A SIMULATION STUDY FOR EVALUATING THE AREA UNDER THE ROC CURVE AND THE ERROR RATE IN BINARY CLASSIFICATIONS

By

Qinhua Huang

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biostatistics–Master of Science

2015

ABSTRACT

A SIMULATION STUDY FOR EVALUATING THE AREA UNDER THE ROC CURVE AND THE ERROR RATE IN BINARY CLASSIFICATIONS

By

Qinhua Huang

The area under the ROC curve (AUC) and the error rate are two important criteria designed to measure the performance of classifiers. The maximum AUC and the minimum error rate indicates the best classification. However, one cannot get the minimum error rate and the maximum AUC simultaneously under the same classifier. It is thus of interest to investigate the relationship between the AUC and the error rate. Studying the relationship between the AUC and error rate, Cortes and Mehryar (2004) have provided an expression of the expected value of the AUC for a given error rate. In this thesis, I first study the validity of the expression given by Cortes and Mehryar (2004), after that, I investigate the error rate distribution under a fixed range of AUC.

My results show that Cortes and Mehryar's expression is not valid under some specific situations, and the expected average value of AUC is always smaller than the estimate of AUC from Mote-Carlo samples. When the proportion of positive samples is not close to 0.5, the expected average value of AUC calculated by Cortes and Mehryar's expression deviates largely from the Mote-Carlo samples of AUC. This indicates that the expression of the expected average value of AUC for given error rate may not be accurate and should be caution used. I also provide useful information for the quantiles of the error rate for given fixed range of AUC, with the proportion of positive samples varying in [0.1, 0.5].

ACKNOWLEDGMENTS

I wish to express my sincere thanks to all my graduate and undergraduate professors.I am extremely thankful and indebted to them for sharing expertise, and sincere and valuable guidance and encouragement extended to me. I also thank my parents for the unceasing encouragement, support and attention. I am also grateful to my friends who supported me through this venture. I also place on record, my sense of gratitude to one and all, who directly or indirectly, have lent their hand in this venture.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Background

Classification is a common task in many fields of applications such as healthcare, genetic analysis, and computer science. For example, Planet et al. (2001) proposed a molecular key-based method to classify putative NTPase genes precisely [2]. Another example is given by Pang et al. (2002), who studied how to classify documents by sentiment using machine learning including Navie Bayes,maximum entropy classification and support vector machines [3]. Finally, Gorno-Tempini (2011) studied how to classify primary progressive aphasia and its three main variants [1].

After the classification, it is important to study the accuracy of the classifications. There are two common criteria used to measure the performance of classification: the error rate and the area under receiver operating characteristics (ROC) curve (AUC). Recently, some researchers pointed out that the AUC may be more pertinent measurement for classification than the misclassification error rate[4].

The ROC curve is a plot that tests the performance of a binary classifier system as its discrimination threshold is varied, thus it can select classifiers based on their performance. It has been used for a long time, and have been extended to visualize and analyze the diagnostic systems' behavior [43]. Besides, an increasing number of medical decisions have been made based on the ROC graph, and a growing usage of the ROC curves have been seen in machine learning community because of the realization that the error rate is not accurate enough to measure the classification

performance [22]. Apart from being a mainly performance graphing method, the ROC graph also has properties making it very useful for estimating error costs of skewed class distribution. And these properties have become more and more important because the research about cost-sensitive learning has gained a lot of attention lately [16].

1.2 Motivation

The area under ROC curve (AUC) and the misclassification error rate are two important criteria designed to measure the performance of classifiers. For instance, Simon et al.(2003) used misclassification rate to measure the performance of a class predication for DNA microarray data [42]. Golub et al. (1999) used the cumulative error rate to assess the accuracy of an gene expression based classifier for cancer [39]. Wang et al. (2007) also used the overall error rate to assess their classifier for rapid assignment of rRNA Sequence into higher taxonomy [40]. Another example is given by Krizhevsky et al. (2012), who measured their classifier that designed to classify high-resolution images in the ImageNet LSVRC-2010 contest by error rate [9]. Furthermore, Statnikov et al. (2008) used the AUC to compare the random forests and support vector machines for cancer classification based on microarray [34]. And Lee et al.(2008) used AUC to evaluate the performance of a new method of classification which based on athway activities inferred for each patients [35]. Finally, Ma et al. (2005) proposed a new method used a sigmoid approximation to the AUC as a objective function to select and classify biomarker [8].

The most common methods to measure the performance of classification exercise is the error rate and AUC. However, one cannot get the minimum error rate and the maximum AUC simultaneously under the same classifier. It is thus of interest to investigate the relationship between the AUC and the misclassification rate. Cortes and Mehryar(2004) have provided an expression of the

expected average value and the variance of AUC given a fixed error rate. However, the authors warned that these equations require the classification or rankings with k errors to be equiprobable. By equiprobable, they mean situations in which each test sample has the equal probability of being misclassified [4]. In this thesis, I study the expression provided by Cortes and Mehryar(2004), and point out the expression is inappropriate in some specific situations. And I conduct a simulation experiment to investigate the relationship between the estimates of AUC and the estimate of error rate. I conduct this experiment by simulating a binary distribution, and using logistic regression with threshold as a classifier. I assume that the threshold for the classifier follows an uniform distribution from 0 to 1. Then I calculate the estimate of AUC and the estimate of error rate for each classification. To investigate how the estimate of error rate is distributed under the fixed ranges of value of the estimate of AUC, I draw the cumulative distribution function (CDF) plots and probability distribution function (PDF) plots of the estimate of error rate.

1.3 Classifier performance

In this thesis, I study the binary classification situations. In a binary classification exercise, every sample is assigned to positive or negative class. A classifier is used to predict which class should the sample be assigned to. Different classifiers produce different outcomes to predict the sample's class, some of them produce discrete class labels and others produce continuous outputs to different thresholds. The thresholds can differ from 0 to 1 for a binary classification. If the outputs of a classifier is smaller than the threshold, then the sample is classified as a negative; if the output of a classifier is larger or equal to the threshold, then the sample is classified as a positive.

Given a classification model and a test sample, there may be four different outcomes (Table 1.1). If the test sample is positive and is assigned correctly, it is a true positive; if it is positive but

Table 1.1: Classifier Performance

	Condition Positive	Condition Negative
Test Positive	True positive	False positive
Test Negative	False negative	True negative

is assigned to negative, it is a false negative. If the test sample is negative and be assigned correctly, it is a true negative; if the test sample is negative but is assigned to positive, it is a false positive.

Some classification functions are used to measure the performance of binary classification. The sensitivity (true positive rate) is estimated as

$$TruePositiveRate = \frac{TruePositive}{TotalPositives}$$

The false positive rate is estimated as

$$FalsePositiveRate = \frac{FalseNegative}{TotalNegatives}$$

The specificity is estimated as

$$Specifity = \frac{TrueNegative}{FalsePositives + TrueNegatives} = 1 - FalsePositiveRate$$

1.4 Definition of ROC curve

The ROC curve is a plot that demonstrates how a binary classifier performs. It is a two-dimension graph with the true positive rate on the Y axis and the false positive rate on the X axis. The ROC curve can illustrate the relationship between the false positive and the true positive. Figure (1.1) is a simple example of the ROC curve. Here, the diagonal represents the random classification, in other

words, the classifier is conducted as a fair coin toss, and it is drawn as a reference. Point (0, 0) is located at lower left and demonstrates the situation in which never issuing a positive classification; the classifier commits no false positive errors but also no true positives. Oppositely, the point (1, 1) which is located at the upper right corner demonstrates no issuing negative classifications. The point (0, 1) shows the perfect classification with zero false positive rate and one true positive rate. Intuitively, one point performs a better classification if it is located to the northwest of another because it has a higher true positive rate and a low false positive rate. Usually, a classifier which appears near the X axis and on the left-hand side of a ROC curve would be taken as "conservative" because they make positive classifications only with strong evidence so they make few positive errors; however, the true positive rate doesn't perform well too. And a classifier which appears on the upper right-hand side of an ROC curve always be taken as "liberal" because they make positive classifications with weak evidence to increase the true positive rate, but the high true positive rate always affects the high false positive rate.

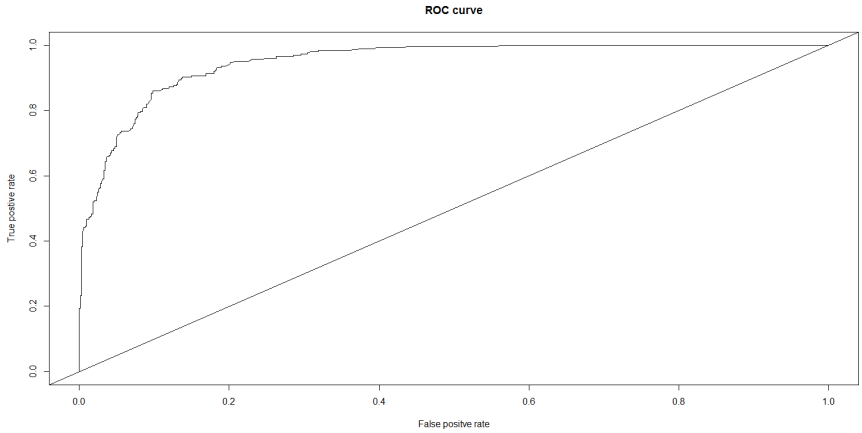

Figure 1.1: ROC curve

1.5 Definition of AUC

As I mentioned in the previous section, an ROC curve is a plot of the true positive rate as a function of the false positive rate. Reducing ROC performance from two dimensions to one single scalar value may be easier to compare the performances of classifiers. AUC, which is defined as the area under the ROC curve, is the most common method to measure the ROC performance. Since it is a portion of area of the unit square, the value of AUC will always between 0 and 1. However, since the random classification produces the diagonal line between (0, 0) and (1, 1) has an area of 0.5, no realistic classifier should have an AUC under 0.5.

The value of AUC could be calculated by the expression given by Mann and Whitney (1947) and Wilcoxon (1945), which is called Wilcoxon-Mann-Whitney statistic. The statistic is given by:

$$W = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} I(x_i > y_j)}{mn} \tag{1.1}$$

It is based on pairwise comparisons between a sample x_i, $i = 1, ..., m$ of random variable X and a sample y_j, $j = 1, ..., n$, of random variable Y. We identify $x_1, x_2, ..., x_m$ as the classifier outputs for m positive samples, and $y_1, y_2, ..., y_n$ as the classifier outputs for n negative samples. The proof of this expression is based on the observation that the AUC value is exactly the probability $P(X > Y)$. So the AUC can be used as a measure of pairwise comparisons between classifications of the two classes. With a perfect ranking, all positive samples are ranked higher than the negative ones and AUC=1.

Chapter 2

Literature Review

2.1 Methods of Classification

As a common task, classification has been studied in many cases. Researchers studied various methods of classifications for different situations. Friedman (1989) studied how to use linear discriminant analysis and Fisher's linear discriminant method to classify multiple classes of samples [23]. Mika et al. (1999) stated that linear discriminant analysis is a appropriate method to classify continuous observations. Oppositely, the discriminant correspondence analysis is more appropriate to classify discrete variable [24]. Murthy (1998) studied how to conduct decision trees method in machine learning area [25]. Decision trees are methods that classify samples by sorting them based on feature values. Each node in a decision tree stands for a feature in a sample to be classified, and each branch stands for a value that the node can assume [28].

Another well known classifier is Beyesian networks. Naive Bayesian networks is one of the simplest Beysian networks. It's is combined by a directed acyclic graphs with one unobserved node and several observed nodes and an assumption that the several observed nodes are independent(Good, 1950). Another statistical methods for classification is instance-based learning . Mitchell (1997) indicated that instance-based learning algorithms delay the generalization process until classification is performed, and thus they are lazy-learning algorithms. [27]. Although lazy-learning algorithms saved time for the training phase, it requires more time on classification process[28].

2.2 Development of ROC Curve Analysis

The first occurrence of ROC curve was during World War II, and it was developed by radar engineers to detect the enemy object. Then ROC curve was used in the field of psychology to account the perceptual detection of stimuli. Since then the ROC analysis has became useful in many fields such as medicine, radiology, biometrics and data mining research.

Metz(1978) discussed the basic principles of ROC analysis. They showed that the ROC analysis could combine the true positive fraction and the false positive fraction, and make it easier to compare hypothetical tests based on basic classification performance [14]. To estimate the value of the AUC, Hanley et al.(1982) stated that the area under ROC curve represents a probability that a randomly chosen positive sample is rated higher than a randomly chosen negative sample. And this probability is the same quality of estimated by the nonparametric Wilcoxon statistic [21]. Moses(1993) proposed a construction to do ROC analysis by four steps to [29]. Bradley(1997) further investigated the use of ROC analysis as a measure of classifier performance in the area of machine learning algorithms. They stated that AUC has many advantages compared to overall accuracy (misclassification rate) as a measure performance [18]. Metz et. al(1998) provided a new generalized method for ROC curve fitting. The new algorithm named ROCKIT conducts all analyses available from previous ROC software and provides 95% confidence interval for each estimates [30].

2.3 Study of Investigating the Relationship between AUC and the Misclassification Rate

In many classification exercise, researchers chose misclassification rate to measure the performance of the classifier. For example, Kim et al.(2003) studied the classification error rate estimation by bootstrap [36]. Another example is given by Och et al. (2003), who provided a new algorithm for unsmoothed error count and studied different training criteria of statistical machine translation models for optimize the minimum error rate [10]. Meanwhile, some researchers proposed that the area under the ROC curve is an alternative measure to evaluate the classification models. Herschtal and Raskutti (2004) introduced a binary classifier called RankOpt that can optimise AUC using gradient descent [7]. Agarwal (2005) studied the generalization bounds for AUC. In their paper, they defined the expected accuracy of ranking function and derive distribution-free probabilistic bounds on the deviation of the empirical AUC of a ranking function. Furthermore, they also derived both a large deviation bounds and a uniform convergence bound [31]. Thus it is of interest to study the misclassification error rate and the AUC.

Cortes and Mehryar (2004) conducted a statistical analysis to investigate how AUC is related to error rate. They derived the expression to compute the expected value of AUC over all classifications with a fixed error rate. Given a fixed error k, they pointed out there are three classification situations: i) samples are classified correctly, ii) positive samples are misclassified to negative and iii) negative samples are misclassified to positive. They further computed the AUC for each situation, and provided an expression to calculate average value of the AUC given k errors and x false positive examples.

$$< AUC >_x = 1 - \frac{\frac{x}{n} + \frac{k-x}{m}}{2} \tag{2.1}$$

Besides, they have provided an expression to calculate the variance of AUC given x false positive samples. One year later, they gave an expression to calculate the confidence interval for AUC given error number k and the number of positive samples and negative samples. Their analysis gave us a good starting point to study the relationship between the error rate and the AUC. However, these expressions given by Cortes and Mehryar are only correct under the assumption that all classifications or rankings with k errors are equiprobable, which means each sample has the same probability to be misclassified. This condition is rarely met in realistic settings.

Chapter 3

Different Approaches to Address the Relationship between the AUC and the Misclassification Rate

3.1 The Expected Value of the AUC under fixed error rate

Cortes and Mehryar(2004) have provided an expression of the expected value of AUC overall classifications with a fixed number of errors and compared that to the error rate.

Assume that the number of error k is fixed,and a binary classification task with m positive samples and n negative samples is given. Under the assumption that all classifications or rankings with k errors are equiprobable, the expected value of the AUC is given by Eq. (3.1) [4].

$$< A >_{m,n,k}= 1 - \frac{k}{m+n} - \frac{(n-m)^2(m+n+1)}{4mn}(\frac{k}{m+n} - \frac{\sum_{x=0}^{k-1}\binom{m+n}{x}}{\sum_{x=0}^{k}\binom{m+n+1}{x}}), \tag{3.1}$$

Which is equivalent to

$$< A >= \frac{\sum_{x=0}^{k}\binom{N}{x}\binom{N'}{x'}(1 - \frac{\frac{x}{n}+\frac{k-x}{m}}{2})}{\sum_{x=0}^{k}\binom{N}{x}\binom{N'}{x'}} \tag{3.2}$$

Where x is the number of false positive samples, x' is the number of false negative samples, N is the number of negative samples, and N' is the number of positive samples. The proof of this expression

is based on weighting the expression (2.1) with the total number of possible classifications for a given x. Thus, there are $\binom{N}{x}$ possible ways of choosing x false positive examples and $\binom{N'}{x'}$ possible ways of choosing x' negative examples. Here, the authors assumed the following condition: $0 \leq x \leq k$, and $x' = k - x$.

However, the authors did not consider a situation where the number of misclassification k is larger than the number of negative samples y, and in that situation, the range of false positive x should be $0 \leq x \leq n$. Similarly, the number of false negative $k - x$ should is less than m, which means $x \leq m$. Thus, the value range of x should be $[0, min(m,n,k)]$. If we still use the expression (3.2) to calculate the expectation of AUC given x when $k > n$ or $k > m$, the expectation of AUC can be less than 0.5 or even negative.

For example, if I have 100 positive samples, 900 negative samples, and 200 misclassified samples, the expectation of AUC is -0.1429524. I can also indicate this issue by plotting the value of the AUC expectation calculated by expression (3.1). The plot of expression (3.1) with 100 positive examples and 900 negative examples is shown in figure (3.1).

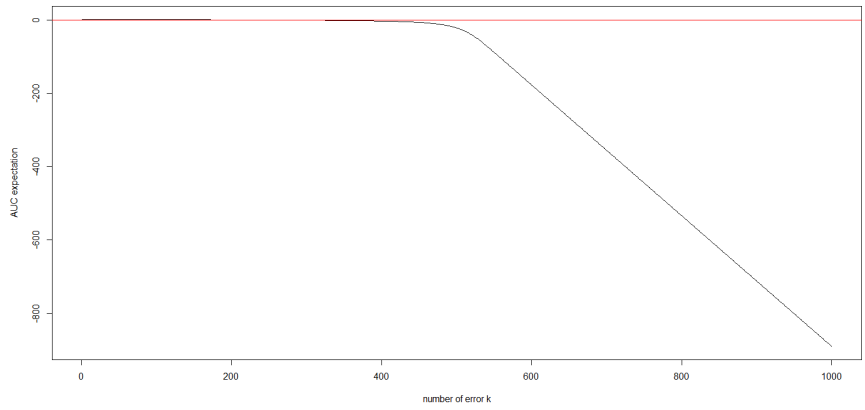

Figure 3.1: AUC Expectation (m=100,n=900)

Here, the red line is shown as a reference line as $AUC = 0$. From the figure we can see that when the number of error k is much larger then the number of positive samples m, the expectation

calculated by expression (3.1) is negative. As I mentioned in previous sections, AUC is a probability of positive sample ranked higher than negative samples correctly, which indicates that AUC should have a positive value. In order to use the expression (3.1) correctly, the assumption that the number of errors is less than the number of negative samples and the number of positive samples should be added.

3.2 Simulating the AUC Distribution under Fixed Error Rate

The pervious section showed that the expression (3.1) is not valid when the error number k is larger than the number of positive samples m or the number of negative samples n. When the error number k is smaller than $min(m,n)$, the derivation of expression (3.1) is correct. Since the assumption that each sample has same probability to be misclassified is hard to be achieved in realistic scenarios, I further conduct an extensive simulation experiment to evaluate the validity of this expression when $k \leq min(m,n)$ for moderate to large deviations of the equiprobable assumption.

3.2.1 Generating Binary Distribution

In order to investigate the validity of expression (3.1), I simulate binary distributed data from logistic regression model. Since expression (3.1) is conditioned on m,n and k, I investigate five situations corresponding to the ratio of positive samples $r = m/(m+n)$ equals to $0.1, 0.2, 0.3, 0.4, 0.5$.

First,I generated 1000 $a_1, a_2, a_3, a_4, a_5 \sim N(0, 0.5)$, and I set $\beta_1 = 3$, $\beta_2 = -5.5$, $\beta_3 = -5$, $\beta_4 = 2.5$, $\beta_5 = -1$.

Then I generated 1000 $u \sim U[0,1]$ independent by a_i.

I define $\theta = 1/1 + exp(-(\beta_0 + \beta_1 \times a_1 + \beta_2 \times a_2 + \beta_3 \times a_3 + \beta_4 \times a_4 + \beta_5 \times a_5))$

I label examples to 0 and 1 by following rules:

if $u \geq \theta$ then $y = 0$

if $u < \theta$ then $y = 1$

To get a binary distributed data with 10% positive samples, I set $beta_0 = -8$, and only use the binary distributed datasets which have 100 positive samples ($y = 1$) to conduct the simulation experiment. To get binary distributed data with 20% positive samples, I set $beta_0 = -4.5$, and only use the binary distributed datasets with 200 positive samples($y = 1$) to conduct the simulation experiment. Similarly, I set $beta_0 = -1.5, 1, 3.5$ corresponding to get the datasets with 30%, 40% and 50% positive samples. The reason I adjust $beta_0$ is to adjust the probability $(y = 1|a)$. By adjusting $P(y = 1|a)$, I can get the dataset with around 10% to 50% positive samples.

3.2.2 Using Logistic Regression as a Classifier

The logistic regression is a direct probability model. It is a special case of generalized linear models. For logistic regression, the conditional distribution of binary data given covariates follows a Bernoulli distribution with success probability bounded between 0 and 1. Thus one can use binary logistic model to predict binary outcomes based on predictor variables.

Given a binary random variable y and a vector of predictors (could be continuous or discrete) a, logistic regression can be used to predict the success probability $P(y|a)$.

$$P(y = 1|a) = \frac{1}{1 + exp(\beta_0 + \sum_{i=1}^{b} \beta_i a_i)} \tag{3.3}$$

Where b is the number of predictors. $P(y|a)$ leads to a simple linear expression for classification. Generally, we assign the label $Y = 1$ if the following condition holds:

$$\frac{P(y = 0|a)}{P(y = 1|a)} < 1,$$

Which is equivalent to

$$exp(\beta_0 + \sum_{i=1}^{b} \beta_i a_i) < 1$$

After taking natural log of both sides, we can assign $y = 1$ if a satisfies

$$\beta_0 + \sum_{i=1}^{b} \beta_i a_i < 0,$$

and assign $y = 0$ otherwise.

In the previous section, I simulate the binary distributed data (y, a). Then I choose logistic regression as a binary classifier to classify the data. I choose a_1, a_3, a_5 as predictors, and use expression (3.3) to get the success probability can compare the probability to thresholds. I assume that the threshold follows a uniform distribution from 0 to 1, and I randomly select one threshold from $U(0,1)$ for each classification. If the predicted probability is less or equal to threshold, then $\hat{y} = 0$, else $\hat{y} = 1$.

3.2.3 Estimate of AUC and Expected AUC Calculation

As I mentioned before, the AUC is the area under ROC curve,a plot of true positive rate as a function of false positive rate. To calculate the area under curve, I can integral under the ROC curve. Thus it is necessary to calculate the true positive rate and the false negative rate at each point of ROC curve first.

From my simulation, I can get the true positive rate and false negative rate by following formula.

$$TPS = \frac{number of (\hat{Y} = Y = 1)}{number of (Y = 1)}$$

$$FPS = \frac{number of (\hat{Y} = 0 and Y = 1)}{number of (Y = 0)}$$

I choose 1000 points on the ROC curve and calculated the true positive rate and false positive rate for each point. Then I calculated the integral under the ROC curve to get the AUC. The Eq.(3.1) given by Cores and Mehryar is the expected average AUC value under fixed error number k, which means a fixed error rate. However, it is difficult to investigate the AUC distribution under every $k \subset [0, min(m,n)]$. So I investigate the AUC distribution under several small ranges of $k = 10$. And I compare the estimate of AUC with the the expected average AUC under each range. To calculate the expected average AUC, the number of error k needs to be known for each iteration. I define the error number k as the count number of $\hat{(Y_i)} \neq (Y_i)$. Then I plug m,n,k into expression (3.1) to get the expected average AUC.

3.2.4 Comparing the Estimate of AUC versus Expected Average AUC

I have 1000 samples in total, and there are m positive examples and n negative samples. I define the ratio of positive samples r as $r = m/(m+n)$. I compare the estimate of AUC and expected average AUC under $r = 0.1, 0.2, 0.3, 0.4, 0.5$. First, I draw the empirical cumulative probability function (CDF) plots of AUC under specific range of k to investigate the AUC distribution. After that, I draw the probability density function (PDF) plots of the estimate of AUC under specific range of k, and add the lower bound and upper bound of expected average AUC calculated by expression (3.1) as reference lines to compare the estimate of AUC and the expected average AUC. Finally, I give the descriptive statistics to show the difference between the estimate of AUC and the expected average AUC under specific ranges of error number k.

First, I looked at the first situation where $r = 10\%$, which means there are 100 positive samples and 900 negative samples. In this situation, I only investigate the estimate of AUC distribution when $k \leq 100$. Here, we plot the CDF and PDF for estimate of AUC with the estimate of error rate from 0.07 to 0.08, 0.08 to 0.09 and 0.09 to 0.1. In each range of error rate, I have 2,924, 70,609

and 312,565 estimates of AUC values correspondingly.

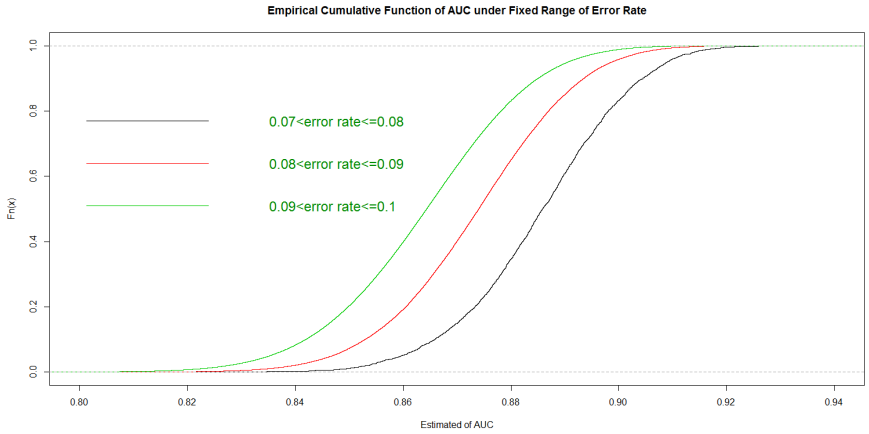

Figure 3.2: CDF of estimate of AUC(r=10%)

Figure (3.2) showes that the estimate of AUC with lower error rate has a higher value. The minimum and maximum of the estimate of AUC with error rate from 0.07 to 0.08 is largest among three estimate of AUC. Then I plot the PDF of the estimate of AUC with error rate from 0.07 to 0.08, 0.08 to 0.09 and 0.09 to 0.1, and add the upper bound and lower bound of expected average AUC calculated by expression (3.1). The PDF plot is shown in figure (3.3). Then I calculate the difference between average estimate of AUC and the bounds(upper and lower) corresponding to each range of error rate (Table3.1) as reference.

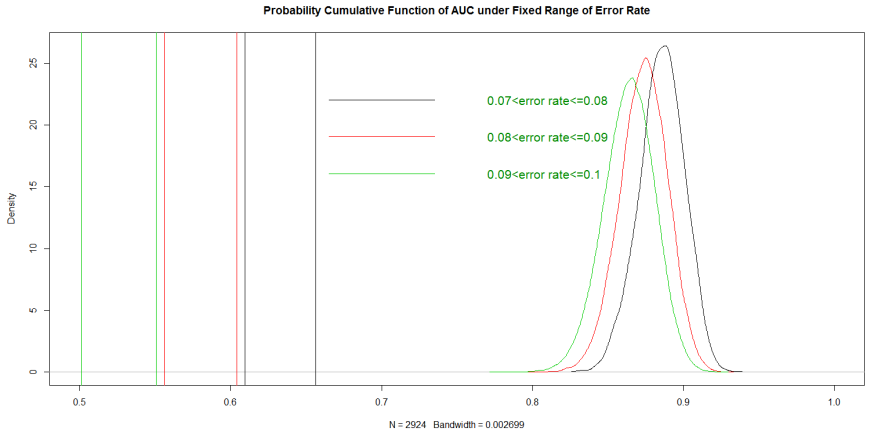

Figure 3.3: PDF of estimate of AUC(r=10%)

Table 3.1: Difference between estimate of AUC and expected average AUC (r=10%)

Error rate	(0.07,0.08]	(0.08,0.09]	(0.09,0.1]
Average estimate of AUC-lower bound	0.27612	0.31761	0.36258
Average estimate of AUC-upper bound	0.22388	0.2642	0.3079

Figure (3.3) indicates that the expected average AUC value is always lower than the estimate of AUC. And from table(3.1) I find that when the error rate gets larger, the difference between estimate of AUC and expected average AUC get larger too.

When $r = 20\%$, there are 200 positive samples and 800 negative samples. In this case, I only investigate the estimate of AUC distribution when $k \leq 200$. I plot the CDF and PDF of the estimate of AUC with the estimate of error rate from 0.15 to 0.16, 0.17 to 0.18 and 0.19 to 0.2. In each range of error rate, I have 39,892, 100,487 and 110,114 estimates of AUC values correspondingly. The PDF is shown in figure (3.4).

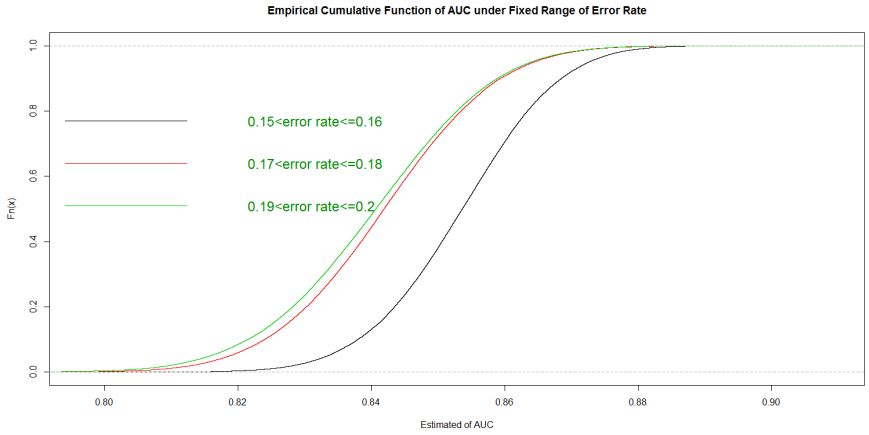

Figure 3.4: CDF of estimate of AUC(r=20%)

From figure (3.4) I can see that the estimate of AUC has similar distribution under the range of error rate from 0.17 to 0.18 and 0.19 to 0.20. Then I plot the PDF plot of the estimates of AUC to investigate the difference between the estimates of AUC and the expected average AUC (figure(3.5)).

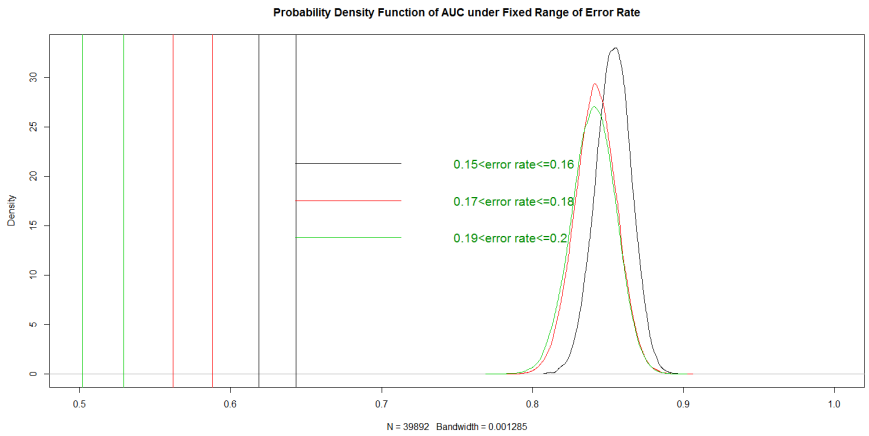

Figure 3.5: PDF of estimate of AUC(r=20%)

From figure (3.5) I find that the expected average AUC is always lower than the estimates of AUC. And the estimates of AUC has the similar one mode distribution under error rate within range of $(0.17, 0.18]$ and $(0.19, 0.20]$. The estimate of AUC with error rate from 0.15 to 0.16 is higher than the other two. The difference between average estimate of AUC and the bounds(upper and lower) corresponding to each range of error rate is shown in table (3.2). The table shows that when error rate gets larger, the difference between estimate of AUC and the expected average AUC gets larger too.

Table 3.2: Difference between estimate of AUC and expected average AUC (r=20%)

Error rate	(0.15,0.16]	(0.17,0.18]	(0.19,0.20]
Average estimate of AUC-lower bound	0.23484	0.28004	0.30864
Average estimate of AUC-upper bound	0.20752	0.25125	0.27902

When $r = 30\%$, there are 300 positive samples and 700 negative samples. In this situation, expression (3.1) only valid when $k \leq 300$. We draw the CDF and PDF plot of the estimates of AUC with the estimate of error rate from 0.19 to 0.20, 0.24 to 0.25 and 0.29 to 0.3. In each range of error rate, I have 7,060, 48,399 and 55,015 estimates of AUC value. The plot of the CDF of the estimates of AUC under each error rate range is shown in 3.6.

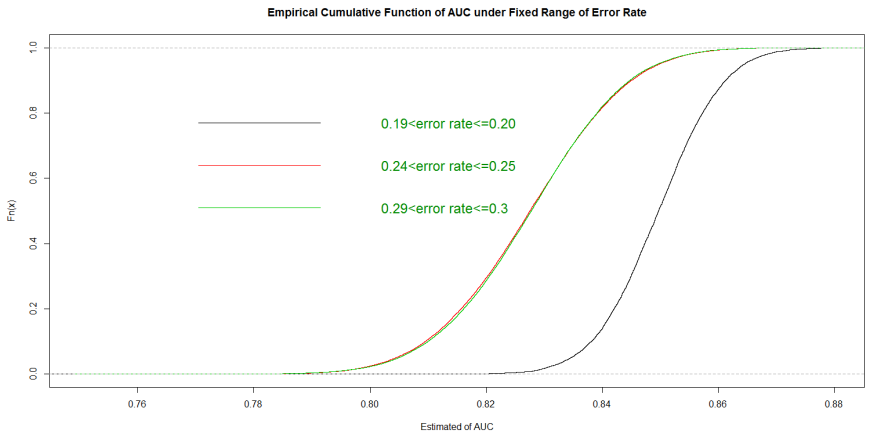

Figure 3.6: CDF of estimate of AUC(r=30%)

From figure 3.6 I notice that the estimate of AUC under error rate from 0.19 to 0.20 has the highest value. The distribution of the estimate of AUC under error rate from 0.24 to 0.25 an 0.29 to 0.30 are almost same. To further investigate the difference between the estimate of AUC and the expected average AUC with same range of error rate, I plot the PDF plot of the estimate of AUC with bounds of expected average AUC as reference (figure (3.7)).

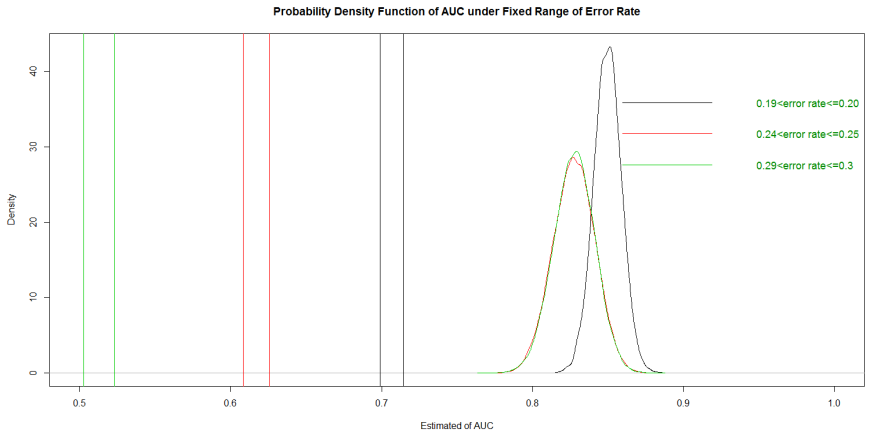

Figure 3.7: PDF of estimate of AUC(r=30%)

From figure (3.7) I find that the value of the expected average AUC is much smaller than the estimates of AUC under each range of error rate. And the table(3.3) of difference between average estimate of AUC and the bounds(upper and lower) corresponding to each range of error rate shows

that when the error rate gets larger, the difference gets larger simultaneously.

Table 3.3: Difference between estimate of AUC and expected average AUC (r=30%)

Error rate	(0.19,0.20]	(0.24,0.25]	(0.29,0.3]
Average estimate of AUC-lower bound	0.1507	0.21916	0.32513
Average estimate of AUC-upper bound	0.13375	0.20008	0.30225

For the situation where $r = 40\%$, there are 400 positive samples and 600 negative samples. In this situation, I only investigate the distribution of estimate of AUC when $k \leq 400$. I plot the CDF and PDF of the estimate of AUC with the estimate of error rate from 0.24 to 0.25, 0.30 to 0.31, 0.34 to 0.35 and 0.39 to 0.4. In each range of error rate, I have 41,575, 26,702, 20,451 and 31,572 estimates of AUC value. The plot of the CDF of estimates of AUC under each error rate range is shown in 3.8.

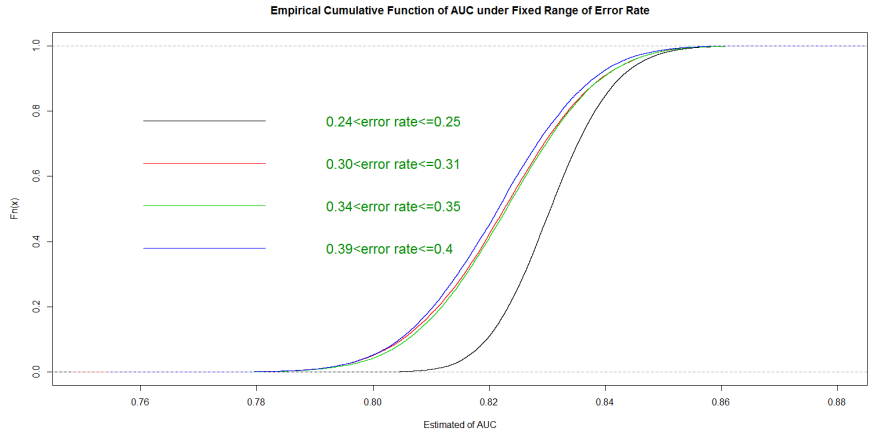

Figure 3.8: CDF of estimate of AUC(r=40%)

From this plot I notice that when the error rate is from 0.24 to 0.25, the estimate of AUC value is significant higher than others. When error rate is from 0.39 to 0.4, the estimate of AUC value is smallest. And the distribution of the estimate of AUC with error rate from 0.3 to 0.31, 0.34 to 0.35 and 0.39 to 0.4 is very similar with each other. The PDF plot of the estimate of AUC is shown in figure (3.9).

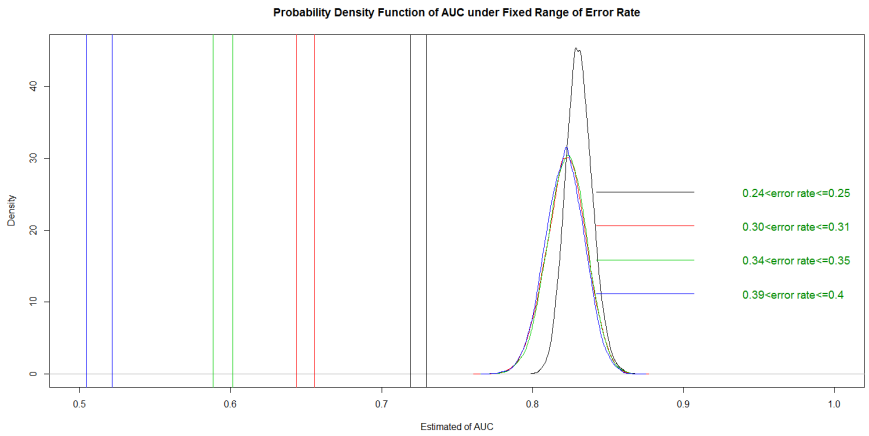

Figure 3.9: PDF of estimate of AUC(r=40%)

From figure (3.9) I can clearly see that the estimate of AUC with error rate from 0.24 to 0.25 has the largest value, and the difference between it and the expected average AUC is smallest. For error rate from 0.3 to 0.31, 0.34 to 0.35 and 0.39 to 0.4, the estimate of AUC has similar distribution. However, the expected average AUC value differs a lot for these three ranges of error rate. The table contains the difference between average estimate of AUC and the bounds(upper and lower) corresponding to each range of error rate is shown in table (3.4).

Table 3.4: Difference between estimate of AUC and expected average AUC (r=40%)

Error rate	(0.24,0.25]	(0.3,0.31]	(0.34,0.35]	(0.39,0.4]
Average estimate of AUC-lower bound	0.11185	0.17863	0.23464	0.31688
Average estimate of AUC-upper bound	0.09987	0.16557	0.22012	0.29808

This table also shows that when error rate gets larger, the difference between the estimate of AUC and the expected average AUC gets larger.

When $r = 50\%$, there are 500 positive samples and 500 negative samples. So I investigate the situation that $k \leq 500$. I plot the CDF and PDF for estimate of AUC with the estimate of error rate from 0.24 to 0.25, 0.30 to 0.31, 0.34 to 0.35, 0.39 to 0.4, 0.44 to 0.45 and 0.49 to 0.50. In each range of error rate, I have 20,467, 28,456, 19,620, 15,038, 13,536 and 29,677 estimates of AUC

value. The CDF plot of estimate of AUC is shown in figure (3.10).

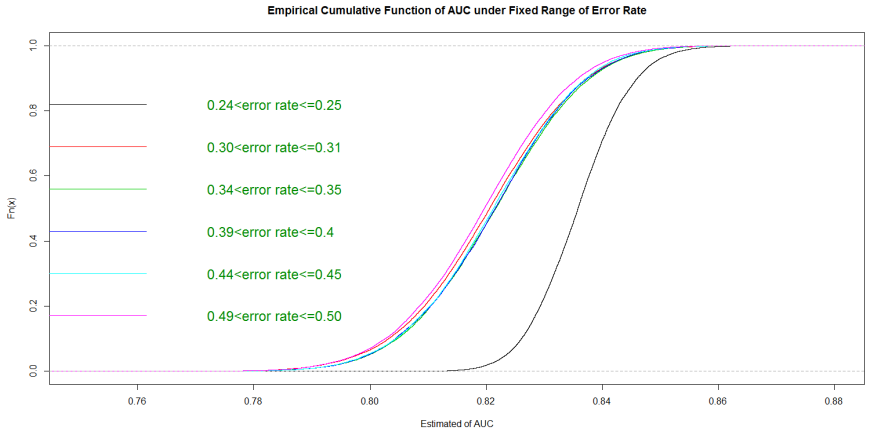

Figure 3.10: CDF of estimate of AUC(r=50%)

Figure (3.10) shows that the value of estimate of AUC with error rate from 0.24 to 0.25 is significantly higher than other estimate of AUC with higher error rate. The estimate of AUC with error rate from 0.3 to 0.31, 0.34 to 0.35, 0.39 to 0.4, 0.44 to 0.45 and 0.49 to 0.5 has similar distribution. The PDF plot of those estimates of AUC is shown in figure (3.11).

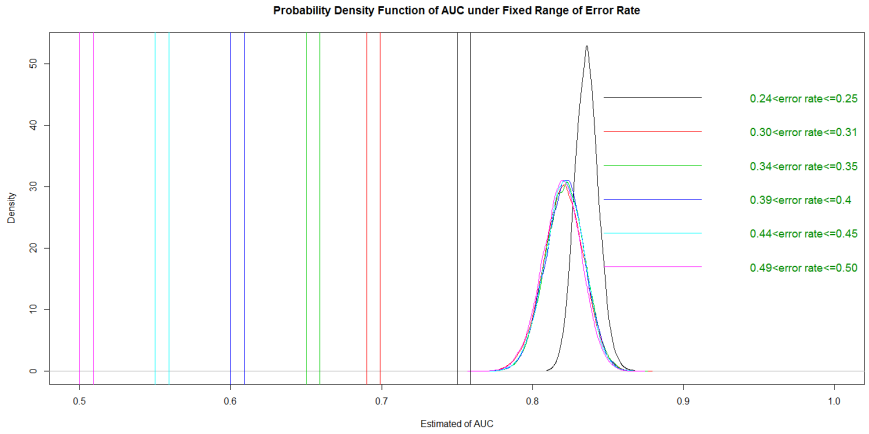

Figure 3.11: PDF of estimate of AUC(r=50%)

The PDF plot (figure (3.11) also shows that estimate AUC with error rate from 0.24 to 0.25 has the highest value, and the range of it is much narrow than others. For the estimate of AUC with other ranges of error rate, the distribution is similar. However, the expected average AUC with

those ranges of error rate differs a lot. And I calculate the difference between average estimate of AUC and the bounds(upper and lower) corresponding to each range of error rate (Table3.5) as reference. The table shows that the difference between the estimate of AUC and the expected average AUC gets larger when error rate gets larger. However, in this situation, the difference is smaller than pervious situation with smaller proportion of positive samples.

Table 3.5: Difference between estimate of AUC and expected average AUC (r=50%)

Error rate	(0.24,0.25]	(0.3,0.31]	(0.34,0.35]	(0.39,0.4]	(0.44,0.45]	(0.49,0.5]
Avg Est. of AUC-lower bound	0.0859	0.1304	0.171	0.2212	0.271	0.3194
Avg Est. of AUC-upper bound	0.0759	0.1204	0.161	0.2112	0.261	0.3094

In order to compare the difference between estimate of AUC and expected average AUC in each situation, I calculated the mean of difference between average estimate of AUC bounds (upper and lower). The results are showed in table 3.6.

Table 3.6: The Difference between estimate of AUC and Expected Average AUC

r=m/(m+n)	0.1	0.2	0.3	0.4	0.5
mean of difference(with lower bound)	0.318771	0.274505	0.231665	0.2105	0.199911
mean of difference(with upper bound)	0.265325	0.245932	0.212028	0.195908	0.189911

Table 3.6 indicates that when r is gets larger, the difference between the expected average AUC and estimate of AUC is gets smaller. That means when the positive samples and negative samples are even distributed, the deviation of expression (3.1) is smallest though the equiprobable assumption is not satisfied. But when the positive samples and negative samples are not evenly distributed, we cannot use equation 3.1 as a reference.

3.3 Study the Error Rate Distribution Under the Fixed AUC

Another objective of this thesis is to study how the error rate distributed under a fixed range of AUC. To observe the distribution clearly, I draw the CDF plots of error rate for the fixed range AUC of (0.49,0.51), (0.59,0.61), (0.69,0.71), (0.79,0.81) and (0.89,0.91). And I study situations that the ratio of positive samples $r = m/(m+n)$ vary from 0.1 to 0.5 by 0.1 to investigate whether the distribution conditional on the ratio of positive examples.

In this section, I use the same Monte Carlo datasets with the previous section. In order to get the large range of estimate AUC, I choose different predictors for logistic regression classifier for each range of estimate of AUC. The threshold of classifier is randomly chosen from a uniform distribution $U(0,1)$.

3.3.1 CDF and PDF Plots for Error Rate Under the Fixed Range AUC

First, I study the situation that $r = 10\%$. The CDF plot and PDF plot are shown in figure (3.12) and figure (3.13). For estimate of AUC from 0.49 to 0.51, I chose a_5 as predictor, and I have 3,444 estimates of AUC in this range. For estimate of AUC from 0.59 to 0.61, I chose a_1 as predictor, and I have 15,620 estimates of AUC in this range. For estimate of AUC from 0.69 to 0.71, I chose a_1, a_4, a_5 as predictor, and I have 35,709 estimates of AUC in this range. For estimate of AUC from 0.79 to 0.81, I chose a_1, a_3, a_5 as predictor, and I have 7,982 estimates of AUC in this range. And for estimate of AUC from 0.89 to 0.91, I chose a_2, a_4, a_5 as predictor, and I have 31,450 estimates of AUC in this range.

Figure 3.12 shows that when the estimate of AUC from 0.49 to 0.51, the majority values of estimate of error rate is around 0.1. When AUC gets larger, the range of error rate becomes large too, however, the maximum error rate is always around 0.9. The larger the AUC, the smaller the

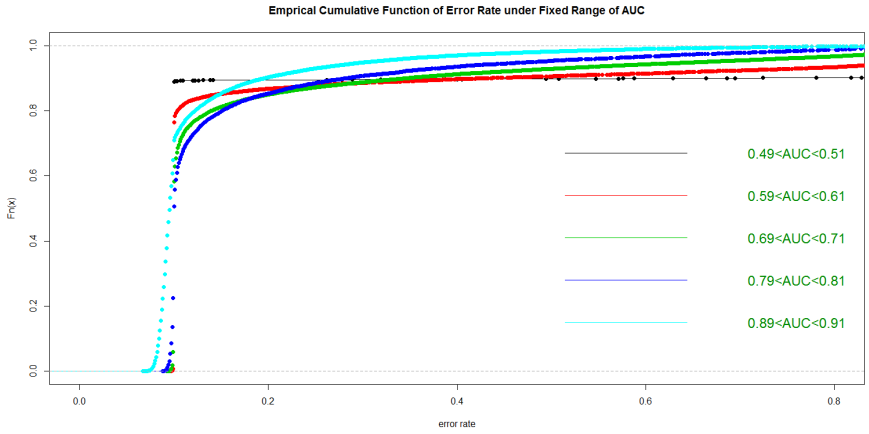

Figure 3.12: Error Rate Empirical Cumulative Distribution Under fixed AUC(r=10%)

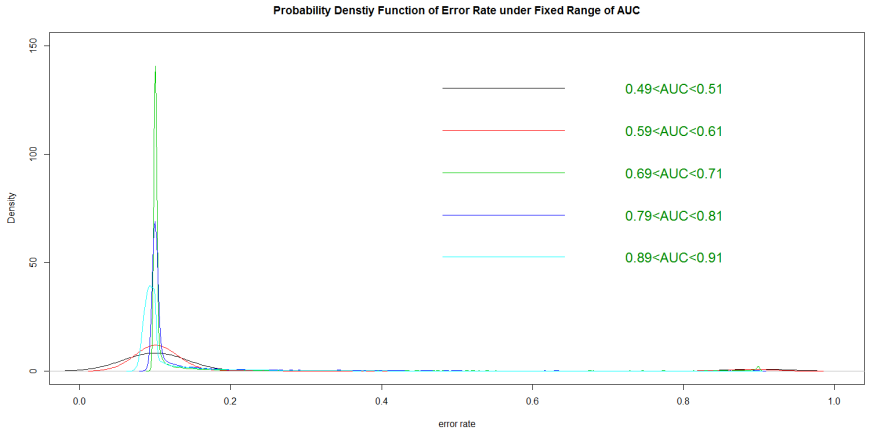

Figure 3.13: Error Rate Distribution Under fixed AUC(r=10%)

minimum error rate. From figure 3.13 I find that the mode of error rate is around 0.1, and the when AUC gets large, the mode of the density gets closer to 0. I also calculate the mean, median, and quantile of the error rate under different AUC (table3.7).

Table 3.7: Descriptive Statistics of Error Rate under Fixed AUC(r=10%)

AUC	0.49-0.51	0.59-0.61	0.69-0.71	0.79-0.81	0.89-0.91
Minimum	0.1	0.095	0.092	0.088	0.067
1st Quantile	0.1	0.1	0.1	0.1	0.089
Median	0.1	0.1	0.1	0.101	0.096
Mean	0.1817	0.1787	0.1691	0.156	0.1262
3rd Quantile	0.1	0.1783	0.116	0.127	0.107
Maximum	0.901	0.901	0.901	0.9	0.9

The descriptive statistic shows that the mean of the error rate gets smaller when the AUC gets larger. The maximum error rate for each AUC around 0.9, and the the median of the error rate is always around 0.1.

When $r = 20\%$, there are 200 positive samples and 800 negative samples. For estimate of AUC from 0.49 to 0.51, I chose a_5 as predictor, and I have 1,968 estimates of AUC in this range. For estimate of AUC from 0.59 to 0.61, I chose a_1 as predictor, and I have 7,798 estimates of AUC in this range. For estimate of AUC from 0.69 to 0.71, I chose a_1, a_4, a_5 as predictor, and I have 31,384 estimates of AUC in this range. For estimate of AUC from 0.79 to 0.81, I chose a_3, a_4, a_5 as predictor, and I have 11,826 estimates of AUC in this range. And for estimate of AUC from 0.89 to 0.91, I chose a_1, a_2, a_5 as predictor, and I have 60,867 estimates of AUC in this range.

I get the CDF plot and PDF plot in figure 3.14 and figure 3.15.

From the empirical cumulative function plot (3.14) I find that when estimate of AUC from 0.49 to 0.51, the majority of error rate is 0.2. When AUC gets larger, the minimum error rate gets smaller, but the maximum error rate is always around 0.8. The probability density plot (3.15) shows that there are two modes of probability in this plot. The higher one is around 0.2, and the

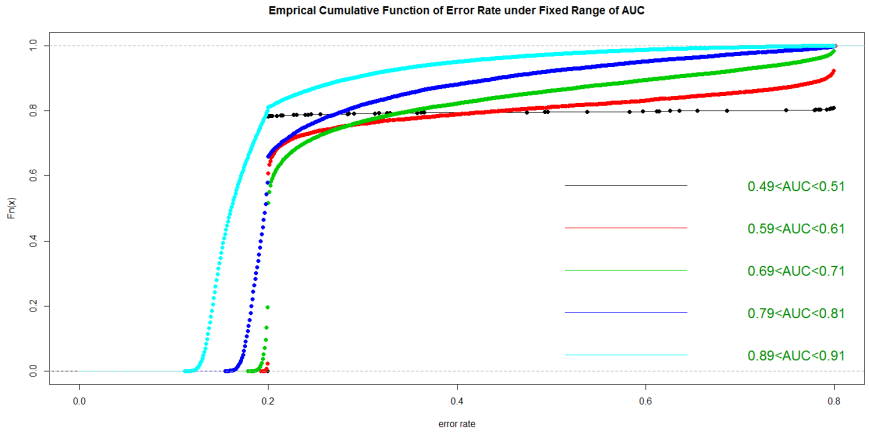

Figure 3.14: Error Rate Empirical Cumulative Distribution Under fixed AUC(r=20%)

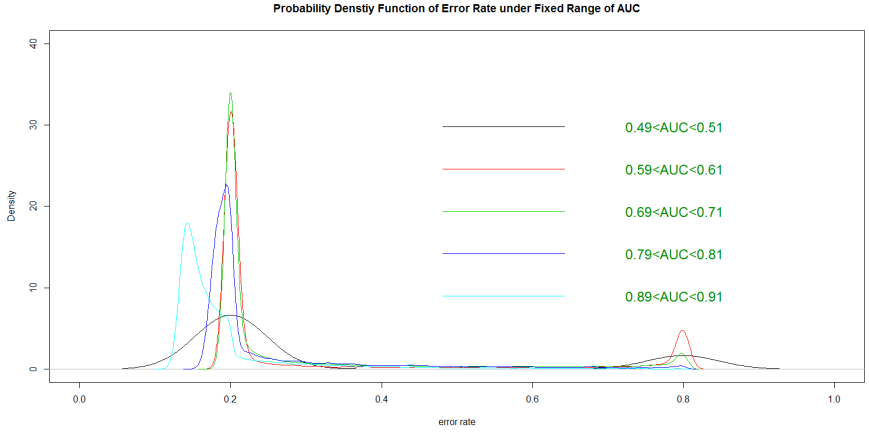

Figure 3.15: Error Rate Distribution Under fixed AUC(r=20%)

lower one is around 0.8. When the AUC becomes larger, the higher mode is more closer to 0.1. The descriptive statistics are shown in table (3.8).

Table 3.8: Descriptive Statistics of Error Rate under Fixed AUC(r=20%)

AUC	0.49-0.51	0.59-0.61	0.69-0.71	0.79-0.81	0.89-0.91
Minimum	0.199	0.192	0.178	0.154	0.111
1st Quantile	0.2	0.2	0.2	0.185	0.144
Median	0.2	0.2	0.2	0.197	0.161
Mean	0.3226	0.3157	0.2917	0.2524	0.1925
3rd Quantile	0.2	0.276	0.28	0.239	0.192
Maximum	0.8	0.801	0.8	0.8	0.8

This table indicates that the maximum error rate under each AUC is almost the same, which around 0.8. The mean error rate gets smaller when AUC gets larger. For the first four range of estimate of AUC, the first quantile of error rate is around 0.2. But for estimate of AUC from 0.89 to 0.91, the first quantile of error rate is munch more smaller, which is around 0.14. The minimum and mean of error rate gets smaller when the range of AUC estimate gets larger.

When $r = 30\%$, I have 300 positive samples and 700 negative samples. For estimate of AUC from 0.49 to 0.51, I chose a_5 as predictor, and I have 1,488 estimates of AUC in this range. For estimate of AUC from 0.59 to 0.61, I chose a_1 as predictor, and I have 4,156 estimates of AUC in this range. For estimate of AUC from 0.69 to 0.71, I chose a_1, a_4, a_5 as predictor, and I have 32,683 estimates of AUC in this range. For estimate of AUC from 0.79 to 0.81, I chose a_3, a_4, a_5 as predictor, and I have 35,089 estimates of AUC in this range. And for estimate of AUC from 0.89 to 0.91, I chose a_1, a_2, a_4 as predictor, and I have 20,490 estimates of AUC in this range.I get the CDF plot and PDF plot in figure 3.16 and figure 3.17.

Based on empirical cumulative function plot (3.14) I notice that when the range of AUC estimate is 0.49 to 0.51, there are two jump of CDF plot, one is around 0.3 and the other is around 0.7. When the AUC estimate gets larger, the minimum error rate gets smaller, but the maximum AUC

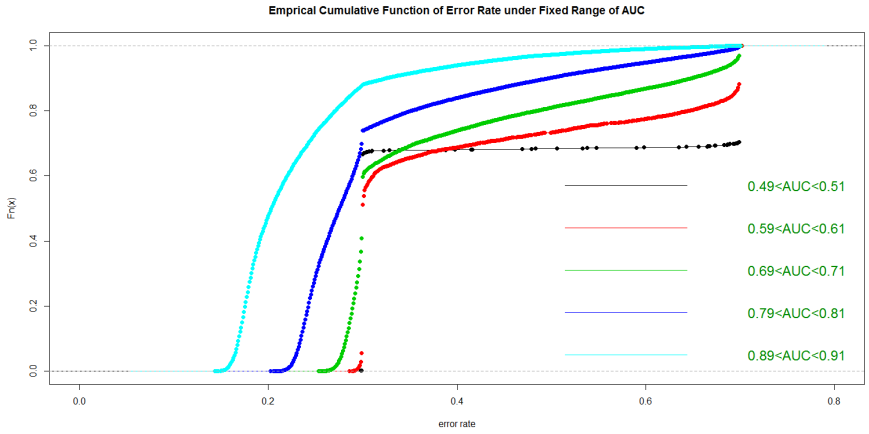

Figure 3.16: Error Rate Empirical Cumulative Distribution Under fixed AUC(r=30%)

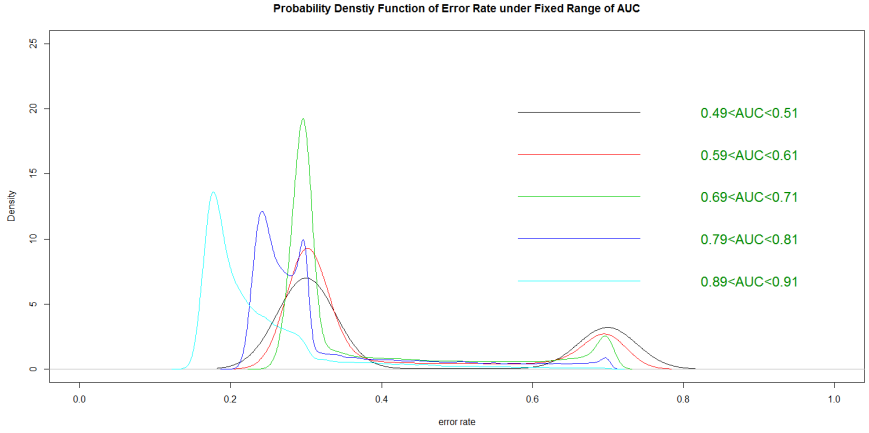

Figure 3.17: Error Rate Distribution Under fixed AUC(r=30%)

is always around 0.7. From the probability density plot I can see that when AUC estimate is small (range of (0.49,0.51),(0.59,0.61),(0.69,0.71)) ,there are two modes of error rate, one is around 0.7 and the other is around 0.3. When AUC estimate gets larger,there are multiple modes. Both of them have one mode around 0.7. For AUC estimate range from 0.79 to 0.81, it has one mode around 0.3, one mode around 0.2 and another around 0.7. And for estimate of AUC from 0.89 to 0.91, it has one mode around 0.25 and another around 0.19. The descriptive statistics are shown in table (3.9).

Table 3.9: Descriptive Statistics of Error Rate under Fixed AUC(r=30%)

AUC	0.49-0.51	0.59-0.61	0.69-0.71	0.79-0.81	0.89-0.91
Minimum	0.298	0.286	0.253	0.202	0.143
1st Quantile	0.3	0.3	0.293	0.247	0.179
Median	0.3	0.3	0.3	0.276	0.204
Mean	0.4264	0.4078	0.375	0.3157	0.2339
3rd Quantile	0.7	0.534	0.414	0.309	0.256
Maximum	0.7	0.702	0.701	0.7	0.7

This table shows that the maximum error rate for each value of the AUC is around 0.7. The minimum and mean error rate gets smaller when the AUC gets larger. The median and first quantile of error rate for first three ranges of AUC estimate are around 0.3. For the AUC estimate from 0.89 to 0.91, the first quantile and median error rate is much more smaller than 0.3.

For $r = 40\%$, I have 400 positive samples and 600 negative samples. For estimate of AUC from 0.49 to 0.51, I chose a_5 as predictor, and I have 1,488 estimates of AUC in this range. For estimate of AUC from 0.59 to 0.61, I chose a_1 as predictor, and I have 5,870 estimates of AUC in this range. For estimate of AUC from 0.69 to 0.71, I chose a_1, a_4, a_5 as predictor, and I have 34,875 estimates of AUC in this range. For estimate of AUC from 0.79 to 0.81, I chose a_1, a_3, a_5 as predictor, and I have 16,566 estimates of AUC in this range. And for estimate of AUC from 0.89 to 0.91, I chose a_2, a_3, a_5 as predictor, and I have 9,290 estimates of AUC in this range.

The CDF plot and PDF plot are shown in figure 3.18 and figure 3.19.

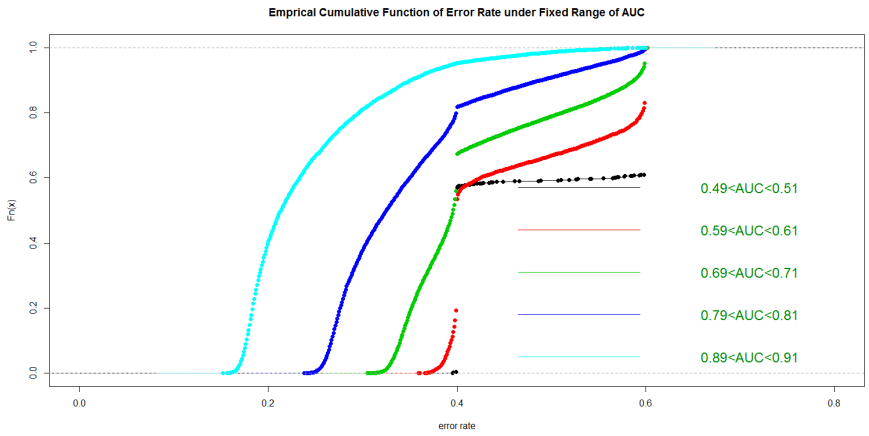

Figure 3.18: Error Rate Empirical Cumulative Distribution Under fixed AUC(r=40%)

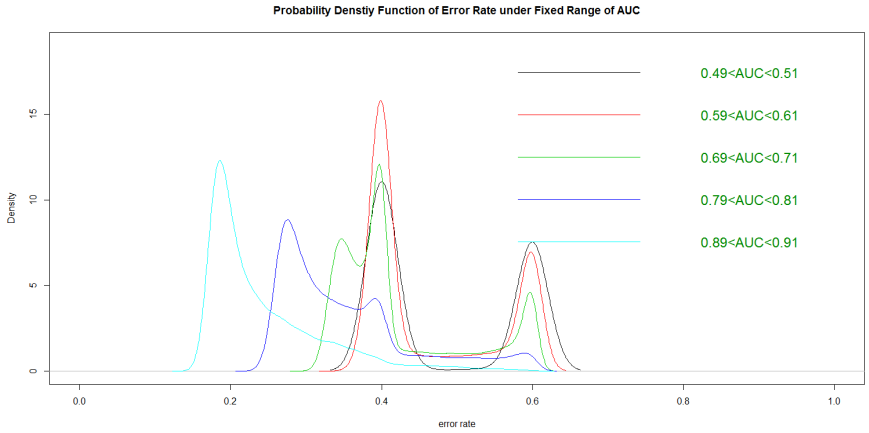

Figure 3.19: Error Rate Distribution Under fixed AUC(r=40%)

From the CDF plot I can see that when the estimate of AUC is from 0.49 to 0.51, there are two jump in the CDF plot, one around 0.4 and the other around 0.6. When AUC estimate gets larger, the minimum error rate getting smaller. For AUC estimate from 0.89 to 0.91, the minimum error rate is less than 0.2. However, the largest error rate still around 0.6. From the probability density plot I notice that for estimate of AUC from 0.49 to 0.51 and 0.59 to 0.61, the error rate has two modes, one around 0.6 and another around 0.4. For estimate of AUC from 0.69 to 0.71, the error rate has three modes, one around 0.4, one around 0.35 and the other around 0.6. For AUC estimate

from 0.79 to 0.81, the error rate has three modes, one around 0.3, one around 0.4 and the other around 0.6. However, for estimate of AUC from 0.89 to 0.91, the error rate only has one mode and it is around 0.2. The descriptive statistics are s shown in table (3.10).

Table 3.10: Descriptive Statistics of Error Rate under Fixed AUC(r=40%)

AUC	0.49-0.51	0.59-0.61	0.69-0.71	0.79-0.81	0.89-0.91
Minimum	0.395	0.359	0.305	0.238	0.152
1st Quantile	0.4	0.4	0.359	0.282	0.187
Median	0.4	0.4	0.396	0.325	0.214
Mean	0.4815	0.4643	0.4222	0.3505	0.2431
3rd Quantile	0.6	0.579	0.463	0.391	0.276
Maximum	0.6	0.602	0.601	0.6	0.6

This table shows that the maximum error rate for all ranges of the estimate of AUC is around 0.6. The mean and third quantile error rate gets smaller along with the AUC gets larger. The first quantile and minimum error rate is around 0.4 for estimate of AUC from 0.49 to 0.51 and 0.59 to 0.61. And when AUC gets larger, the first quantile and minimum error rate gets smaller.

For $r = 50\%$, there are 500 positive samples and 500 negative samples. For estimate of AUC from 0.49 to 0.51, I chose a_5 as predictor, and have 1,089 estimates of AUC in this range. For estimate of AUC from 0.59 to 0.61, I chose a_4 as predictor, and have 14,419 estimates of AUC in this range. For estimate of AUC from 0.69 to 0.71, I chose a_1, a_4, a_5 as predictor, and have 37,195 estimates of AUC in this range. For estimate of AUC from 0.79 to 0.81, I chose a_1, a_3, a_5 as predictor, and I have 19,311 estimates of AUC in this range. And for estimate of AUC from 0.89 to 0.91, I chose a_1, a_3, a_4 as predictor, and have 5,969 estimates of AUC in this range.

Since the probability density function for estimate of AUC from 0.49 to 0.51 and for other ranges of estimate of AUC differs a lot, I draw the PDF for these two situations separately. The CDF plots and PDF plot are showen in figure 3.20, figure 3.21,and figure 3.22.

From the CDF plot I notice that when the estimate of AUC is from 0.49 to 0.51, the majority

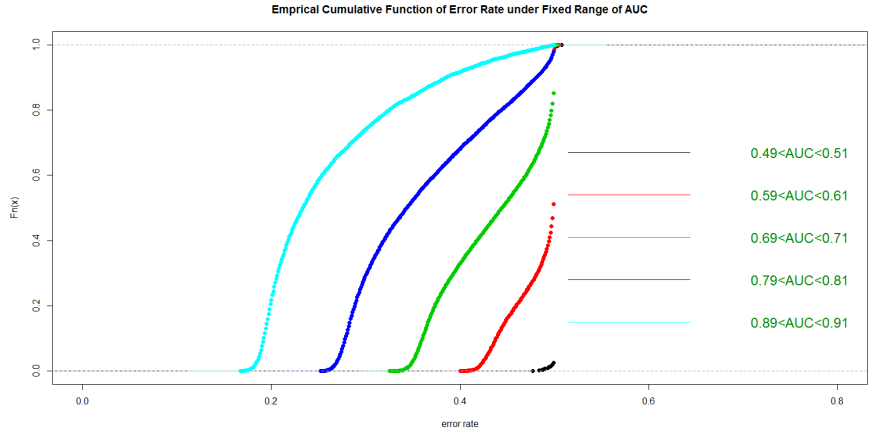

Figure 3.20: Error Rate Empirical Cumulative Distribution Under fixed AUC(r=50%)

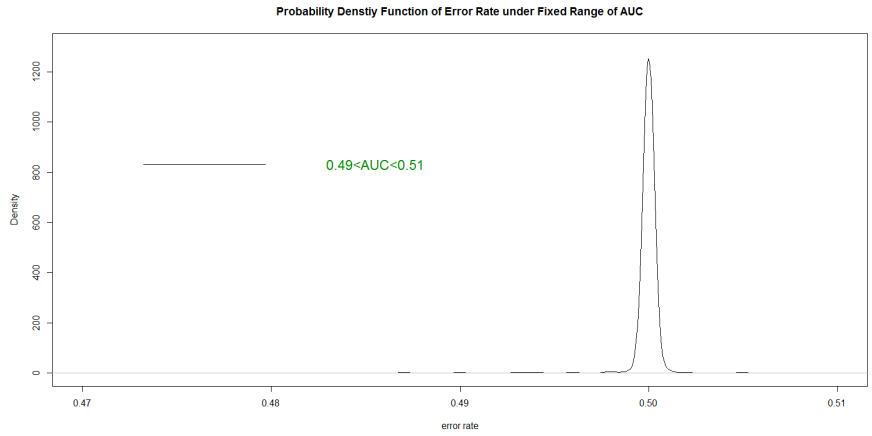

Figure 3.21: Error Rate Distribution Under fixed AUC(r=50%)

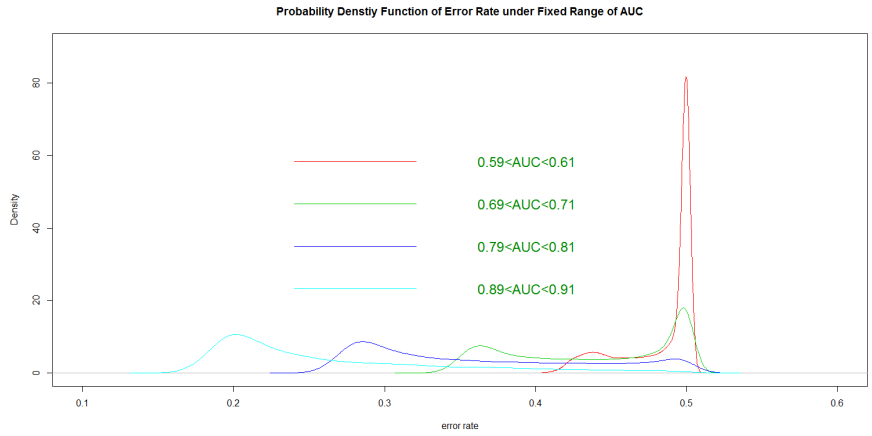

Figure 3.22: Error Rate Distribution Under fixed AUC(r=50%)

of error rate is equal to 0.5. And when AUC gets larger, the error rate varies from 0.16 to 0.5. When the estimate of AUC is from 0.89 to 0.91, the minimum error rate is less than 0.2. From the figure 3.21 I find that when $AUC = 0.5$, the probability that error rate equals to 0.5 is extremely large. Based on figure 3.22 I find that when the estimate of AUC is from 0.59 to 0.61, the error rate value has two modes, one is around 0.5 and the other is around 0.45. And for estimate of AUC is from 0.69 to 0.71 , the error rate has two modes around 0.5 and 0.3. For estimate AUC from 0.79 to 0.81 and from 0.89 to 0.91, the error rate only has one mode which is around 0.25 and 0.2 correspondingly. The descriptive statistics are shown in table (3.11).

Table 3.11: Descriptive Statistics of Error Rate under Fixed AUC(r=50%)

AUC	0.49-0.51	0.59-0.61	0.69-0.71	0.79-0.81	0.89-0.91
Minimum	0.477	0.4	0.325	0.252	0.167
1st Quantile	0.5	0.472	0.383	0.295	0.203
Median	0.5	0.499	0.445	0.344	0.234
Mean	0.4999	0.4836	0.437	0.3622	0.2631
3rd Quantile	0.5	0.5	0.494	0.425	0.304
Maximum	0.508	0.505	0.505	0.5	0.5

From this table I find that the maximum error rate is around 0.5. The mean error rate gets smaller when AUC gets larger. Only for estimate of AUC from 0.49 to 0.51, the first quantile and the minimum error rate is around 0.5. For other larger AUC, the first quantile and minimum error rate is much smaller than 0.5.

Chapter 4

Conclusion

4.1 Summary

The objective of this thesis was to investigate the relationship between AUC and misclassification rate. Cortes and Mehryar (2004) have provided expression for the expected value of AUC given error number k that is only valid when all classification or rankings with k errors are equiprobable. The assumption of this equation is too strong to met the real life scenarios. And I found that Cortes and Mehryar's expression is not valid in the situation where error number k is larger than $min(m,n)$. For their expression to be valid, the constraint $k > min(m,n)$ needs to be imposed.

I simulated a binary distribution using logistic regression and used logistic regression model as a classifier to study the relationship between misclassification rate and AUC. First, I compared the estimate of AUC value to the expected average value of AUC calculated by equation 3.1 only for situation that $k \leq min(m,n)$. The results showed the expected average value of AUC is always lower than the estimate of AUC. When the positive samples and negative samples were evenly distributed, the difference between estimate of AUC and expected average value of AUC are smallest. Thus, one can use equation 3.1 as a reference to learn the relationship between the AUC and the error rate when I have same proportion of positive examples and negative examples. But when the proportions of positive samples are extreme close to 0 or 1, this expression is very questionable to be used.

Furthermore, I studied the error rate distribution under a fixed range of AUC value when r

varies from 0.1 to 0.5. The results showed that when $r = 0.1, 0.2$, the mode of error rate is always around r or $1 - r$. When AUC becomes larger, the distribution of error rate becomes to the right skewed, and the mean error rate becomes smaller.

4.2 Limitation

In this thesis, I did a simulation study to investigate the relationship between AUC and misclassification rate for binary distribution. Specifically, I tested the validity of the expression given by Cortes and Mehryar (2004) and studied the distribution of error rate under the estimate of AUC.

In the first analysis part, I calculated the estimate of AUC given fixed range of estimate of error rate to validate the expression given by Cortes and Mehryar (2004). I got the estimate of error rate by a logistic regression classifier. Because the threshold for the classifier is unknown, I just simply assumed that the threshold follows a uniform distribution from 0 to 1. This assumption may be uncorrect. In the second analysis part, I studied the distribution of estimate of error rate under a fixed range of estimate of AUC. Because AUC is a continuous variable and I can not get the error rate distribution under every single value of AUC, I just specifically looked into 5 intervals of AUC. I did not use the same classifier the get these 5 intervals of AUC value because it needs an extremely large number of Mote Carlo samples. Instead, I used 5 different classifier to get these 5 intervals of AUC value. This is not appropriate and the distribution of error rate under AUC would be more accurate if I had more Mote Carlo samples.

4.3 Discussion

In this thesis, I evaluated the validity of the expression provided by Cortes and Mehryar (2004) for moderate to large deviations of the equiprobable. My results showed that when the positive samples and negative samples are not evenly distributed, the expression is questionable. Based on my work, people can have a brief idea of how error rate is distributed under a fixed range of AUC. To investigate the relationship more precise, I can apply the Bayesian inference methods. The difficulties of Bayesian inference is that both the error rate and the AUC are random variables, and it is hard to define the distribution of AUC and error rate.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Gorno-Tempini, M. L., et al. "Classification of primary progressive aphasia and its variants." Neurology 76.11 (2011): 1006-1014.

[2] Planet, Paul J., et al. "Phylogeny of genes for secretion NTPases: identification of the widespread tadA subfamily and development of a diagnostic key for gene classification." Proceedings of the National Academy of Sciences 98.5 (2001): 2503-2508.

[3] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[4] Cortes, Corinna, and Mehryar Mohri. "AUC optimization vs. error rate minimization." Advances in neural information processing systems 16.16 (2004): 313-320.

[5] Mohri, C. "Confidence intervals for the area under the ROC curve." Advances in neural information processing systems 17 (2005): 305.

[6] Hand, David J., and Robert J. Till. "A simple generalisation of the area under the ROC curve for multiple class classification problems." Machine learning 45.2 (2001): 171-186.

[7] Herschtal, Alan, and Bhavani Raskutti. "Optimising area under the ROC curve using gradient descent." Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.

[8] Ma, Shuangge, and Jian Huang. "Regularized ROC method for disease classification and biomarker selection with microarray data." Bioinformatics 21.24 (2005): 4356-4362.

[9] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[10] Och, Franz Josef. "Minimum error rate training in statistical machine translation." Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003.

[11] Ben-David, Shai, et al. "Minimizing the misclassification error rate using a surrogate convex loss." arXiv preprint arXiv:1206.6442 (2012).

[12] Murthy, Sreerama K. "Automatic construction of decision trees from data: A multi-disciplinary survey." Data mining and knowledge discovery 2.4 (1998): 345-389.

[13] Joachims, Thorsten. "A support vector method for multivariate performance measures." Proceedings of the 22nd international conference on Machine learning. ACM, 2005.

[14] Metz, Charles E. "Basic principles of ROC analysis." Seminars in nuclear medicine. Vol. 8. No. 4. WB Saunders, 1978.

[15] Ferri, Cesar, Jose Hernandez-Orallo, and R. Modroiu. "An experimental comparison of performance measures for classification." Pattern Recognition Letters 30.1 (2009): 27-38.

[16] Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.

[17] Ling, Charles X., Jin Huang, and Harry Zhang. "AUC: a statistically consistent and more discriminating measure than accuracy." IJCAI. Vol. 3. 2003.

[18] Bradley, Andrew P. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." Pattern recognition 30.7 (1997): 1145-1159.

[19] Mann, H.B., Whitney, D.R. (1947) On a test whether one of two random variables is stochastically larger than the other. Ann. Math. Statist., 18, pp. 50-60.

[20] Wilcoxon, F. (1945) Individual comparisons by ranking methods. Biometrics, 1, pp. 80-83.

[21] Hanley, James A., and Barbara J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." Radiology 143.1 (1982): 29-36.

[22] Huang, Jin, and Charles X. Ling. "Using AUC and accuracy in evaluating learning algorithms." Knowledge and Data Engineering, IEEE Transactions on 17.3 (2005): 299-310.

[23] Breiman L., Friedman J.H., Olshen R.A., Stone C.J.(1984) Classification and Regression Trees, Wadsforth International Group.

[24] Mika, S., Ratsch, G., Weston, J., Scholkopf, B. and Muller, K.-R. (1999), Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, Neural Networks for Signal Processing IX, pages 41-48. IEEE.

[25] Murthy, Sreerama K. "Automatic construction of decision trees from data: A multi-disciplinary survey." Data mining and knowledge discovery 2.4 (1998): 345-389.

[26] Good I.J. (1950), Probability and the Weighing of Evidence, London, Charles Grin.

[27] Mitchell, T. (1997). Machine Learning. McGraw Hill.

[28] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." (2007): 3-24.

[29] Moses, Lincoln E., David Shapiro, and Benjamin Littenberg. "Combining independent studies of a diagnostic test into a summary roc curve: Data-analytic approaches and some additional considerations." Statistics in medicine 12.14 (1993): 1293-1316.

[30] Metz, Charles E., Benjamin A. Herman, and Cheryl A. Roe. "Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets." Medical Decision Making 18.1 (1998): 110-121.

[31] Agarwal, Shivani, et al. "Generalization bounds for the area under the ROC curve." Journal of Machine Learning Research. 2005.

[32] Dreiseitl, Stephan, and Lucila Ohno-Machado. "Logistic regression and artificial neural network classification models: a methodology review." Journal of biomedical informatics 35.5 (2002): 352-359.

[33] Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

[34] Statnikov, Alexander, Lily Wang, and Constantin F. Aliferis. "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification." BMC bioinformatics 9.1 (2008): 319.

[35] Lee, Eunjung, et al. "Inferring pathway activity toward precise disease classification." PLoS computational biology 4.11 (2008): e1000217.

[36] Kim, Ji-Hyun. "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap." Computational Statistics and Data Analysis 53.11 (2009): 3735-3745.

[37] Lagreid, Astrid, et al. "Predicting gene ontology biological process from temporal gene expression patterns." Genome research 13.5 (2003): 965-979.

[38] Deselaers, Thomas, Daniel Keysers, and Hermann Ney. "Classification error rate for quantitative evaluation of content-based image retrieval systems." Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Vol. 2. IEEE, 2004.

[39] Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science 286.5439 (1999): 531-537.

[40] Wang, Qiong, et al. "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." Applied and environmental microbiology 73.16 (2007): 5261-5267.

[41] Kuncheva, Ludmila I., and Christopher J. Whitaker. "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy." Machine learning 51.2 (2003): 181-207.

[42] Simon, Richard, et al. "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification." Journal of the National Cancer Institute 95.1 (2003): 14-18.

[43] Swets, John A. "Measuring the accuracy of diagnostic systems." Science 240.4857 (1988): 1285-1293.