7: False positive rate comparison between the optimal linear classifiers by the PR curve and the ROC curve on data with different proportions of positives π.

ROC classifier increases in a somewhat linear fashion, while the FDR by the PR classifier stays at zero until 25% of the positives enter. The optimal ROC linear classifiers always lead to a higher FDR value than the optimal PR classifiers until reaching just above 50% of the true positives. This occurs at a high FDR, somewhere between FDR of 0.3 and 0.8 depending on the proportion of positives. Besides, more skewed the dataset is, more obvious the disparity between two FDR lines can be observed.

In addition, the FPR comparisons between the optimal PR and ROC linear classifiers are also examined in Figure 2.7. Each graph plots the averaged FPR against the number of true positives correctly classified. We see that the shapes of averaged FPR curves are nearly the same under 6 different scenarios of positive proportions. It is consistent with the fact that FPR is ignorant to the class label distribution. Although the PR curve doesn't focus on controlling the metric FPR, the optimal PR linear classifiers lead to a lower FPR value than the optimal ROC classifiers until FPR reaches 0.2. Therefore, we can reach the conclusion that, the score ranking by the optimal PR linear classifier outperforms the optimal classifier by the optimal ROC linear classifier in terms of FDR as well as FPR, especially under highly unbalanced situation.

2.6 Conclusion

In this paper, we provide examples to support the statement that the classifier optimizing the ROC curve does not always optimize the PR curve, and further demonstrate that the linear classifier obtained by the PR curve outperforms the ROC curve under unbalanced data. There is no global rule to determine the superiority of classifiers. In general, it depends on the data and potential misclassification costs trading off Type I vs Type II error. In the case of the unbalanced situation, where the number of positive objects is relatively fewer comparing to negative ones, FDR is a common measure on the number of misclassified negatives. In that case, the number of misclassified negatives at the early part of a score ranking is minimized, and consequently the number of correctly labeled positives is maximized. Due to the ROC curve's insensitivity to class label distribution and preference on avoiding positives left behind

at the tail, negative objects are often positioned ahead of positives under unbalanced data, which could lead to high FDR even early on. Although the partial ROC curve is able to meet low FDR demand, it involves specification of threshold for FPR. On the other hand, the PR curve is sensitive to the proportion of positive, which meets the low FDR demand and if free of choice of threshold. Therefore, the PR curve is shown to be a worthwhile approach to construct a linear classifier under unbalanced cases.

Chapter 3

Forward Selection via Maximizing the Area under the ROC Curve for Classification

3.1 Introduction

Classification and variable selection are two widely discussed topics in various research areas and statistical applications. Consider a binary classification problem with outcome $Y \in \{0, 1\}$ indicating the positive group and the negative group respectively, and covariate vector $X \in \mathbb{R}^p$. Classification methods intend to construct a classification rule that maps each object to a predicted class, e.g. positive or negative, based on a set of training objects whose memberships are known. The classification rule, in general, takes the form as:

$$\hat{Y} = \mathbb{I}\{g(X) > c\}, \tag{3.1}$$

where $\mathbb{I}$ is the indicator function, $g(.)$ generates a score value for each object, and the threshold c determines the class label assignments. Variable selection is to select relevant variables among

the pool, while ignoring irrelevant ones. The objective of variable selection is four-fold: providing a better understanding of the underlying relationship between outcomes, improving the accuracy to the estimation of the interested quantities, enhancing the model's generalization and avoiding overfitting, and reducing time and economic cost on data collection in the future. Therefore, the selection of variables becomes an important and essential task as hundreds to tens of thousands of variables are currently available.

Many variable selection methods in the classification context are explored. In general, there are three types of variable selection methods (Guyon and Elisseeff (2003)): filter, wrapper, and embedded. The filter approach is a preprocessing step to score the importance of each variable independently according to a particular metric, e.g. correlation coefficient, individual predictive power, and etc. It is favored because of its simplicity, scalability, and computational efficiency. Meanwhile, the filter method is limited because the measured score is ignorant to the other given variables. Thus the filter method, e.g. variable ranking, is normally used as a baseline method prior to using the variable selection metrics (Weston et al. (2003), Forman (2003)). The wrapper approach utilizes an objective function to assess the usefulness of subsets of variables according to their prediction performances (Kohavi and John (1997)). It is a powerful and robust tool to find the optimal subset. Although the wrapper method is often criticized for requiring massive of amounts of computation, forward selection and backward elimination have computational advantages. This is because both two greedy search strategies yield nested subsets of variables. Forward selection starts with an empty set of variables and adds one variable at each iteration until the full set is reached; on the other hand, backward elimination begins at a full set and removes one variable at a time. The wrapper approach is widely used and can be incorporated with any classification algorithm, e.g. logistic regression, random forest, CART, etc. (In Lee and Koval (1997), Peng et al. (2010), Zellner et al. (2004)). The embedded method is an approach that combines two terms together: the objective function to be optimized and the number of selected variables to be minimized. Regularization method is one of the typical embedded approaches. Efron et al. (2004) proposed the least angle regression combined with LASSO

(Tibshirani (1996)) and forward stagewise selection to construct the optimal linear classifier; Zhu et al. (2003) and Wang et al. (2006) introduced 1-norm and 2-norm support vector machines in two-class classification problems respectively; and Friedman et al. (2010) developed fast algorithms for generalized linear models with penalties of LASSO, ridge regression, and elastic.

In addition to the aforementioned methods, the receiver operating characteristic (ROC) curve, a gold criterion to evaluate classification performance, also has wide applications dealing with tasks for variable selection. The ROC curve assesses classification accuracy by true positive rate (TPR) and false positive rate (FPR) of class label assignments. TPR measures the fraction of positive objects that are correctly classified; and FPR measures the fraction of negative objects that are misclassified as positive. The area under the curve (AUCROC) is a threshold-free metric for performance evaluation, which is defined as:

$$\text{AUCROC} = \int_0^1 \text{TPR} \quad d\text{FPR}. \tag{3.2}$$

Bamber (1975) investigated the integral area in (3.2) and gave an important statistical property by:

$$\text{AUCROC} = \Pr(g(\tilde{X}_{11}) > g(\tilde{X}_{01})), \tag{3.3}$$

where $\tilde{X}_{11}$ denotes a random object from the positive group and $\tilde{X}_{01}$ is a random object from the negative group. Thus it is the probability of the score for a randomly chosen positive object higher than the score for a random chosen negative object. Pepe et al. (2003) adopted the ROC curve to rank diagnostic performance of genes in microarray experiments. Ma and Huang (2005) implemented the threshold gradient decent algorithm to obtain the coefficients for selected biomarkers by maximizing AUCROC. They then applied cross-validation approach to determine the optimal classifier. In order to determine the number of variables to be chosen, Graf and Bauer (2009) investigated the relationship between false discovery rate threshold for multiple tests control and the optimal AUCROC value for future independent population prediction.

This work proposed an ROC-based algorithm for variable selection to construct a linear classifier, which takes linear functionality $\beta^T X$ in the classification rule (3.1). The proposed AUCROC forward selection (AUCFW) algorithm is able to find important variables, even under the situations with highly correlated covariates. The subset of selected variables also has good predictive performance. Thus it is a desired algorithm under the situation where covariate candidates are highly correlated. Besides, the algorithm is a non-parametric approach, hence it is free of data distribution assumptions and has stable classification performance.

The AUCFW algorithm aims to select the optimal subset of variables by maximizing AUCROC. The algorithm adopts forward selection as the selection approach and score statistics as the selection criterion. Likelihood ratio, Wald, and score statistics are the three most commonly used hypothesis statistics under the likelihood theory. Score statistics only needs to evaluate maximal likelihood estimator (MLE) at the null hypothesis, hence it stands out due to its lower computational requirements when using forward selection. Following the derivation of score statistics under the likelihood theory, we calculate the pseudo score equation by taking the first derivative of AUCROC, which is a two-sample U-statistics. Due to the asymptotical property of the two-sample U-statistics, which converges to a normal distribution, we are able to adopt the pseudo score statistics as the selection criteria to include a variable at each iteration. Furthermore, the p-value associated with the U-statistics can be set up as a stopping rule for iteration termination.

The rest of this paper is organized as follows. We give a brief introduction on the ROC curve, forward selection and score statistics in section 2, and move on to the pseudo score statistics calculation based on AUCROC in section 3. Section 4 describes the implementation of the AUCFW algorithm in details. Simulation studies are conducted in section 5 and a real data application is analyzed in section 6. Conclusion and discussion can be found in section 7.

3.2 Background

3.2.1 The Receiver Operating Characteristic Curve

The ROC curve was first developed as a signal detection technique to assess the accuracy of radar equipments in World War II (Egan (1975)). Later the ROC curve gained great amount of attention in solving medical decision problems due to its assessment on both false positive and false negative errors (Lusted (1978)), and recently have been widely used for model evaluation and selection in many applications of machine learning (Spackman, Kent A. (1989)). Fawcett (2004) gave a tutorial introduction and a practical guide on the researches and applications of the ROC curve. The curve itself describes a tradeoff between FPR (as x-axis) and TPR (as y-axis) based on an ordered object ranking as the threshold c varies in the classification rule (3.1). Each classification decision determined by a threshold c can be presented in a confusion matrix as in Table 1.1, which is composed of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). Then the metrics in ROC space can be written as:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \quad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{3.4}$$

Each confusion matrix produces a single point in ROC space. As the threshold c varies, a series of confusion matrices are generated and the corresponding one-to-one points in ROC space are obtained.

Area under the Curve

The ROC curve is useful for visualizing a series of confusion matrices and evaluating the overall classification performance. Its AUCROC is a scalar measure to summarize the whole curve as well as the discriminatory accuracy of a classifier. Given a linear classifier $\beta^T X$, the AUCROC expression in probability as in Equation (3.3) is given by:

$$\mathrm{AUCROC}(\beta) = \Pr(\beta^T \tilde{X}_{11} > \beta^T \tilde{X}_{01}). \tag{3.5}$$

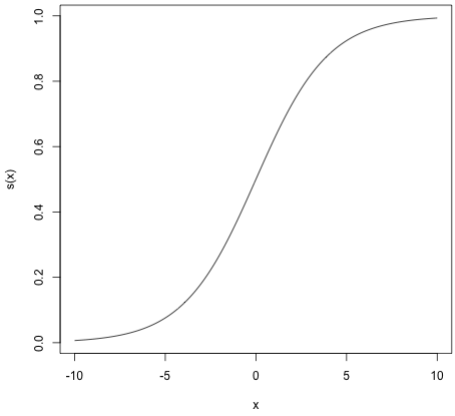
Replacing the probability in (3.5) by its empirical version, we obtain the estimated AUCROC value as equivalent to the Mann-Whitney statistic for a two-sample problem(Hanley and McNeil (1982)), which is expressed as:

$$\hat{\text{AUCROC}}_{\text{MW}}(\beta) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{I}(\beta^T X_{1i} - \beta^T X_{0j} > 0), \tag{3.6}$$

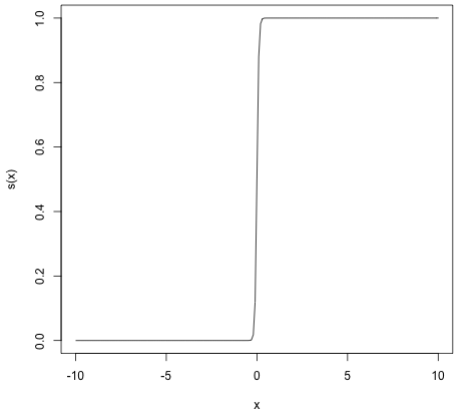
where $X_{1i}$is the covariate vector for the i^{th} object in the positive group, and X_{0j} is the covariate vector for the j^{th} object belonging to the negative group.

The Sigmoid Approximation

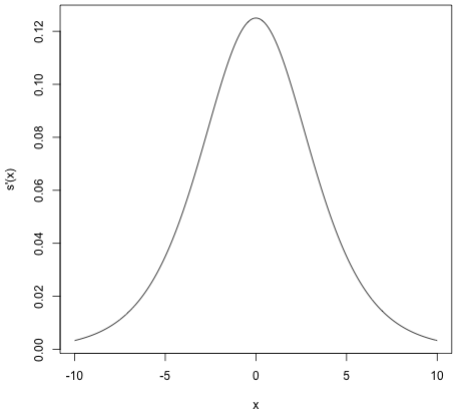
Although the empirical estimated AUCROC in Equation (3.6) has a nice interpretation and is easy to compute, it is a neither continuous nor differentiable function of β. Thus it is a challenging task to find the maximizer for $\hat{\text{AUCROC}}_{\text{MW}}$. The primary difficulty comes from the indicator function, $\mathbb{I}(x) = \mathbb{I}(x > 0)$. We proposed to use the sigmoid function, $s(x) = \dfrac{1}{1+\exp(-x/\sigma_N)}$, to approximate the indicator function. The parameter σ_N in $s(x)$ is a scale parameter, which determines the magnitude of the sigmoid function's approximation to the indicator function. In order to illustrate the important role of σ_N in approximation, we gave plots of the sigmoid function and its first derivative, $s'(x) = \dfrac{\exp(-x/\sigma_N)}{\sigma_N(1+\exp(-x/\sigma_N))^2}$, under two extreme values of σ_N shown in Figure 3.1. The plot in Figure 3.1(a) displays a somewhat linear fashion under $\sigma_N = 2$, thus it is a poor approximation to the indicator function. The plot for its corresponding first derivative function, shown in Figure 3.1(c), is a also smooth curve and doesn't reach the value of 0 until $\mid x \mid$ goes beyond the range of 10. On the other hand, Figure 3.1(b) shows a nearly vertical jump from value of 0 to 1 under $\sigma_N = 0.05$, so it can be considered a perfect approximation. Its first derivative, in Figure 3.1(d), reaches and keeps at the value of 0 beyond a tiny neighbor of x around 0. Therefore, in order to obtain an excellent approximation to the indicator function, a small value of σ_N is needed. In order words, we require large values of $\mid x/\sigma_N \mid$ to assure accuracy.



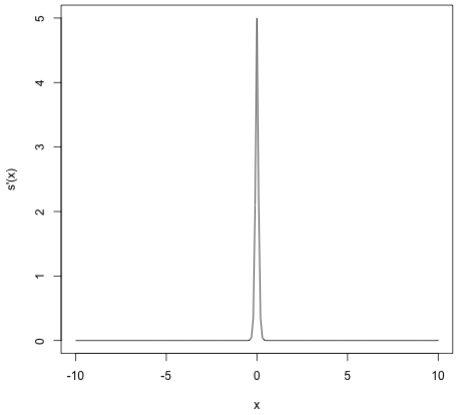
(a) The sigmoid function $s(x)$
$\sigma_N = 2$



(b) The sigmoid function $s(x)$
$\sigma_N = 0.05$



(c) The first derivative $s'(x)$
$\sigma_N = 2$



(d) The first derivative $s'(x)$
$\sigma_N = 0.05$

Figure 3.1: The approximation of the sigmoid function, $s(x) = \dfrac{1}{1+\exp(-x/\sigma_N)}$, to the indicator function, and its first derivative with respect to x, $s'(x) = \dfrac{\exp(-x/\sigma_N)}{\sigma_N(1+\exp(-x/\sigma_N))^2}$, under different values of the scale parameter σ_N.

By applying the sigmoid function to approximate $\widehat{\text{AUCROC}}_{\text{MW}}$ in Equation (3.6), we obtain a continuous and differentiable AUCROC expression as:

$$\widehat{\text{AUCROC}}_{\text{MWS}}(\beta) = R_N(\beta) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{1+\exp(-\beta^T(X_{1i}-X_{0j})/\sigma_N)}. \tag{3.7}$$

This is the target function to be maximized in this article. As discussed before, small value of $\mid \beta^T(X_{1i}-X_{0j})/\sigma_N) \mid$ leads to poor accuracy of approximation and consequently introduces systematic bias in β estimation. To improve the accuracy, we choose σ_N to guarantee $\mid \beta^T(X_{1i}-X_{0j})/\sigma_N \mid > 5$ (Ma and Huang (2005)), so that $R_N(\beta)$ estimator could be close to $\widehat{\text{AUCROC}}_{\text{MW}}$ estimator. The approach for determining σ_N is given in section 4. The first derivative of $R_N(\beta)$ with respect to the q^{th} variable, $\beta_{q^{\text{th}}}$, is represented as

$$\frac{\partial R_N}{\partial \beta_{q^{\text{th}}}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\phi(X_{1i},X_{0j})(x_{1iq^{\text{th}}}-x_{0jq^{\text{th}}}), \tag{3.8}$$

where

$$\phi(X_{1i},X_{0j}) = \frac{\exp(-\beta^T(X_{1i}-X_{0j})/\sigma_N)}{\sigma_N(1+\exp(-\beta^T(X_{1i}-X_{0j})/\sigma_N))^2}, \tag{3.9}$$

and $(x_{1iq^{\text{th}}}-x_{0jq^{\text{th}}})$ is difference of the q^{th} variable between the object i in the positive group and the object j in the negative group.

There are two remarks about the functionality of the classifiers when using the estimated AUCROC in either (3.6) or (3.7). First, any constant adding to the score value doesn't alter the ranking among objects or the empirical estimated AUCROC value. So it is fair to assume that no intercept is included in the linear classifier $\beta^T X$. Then we use the notation β to denote the coefficient vector without intercept: $\beta = (\beta_1,\ldots,\beta_p)^T$. Second, any multiplier to the current score value does not alter the object ranking either. In other words, β is only identified up to a scalar-multiple. In order to make the coefficient identifiable, we set $\mid \beta_{(1)} \mid = 1$, where $\beta_{(1)}$ is the first selected variable and serves as the 'anchor'.

3.2.2 Forward Selection

Both forward selection and backward elimination are the two useful greedy search strategies to find the optimal subset of variables out of the pool candidates. Given a pre-specified selection criterion, forward selection starts with no variables in the model, tests the addition of each variable, adds the one that has the most significant improvement to the model, and repeats this process until none meets the criterion. Backward elimination, on the other hand, starts with the complete set of variables, and removes one variable at each iteration that makes the least contribution to the model until none is in the model. The pros and cons for these two search approaches were discussed by Mantel (1970) and Hocking (1976). In general, forward selection is preferred to backward elimination owing to: (1) it generates a solution path describing the sequence of variables entering the model in an order of importance; (2) it provides an efficient approach to obtain the important variables in the situation where the proportion of important variables is small; (3) it can deal with the situation where the number of covariates exceeds the number of observations. For the proposed algorithm, we adopt forward selection as the search strategy to obtain the optimal subset.

3.2.3 Hypothesis Testing

In this section, we aim to demonstrate that the first derivative of AUCROC target function can be treated as a score statistic. It is an appropriate selection criteria and test statistic for obtaining a solution path as well as setting a stopping rule. First, we need to specify the hypothesis testing at the q^{th} iteration of variable selection. Given $q-1$ covariates selected in the model, the hypothesis testing is interested to investigate if there exists a q^{th} covariate making significant improvement, and is given by:

$$H_0 : \beta_{(q)|q-1} = 0 \quad \text{vs.} \quad H_a : \beta_{(q)|q-1} \neq 0, \tag{3.10}$$

where $\beta_{(q)}$ indicates the q^{th} variable entering the model and the notation $\beta_{(q)|q-1}$ denotes the efficient for the q^{th} variable given the previous $q-1$ variables.

Hypothesis Test Statistics

Under likelihood theory, likelihood ratio, Wald and score statistics are the three most commonly used statistics assessing the magnitude of model improvement and statistics significance. Suppose we have a log-likelihood function $l(\beta; \mathbf{X})$, and the null hypothesis testing is $H_0 : \beta = \beta_0$. The three test statistics can be calculated as:

- LR $= 2l(\hat{\beta}_{H_0}; \mathbf{X}) - 2l(\hat{\beta}_{H_a}; \mathbf{X})$,
  where $\hat{\beta}_{H_0}$ is the maximal likelihood estimator (MLE) under the null and $\hat{\beta}_{H_a}$ is the MLE without any constraints;
- W $= (\hat{\beta}_{H_a} - \beta_0)^T \mathbf{\Sigma}^{-1} (\hat{\beta}_{H_a})(\hat{\beta}_{H_a} - \beta_0)$,
  where $\mathbf{\Sigma}(\beta)$ is the covariance matrix for β, and $\mathbf{\Sigma}^{-1}$ denotes the inverse matrix;
- S $= \mathbf{u}^T(\hat{\beta}_{H_0}) \mathbf{I}^{-1}(\hat{\beta}_{H_0}) \mathbf{u}(\hat{\beta}_{H_0})$,
  where $\mathbf{u}(\beta)$ is the score vector of $l(\beta; \mathbf{X})$ and $\mathbf{I}(\beta)$ is the information matrix.

Although the above three hypothesis test statistics are derived differently, all three are asymptotically equivalent. They all have approximately a chi-squared distribution with p_0 degrees of freedom in large samples, where p_0 is the number of parameters investigated. Meanwhile, they have different performances on assessing statistics significance. Rao and Mukerjee (1997), Yi and Xikui (2011) and Sutradhar and Bartlett (1993) discussed and compared the three test statistics under different problem settings and scenarios. In terms of calculation, the main differences among the three statistics are summarized as:

- Likelihood Ratio statistics needs to evaluate MLE under both the null and alternative hypotheses;
- Wald statistics compares the difference between MLE under the alternative and the null value in relation to its variance;

- Score statistics tests the vector of gradients of the log-likelihood and evaluated at the null hypothesis.

Comparing to either Likelihood Ratio or Wald statistics, score statistics only needs to evaluate MLE under the null. So there is significant computational benefit of using score statistics when implementing forward selection, especially when investigating high dimensional data. The coefficient β plugged in to calculate the score statistic for the current iteration is the coefficient estimation obtained from the last iteration by maximizing the target function. On the other hand, if backward elimination is implemented, then Wald statistic would be a better option among three.

Score Statistics

Recall that, under likelihood theory, score statistics for testing the overall effect of β is presented as

$$\mathrm{S} = \mathbf{u}^T(\hat{\beta}_{H_0})\mathbf{I}^{-1}(\hat{\beta}_{H_0})\mathbf{u}(\hat{\beta}_{H_0}), \tag{3.11}$$

where the score vector $\mathbf{u}(\beta)$ is the first derivative of log-likelihood function with respect to β, and the information matrix $\mathbf{I}(\beta)$ is the covariance matrix of the score vector. Under some regularity conditions, it can be proved that $E(\mathbf{u}(\beta)) = \mathbf{0}$. And the score statistics itself has an asymptotic chi-square distribution. Equivalently, the score vector $\mathbf{u}(\beta)$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}(\beta)$.

Rather than investigating the overall effect, we are interested to see the conditional effect of each candidate variable in forward selection. The corresponding hypothesis testing is specified in (3.10). The distribution for a conditional score equation, $\mathrm{u}_{(q)|q-1}$, can be directly derived from the joint normal distribution of the score vector $\mathbf{u}(\beta)$ and is given as:

$$\mathrm{u}_{(q)|q-1} \sim \mathrm{N}(0, \sigma^2_{(q)|q-1}), \tag{3.12}$$

$$\sigma^2_{(q)|q-1} = \sigma^2_{(q)} - \mathbf{\Sigma}^T_{(q),q-1}\mathbf{\Sigma}_{q-1}\mathbf{\Sigma}_{(q),q-1}, \tag{3.13}$$

where $\sigma^2_{(q)}$ is the variance of the score function, $\mathrm{u}_{(q)}$, for the q^{th} variable entering the model, $\mathbf{\Sigma}_{q-1}$ is the covariance matrix of the score vector, $\mathbf{u}_{q-1}$, for the $q-1$ variables already in the model, and $\Sigma_{(q),q-1}$ is the covariance vector of $\mathrm{u}_{(q)}$ and $\mathbf{u}_{q-1}$. Then the standardized conditional score statistics is

$$S_{(q)|q-1} = \frac{\mathrm{u}_{(q)|q-1}}{\sigma_{(q)|q-1}}, \tag{3.14}$$

where $S_{(q)|q-1}$ converges to a standard normal distribution. The covariate, whose absolute conditional score statistic in Equation (3.14) has the maximum value among candidates, is considered to have the most potential to improve the model by the optimal β_q estimator under the alternative hypothesis. Besides, the associated p-value can be calculated from the asymptotic standard normal distribution, and used as a stopping rule to determine the statistics significance.

The next step is to construct a pseudo score statistics for our own problem. The empirical estimated AUCROC in Equation (3.7) is the objective function to be maximized, and then its partial derivative in Equation (3.8) is the corresponding pseudo score equation with respect to the q^{th} covariate. The derived pseudo score equation is a weighted sum of the differences of the q^{th} covariate between all possible pairs of objects from the positive and negative groups. We notice that the weight, displayed in Equation (3.9), for each pair only depends on the linear difference between the covariates already chosen in the model, $\beta^T(X_{1i} - X_{0j})$, and the scale parameter σ_N. From Figure 3.1(a) and 3.1(c), we say that, given a scale parameter σ_N, the further the approximated value by the sigmoid function is away from 0 or 1, the more weight the pair would have for the next candidate covariate. However, if a extremely small value of σ_N is chosen, e.g. $\sigma_N = 0.05$ in Figure 3.1(d), we can tell that the weight function ϕ for a great portion of pairs would be 0, since there is only a small range in $\beta^T(X_{1i} - X_{0j})$ for non-zero weight. In that case, the information from the pairs whose weights ϕ are 0 would be lost. Consequently, the potential contribution from the candidate covariate is diluted. Furthermore, smaller the value of σ_N is, more information would be lost. Thus, small σ_N would make the contributions among candidates non-differentiable, and consequently veil the potential significant improvement to the model.

Besides, in order to apply the score statistics to the proposed AUCFW algorithm, we need to

1. derive covariance matrix of the score vector $\frac{\partial R_n}{\partial \beta}$ defined in Equation (3.8), and the conditional variance $\sigma^2_{(q)|q-1}$;
2. obtain the asymptotic distribution of the score vector $\frac{\partial R_n}{\partial \beta}$ as well as the conditional score statistics $S_{(q)|q-1}$.

3.3 Statistics Inference Using Two-sample U-statistics

We examine the relationship between U-statistics and the pseudo score equation in Equation (3.8) on the basis of a two-sample U-statistics. Suppose we have $\tilde{X}_{11}, \ldots, \tilde{X}_{1n}$, $\tilde{X}_{01}, \ldots, \tilde{X}_{0m}$ independently and identically distributed with cumulative distribution functions F_1 and F_0, and a kernel function ψ is defined with 2 arguments. Given two sample sets $X_{11}, \ldots, X_{1n}$ and $X_{01}, \ldots, X_{0m}$, we obtain a two-sample U-statistics as

$$U_{n,m} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \psi(X_{1i}, X_{0j}). \tag{3.15}$$

It represents the average of the kernel function $\psi(X_1, X_0)$ under all possible paris from the two sample sets. Moreover, it can be shown that $U_{n,m}$ is an unbiased estimator for

$$\theta = \int \int \psi(u, v) dF_1(u) dF_0(v).$$

More generally, a kernel function can be defined with $k + l$ arguments and is symmetric in each set of arguments, which is given as $\psi(X_{11}, \ldots, X_{1k}; X_{01}, \ldots, X_{0l})$. In this article, the primary interest is the U-statistics in Equation (3.15), for the pseudo score equation in (3.8) is a two-sample U-statistics with a bivariate kernel. We use U to denote a two-sample U-statistics and re-write the score equation in (3.8) as the U-statistics presentation to simplify the notation in the remaining sections:

$$U^{(q)} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \psi^{(q)}(X_{1i}, X_{0j}). \tag{3.16}$$

where $\psi^{(q)}(X_{1i}, X_{0j}) = \phi(X_{1i}, X_{0j})(x_{1iq^{\text{th}}} - x_{0jq^{\text{th}}})/\sigma_N$, and $\phi(X_{1i}, X_{0j})$ is presented in Equation (3.9). Furthermore, since the optimal β_q is expected to reach the maximum value of empirical estimated AUCROC in (3.7), it can be obtained that $E(U^{(q)}) = 0$.

3.3.1 Covariance Matrix for a Vector of Two-sample U-Statistics

In order to derive the conditional variance $\sigma^2_{(q)|q-1}$ for the score statistics $S_{(q)|q-1}$ in (3.14), it is useful to investigate the variance-covariance matrix for a vector of two-sample U-statistics based on the same two independent samples $X_{11}, \ldots, X_{1n}$ and $X_{01}, \ldots, X_{0m}$. We first give the variance expression for a single two-sample U-statistics, $U^{(q)}$, and then extend that expression to the covariance derivation between $U^{(s)}$ and $U^{(t)}$.

Based on the definition of two-sample U-statistics in Equation (3.15), we see that the variance of U is composed of the sum of the covariances between kernels under all possible pairs $\{X_{1i}, X_{0j}\}$. Thus the variance for one single U-statistics is given by:

$$\begin{aligned} Var(U) &= \binom{n}{1}^{-2}\binom{m}{1}^{-2} \sum_{i,i'=1}^{n}\sum_{j,j'=1}^{m} Cov[\psi(X_{1i}, X_{0j}), \psi(X_{1i'}, X_{0j'})] \\ &= \binom{n}{1}^{-2}\binom{m}{1}^{-2} \left[\sum_{i=i'}\sum_{j\neq j'} \varsigma^2_{1,0} + \sum_{i\neq i'}\sum_{j=j'} \varsigma^2_{0,1} + \sum_{i\neq i'}\sum_{j\neq j'} \varsigma^2_{1,1}\right], \\ &= \frac{1}{(mn)^2}\left[n(m-1)\varsigma^2_{1,0} + (n-1)m\varsigma^2_{0,1} + nm\varsigma^2_{1,1}\right] \end{aligned} \tag{3.17}$$

where $\varsigma^2_{1,0}$ is the covariance value when the element from the set $X_{11}, \ldots, X_{1n}$ is identical, $\varsigma^2_{0,1}$ is the covariance value when the element from the set $X_{01}, \ldots, X_{0m}$ is identical, $\varsigma^2_{1,1}$ is the covariance value when two elements are identical, and the coefficient in front denotes the number of $Cov[\psi(X_{1i}, X_{0j}), \psi(X_{1i'}, X_{0j'})]$ in each scenario. Furthermore, the derivation for $\varsigma^2_{1,0}$,

as an example, is written as:

$$\begin{aligned}
\varsigma_{1,0}^2 &= Cov[\psi(\tilde{X}_{1i}, \tilde{X}_{0j}), \psi(\tilde{X}_{1i}, \tilde{X}_{0j'})] \\
&= E(\psi(\tilde{X}_{1i}, \tilde{X}_{0j})\psi(\tilde{X}_{1i}, \tilde{X}_{0j'})) - E(\psi(\tilde{X}_{1i}, \tilde{X}_{0j}))E(\psi(\tilde{X}_{1i}, \tilde{X}_{0j^j})) \\
&= E(E(\psi(\tilde{X}_{1i}, \tilde{X}_{0j})\psi(\tilde{X}_{1i}, \tilde{X}_{0j'}) \mid \tilde{X}_{1i})) - E(E(\psi(\tilde{X}_{1i}, \tilde{X}_{0j}) \mid \tilde{X}_{1i}))E(E(\psi(\tilde{X}_{1i}, \tilde{X}_{0j^j}) \mid \tilde{X}_{1i})) \\
&= E(\psi_{1,0}^2(\tilde{X}_{1i})) - E^2(\psi_{1,0}(\tilde{X}_{1i})) \\
&= Var(\psi_{1,0}(\tilde{X}_{1i})),
\end{aligned} \tag{3.18}$$

where we use the notation $\psi_{1,0}(\tilde{X}_{1i})$ to indicate the conditional expectation of the kernel $E(\psi(\tilde{X}_{1i}, \tilde{X}_{0j^j}) \mid \tilde{X}_{1i})$. The similar derivation can be given to $\varsigma_{0,1}^2$ and $\varsigma_{1,1}^2$.

Applying the variance calculation in Equations (3.17) and (3.18), we obtain the variance for the score function with respect to the q^{th} covariate as:

$$Var(U^{(q)}) = \frac{1}{nm}\big[(m-1)\varsigma_{1,0}^{(q)2} + (n-1)\varsigma_{0,1}^{(q)2} + \varsigma_{1,1}^{(q)2}\big], \tag{3.19}$$

where

$$\begin{aligned}
\varsigma_{1,0}^{2(q)} &= Var\big(\psi_{1,0}^{(q)}(\tilde{X}_{11})\big), \\
\varsigma_{0,1}^{2(q)} &= Var\big(\psi_{0,1}^{(q)}(\tilde{X}_{01})\big), \\
\varsigma_{1,1}^{2(q)} &= Var\big(\psi^{(q)}(\tilde{X}_{11}, \tilde{X}_{01})\big).
\end{aligned} \tag{3.20}$$

The derivation for a two-sample U-statistics variance can be generalized to the case of p-dimensional U-statistics. Suppose we have two two-sample U-statistics, $U^{(s)}$ and $U^{(t)}$, both based on a common sample $X_{11}, \ldots, X_{1n}, X_{01}, \ldots, X_{0m}$ but having different bivariate kernels $\psi^{(s)}$ and $\psi^{(t)}$. Similar to the variance derivation for the pseudo score equation in (3.19), the

covariance can be obtained by

$$Cov(U^s, U^{(t)}) = \frac{1}{nm}\big[(m-1)\varsigma_{1,0}^{2((s),(t))} + (n-1)\varsigma_{0,1}^{2((s),(t))} + \varsigma_{1,1}^{2((s),(t))}\big], \tag{3.21}$$

where

$$\begin{aligned}
\varsigma_{1,0}^{2((s),(t))} =& Cov\big[\psi_{1,0}^{(s)}(\tilde{X}_{11}), \psi_{1,0}^{(t)}(\tilde{X}_{11})\big], \\
\varsigma_{0,1}^{2((s),(t))} =& Cov\big[\psi_{0,1}^{(s)}(\tilde{X}_{01}), \psi_{0,1}^{(t)}(\tilde{X}_{01})\big] \\
\varsigma_{1,1}^{2((s),(t))} =& Cov\big[\psi^{(s)}(\tilde{X}_{11}, \tilde{X}_{01}), \psi^{(t)}(\tilde{X}_{11}, \tilde{X}_{01})\big].
\end{aligned} \tag{3.22}$$

By now, we know that a p-vector of two-sample U-statistics $\mathbf{U}$ has the mean θ and the variance-covariance matrix $\mathbf{\Sigma} = (\sigma^{2((s),(t))})$ defined in (3.19) and (3.21). Applying Equation (3.13) on the variance-covariance matrix $\mathbf{\Sigma}$ to get the conditional variance $\sigma^2_{(q)|q-1}$, we are able to obtain the conditional score statistics using the expression in (3.14).

3.3.2 Asymptotic Distribution

The asymptotic distribution for a single two-sample U-statistics can be obtained based on Hoeffding's decomposition and central limit theorem (Hoeffding (1961), Lee (1990)). Given the two-sample U-statistics in (3.16), we have its asymptotic distribution as

$$\sqrt{N}(U^{(q)} - \theta^{(q)}) \xrightarrow{D} N(0, p_1^{-1}\varsigma_{1,0}^{(q)2} + (1-p_1)^{-1}\varsigma_{0,1}^{(q)2}), \tag{3.23}$$

where $n/N \to p_1 \in (0,1)$ as n and $m \to \infty$.

It is easy to know that the joint limiting distribution for $\mathbf{U} = (U^{(1)}, \ldots, U^{(p)})^T$ is a multivariate normal distribution. If F_1 and F_0 are continuous and both of their second moments exist, then as $n \to \infty$ and $m \to \infty$, the joint distribution of

$$\big[\sqrt{N}(U^{(1)} - \theta^{(1)}), \sqrt{N}(U^{(2)} - \theta^{(2)}), \ldots, \sqrt{N}(U^{(p)} - \theta^{(p)})\big]$$

converges to a multivariate normal distribution with zero vector mean and the limiting variance-

covariance matrix $\mathbf{\Sigma}_{lim}$. The asymptotic variance-covariance matrix can be obtained as $\mathbf{\Sigma}_{lim} = (\sigma_{lim}^{2((s),(t)}$), where

$$\sigma_{lim}^{2((s),(t)} = \frac{1}{p_1}\varsigma_{1,0}^{2((s),(t))} + \frac{1}{1-p_1}\varsigma_{0,1}^{2((s),(t))},$$

with $n/N \to p_1 \in (0,1)$.

Knowing the asymptotic distribution for the U-statistics vector $\mathbf{U}$ is a multivariate normal distribution, it is obvious that the asymptotic conditional distribution for the two-sample statistics $\mathrm{u}_{(q)|q-1}$ is also a normal distribution with mean zero and variance $\sigma^2_{(q)|q-1}$ given by (3.13). Equivalently, the conditional score statistics $S_{(q)|q-1}$ represented in (3.14) converges to a standard normal distribution.

3.4 Implementation of the AUCROC Forward Selection Algorithm

Recall that no intercept is included in the linear classifier $\beta^T X$, we need to set $\mid \beta_{(1)} \mid = 1$ in order to make β identifiable. The covariate candidate, which leads to the maximal marginal AUCROC value, is selected as the anchor variable $\beta_{(1)}$. The coefficient sign assignment depends on the object ranking in the increasing or decreasing order by the anchor variable. Positive sign indicates that the decreasing order has the largest AUCROC value; and the increasing order results in negative sign.

3.4.1 Choice of the Scale Parameter σ_N

After the anchor is determined, we need to identify the value the scale parameter σ_N. Based on the previous discussion, the choice of σ_N is supposed to meet two objectives: obtaining the accuracy estimation for coefficient β in optimization and unveiling the significant model improvement by the candidates in variable selection. On one hand, we need σ_N to be small so that the empirical estimated AUCROC in (3.7) is an excellent approximation to the AUCROC expression in (3.6). On the other hand, the scale parameter σ_N cannot be too small, otherwise

a great portion of weights in (3.9) would go to 0 and consequently the candidate covariate difference by an important variable becomes hard to be detected. Hence, there is a trade off between assuring the accuracy of β and selecting the important variables. By only taking the optimization aspect into consideration, Ma and Huang (2005) chose the σ_N to be the largest number that satisfies the minimum of $\mid x_{1i(1)} - x_{0j(1)} \mid$ larger than 5, where $x_{1i(1)}$ and $x_{0i(1)}$ denote the covariate values of the anchor from the positive and negative groups. Therefore, by considering both two aspects, we choose the largest σ_N value that makes the 5-quantiles (Q5), 20-quantiles (Q20) and average (AVG) of the $\mid x_{1i(1)} - x_{0j(1)} \mid$ instead of the minimum value to be larger than 5. In other words, we set σ_N to be Q5/5, Q20/5, and AVG/5 respectively.

3.4.2 The AUCROC Forward Selection Algorithm

The AUCFW algorithm first establishes a variable selection solution path describing a sequence of variables entering the model, and then identifies the best subset by the p-value at each iteration. After choosing and identifying β_{q-1}, we implement the algorithm for the q^{th} iteration following the iterative steps:

(i) Compute the score function by plugging $\beta_{\mathbf{q-1}}$ using U-Statistics expression with respect to each covariate: $U^{(i)}, i = 1, \ldots, p$;

(ii) Derive the variance -covariance matrix $\mathbf{\Sigma}$ for the U-statistics vector $\mathbf{U} = (U^{(1)}, \ldots, U^{(p)})^T$;

(iii) Obtain the conditional variance $\sigma^2_{(i)|q-1}$ and the conditional score statistics $S_{(i)|q-1}$ for the remaining covariate candidates, $i = q, \ldots, p$;

(iv) Choose the candidate with the largest absolute value of $S_{(q)|q-1}$ and compute its p-value from a standard normal distribution;

(v) Estimate $\beta_{\mathbf{q}}$ by maximizing AUCROC in Equation (3.7).

When the associated p-value is less than a pre-specified Type I error α, the iteration can be stopped. Some software adopts 0.5 as the default cutoff when conducting automated forward

selection algorithm. For AUCFW algorithm, we discuss the relationship between p-value and the selected solution in the simulation section.

By ignoring the p-value stopping rule, we are able to obtain a complete solution path via forward selection. However, when problems are in hard settings, e.g. highly skewed data with $\pi = 0.01$, it is possible that no more additional variables is able to improve the model after certain iterations. In that case, we check the coefficient for the newly added variable. If it keeps at value of 0, then the iteration needs to be terminated by knowing that the optimal AUCROC value has been reached at the last iteration. Besides, if a problem runs into complete separation or quasi-complete separation, then we need to stop the iteration as well. In theory, the corresponding score function value is zero for any of the covariate candidates if separation is reached. In AUCFW algorithm implementation, since the absolute zero value is hard to be reached, we set a threshold for the score function U to detect separation, e.g. 10^{-50}.

For finding the maximizer β_q in step (v), we apply Newton's method algorithm for unconstrained optimization (nlm() function in R). The coefficient estimator β_{q-1} from the last iteration in AUCFW algorithm and the coefficient estimator from a linear logistic regression model are adopted as the starting values for β. Knowing the linear classifier by maximizing AUCROC has no intercept and a one-normed coefficient $\beta_{(1)}$, we rescale the estimates from logistic regression according to the anchor.

3.4.3 Empirical Estimator for Variance-Covariance Matrix

Knowing the distributions for the two independent random samples $\{\tilde{X}_1\}$ and $\{\tilde{X}_0\}$, F_1 and F_0, we are able to derive the expressions for the variance-covariance matrix $\mathbf{\Sigma}$ for a vector of U-statistics using Equations (3.19) and (3.21). However, when dealing with a real problem, the distributions F_1 and F_0 are normally unknown. One approach is to derive the empirical cumulative density functions, $\hat{F}_1$ and $\hat{F}_0$, based on the sample data. Alternatively, we can compute the empirical estimates for the components of the variance-covariance matrix directly using the variance and covariance of the conditional expectations shown in Equations (3.20)

and (3.22). Therefore, when implementing the step (ii) of AUCFW algorithm, we compute the empirical estimates using the following steps:

(a) Compute the empirical version of the conditional expectations, $\hat{\psi}^{(q)}_{1,0}(X_{1i})$ and $\hat{\psi}^{(q)}_{0,1}(X_{0j})$ for $q = 1, \ldots, p$, $i = 1, \ldots, n$ and $j = 1, \ldots, m$, as

$$\hat{\psi}^{(q)}_{1,0}(X_{1i}) = \frac{1}{m}\sum_{j=1}^{m}\psi^{(q)}(X_{1i}, X_{0j}), \quad \hat{\psi}^{(q)}_{0,1}(X_{0j}) = \frac{1}{n}\sum_{i=1}^{n}\psi^{(q)}(X_{1i}, X_{0j}).$$

(b) Estimate the variance-covariance components by plugging $\hat{\psi}^{(q)}_{1,0}(X_{1i})$ and $\hat{\psi}^{(q)}_{0,1}(X_{0j})$.

Then we obtain the empirical estimates for variance-covariance matrix $\hat{\mathbf{\Sigma}}$ based the observed sample data. Then the empirical estimates for the conditional covariance $\hat{\sigma}^2_{(q)|q-1}$ can also be easily computed.

3.4.4 Empirical Estimator Using A Random Sample

In order to improve the algorithm efficiency, we use a random sample of the objects rather than the whole N number of objects. In AUCROC estimation, it is known that the sampling unit is a pair of objects from the positive and negative groups respectively. Each pair has the equal chance to be selected under random sampling scheme. We choose $N^{'}$ pairs of random sampling (without replacement), where $N^{'}$ is proportional to the total number of pairs, e.g. $N^{'} = 0.2 * m * n$.

Knowing the interpretations for components in variance-covariance matrix of U-statistics, we make modifications in their calculations to accommodate to the case where a portion of sample pairs are used. Suppose we randomly select $N^{'}$ pairs $\{X^{'}_{1i'}, X^{'}_{0j'}\}$ from the original sample data, and $n^{'}$ and $m^{'}$ are the numbers of unique objects in the positive and negative groups respectively. Here are the modifications we need to make:

- The U-statistics using the sampled data in pairs is written as:

$$U^{'(q)} = \frac{1}{N'}\sum_{\{i',j'\}\in A}\psi^{(q)}(X^{'}_{1i'}, X^{'}_{0j'}), \quad \text{for } q = 1, \ldots, p,$$

where A is the set of paired index for the randomly selected pairs.

- The empirical version of the conditional expectation is modified as:

$$\hat{\psi}^{'(q)}_{1,0}(X^{'}_{1i'}) = \frac{1}{m_{i'}} \sum_{j'=1}^{m_{i'}} \psi^{(q)}(X^{'}_{1i'}, X^{'}_{0j'}), \quad \hat{\psi}^{'(q)}_{0,1}(X^{'}_{0j'}) = \frac{1}{n_{j'}} \sum_{i'=1}^{n_{j'}} \psi^{(q)}(X^{'}_{1i'}, X^{'}_{0j'}),$$

where $n^{'}$ and $m^{'}$ are the numbers of unique objects in the positive and negative groups respectively.

It is easy to find that we have $n^{'}$ number of empirical estimates for $\psi_{1,0}$ and $m^{'}$ number of empirical estimates for $\psi_{0,1}$ for each covariate.

3.5 Simulation Study

A simulation study is conducted to investigate how AUCFW algorithm performs to select important variables among the variable-pool. The comparisons between AUCFW algorithm and the regular forward selection in logistic regression (LogFW) are reported under the datasets with correlated covariates. We use an R package logi to perform LogFW, which is developed by Boos and Holloway (2013) and presented at IMPACT. The simulation results are summarized from three aspects: (1) variable selection complete solution path describing the sequence of variables entering the model; (2) the relationship between the selected solution and the associated p-value; (3) predictive performance using predicted AUCROC.

3.5.1 Data Generation

The simulation study aims to investigate AUCFW algorithm's performance on selecting variables to construct an optimal linear classifier. We perform the variable selection algorithms on the data pattern where the covariates are highly correlated. The covariate matrix $\mathbf{X}_{p\times N}$ is generated from a multivariate normal distribution with mean zero and covariance matrix using the autoregressive(1) (AR(1)) covariance structure. The $(i,j)^{\text{th}}$ element in AR(1) covariance structure is presented as $\sigma^2\rho^{|i-j|}$, where σ^2 is set to be 1 in the simulation for simplicity. The

covariate matrix $\mathbf{X}_{p\times N}$ indicates N number of objects in total and p dimensional of covariate vector for each object, where $p = 50$ (5 out of 50 variables are important) and $N = 500$ in the simulation. We assign the class label to each object based on the linear score $W = \beta^T X$. It is easy to derive that:

$$\tilde{W} \sim N(0, \sigma_W^2), \tag{3.24}$$

where $\sigma_W^2 = \beta^T Var(\mathbf{X})\beta$. Then a bernoulli distribution with parameter p_W is adopted for each object to determine the class label assignment, where p_W describes the probability to be assigned to the positive group given the score value W. This conditional probability can be represented using a logistic model written by

$$p_W = \Pr(Y = 1 \mid W) = \frac{\exp(W + a)}{1 + \exp(W + a)},$$

where a is a shifting parameter and its value can be derived from the pre-specified the proportion of positives π.

In addition to the settings described above, there are three primary simulation settings's effects to be investigated for the variable selection performance:

- Skewness of class label distribution, e.g. $\pi = 0.5$ (balanced data) and $\pi = 0.1$ (skewed data);
- Magnitude of correlation between covariates, e.g. $\rho = 0.7$ (weak correlation), $\rho = 0.8$ (moderate correlation) and $\rho = 0.9$ (strong correlation) in AR(1) covariance structure;
- Signal of covariates' coefficients: e.g. $\beta = (0.8, 0.8, 0.8, 0.8, 0.8)^T$ (equal signal) and $\beta = (0.4, 0.4, 0.8, 0.8, 0.8)^T$ (unequal signal).

It is know that the correlation between the covariates by AR(1) structure depends on the distance-between, so the farther the two covariates are the weaker the correlation they have. If the important covariates are set to be far from each others, then the correlation pattern fail to play the role. Thus the important covariates are generated in-a-row. Based on the interested simulation factors, we list four scenarios to be investigated as follows:

- Data#1: $p = 50$, $\pi = 0.5$, $\beta = (0.8, 0.8, 0.8, 0.8, 0.8, \overbrace{0, \ldots, 0}^{45})^T$;
- Data#2: $p = 50$, $\pi = 0.5$, $\beta = (0.4, 0.4, 0.8, 0.8, 0.8, \overbrace{0, \ldots, 0}^{45})^T$;
- Data#3: $p = 50$, $\pi = 0.1$, $\beta = (0.8, 0.8, 0.8, 0.8, 0.8, \overbrace{0, \ldots, 0}^{45})^T$;
- Data#4: $p = 50$, $\pi = 0.1$, $\beta = (0.4, 0.4, 0.8, 0.8, 0.8, \overbrace{0, \ldots, 0}^{45})^T$.

Within each scenario, we set ρ to be 0.7 / 0.8 / 0.9 in AR(1) covariance structure for examining the affect of correlation. Besides, 200 datasets are generated for each scenario, and the outcome presented are summarized based on the average performance of these 200 datasets. For each dataset, we randomly select 2500 pairs of positive and negative objects to conduct variable selection in the simulation study.

3.5.2 Variable Selection Complete Solution Path

One approach to examine the performance of a variable selection algorithm is the sequence of variables selected in the model. Thus the whole solution paths obtained by AUCFW algorithm and LogFW algorithm are compared. If we treat the true important variables as positives and the bad variables as negatives, the variable label assignment itself is a classification problem. Since we are interested in the ranking between important and unimportant variables rather than the ranking within each group, the ROC curve is a good tool to visualize and describe the solution path.

However, occasionally, either AUCFW algorithm or LogFW algorithm would result in an incomplete solution path owing to early termination, and then only partial important variables are selected. Thus we propose an approach to complete the solution path and the corresponding ROC curve, which is able to imitate the chance of selection among the remaining candidates. The proposed approach is conducted based on an assumption that the remaining variables, regardless of important or unimportant, have the same probability to be selected. Equivalently, in ROC space, this approach generates a straight line connecting the terminating point and

the ending point $(1, 1)$. The selection rate or the slope is exactly the fraction of important ones over the remaining candidates. The proposed approach might not be the best estimation or approximation to the true complete solution path. It can lead to underestimation by knowing that no matter how weak the magnitude of coefficient values the important variables have, the important one are supposed to have stronger signal to be selected than the unimportant ones do. Despite of this, it is the most applicable approach to complete the solution path.

The comparisons under different data settings for solution paths are displayed in Figure 3.3 - Figure 3.6. For each data setting, three subfigures (a), (b) and (c) are shown for different magnitudes of correlation between covariates, where the parameter ρ for AR(1) is 0.7, 0.8 and 0.9 respectively. On each subfigure, four ROC curves are generated. The ROC curve in red solid line is generated by LogFW algorithm, the curve in blue longdashed line is generated by AUCFW algorithm with σ_N = Q5/5 (AUCFWQ5), the curve in pink dashed line is generated by AUCFW algorithm with σ_N = Q20/5 (AUCFWQ20), and the curve in grey dotdashed line is generated by AUCFW algorithm with σ_N = AVG/5 (AUCFWAVG). The summarized curve is point-wise averaged at each FPR value over the complete solution paths from 200 simulated datasets. The averaged AUCROC and its standard error for the variable selection algorithms are also provided on each subfigure. The comparisons can be examined and summarized in two aspects:

- TPR at the early stage of variable ranking where FPR is low;
- AUCROC for the overall variable selection performance.

Comparing subfigures in Figure 3.3 for Data #1, where the class label is balanced and coefficients' signals are equally strong, we have the following observations. In Figure 3.3(a) where the correlation is weak, the ROC curve by LogFW algorithm outperforms those by AUCFW algorithms with different σ_N values. It reaches TPR of 0.9 when FPR is 0, which indicates at least 4 out of 5 good variables are selected before any bad ones enter the model. The overall AUCROC by LogFW also has the largest the values. On the other hand, among

the three ROC curves by AUCFW algorithms, AUCFWAVG has more power to select the important variables than the other two. This observation is consistent with the conclusion drawn from the previous section that small value of σ_N tends to veil the contributions of the important variables. In Figure 3.3(b) where the correlation is moderate, LogFW still has better overall performance than AUCFW in terms of AUCROC value. However, we observe that the disparity between LogFW and AUCFW at the early part decreases. Moreover, the ROC curves by AUCFWQ5 and AUCQ20 cross the curve by LogFW at the late stage of variable selection. It indicates that the accuracy of the coefficient estimation tends to iteratively optimize AUCROC by including more variables. Under strong correlation situation in Figure 3.3(c), LogFW and AUCFWAVG have similar performance, and AUCFWQ5 and AUCFWQ20 beat them in terms of the overall AUCROC. This outperformance in AUCROC is because that both LogFW and AUCFWAVG terminate the iteration early due to the high correlation while AUCFWQ5 and AUCFWQ20 continue to include more variables. Thus, the choice of σ_N depends on both data features and problem requirements. If the accuracy of the selection at the early stage is the primary interest, the value of AVG/5 is recommended to the scale parameter σ_N. If some large value of FPR is tolerated, AUCFWQ5 might be a better choice since it can provide more accurate estimates. Throughout the three subfigures in Figure 3.3, we can conclude in general that LogFW algorithm performs better than AUCFW algorithms under this easy setting, but the performances of AUCFW algorithms are much more stable as the magnitude of correlation varies. The affect of correlation between covariates on LogFW algorithm is stronger than the affect on AUCFW algorithms.

Looking into problems with harder settings, e.g. the datasets with unequal signal for coefficients in Figure 3.4, we see that the advantage of LogFW gets less obvious as the correlation changes from weak to strong. In Figure 3.4(c), the compared results are turned over that AUCFWAVG outperforms LogFW for the overall AUCROC value as well as the TPR value at the early stage. In addition, AUCFWQ20 performs as well as LogFW when FPR value is small. Due to the fact that the variables with smaller coefficient value are easy to be neglected during

variable selection process, all the algorithms have worse performance comparing to the datasets with equal signal. However, the impact on LogFW algorithm seems to be much stronger. Figure 3.5 and Figure 3.6 present the simulation results for the unbalanced class label data. Even in the case with weak correlation, AUCFW algorithms outperform LogFW at different stages of variable selection depending on the tolerance of FPR. Therefore, we can conclude that, AUCFW algorithm is more desirable and flexible variable selection algorithm when we are faced with hard problems, e.g. highly correlated data, skewed class label distribution, etc.

3.5.3 Stopping Rule Using P-value

As a variable selection algorithm, having a stopping rule that terminates the iteration and determines the final solution is essential. Taking advantage of the two-sample U-statistics, we are able to obtain the p-value associated with the pseudo score statistics from its asymptotic standard normal distribution. In this part, we are interested to investigate the relationship between the variable selection solution and its associated p-value, as well as to compare the p-value determined solutions between LogFW and AUCFW algorithms.

We still use curve presentation to compare the solutions between AUCFW and LogFW algorithms. Different from the variable selection complete solution paths, which is point-wise averaged at the value of FPR, the solution paths presented in this section are grid on continuous varied cutoff p-value. Thus, the p-value-based solution path is determined not only by the sequence of selected variables, but also by the significance of each variable's contribution to the model improvement. We summarize the solutions with p-value < 0.3 and present the outcome by two curves:

- the partial ROC curve with the specified range of FPR from 0 to 0.5;
- the precision-recall (PR) curve, which plots recall against precision.

The PR curve is an alternative to the ROC curve for classification performance evaluation under highly unbalanced data, where recall is the same as TPR and precision measures the fraction

of truly positives out of the classified positives. Because of the investigated range of p-value, we believe that the included solutions might not involve too many unimportant variables. In that case, the partial ROC curve would be displayed in a very limited range of FPR. On the other hand, the PR curve is more complete and is displayed in a wider range of recall, because the included solutions might contain the majority of the important variables. Besides, the PR curve is able to provide a clearer picture describing the early stage of variable selection.

Figure 3.7 - Figure 3.10 are the partial ROC curves and the PR curves for the summarized p-value based solution paths on different simulation settings. Each figure is composed of six subfigures. The three subfigures at top are the partial ROC curves for $\rho = 0.7, 0.8$ and 0.9 in AR(1) correlation structure respectively; and the three subfigures at bottom are the PR curves for the corresponding three magnitudes in correlation. In each subfigure, we use the same symbols to present the p-value-based solution paths generated by LogFW, AUCFWQ5, AUCFWQ20, and AUCFWAVG algorithms. One normally used criterion, which determines one solution is more desirable than the other, is the portion of important variables among the selected ones at the same level of TPR. In that case, we compare the precision values corresponding to one specified TPR value on the PR curves, and prefer the one with higher precision value. On the other hand, we also compare the partial ROC curve grid on p-value with the curve by the whole solution path to examine the impact of p-value.

For Data #1 displayed in Figure 3.7, we observe that, under weak correlation situation in Figure 3.7(a), the p-value-based solution path by LogFW is still superior to those by AUCFW algorithms. However, the TPR values at the point with no unimportant variables drop comparing to the complete solution paths. It indicates that both LogFW and AUCFW algorithms fail to show strong significance to include the next important variable owing to the correlation between covariates. When the correlation between covariates gets moderate in Figure 3.7(b) and 3.7(e), the solutions selected by AUCFWQ20 and AUCFWAVG become superiority to those by LogFW at the early stage of the partial ROC curve, which is a totally opposite story told from Figure 3.3(b); and the PR curve presentation confirms the superiority that AUCFWQ20

and AUCFWAVG have better performances than LogFW does when averagely 3 or 4 out of 5 important variables are selected respectively. This phenomenon indicates that as the correlation gets stronger, LogFW becomes more conservative comparing to AUCFW algorithm to include the next variable into the model, despite knowing the next one is the truly important one. Under the highly correlated situation in Figure 3.7(c) and 3.7(f), since AUCFWAVG already shows its superiority in the whole solution path presentation in Figure 3.3(c), AUCFWAVG still outperforms LogFW uniformly throughout the whole curve and the disparity between the two curves become more obvious. The same is applied to AUCFWQ20.

Furthermore, as the problems get harder, e.g. the unbalanced class label or unequal strong signal of coefficients, the benefit of implementing AUCFW algorithms becomes more obvious, that even the performance of variable selection by AUCFWQ5 is better than the performance by LogFW. Therefore, we can conclude that the correlation between covariates prevents LogFW algorithm from including the additional important variables, despite we know the next one in line is a good one. Although the correlation also masks the significance on AUCFW algorithms, the impact is much less obvious comparing to LogFW algorithm.

3.5.4 Predictive Performance

In addition to the variable selection solution path, we also investigate and compare the predictive performance for LogFW and AUCFW algorithms. Given the test sets with the same number of objects, Figure 3.11 - Figure 3.14 present the predictive performance in the graphs of 'predicted AUCROC vs. the number of selected variables' under different simulated scenarios. All the curves rise up as more variables are added to the model, reach its peak when around 5 variables are included, and drop down till the whole set of variables are selected. As problems get harder, the curve peak is achieved from 6 variable are need (see Figure 3.11) to 3 variables (see Figure 3.14). Besides, we see that the drop of the curve at the late stage is smooth and weak although many unimportant variables are added. It seems that the complexity of model doesn't put too much penalty on model generalization. This is because the predicted AUCROC

value depends on the object ranking in the test sets, and the addition of the bad variables can only alter the ranking within a very small region. LogFW and AUCFWAVG have quite close predictive performance when less than 10 variables are selected. The predicted AUCROC value for AUCFWAVG is more stable and flat in the late stage. Both AUCFWQ5 and AUCFWQ20 have worse predictive performance than AUCFWAVG. Although they might have more accurate estimates for the coefficient given the selected variables, it is obvious that the selection of important variables plays more important role in prediction.

3.6 Real Data Analysis

In this section we experiment the proposed AUCFW algorithms on the real classification problems, and then examine and compare its classification performance with LogFW algorithm using predicted AUCROC over test sets. AUCFWAVG algorithm is performed, since it has uniformly better performance in variable selection and prediction than AUCFWQ5 and AUCFWQ20. The prediction results are presented in graphs of 'predicted AUCROC vs. the number of selected variables' and 'predicted AUCROC vs. p-value'. All the real data sets are from the University of California Irvine UCI Repository of Machine Learning Databases.

3.6.1 Nominal Data

For the data with numerical attributes only, we determine the anchor variable by the marginal AUCROC. The sign of the anchor variable, positive or negative, indicates the direction that the classifier points towards. However, when there are nominal attributes involved, the identification of the anchor is complicated, which is composed of two steps: (1) identify the anchor attribute; (2) if a nominal attribute is chosen, then pick one category as the anchor variable. After creating dummy variables for each nominal attribute, the anchor attribute can be determined by the score statistics in quadratic form as in Equation (3.11). The interpretation for the coefficients in front of nominal attribute is totally different from that for nominal attribute. Consider a object ranking produced by a nominal attribute, the objects within each category are tied,

so the object ranking is only determined by the order among the coefficients. In that case, a positive coefficient indicates that the objects in the corresponding category position higher than those in the pre-assigned baseline category in a object ranking, and a negative coefficient denotes the opposite situation. We identify the optimal category order corresponding to the decreasing order of the proportions of positives within each category. The rationale for this is as follows. We require a category order that is able to achieve the maximum AUCROC value based on the sample data. Thus, we want TPR to reach its maximal value of 1 as early possible in ROC space. The shortest routine can be achieved if a convex ROC curve is obtained. In that case, we set the category having the steepest slope connecting with $(0, 0)$ to be the first point, and place the second point that corresponds to the the steepest slope connecting with the first point, and keep this procedure until the last category. It is easy to know that the slope between the two adjacent points is equivalent to the proportion of positives within the first category. We choose the category having the least and largest proportion of positives to be the baseline and the anchor variable respectively. Then the coefficient for the anchor variable is 1, and the coefficients for the remaining categories would be within the range of (0,1). Knowing the coefficient estimator $\beta_{(1)}$ for the anchor attribute, we choose the scale parameter σ_N based on the value of $\mid \beta_{(1)}^T(X_{1i(1)} - X_{0j(1)}) \mid$.

3.6.2 Experiment on Real Data

We apply the AUCFWAVG algorithm on three real datasets. Pima Indians Diabetes Data is a data set on a diabetes diagnosis on the population of females at least 21 years old of Pima Indian heritage (Smith et al. (1988)). It has been investigated through many classification approaches, e.g. neutral network, support vector machine and etc. (Smith et al. (1988), Bennett and Blue (1998)). SPECTF Heart Data is a data set on cardiac Single Proton Emission Computed Tomography (SPECT) images, which is used for application to cardiac SPECT diagnosis by Kurgan et al. (2001). The class label distribution is highly unbalanced. Adult Data is problem ,with both numeric and nominal attributes, to predict whether a person makes over 50K

a year. It is a data set extracted from the 1994 Census database, and was first analyzed using Naive-Bayes classifiers and decision-trees by Kohavi (1996). In our experiment, we randomly select 2/3 of the observations to be a training set and treat the remaining 1/3 as a test set. The description of datasets is shown in Table 3.1.

Table 3.1: Real Datasets Description

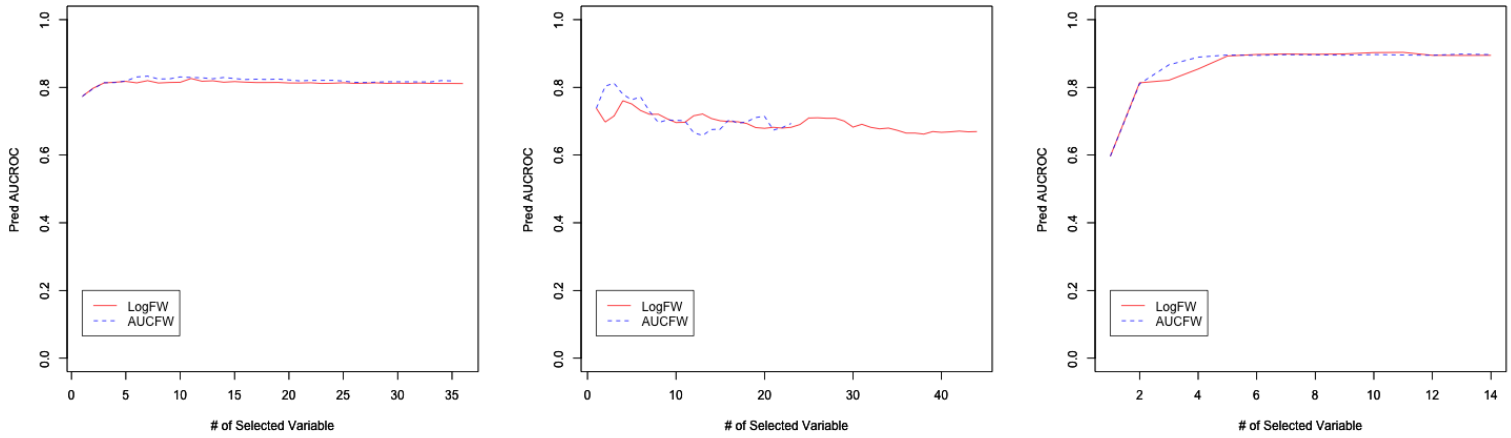
Dataset name	Number of attributes	Number of samples	Sample distribution in training set	Sample distribution in test set
Pima Indians Diabetes Data	36	768	178 : 333	90 : 167
SPECTF Heart Data	44	267	36 : 141	19 : 71
Adult Data	14	901	149 : 442	86 : 224

Figure 3.2 displays the comparisons of predication performance for LogFW and AUCFWAVG algorithms on the three real classification datasets. The predicted AUCROC values by the two algorithms are quite close to each other, but the one by AUCFWAVG in blue dashed line slightly outperforms the one by LogFW in red solid line. Figure 3.2(a) and 3.2(d) present the results for Diabetes Data. From Figure 3.2(a), we can tell that two algorithms select the same first three variables, but lead to different sequential variables afterwards. The predicted AUCROC by AUCFWAVG reaches the maximum value when 6 attributes are selected; on the other hand, the maximum of the predicted AUCROC by LogFW is achieved when 12 attributes are chosen. Further, the maximum value by AUCFWAVG is still higher than that by LogFW. If using p-value as the stopping criterion, the optimal subsets for both two algorithms can be achieved by p-value around 0.05. For SPECTF Data in Figure 3.2(b) and 3.2(e), the curves are relatively wiggled due to the unbalanced data. The sequences of selected variables are different after the first one is chosen. Besides, the maximum value of the predicted AUCROC by AUCFWAVG is reached when 2 attributes are selected, which cannot be transcended by LogFW. Adult Data is a problem with both numeric and nominal attributes. Since Logi function cannot solve the

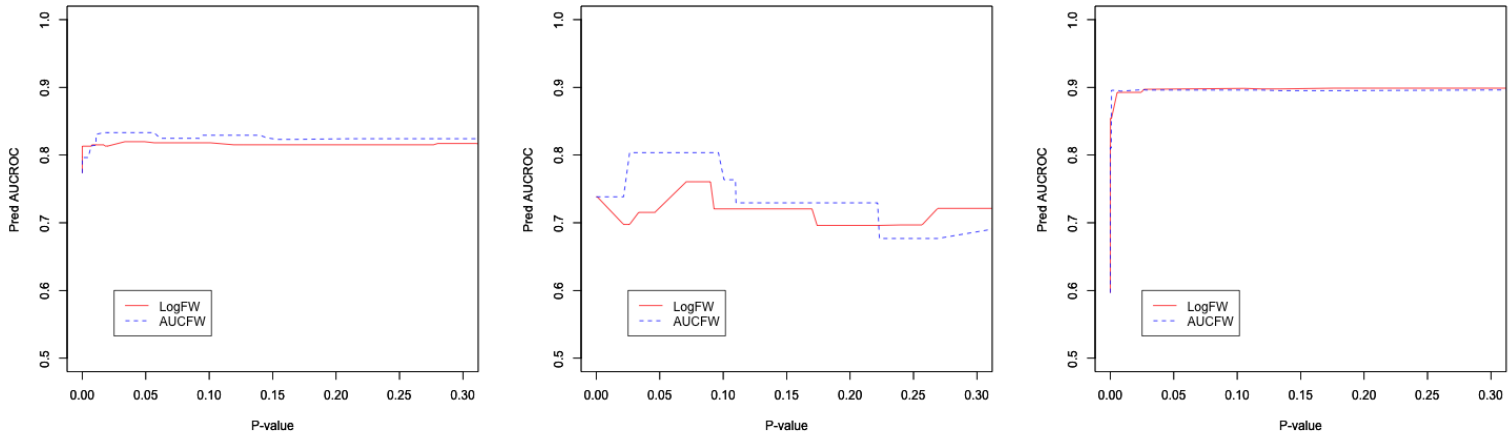
problem with nominal attributes, a new function is written to perform logistic regression in forward selection based on score statistics for nominal attributes. In Figure 3.2(c) and 3.2(f), we find that the prediction performance and the sequences of variable selection are quite close, except for the interval when the 3^{th} and 4^{th} attributes are chosen. Therefore, AUCFWAVG algorithm can perform as well as LogFW algorithm does in prediction.

3.7 Conclusion

In this chapter, we develop a non-parametric ROC-based forward variable selection algorithm by maximizing AUCROC to obtain an optimal subsets from variable-pool, and further present that AUCFW algorithm outperforms LogFW algorithm in terms of the variable selection solution path, especially under the data pattern with highly correlated covariates. Different from other variable selection approaches, the ROC-based method focuses on controlling false positive and false negative errors. We use the sigmoid function as an approximation to the indicator function, so that the empirical AUCROC expression is equivalent to the two-sample Mann-Whitney statistics. Due to its continuous and differentiable property, we derive the pseudo score statistics for the AUCROC objective function by taking the first derivative. Besides, take advantage of the theory of the two-sample U-statistics, we obtain the asymptotic distribution for the score statistics as well as the associated p-value as a stopping rule for forward selection procedure. This algorithm also extends naturally for both numerical and nominal attributes.

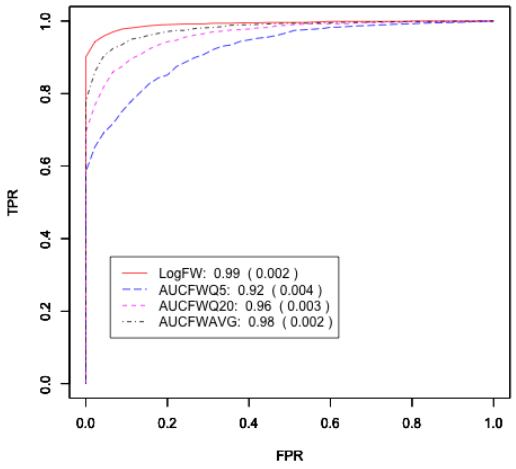


(a) Pima Indians Diabetes Data (b) SPECTF Heart Data (c) Adult Data

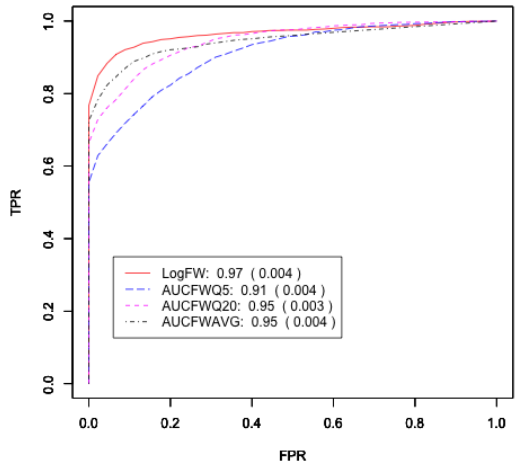


(d) Pima Indians Diabetes Data (e) SPECTF Heart Data (f) Adult Data

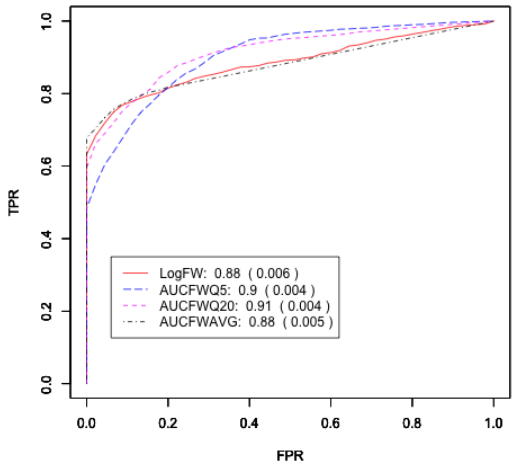
Figure 3.2: The comparisons between AUCFW and LogFW algorithms on the predicted AUCROC value over the test sets. Two graphs are generated for each real data: one plots "the predicted AUCROC vs. the number of selected variables" and the other plots "the predicted AUCROC vs. p-value".



(a) AR(1): $\rho = 0.7$

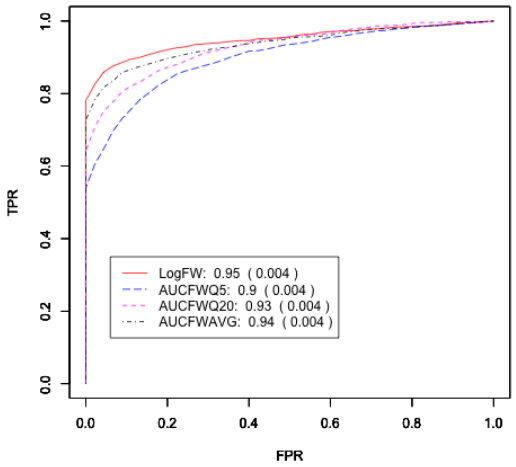


(b) AR(1): $\rho = 0.8$

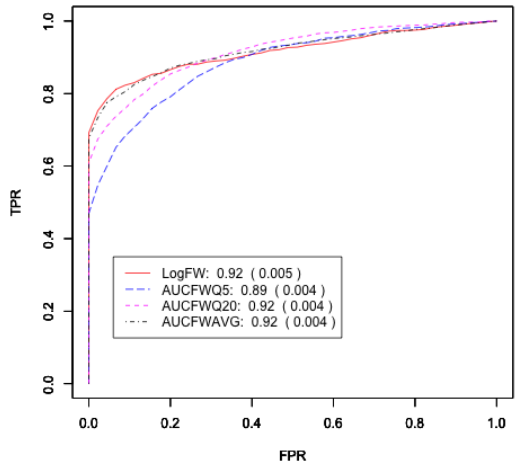


(c) AR(1): $\rho = 0.9$

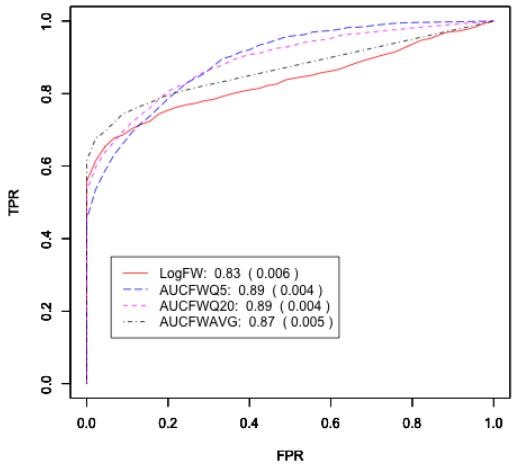
Figure 3.3: Solution path presented in the ROC curves for Data #1, where $N = 500$, $p = 50$, $\pi = 0.5$, and $\beta = (0.8, 0.8, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.



(a) AR(1): $\rho = 0.7$

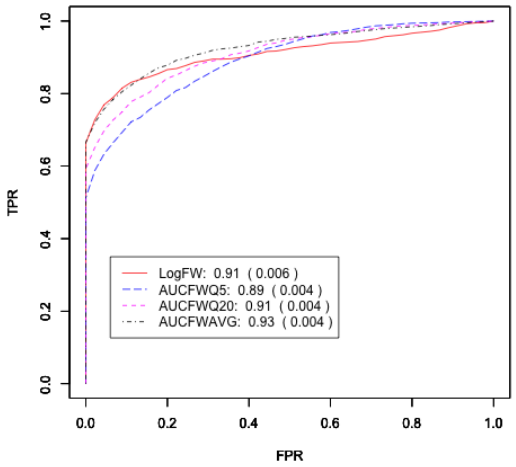


(b) AR(1): $\rho = 0.8$

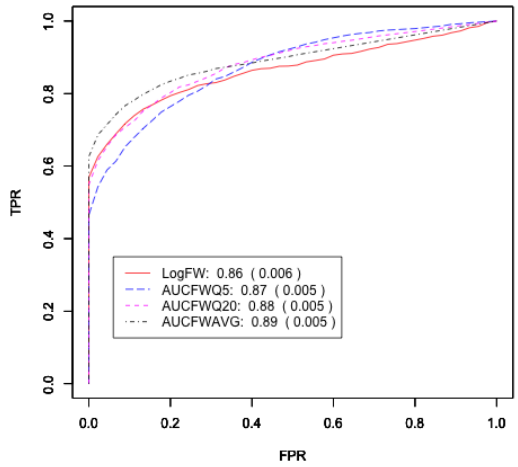


(c) AR(1): $\rho = 0.9$

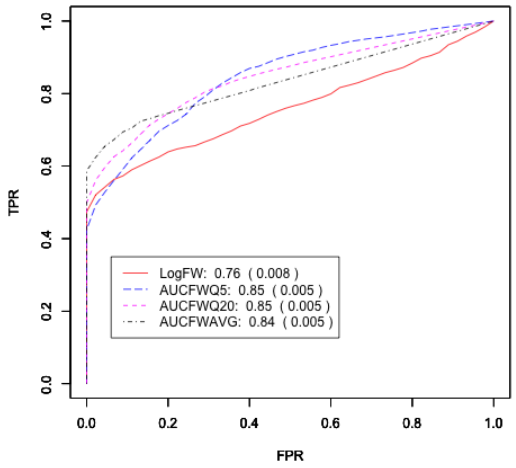
Figure 3.4: Solution path presented in the ROC curves for Data #2, where $N = 500$, $p = 50$, $\pi = 0.5$, and $\beta = (0.4, 0.4, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.



(a) AR(1): $\rho = 0.7$

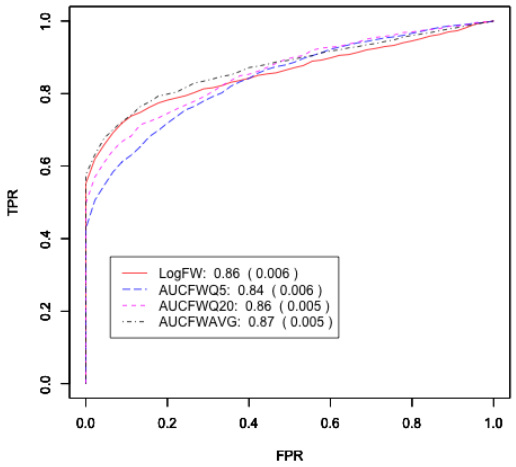


(b) AR(1): $\rho = 0.8$

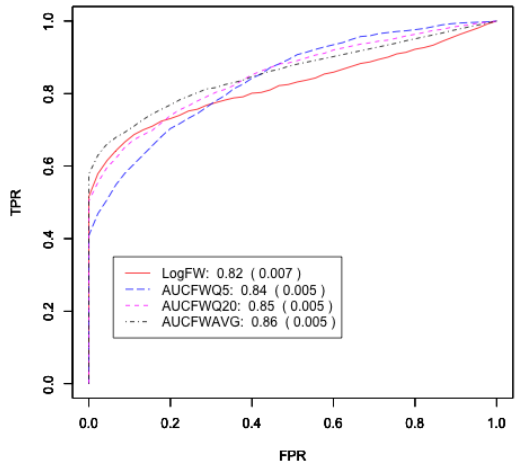


(c) AR(1): $\rho = 0.9$

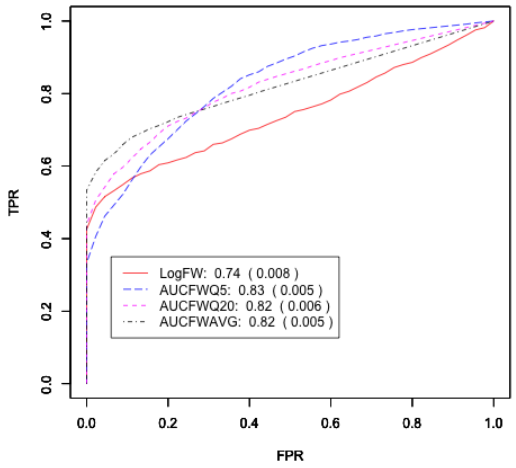
Figure 3.5: Solution path presented in the ROC curves for Data #3, where $N = 500$, $p = 50$, $\pi = 0.1$, and $\beta = (0.8, 0.8, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.



(a) AR(1): $\rho = 0.7$

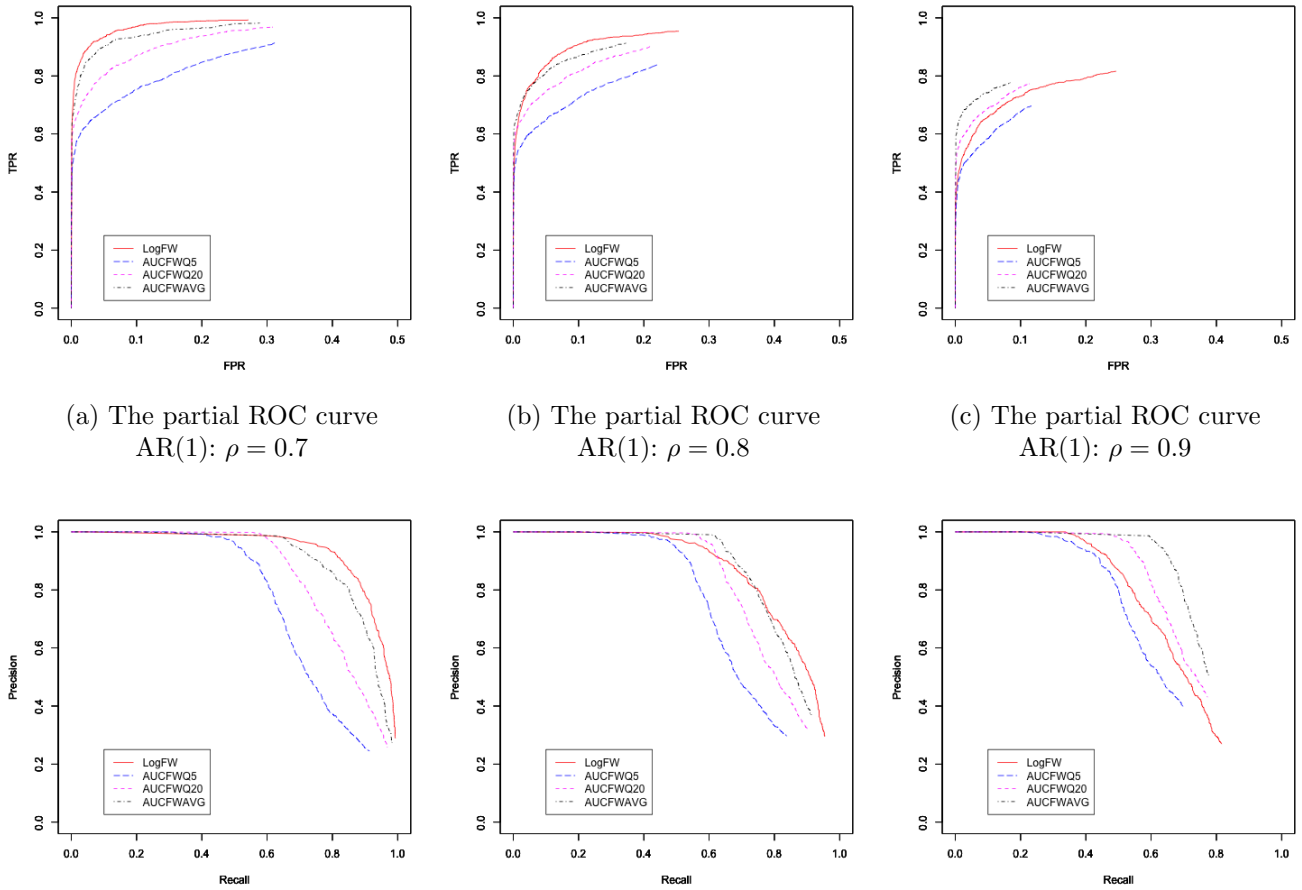


(b) AR(1): $\rho = 0.8$



(c) AR(1): $\rho = 0.9$

Figure 3.6: Solution path presented in the ROC curves for Data #4, where $N = 500$, $p = 50$, $\pi = 0.1$, and $\beta = (0.4, 0.4, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.



(a) The partial ROC curve AR(1): $\rho = 0.7$

(b) The partial ROC curve AR(1): $\rho = 0.8$

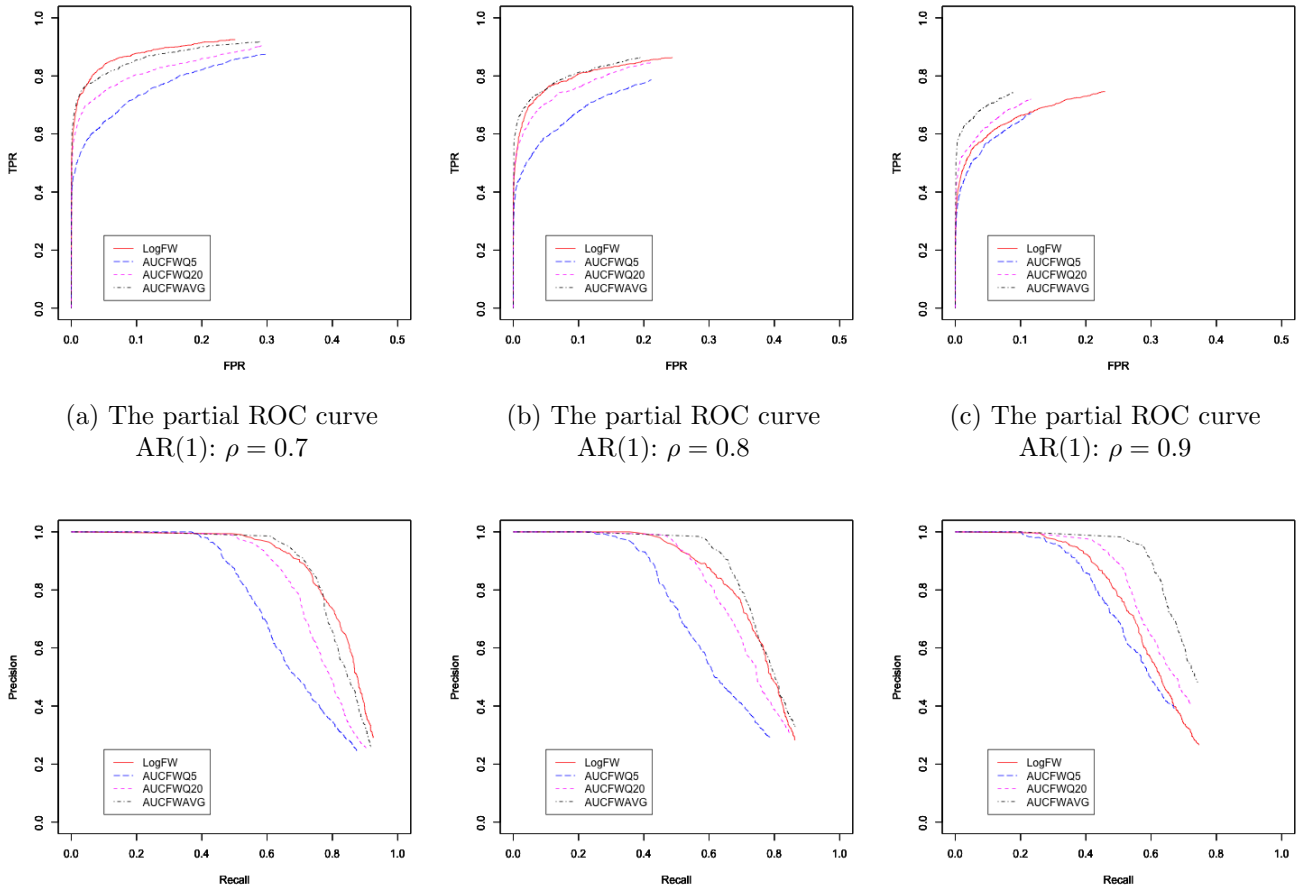
(c) The partial ROC curve AR(1): $\rho = 0.9$

(d) The PR curve AR(1): $\rho = 0.7$

(e) The PR curve AR(1): $\rho = 0.8$

(f) The PR curve AR(1): $\rho = 0.9$

Figure 3.7: Solution path grid on p-value presented in the partial ROC curves and the PR curves for Data #1A, where $N = 500$, $p = 50$, $\pi = 0.5$, and $\beta = (0.8, 0.8, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.



(a) The partial ROC curve AR(1): $\rho = 0.7$

(b) The partial ROC curve AR(1): $\rho = 0.8$

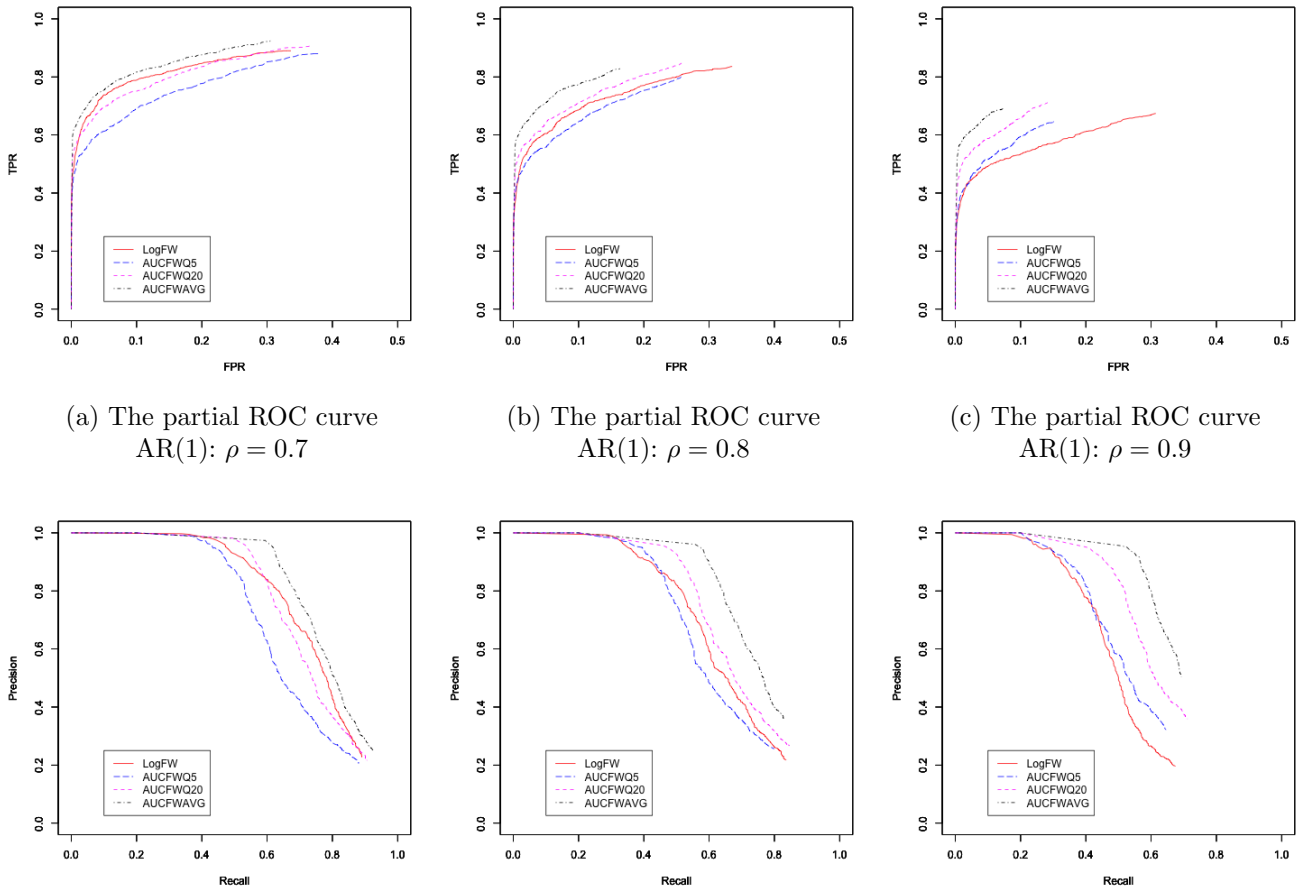
(c) The partial ROC curve AR(1): $\rho = 0.9$

(d) The PR curve AR(1): $\rho = 0.7$

(e) The PR curve AR(1): $\rho = 0.8$

(f) The PR curve AR(1): $\rho = 0.9$

Figure 3.8: Solution path grid on p-value presented in the partial ROC curves and the PR curves for Data #1B, where $N = 500$, $p = 50$, $\pi = 0.5$, and $\beta = (0.4, 0.4, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.



(a) The partial ROC curve AR(1): $\rho = 0.7$

(b) The partial ROC curve AR(1): $\rho = 0.8$

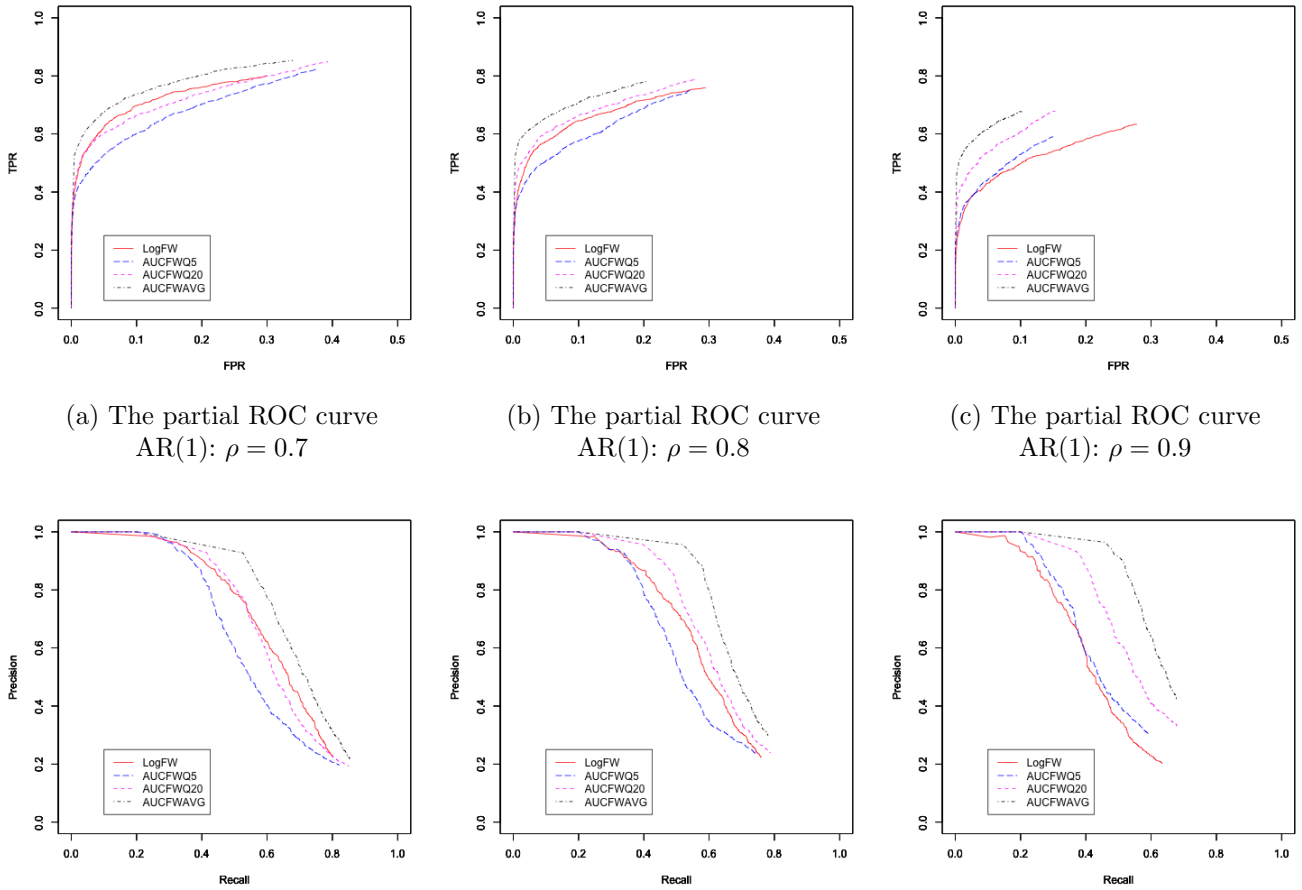
(c) The partial ROC curve AR(1): $\rho = 0.9$

(d) The PR curve AR(1): $\rho = 0.7$

(e) The PR curve AR(1): $\rho = 0.8$

(f) The PR curve AR(1): $\rho = 0.9$

Figure 3.9: Solution path grid on p-value presented in the partial ROC curves and the PR curves for Data #1As, where $N = 500$, $p = 50$, $\pi = 0.1$, and $\beta = (0.8, 0.8, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.



(a) The partial ROC curve AR(1): $\rho = 0.7$

(b) The partial ROC curve AR(1): $\rho = 0.8$

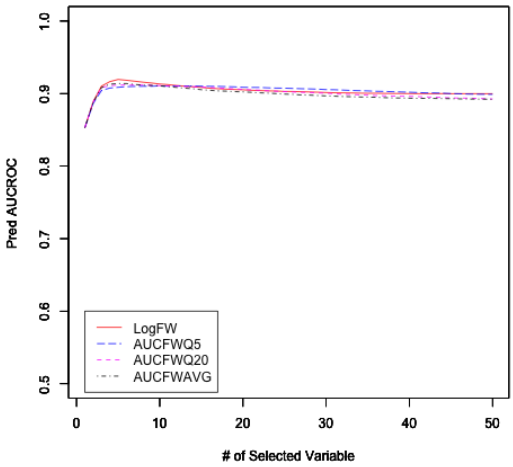
(c) The partial ROC curve AR(1): $\rho = 0.9$

(d) The PR curve AR(1): $\rho = 0.7$

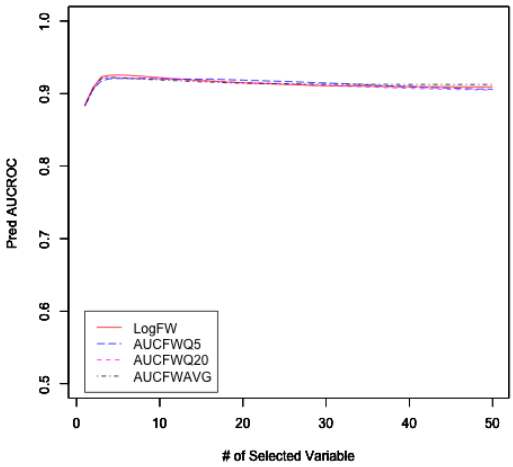
(e) The PR curve AR(1): $\rho = 0.8$

(f) The PR curve AR(1): $\rho = 0.9$

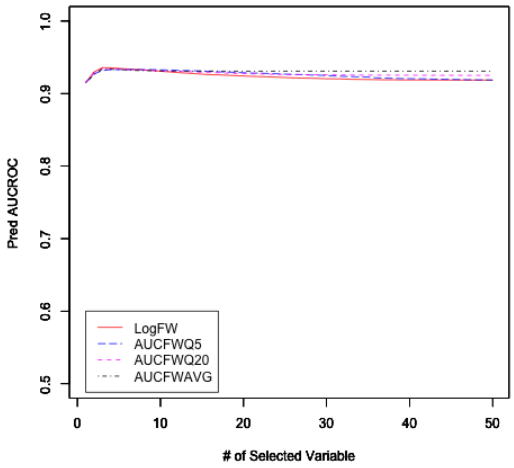
Figure 3.10: Solution path grid on p-value presented in the partial ROC curves and the PR curves for Data #1Bs, where $N = 500$, $p = 50$, $\pi = 0.1$, and $\beta = (0.4, 0.4, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.



(a) AR(1): $\rho = 0.7$

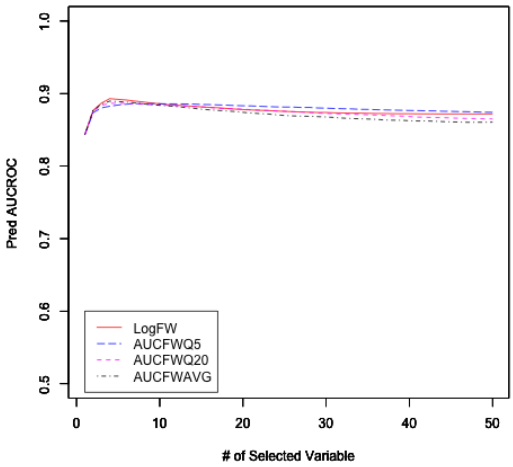


(b) AR(1): $\rho = 0.8$

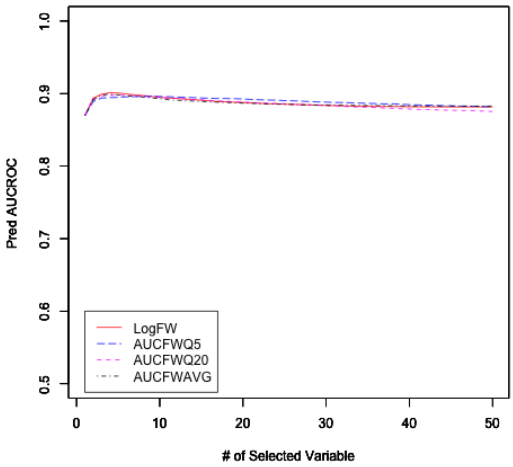


(c) AR(1): $\rho = 0.9$

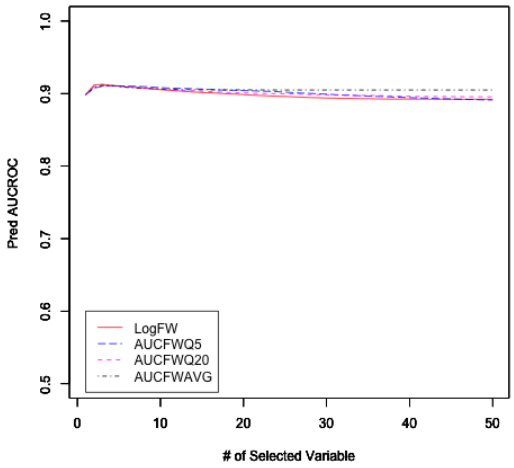
Figure 3.11: Investigation in prediction: predicted AUCROC vs. the number of variable selected for Data #1, where $p = 50$, $\pi = 0.5$, and $\beta = (0.8, 0.8, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.



(a) AR(1): $\rho = 0.7$

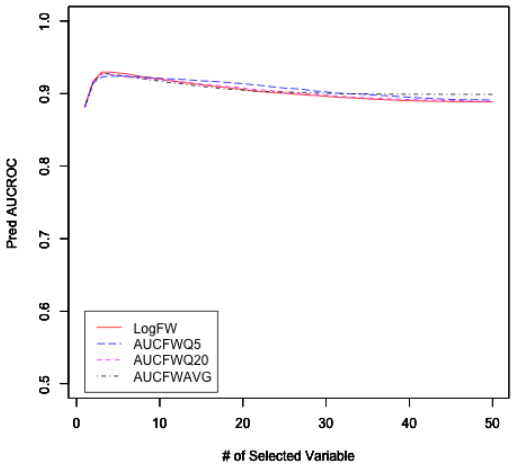


(b) AR(1): $\rho = 0.8$

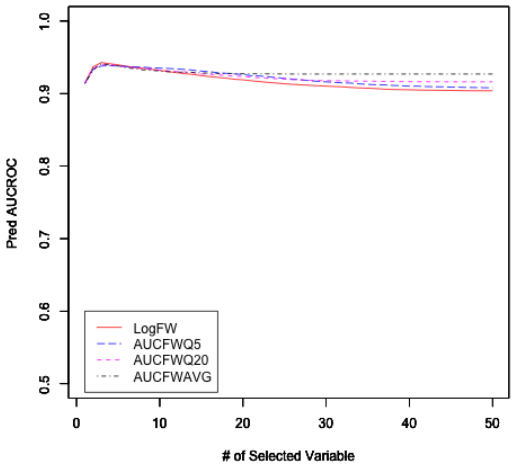


(c) AR(1): $\rho = 0.9$

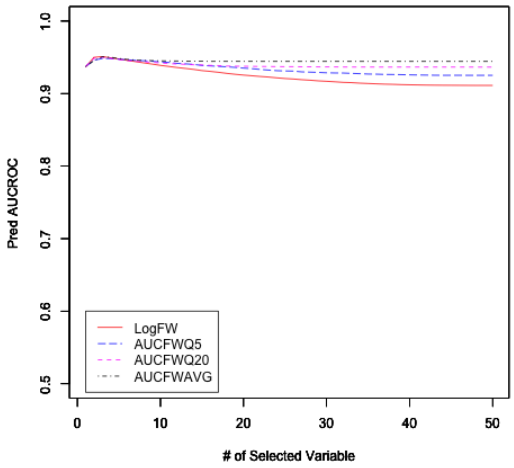
Figure 3.12: Investigation in prediction: predicted AUCROC vs. the number of variable selected for Data #2, where $p = 50$, $\pi = 0.5$, and $\beta = (0.4, 0.4, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.



(a) AR(1): $\rho = 0.7$

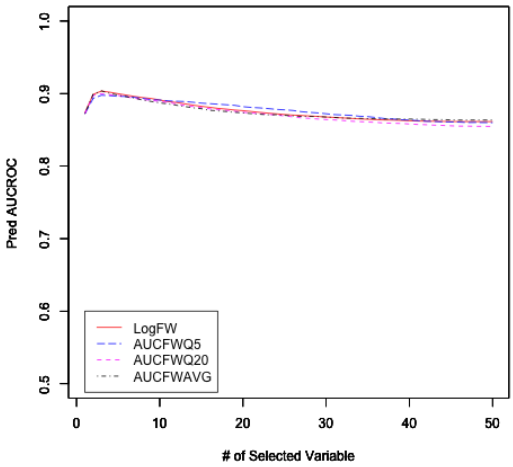


(b) AR(1): $\rho = 0.8$

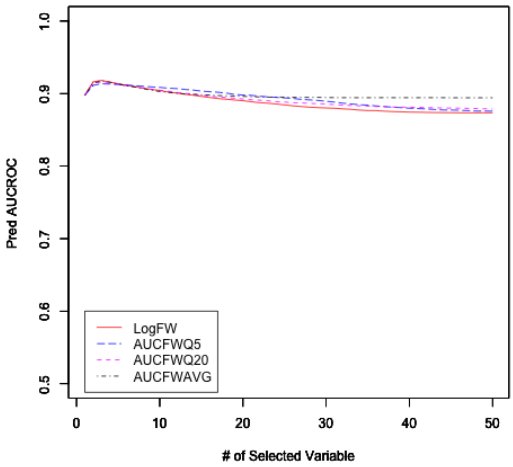


(c) AR(1): $\rho = 0.9$

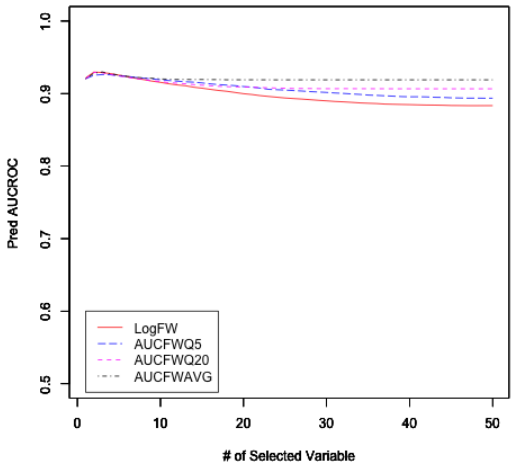
Figure 3.13: Investigation in prediction: predicted AUCROC vs. the number of variable selected for Data #3, where $p = 50$, $\pi = 0.1$, and $\beta = (0.8, 0.8, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.



(a) AR(1): $\rho = 0.7$



(b) AR(1): $\rho = 0.8$



(c) AR(1): $\rho = 0.9$

Figure 3.14: Investigation in prediction: predicted AUCROC vs. the number of variable selected for Data #4, where $p = 50$, $\pi = 0.1$, and $\beta = (0.4, 0.4, 0.8, 0.8, 0.8, 0, \ldots, 0)^T$.

REFERENCES

Abt, K. (1967). On the identification of the significant independent variables in linear models. *Metrika*, 12(1):1–15.

Ahmad, I. A. (1981). On Some Asymptotic Properties of U-Statistics. *Scandinavian Journal of Statistics*, 8(3):175–182.

Bamber, D. (1975). The Area above the Ordinal Dominance Graph And the Area below the Receiver Operating Characteristic Graph. *Journal of Mathematical Psychology*, 12(4):387–415.

Beale, E. (1970). Note on Procedures for Variable Selection in Multiple Regression. *Technometrics*, 12:909–914.

Bennett, K. P. and Blue, J. A. (1998). A Support Vector Machine Approach to Decision Trees. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 3, pages 2396–2401.

Boos, D. D. and Holloway, S. T. (2013). logi: Logistic regression - forward selection.

Boyd, K., Eng, K., and Page, C. (2013). Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In Blockeel, H., Kersting, K., Nijssen, S., and Železný, F., editors, *Lecture Notes in Computer Science*, pages 451–466. Springer Berlin Heidelberg.

Bradley, A. P. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, 30:1145–1159.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The Binormal Assumption on Precision-Recall Curves . In *2010 20th International Conference on Pattern Recognition (ICPR)*, pages 4263–4266. IEEE.

Castro, C. L. and Braga, A. P. (2008). Optimization of the Area Under the ROC Curve . In *2008 10th Brazilian Symposium on Neural Networks (SBRN)*, pages 141–146. IEEE.

Clémençon, S. and Vayatis, N. (2009). Nonparametric Estimation of the Precision-Recall Curve. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 185–192, New York, NY, USA. ACM.

Davis, J. and Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the International Conference on Machine Learning (ICML)*, pages 233–240.

Dehling, H. and Fried, R. (2012). Asymptotic distribution of two-sample empirical -quantiles with applications to robust tests for shifts in location. *Journal of Multivariate Analysis*, 105(1):124 – 140.

Dehling, H., Fried, R., Garca, I., and Wendler, M. (2013). Change-point detection under dependence based on two-sample u-statistics. *ArXiv e-prints*.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics*, 32:407–499.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. Series in Cognition and Perception. Academic Press, New York, NY.

Faraggi, D. and Reiser, B. (2002). Estimation of the Area under the ROC Curve. *Statistics in Medicine*, 21:3093–3106.

Fawcett, T. (2004). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. *ReCALL*, 31:1–38.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33.

Fu, K.-A. (2012). An Application of U-Statistics to Nonparametric Functional Data Analysis. *Communications in Statistics - Theory and Methods*, 41(9):1532–1542.

Gonçalves, L., Subtil, A., Oliveira, M. R., and Bermudez, P. D. Z. (2014). ROC Curve Estimation: An Overview. *Revstat - Statistical Journal*, 12:1–20.

Graf, A. C. and Bauer, P. (2009). Model Selection Based on FDR-Thresholding Optimizing the Area under the ROC-Curve. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–20.

Gu, J. and Ghosal, S. (2009). Bayesian ROC Curve Estimation under Binormality Using A Rank Likelihood. *Journal of Statistical Planning and Inference*, 139(6):2076–2083.

Guvenir, H. A. and Kurtcephe, M. (2013). Ranking Instances by Maximizing the Area under ROC Curve. *Knowledge and Data Engineering, IEEE Transactions on*, 25(10):2356–2366.

Guyon, I. and Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hanley, J. A. (1988). The Robustness of the "Binormal" Assumptions Used in Fitting ROC Curves. *Medical Decision Making*, 8:197–203.

Hanley, J. A. (1998). The Robustness of the "Binormal" Assumptions Used in Fitting ROC Curves. *Medical Decision Making*, 8:197–203.

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.

Hocking, R. R. (1976). The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32(1):1– 49.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19:293–325.

Hoeffding, W. (1961). The Strong Law of Large Numbers for U-Statistics. *Institute of Statistics Mimeo Series No. 302*.

Hsu, M.-J. and Hsueh, H.-M. (2013a). The Linear Combinations of Biomarkers Which Maximize the Partial Area under the ROC Curves. *Computational Statistics*, 28(2):647–666.

Hsu, M.-J. and Hsueh, H.-M. (2013b). The Linear Combinations of Biomarkers Which Maximize the Partial Area under the ROC Curves. *Computational Statistics*, 28:647–666.

In Lee, K. and Koval, J. J. (1997). Determination of the best significance level in forward stepwise logistic regression. *Communications in Statistics - Simulation and Computation*, 26(2):559–575.

Keilwagen, J., Grosse, I., and Grau, J. (2014a). Area under Precision-Recall Curves for Weighted and Unweighted Data. *PLoS ONE*, 9(3):e92209.

Keilwagen, J., Grosse, I., and Grau, J. (2014b). Area under Precision-Recall Curves for Weighted and Unweighted Data . *PLoS ONE*, 9(3):e92209.

Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 202–207.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324.

Komori, O. and Eguchi, S. (2010). A Boosting Method for Maximizing the Partial Area Under the ROC Curve. *BMC Bioinformatics*, 11(1):1–17.

Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M. R., and Goodenday, L. S. (2001). Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis. *Artificial Intelligence in Medicine*, 23:149.

Lee, J. (1990). U-statistics: Theory and practice.

Lenth, R. V. (1983). Some Properties of U Statistics. *The American Statistician*, 37:311–313.

Lusted, L. B. (1960). Logical analysis in roentgen diagnosis. *Radiology*, 74(2):178.

Lusted, L. B. (1978). General Problems in Medical Decision Making with Comments on ROC Analysis. *Seminars in Nuclear Medicine*, 8(4):299–306.

Ma, H., Bandos, A. I., Rockette, H. E., and Gur, D. (2013). On Use of Partial Area Under the ROC Curve for Evaluation of Diagnostic Performance. *Statistics in Medicine*, 32(20):3449–3458.

Ma, S. and Huang, J. (2005). Regularized ROC Method for Disease Classification and Biomarker Selection with Microarray Data. *Bioinformatics*, 21:4356–4362.

Mantel, N. (1970). Why Stepdown Procedures in Variable Selection. *Technometrics*, 12:621–625.

Marrocco, C., Duin, R. P. W., and Tortorella, F. (2008). Maximizing the Area Under the ROC Curve by Pairwise Feature Combination. *Pattern Recognition*, 41(6):1961–1974.

McClish, D. K. (1989). Analyzing a Portion of the ROC Curve. *Medical Decision Making*, 9(3):190–195.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283 – 298.

Metz, C. E. (1986). ROC Methodology in Radiologic Imaging. *Investigative Radiology*, 21(9):720 – 733.

Nelder, J. A. and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313.

Oliphant, L., Burnside, E., and Shavlik, J. (2010a). Boosting First-Order Clauses for Large, Skewed Data Sets. In De Raedt, L., editor, *Inductive Logic Programming*, pages 166–177. Springer Berlin Heidelberg.

Oliphant, L., Burnside, E., and Shavlik, J. (2010b). Boosting First-Order Clauses for Large, Skewed Data Sets. In De Raedt, Luc, editor, *Inductive Logic Programming*, volume 5989 of *Lecture Notes in Computer Science*, pages 166–177. Springer Berlin Heidelberg.

Peltonen, J. and Kaski, S. (2011). Generative Modeling for Maximizing Precision and Recall in Information Visualization. *Journal of Machine Learning Research*, 15:597–587.

Peng, Y., Wu, Z., and Jiang, J. (2010). A Novel Feature Selection Approach for Biomedical Data Classification. *Journal of Biomedical Informatics*, 43(1):15 – 23.

Pepe, M. S., Cai, T., and Longton, G. (2006). Combining Predictors for Classification Using the Area under the Receiver Operating Characteristic Curve. *Biometrics*, 62(1):221–229.

Pepe, M. S., Longton, G., Anderson, G. L., and Schummer, M. (2003). Selecting Differentially Expressed Genes from Microarray Experiments. *Biometrics*, 59:133–142.

Raghavan, V., Bollmann, P., and Jung, G. S. (1989a). A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. *ACM Trans. Inf. Syst.*, 7:205–229.

Raghavan, V., Jung, G. S., and Bollmann, P. (1989b). A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. *ACM Transactions on Information System*, 7(3):205–229.

Rao, C. and Mukerjee, R. (1997). Comparison of lr, score, and wald tests in a non-iid setting. *Journal of Multivariate Analysis*, 60(1):99 – 110.

Ricamato, M. T. and Tortorella, F. (2011). Partial AUC maximization in a linear combination of dichotomizers. *Pattern Recognition*, 44(10-11):2669–2677.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., and Johannes, R. S. (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 261–265. American Medical Informatics Association.

Song, X. and Ma, S. (2010). Penalized variable selection with u-estimates. *Journal of nonparametric statistics*, 22:499–515.

Spackman, Kent A. (1989). Signal Detection Theory: Valuable Tools for Evaluating Inductive Learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 160–163.

Sproule, R. N. (1974). Asymptotic Properties of U-Statistics . *Transactions of the American Mathematical Society*, 199:55–64.

Sutradhar, B. C. and Bartlett, R. F. (1993). Monte carlo comparison of wald's, likelihood ratio and rao's tests. *Journal of Statistical Computation and Simulation*, 46(1-2):23–33.

Swets, J. A. (1973). The Relative Operating Characteritics in Psychology. *Science*, 182(4116):990–1000.

Swets, J. A. (1986). Indices of Discrimination or Diagnostic Accuracy: Their ROCs and Implied Models. *Psychological Bulletin*, 99(1):100–117.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, New York.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Tosteson, A. and Begg, C. (1988). A General Regression Methodology for ROC Curve Estimation. *Med Decis Making*, 8:204–215.

Wang, L., Zhu, J., and Zou, H. (2006). The Doubly Regularized Support Vector Machine. *Statistica Sinica*, 16:589–615.

Wang, Z. and Chang, Y.-C. I. (2011). Marker Selection via Maximizing the Partial Area Under the ROC Curve of Linear Risk Scores. *Biostatistics*, 12(2):369–385.

Wang, Zhanfeng and Chang, Yuan-Chin Ivan (2011). Marker Selection via Maximizing the Partial Area under the ROC Curve of Linear Risk Scores. *Biostatistics*, 12(2):369–385.

Weston, J., Elisseeff, A., Schölkopf, B., and Tipping, M. (2003). Use of the zero norm with linear models and kernel methods. *J. Mach. Learn. Res.*, 3:1439–1461.

Xu, K. (2007). U-Statistics and Their Asymptotic Results for Some Inequality and Poverty Measures. *Econometric Reviews*, 26:567–577.

Yi, Y. and Xikui, W. (2011). Comparison of wald, score, and likelihood ratio tests for response adaptive designs. *Journal of Statistical Theory and Applications*, 10(4):553–569.

Yousef, Waleed A. (2013). Assessing Classifiers in Terms of the Partial Area under the ROC Curve. *Computational Statistics and Data Analysis*, 64:51 – 70.

Zellner, D., Keller, F., and Zellner, G. E. (2004). Variable Selection in Logistic Regression Models. *Communications in Statistics - Simulation and Computation*, 33(3):787–805.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2003). 1-Norm Support Vector Machines. In *Neural Information Processing Systems*, page 16. MIT Press.