Boosting Methods for Maximization of the Area under the ROC Curve and their Applications to Clinical Data

Osamu Komori

Doctor of Statistics

Department of Statistical Science

School of Multidisciplinary Sciences

The Graduate University for Advanced Studies

2010

Preface: Motivation and outline of this thesis

With the advent of information age, huge amount of data has been collected in laboratories and hospitals. It includes not only clinical data such as age, laboratory test values, the size of internal organ; but also genomic data such as gene expression patterns, single nucleotide polymorphism (SNP) and proteome. Based on the information, we want to predict as accurately as possible the condition of the subject (diseased or non-diseased), who comes to a hospital and has gone through some clinical tests. However, it is often difficult to analyze these variety of medical data within a traditional statistical framework. Moreover, there exist criteria that are suitable for medical and clinical sciences. Hence, we have tried to develop a new statistical method that can deal with these data and provide us with a useful information for the discrimination, based on a criterion that is widely used by medical doctors or clinical researchers.

In medical and biological sciences, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) have gained in popularity. The ROC curve originated from the signal detection theory, where the performance of the radar operator who monitors enemy warplanes is measured or compared using the curve. It is also applied in psychology, and now is used in a variety of discrimination problems. Its appealing points are that the false positive rate (FPR) and the true positive rate (TPR) are both measured in the ROC curve, and that the curve is independent of the population prevalence of disease. FPR and 1-TPR express different aspects of the classification performance, so it is important to report the values separately, when evaluating the goodness of the classification. The independence also is suitable for quantifying the inherent accuracy of classification, and this property makes the AUC different from other accuracy measures such as the error rate, the relative risk or the odds ratio.

In this thesis, we have developed a new statistical method that is designed to optimize the AUC based on a boosting technique, which is widely used in the machine learning community. The method can deal with both usual low dimensional settings as well as high dimensional settings. The main concept of boosting is that a strong classifier (score function) is constructed by combining many various “weak classifiers”. The weak classifier means that its discriminant ability is slightly better than random guessing. The method includes an implicit procedure of marker selection in its boosting algorithm, and produce a score function after an appropriate number of iterations. The resulting score plots are shown to be useful for understanding how each marker is associated with the outcome variable, say, the status of the subjects (non-diseased or diseased). Hence, our method put importance on the classification accuracy as well as the interpretation of the result. We also have extended this AUC-based boosting method to pAUCBoost, which focuses on the partial area under the ROC curve (pAUC) that is often more relevant in some clinical or medical situations.

In Chapter 1, we review other accuracy measures than the AUC and pAUC, which are also important in clinical evaluation of markers; we investigate the properties and consider why the AUC and pAUC are getting popular in recent years. In Chapter 2, we also review the status of progress and development in machine learning community, and characterize the property of boosting from an objective viewpoint. We propose a new statistical method, termed AUCBoost, in Chapter 3 and discuss the statistical properties and demonstrate its utility. In Chapter 4, we focus on PSA data analysis. This is a collaborative research with medical doctors in Keio University Hospital. PSA is an abbreviation of prostate specific antigen, and is a primary marker for prostate cancers. The subject with PSA larger than 4 ng/ml is usually recommend to undergo biopsy; however, the value is affected by the age and the size of the prostate gland and other clinical covariates. Hence, we consider a optimal combination of these markers as well as the association to the prostate cancer, using AUCBoost. As a result, we present a "nomogram", by which medical doctors determine whether they perform biopsy in consideration of PSA, age, the volume of prostate gland and the number of biopsy undergone. The point of this nomogram is that the cutoff points are determined so that the sensitivity is at least 95 percent. This idea is quite different from existing nomograms that are based on a probability of having the cancer, and much more suitable for practical medical diagnosis. In Chapter 5, we extend AUCBoost to pAUCBoost, which focuses on the partial area under the ROC curve. We show that pAUCBoost is preferable to AUCBoost in some clinical situations. In Chapter 6, we mention ongoing and future work that I am engaged in now. Finally, we close this thesis with acknowledgements to all persons who supported me during my hard and pleasant doctor course.

Contents

Chapter 1

Classification in medical sciences

The purpose of medical data analysis is to detect useful markers or diagnostic tests, and properly combine them to increase classification performance between diseased subjects and non-diseased ones. It leads to improvement of quality of medical care and alleviation of the mental or financial burden of patients; hence, it is needed to develop a statistical method that not only has a good classification performance but also suits for medical and clinical sciences. In this chapter, we review fundamental points or terms that we should know before analyzing real data actually.

1.1 Medical diagnostic tests

1.1.1 Several types

Medical doctors diagnosis a subject or a patient by checking his temperature or listening to the heart with a stethoscope, which we call simple physical examinations. On the other hand, more sophisticated medical treatments are often needed such as X-rays for lung cancers or kidney stones, MRI (Magnetic Resonance Imaging) for brain diseases and muscle abnormalities. Originally, diagnostic tests are conducted for detecting disease; however, it includes tests for prognosis in a broad sense. In this case, the condition to be detected is not disease but a clinical outcome several months after diagnosis. Tests of disease screening are also included in this category. Usually, screening tests are performed on subjects who have no symptoms of disease; hence, they require high specificity with acceptable sensitivity to avoid adverse effects such as unnecessary follow-up treatment and over-diagnosis. We will refer to the importance of specificity and sensitivity later.

1.1.2 Necessary conditions for medical diagnostic tests

Pepe (2003) and Obuchowski *and others* (2001) suggests several important conditions that medical diagnostic tests should satisfy as follows.

1. The target disease should be mortal or severe.

   - If the target disease is not severe, nobody comes to a hospital to be examined. This is from a cost-effectiveness standpoint.

2. The prevalence rate of the disease should be relatively high.

   - Even if the diagnostic test has high sensitivity and specificity, say, 95% both of them, the probability of disease conditioned on positive test result is just about 16% if the prevalence rate is 1%. On the other hand, we have 50% probability if the prevalence rate is 5%. These are easily calculated by Bayes' theorem.

3. The medical diagnostic tests, especially, screening tests should discriminate disease from pseudo-disease.

   - Pseudo-disease means a disease that never progress or progress so slowly that it does not affect negatively the patient's condition. It is common in diseases that has a long period between onset of disease and the appearance of the signs or symptoms the patient has, or we can see it among patients with short life expectancies. This is the case for prostate cancer screening, in which the progress of the cancer is relatively slow and most of the patients are elderly adults. We address this problem by proposing a new medical tool termed a PSA cut-off nomogram in Chapter 4.

4. Screening should be performed before critical point.

   - A critical point is a boundary point, after which the patient need medical care. For example, the point of metastasis of primary tumor. Hence, a effective treatment is possible before the critical point.

5. The medical test should be harmless.

   - The diagnostic test must not inflict mortality (death due to the disease) or morbidity (being sick with the disease) on those screened.

6. The charge for the diagnostic test should be affordable and available to the patients.

   - More patients are examined, more beneficial and effective the diagnostic test are.

7. Treatment for the disease should be already established.

   - The diagnostic test is meaningful only if the target disease is curable. Note that Parkinson's disease or Alzheimer's disease are well known, but there is no treatment for these diseases.

8. Treatment after the diagnostic test should not be life-threatening nor fatal.

   - In the case that false positive rate is high, this requirement is indispensable. Moreover, note that earlier treatment means that the patient suffers the detrimental effects of the treatment earlier and for a longer time than usual.

9. The accuracy of the diagnostic test should be as high as possible.

   - The patient's burden is alleviated and the benefit is increased if we can grasp and understand the patient's condition accurately and appropriately. This last necessary condition of diagnostic test is the most important in implementing effective treatment for the patients. In Chapter 3 and 5, we propose a new statistical method that is designed to combine various diagnostic tests in order to increase the total accuracy of classification performance.

1.1.3 Case control study and cohort study

There are two types of study designs: *case-control study* and *cohort study*. The first one is also called retrospective study, because the subjects are selected on the basis of known true disease status. Usually, we collect a number of diseased subjects under investigation: the cases; then, we collect the counterparts: the controls who are healthy and free of disease. The latter study is also called prospective study, because we fix a target population and observe what happens during a specific period for the selected subjects. The status of the subject is determined by a gold standard definitive test, which is often invasive such as surgery or biopsy. We next consider the advantages and disadvantages of the two studies.

Advantages of case-control study

1. Case-control study is easily executable and inexpensive in comparison with cohort study, because we can use existing data and collect them much quicker than the follow-up study. This is very suitable for rare diseases or those that have long incubation periods.

2. We can easily keep the balance of the two groups: the controls and the cases. This leads to much smaller sample size needed for accurate results, especially when the prevalence rate is very low. With balanced design, we can also evaluate confounding and interaction more precisely (see Subsection 1.1.5).

Disadvantages of case-control study

1. Case-control studies do not meet one of conditions of *causality principle* (see Subsection 1.1.4). For example, consider a causal effect of drinking upon stomach cancer. We may assume that the habit of drinking causes the stomach cancer. However, there is a possibility that the stomach cancer patients have begun drinking to be comforted and relaxed. We can not take time factor into consideration in case-control studies.

2. The cases in case-control studies may not be appropriate samples that do not represent the targeted population. If there exists a strong association between drinking and a

heavy disease, the collected cases may have tendency to be less drinking-associated because the most of them have already died of its severity of the disease. This gives major impact on the results of the case-control study.

3. Since we start a case control study after fixing a target disease, we can get results regarding only the disease. We can rarely obtain other epidemiological evidences that leads to a further expansion of the study.

4. Case-control studies easily suffer from bias error, because of its way of collecting samples and the accuracy of reports from the two groups is different. The information about the case is more accurate in general, because it is researched more thoroughly. This disadvantage often quoted in the criticism of the case-control approach.

1.1.4 Principles of causality

These principles are suggested by Sir Austin Bradford Hill and cited by Woodward (2005).

1. There should be evidence of a strong association between the risk factor and the disease. Weak relationships may be due to chance occurrence and are more likely to be explained by confounding.

2. There should be evidence that exposure to the risk factor preceded the onset of disease.

3. There should be a plausible biological explanation.

4. The association should be supported by other investigations in different study settings. This is to protect against chance findings and bias caused by a particular choice of study population or study design.

5. There should be evidence of reversibility of the effect. That is, if the cause is removed, the effect should also disappear, or at least less likely.

6. There should be evidence of a dose-response effect. That is, the greater the amount of exposure to the risk factor is, the greater is the chance of disease.

7. There should be no convincing alternative explanation. For instance, the association should not be explained by confounding.

1.1.5 Confounding and interaction

When the relation of the risk and the disease can be *explained* by the third factor, it is called confounding factor. A typical example is the age of the subjects. When the relation of the risk and the disease can be *modified* by the third factor, it is called interaction factor. A typical example is difference of sex: men or women. It is widely known that some diseases are closely related to sex.

1.2 Criteria for diagnostic accuracy

The diagnostic accuracy can be measured by sensitivity, specificity, odds ratio and likelihood ratio when the test result is binary such as positive or negative. On the other hand, if it takes ordered or continuous values, it is more appropriate to use the receiver operating characteristic curve (ROC).

1.2.1 Sensitivity and specificity

Let $x \in \mathbf{R}$ be a marker or test result, y be a class label indicating non-diseased ($y = 0$) or diseased ($y = 1$), and $F(x)$ be a score function. Given a value of score function calculated from a subject having x, we classify him to be positive (diseased) or negative (non-diseased) as follows:

$$\begin{aligned} \text{if } F(x) \geq c &\;\Rightarrow\; \text{positive} \\ \text{else } F(x) < c &\;\Rightarrow\; \text{negative}, \end{aligned}$$

where c is a threshold value. Then we have two resulting probabilities

$$\begin{aligned} \text{sensitivity} &= P(F(x) > c | y = 1) \\ \text{specificity} &= P(F(x) < c | y = 0). \end{aligned}$$

Table 1.1: *Display of result of PSA test*

patient status	positive (PSA≥4)	negative (PSA<4)	total
diseased	127	3	130
non-diseased	251	19	270
total	378	22	400

They are also called true positive rate and false positive rate, respectively. Table 1.1 shows a summary table about prostate specific antigen (PSA) data provided by Keio University Hospital. The total sample number is 400, where the number of diseased and non-diseased subjects are $n_1 = 130$ and $n_0 = 270$, respectively. In this case x is a value of PSA, $F(x) = x$ and $c = 4$ ng/ml, where 4 ng/ml is widely used in urology. The sensitivity and specificity of this PSA data are

$$\begin{aligned} \text{sensitivity} &= 127/130 = 0.977, \\ \text{specificity} &= 19/270 = 0.07. \end{aligned}$$

The confidence interval for sensitivity proposed by (Agresti and Coull, 1998) is

$$\frac{\text{sen} + z^2_{1-\alpha/2}/(2n_1) \pm z_{1-\alpha/2}\sqrt{[\text{sen}(1-\text{sen}) + z^2_{1-\alpha/2}/(4n_1)]/n_1}}{1 + z^2_{1-\alpha/2}/n_1},$$

where sen is the estimate of sensitivity; $z^2_{1-\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution. The confidence interval for specificity is calculated in the same way. In this case with $\alpha = 0.95$, they are (0.934,0.992) for sensitivity and (0.045, 0.107) for specificity, respectively.

1.2.2 The likelihood ratio

There is another index for diagnostic accuracy called the likelihood ratio. The definition for positive result is

$$LR_P = \frac{P(F(x) \geq c|y = 1)}{P(F(x) \geq c|y = 0)},$$

and that for negative result is

$$LR_N = \frac{P(F(x) < c|y = 1)}{P(F(x) < c|y = 0)},$$

The likelihood ratio reflects the magnitude of the test's evidence indicating disease compared to non-disease. If we have $LR_P > 1$, then it means that positive results are more likely for diseased subjects than non-diseased subjects. On the other hand, if $LR_D < 1$, then negative results are more likely observed for non-diseased subjects than the others. Based on Table 1.1, we have

$$\begin{aligned} LR_P &= 127/251 = 0.51, \\ LR_N &= 3/19 = 0.16. \end{aligned}$$

Bayes' theorem gives us post-test probability called *positive predictive value* (PPV) and *negative predictive value* (NPV) as follows:

$$\begin{aligned} PPV &\equiv P(Y = 1|F(x) \geq c) = \frac{\text{sen} \times P(Y = 1)}{\text{sen} \times P(Y = 1) + (1 - \text{spe}) \times P(Y = 0)} \\ NPV &\equiv P(Y = 0|F(x) < c) = \frac{\text{spe} \times P(Y = 0)}{\text{spe} \times P(Y = 0) + (1 - \text{sen}) \times P(Y = 1)} \end{aligned}$$

They are interpreted as the probability of the subject with positive result to be diseased and the probability of the subject with negative result to be non-diseased. They are clinically meaningful; however, note that they are not measures of the intrinsic accuracy of the test because they include the prevalence rates $P(Y = 1)$ and $P(Y = 0)$. We can also calculate post-test odds from pre-test odds using likelihood ratio:

$$\begin{aligned} \frac{PPV}{1 - PPV} &= \frac{P(Y = 1)}{1 - P(Y = 1)} \times LR_P \\ \frac{NPV}{1 - NPV} &= \frac{P(Y = 0)}{1 - P(Y = 0)} \times 1/LR_N. \end{aligned}$$

Using the PSA data, we have

$$\begin{aligned}
\frac{PPV}{1-PPV} &= 130/270 \times 127/251 = 0.24 \\
\frac{NPV}{1-NPV} &= 270/130 \times 19/3 = 13.2
\end{aligned}$$

Chapter 2

Statistical methods in machine learning deriving from surrogates of the 0-1 objective function

In this chapter, we review several typical boosting methods that originate from approximation of the 0-1 objective function, and investigate the some statistical properties, including Bayes risk consistency. The Figure 2.1 illustrates the several surrogates of the 0-1 objective function. Note that the all functions but the normal cumulative function are convex, and this convexity leads to nice statistical properties (Lugosi and Vayatis, 2004; Bartlett *and others*, 2006). On the other hand, the properties of non-convex approximation function have yet to be investigated fully. In the next chapter, we investigate it and propose a new boosting method based on the result.

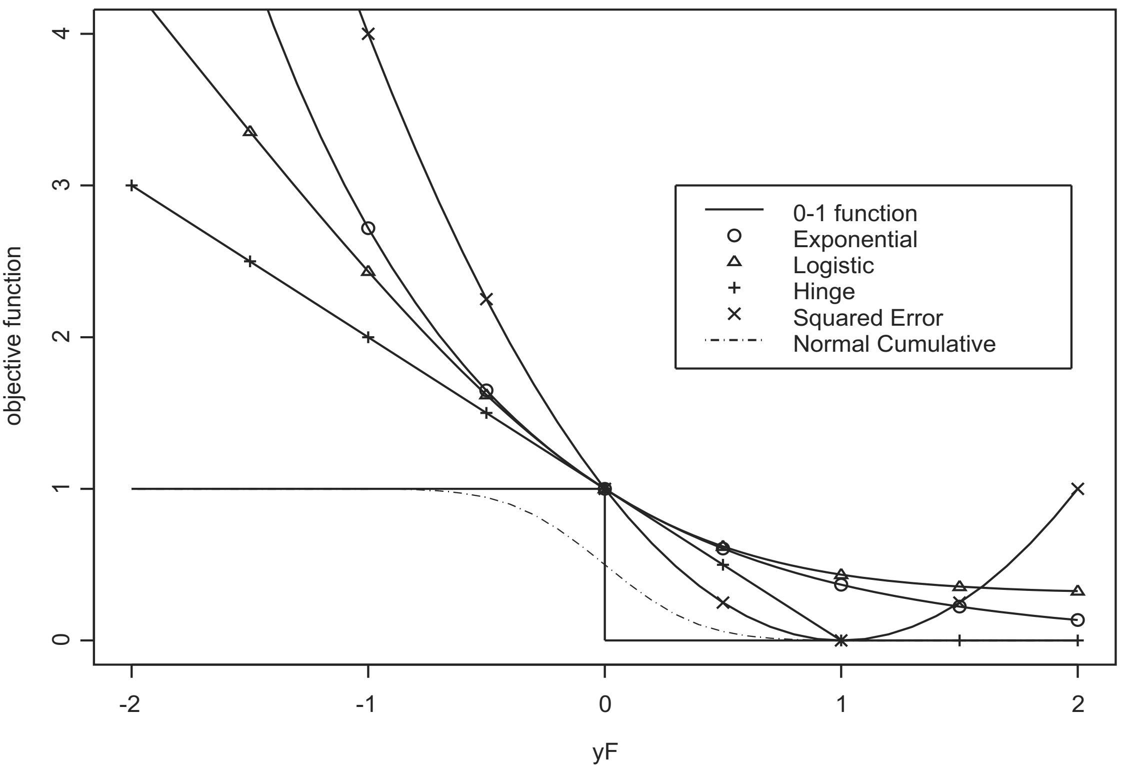


Figure 2.1: Plots of the 0-1 objective function and its various surrogates. The curve labeled "Exponential" is the exponential loss, $\exp(-yF)$; "Logistic" is the negative scaled binomial log-likelihood, $\log(1+\exp(-2yF))+1-\log(2)$; "Hinge" is the piecewise-linear loss in SVM, $(1-yF)_+$; "Squared Error" is $(y-F)^2 (=(1-yF)^2)$ and "Normal Cumulative" is the normal cumulative function with variance 1/10. All the functions are monotone in yF; All the surrogates except for "Normal Cumulative" are convex.

2.1 Typical methods

2.1.1 AdaBoost

AdaBoost was proposed by Freund and Schapire (1997), and has become the most popular boosting method in machine learning community. We assume that a sequence of n training examples $(\boldsymbol{x}_1, y_n), \ldots, (\boldsymbol{x}_n, y_n)$ is drawn randomly according to a distribution $\mathcal{P}$ on $\mathbf{R}^p \times \{0, 1\}$. Define D over the training examples, and this distribution is set to be uniform so that $D(i) = 1/n$ for $i = 1, \ldots, n$. The algorithm of AdaBoost is as follows.

1. Initialize the weight vector: $w_i^t = D(i)$
2. For $t = 1, ..., T$
   (a) Set
   $$\boldsymbol{p}^t = \frac{\boldsymbol{w}^t}{\sum_{i=1}^n w_i^t}, \tag{2.1.1}$$
   where, $\boldsymbol{p}$ and $\boldsymbol{w}$ are in $\mathbf{R}^n$.

   (b) Fit a weak classifier $f_t(\boldsymbol{x})$: $\mathbf{R}^p \to [0, 1]$, to the training data using weights w_i^t.

   (c) Compute the error of f_t
   $$\epsilon_t = \sum_{i=t}^n p_i^t |f_t(\boldsymbol{x}_i) - y_i| \tag{2.1.2}$$

   (d) Set $\beta_t = \epsilon_t/(1 - \epsilon)$

   (e) Set the new weights vector to be
   $$w_i^{t+1} = w_i^t \beta_t^{1-|h_t(x_i)-y_i|} \tag{2.1.3}$$
3. Finally, output a final score function F:
$$F(\boldsymbol{x}) = \begin{cases} 1, \text{ if } \sum_{t=1}^T (\log 1/\beta) f_t(\boldsymbol{x}) \geq 1/2 \sum_{t=1}^T \log 1/\beta \\ 0, \text{ otherwise.} \end{cases} \tag{2.1.4}$$

The next theorem gives the reason why AdaBoost performs well on the training data.

Theorem 2.1.1 (Freund and Schapire (1997))**.** *Given errors $\epsilon_1, \ldots, \epsilon_T$ in the algorithm of AdaBoost, the training error defined by $\epsilon = P_{i \sim D}(F(\boldsymbol{x}_i) \neq y_i)$ is bounded above by*

$$\epsilon \leq 2^T \prod_{t=1}^T \sqrt{\epsilon_t(1 - \epsilon_t)}. \tag{2.1.5}$$

For the details of the proof, see Freund and Schapire (1997). Since the value of ϵ_t can be taken to be smaller than 0.5 at every step t, the value of ϵ goes to 0 if we take T to infinity.

2.1.2 LogitBoost

Friedman *and others* (2000) showed that AdaBoost can be viewed to approximately maximize the Bernoulli log-likelihood, and derived a new boosting method, called LogitBoost, which aims to directly maximize the Bernoulli log-likelihood. Let $y \in (0, 1)$ be a class label and parametrize the binomial probabilities by

$$\begin{aligned} & \log \frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} = 2F(\boldsymbol{x}) \\ \Leftrightarrow \quad & p(\boldsymbol{x}) = \frac{\exp^{F(\boldsymbol{x})}}{\exp^{F(\boldsymbol{x})} + \exp^{-F(\boldsymbol{x})}}. \end{aligned}$$

Then the binomial log-likelihood is

$$\begin{aligned} l(y, p(\boldsymbol{x})) &= y \log(p(\boldsymbol{x})) + (1 - y) \log(1 - p(\boldsymbol{x})) \\ &= -\log(1 + \exp^{-2y_s F(\boldsymbol{x})}) \\ \Big(&= 2y - \log(1 + \exp^{2F(\boldsymbol{x})})\Big), \end{aligned}$$

where $y_s = 2y - 1 \in (-1, 1)$. Hence, the maximization of the likelihood is equivalent to the minimization of the exponential loss, $\exp^{-y_s F(\boldsymbol{x})}$. The update process of LogitBoost is based on Newton-Raphson method. Let $f(\boldsymbol{x})$ is a weak classifier used for updating, then define the expected log-likelihood:

$$El(F + f) = E\big[2y(F(\boldsymbol{x}) + f(\boldsymbol{x})) - \log(1 + \exp^{2F(\boldsymbol{x}) + 2f(\boldsymbol{x})})\big].$$

The first and second derivative at $f(\boldsymbol{x}) = 0$ are

$$\begin{aligned}
s(\boldsymbol{x}) &= \left.\frac{\partial El(F(\boldsymbol{x}) + f(\boldsymbol{x}))}{\partial f(\boldsymbol{x})}\right|_{f(\boldsymbol{x})=0} \\
&= 2E\left.\left[y - \frac{\exp^{F(\boldsymbol{x})+f(\boldsymbol{x})}}{\exp^{F(\boldsymbol{x})+f(\boldsymbol{x})} + \exp^{-F(\boldsymbol{x})-f(\boldsymbol{x})}}\right]\right|_{f(\boldsymbol{x})=0} \\
&= 2E(y - p(\boldsymbol{x})) \\
H(\boldsymbol{x}) &= \left.\frac{\partial^2 El(F(\boldsymbol{x}) + f(\boldsymbol{x}))}{\partial f(\boldsymbol{x})^2}\right|_{f(\boldsymbol{x})=0} \\
&= -4E\left.\left[\frac{\exp^{F(\boldsymbol{x})+f(\boldsymbol{x})}\exp^{-F(\boldsymbol{x})-f(\boldsymbol{x})}}{(\exp^{F(\boldsymbol{x})+f(\boldsymbol{x})} + \exp^{-F()-f(\boldsymbol{x})})^2}\right]\right|_{f(\boldsymbol{x})=0} \\
&= -4E[p(\boldsymbol{x})(1 - p(\boldsymbol{x}))].
\end{aligned}$$

Hence, the updated score function $F(\boldsymbol{x})$ has the form:

$$\begin{aligned}
F(\boldsymbol{x})_{new} &= F(\boldsymbol{x}) - H(\boldsymbol{x})^{-1}s(\boldsymbol{x}) \\
&= F(\boldsymbol{x}) + \frac{E(y - p(\boldsymbol{x}))}{2E[p(\boldsymbol{x})(1 - p(\boldsymbol{x}))]}
\end{aligned}$$

So, we choose a weak classifier among a predetermined set of weak classifiers that satisfy

$$\min_{f(\boldsymbol{x})} E_w\left(\frac{y - p(\boldsymbol{x})}{2p(\boldsymbol{x})(1 - p(\boldsymbol{x}))} - f(\boldsymbol{x})\right)^2,$$

where $w(x) = p(\boldsymbol{x})(1 - p(\boldsymbol{x}))$ and

$$E_w[\cdot] \equiv \frac{E[w(x)\cdot]}{E(w(x))}.$$

Note that the absolute value of the coefficient for $f(\boldsymbol{x})$ is 1, so it can be regarded as one of ϵ-Boost proposed by Rosset *and others* (2004), in which they recommend a very small value of ϵ for the coefficient rather than the one that is determined by greedy line-search as implemented in AdaBoost.

2.1.3 GAMBoost

Tutz and Binder (2006) proposed a boosting method that extends the general additive model (GAM) to the one that can work well in high-dimensional data setting. It works for all simple exponential family distributions, including binomial, Poisson and normal response variables $(y_1, y_2, \ldots, y_n)$. That is, they consider the following probability density function:

$$g(y_i, \eta_i) = \exp\left[(y_i\eta_i - b(\eta_i))/\phi + c(y_i, \phi)\right],\ i = 1, 2, \ldots, n, \tag{2.1.6}$$

where $y_i \in \mathbf{R}$ is a response variable, not a class label; η_i is the natural (or canonical) parameter and ϕ is a dispersion parameter. Note that

$$\begin{aligned} E(Y)(=\mu) &= \frac{\partial b(\eta)}{\partial \eta}(=h(\eta)) \\ Var(Y)(=\sigma^2) &= \phi\frac{\partial^2 b(\eta)}{\partial \eta^2} = \phi\frac{\partial \mu}{\partial \eta}. \end{aligned}$$

Here, we define a function called a canonical (natural) link:

$$\nu(\mu) \equiv h^{-1}(\mu) = \eta.$$

We call η the natural parameter because it is related naturally to the response variable y in Equation (2.1.6). Tutz and Binder (2006) fitted basis functions of the B-splines to the mean of the j-th marker $(j = 1, 2, \ldots, p)$ in the t-th step of the boosting method:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} h\Big(\hat{\eta}_t(x_{1j}) + \{B_1^{(j)}(x_{1j}), \ldots, B_M^{(j)}(x_{1j})\}\boldsymbol{\gamma}\Big) \\ \vdots \\ h\Big(\hat{\eta}_t(x_{nj}) + \{B_1^{(j)}(x_{1j}), \ldots, B_M^{(j)}(x_{nj})\}\boldsymbol{\gamma}\Big) \end{bmatrix} \tag{2.1.7}$$

$$= \begin{bmatrix} h\Big(\hat{\eta}_t(x_{1j}) + \boldsymbol{z}'_{1j}\boldsymbol{\gamma}\Big) \\ \vdots \\ h\Big(\hat{\eta}_t(x_{nj}) + \boldsymbol{z}'_{nj}\boldsymbol{\gamma}\Big) \end{bmatrix} = \begin{bmatrix} h(\eta_1) \\ \vdots \\ h(\eta_n) \end{bmatrix} \tag{2.1.8}$$

where $\hat{\eta}_t$ is an estimator that is estimated until the t-th step; $\boldsymbol{z}'_{ij} = (B_1^{(j)}(x_{ij}), \ldots, B_M^{(j)})$ is a set of the B-spline basis functions for the j-th element of a marker vector $\boldsymbol{x} \in \mathbf{R}^p$; $\boldsymbol{\gamma}$ is a M dimensional coefficient vector for the B-spline. The log-likelihood to be maximized is given by

$$\begin{aligned} l(\boldsymbol{\gamma}) &= \sum_{i=1}^{n} \log g(y_i, \eta_i) \\ &= \sum_{i=1}^{n} (y_i\eta_i - b(\eta_i))/\phi + c(y_i, \phi). \end{aligned}$$

Hench, the penalized log-likelihood is given as

$$l_p(\boldsymbol{\gamma}) = l(\boldsymbol{\gamma}) - \frac{\lambda}{2}\boldsymbol{\gamma}'\boldsymbol{\Lambda}\boldsymbol{\gamma},$$

where $\boldsymbol{\Lambda}$ is a penalty matrix constructed such that $\boldsymbol{\gamma}'\boldsymbol{\Lambda}\boldsymbol{\gamma}$ penalizes first-order differences $\sum_{k=1}^{M-1}(\gamma_{k+1} - \gamma_k)^2$ or higher order differences of parameters, which correspond to basis functions of adjacent knots. The penalized score function is

$$\begin{aligned} s_p(\boldsymbol{\gamma}) = \frac{\partial l_p(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{n} \frac{\partial l(\boldsymbol{\gamma})}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\gamma}} - \lambda\Lambda\boldsymbol{\gamma} \\ &= \sum_{i=1}^{n} \frac{y_i - b'(\eta_i)}{\phi}\boldsymbol{z}_{ij} - \lambda\Lambda\boldsymbol{\gamma} \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{Var(y_i)}\frac{\partial \mu_i}{\partial \eta_i}\boldsymbol{z}_{ij} - \lambda\Lambda\boldsymbol{\gamma} \\ &= \boldsymbol{Z}'_j D(\boldsymbol{\gamma})\boldsymbol{\Sigma}(\boldsymbol{\gamma})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - \lambda\Lambda\boldsymbol{\gamma}, \end{aligned}$$

where $\boldsymbol{Z}'_j = (\boldsymbol{z}_{1j}, \ldots, \boldsymbol{z}_{nj})$; $D(\boldsymbol{\gamma}) = \mathrm{diag}(\partial\mu_1/\partial\eta_1, \ldots, \partial\mu_\mathrm{n}/\partial\eta_\mathrm{n})$ is the variance function that connects $E(Y)$ to $Var(Y)$ using ϕ; $\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_\mathrm{n}^2)$. With the weight function $W(\boldsymbol{\gamma}) = D(\boldsymbol{\gamma})\boldsymbol{\Sigma}(\boldsymbol{\gamma})^{-1}D(\boldsymbol{\gamma})$, it is rewritten as

$$s_p(\boldsymbol{\gamma}) = \boldsymbol{Z}'_j W(\boldsymbol{\gamma})D(\boldsymbol{\gamma})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - \lambda\Lambda\boldsymbol{\gamma}.$$

The penalized Fisher matrix (the mean Hessian matrix) is

$$\begin{aligned} F_p(\boldsymbol{\gamma}) &= E\Big(-\frac{\partial^2 l_p(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\Big) \\ &= E\Big(-\sum_{i=1}^{n}\frac{-b''(\eta_i)}{\phi}\boldsymbol{z}_{ij}\boldsymbol{z}_{ij}' + \lambda\Lambda\Big) \\ &= \sum_{i=1}^{n} E\Big(\frac{(\partial\mu_i/\partial\eta_i)^2}{Var(y_i)}\boldsymbol{z}_{ij}\boldsymbol{z}_{ij}'\Big) + \lambda\Lambda \\ &= \boldsymbol{Z}_j' W(\boldsymbol{\gamma})\boldsymbol{Z}_j + \lambda\Lambda \end{aligned}$$

Hence, Fisher scoring is given by

$$\hat{\boldsymbol{\gamma}}_{new} = \hat{\boldsymbol{\gamma}} + F_p(\hat{\boldsymbol{\gamma}})^{-1}s_p(\hat{\gamma})$$

So, GAMBoost is different from the method of *iterative reweighted least squares* (IRLS), because it uses only the Newton-Raphson method. That is, the process of the least square approach is not included in GAMBoost. Moreover they actually update the coefficient vector as

$$\hat{\boldsymbol{\gamma}}_{new} = F_p(\mathbf{0})^{-1}s_p(\mathbf{0}).$$

This is because in a boosting algorithm, we add the updated coefficient to the already fitted value; hence, we take $\hat{\boldsymbol{\gamma}}$ to be 0 in each boosting step. As a result, the weak classifier that consists of a set of the B-spline basis function is calculated as

$$f_{j,new} = \boldsymbol{Z}_j\hat{\boldsymbol{\gamma}}_{new}, \qquad j = 1, \ldots, p.$$

Then, set $f_j = f_{old,j} + f_{j,new}$ yielding $\hat{\eta}_{j,new}$. The best j is selected among $\{1, \ldots, p\}$ based on the likelihood, and the j-th component of the score function is updated. This process is iterated in GAMBoost.

2.1.4 SVM

Define a hyperplane by

$$\{\boldsymbol{x}:\ f(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x} + \beta_0 = 0\}.$$

The unit vector normal to the plane is

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}/||\boldsymbol{\beta}||,$$

because $\boldsymbol{\beta}'(\boldsymbol{x}_1 - \boldsymbol{x}_2) = 0$ for any two points $\boldsymbol{x}_1, \boldsymbol{x}_2$ lying in the plane. With any point $\boldsymbol{x}_0$ in the plane, the signed distance of a $\boldsymbol{x}$ is

$$\begin{aligned} \boldsymbol{\beta}^{*\prime}(\boldsymbol{x} - \boldsymbol{x}_0) &= \frac{1}{||\boldsymbol{\beta}||}(\boldsymbol{\beta}'\boldsymbol{x} - \boldsymbol{\beta}'\boldsymbol{x}_0) \\ &= \frac{1}{||\boldsymbol{\beta}||}(\boldsymbol{\beta}'\boldsymbol{x} + \boldsymbol{\beta}_0) \end{aligned}$$

In this setting, consider a optimization problem:

$$\begin{gathered} \max_{\boldsymbol{\beta},\beta_0} C \\ \text{subject to } y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \beta_0)/||\boldsymbol{\beta}|| > C,\ i = 1, \dots, n, \end{gathered}$$

where, $y_i \in \{-1, 1\}$ is a class label; n is a sample size. Note that we can keep $||\boldsymbol{\beta}|| = 1/C$ without loss of generality in the maximization process, because the hyperplane is invariant to the scale constrain. Hence, it can be rewritten as

$$\begin{gathered} \max_{\boldsymbol{\beta},\beta_0} \tfrac{1}{||\beta||}\left(= \min_{\boldsymbol{\beta},\beta_0} ||\beta||\right) \\ \text{subject to } y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \beta_0) > 1,\ i = 1, \dots, n. \end{gathered}$$

In more general setting, we consider the slack variables $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ to relax the constraint condition as follows.

$$\begin{gathered} \min_{\boldsymbol{\beta},\beta_0} ||\beta||^2 \\ \text{subject to } y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \beta_0) > 1 - \xi_i,\ \xi_i \geq 0,\ \sum \xi_i \leq \text{constant},\ i = 1, \dots, n. \end{gathered} \tag{2.1.9}$$

The corresponding Lagrange primal function is

$$L_P = \left\{\frac{1}{2}||\boldsymbol{\beta}||^2 + \gamma \sum_{i=1}^{n} \xi_i\right\} + \sum_{i=1}^{n} \alpha_i\{(1-\xi_i) - y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \beta_0)\} + \sum_{i=1}^{n} \mu_i(-\xi_i). \qquad (2.1.10)$$

The necessary condition for the existence of a local minimum of (2.1.10) (Karush-Kuhn-Tucker condition) is there exist constants α_i and μ_i $(i = 1, \ldots, n)$ such that

- Stationarity

$$\frac{\partial L_P}{\partial \boldsymbol{\zeta}} = \mathbf{0},$$
$$\Leftrightarrow \quad \begin{cases} \boldsymbol{\beta} & = \sum_{i}^{n} \alpha_i y_i \boldsymbol{x}_i \\ 0 & = \sum_{i=1}^{n} \alpha_i y_i \\ \alpha_i & = \gamma - \mu_i, \qquad i = 1, \ldots, n. \end{cases}$$

  where $\boldsymbol{\zeta}' = (\boldsymbol{\beta}', \beta_0, \boldsymbol{\xi}')$.

- Primal feasibility

$$y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \beta_0) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, \qquad i = 1, \ldots, n.$$

- Dual feasibility

$$\alpha_i \geq 0,$$
$$\mu_i \geq 0, \qquad i = 1, \ldots, n.$$

- Complementary slackness

$$\alpha_i\{(1-\xi_i) - y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \beta_0)\} = 0$$
$$\mu_i\xi_i = 0, \qquad i = 1, \ldots, n.$$

The Lagrangian dual objective function to be maximized is

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i'\boldsymbol{x}_j,$$

which gives a lower bound on the objective function (2.1.9). The standard software is available for this simple form of the convex optimization problem. The conditions above uniquely characterize the solution to the primal and dual problem. From the condition of $\boldsymbol{\beta}$ in Stationary condition, the solution for $\boldsymbol{\beta}$ has the form

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{x}_i,$$

where α_i that is non-zero must satisfy the first equation exactly in Primal feasibility condition. That is,

$$y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \beta_0) = 1 - \xi_i.$$

Hence, the important samples that are used to determine the solution to (2.1.9) are on the boundary of classification. They are called the *support vectors.* The objective function (2.1.9) of SVM can be rewritten as

$$\min_{\boldsymbol{\beta},\beta_0} ||\beta||^2 + \gamma\sum_{i=1}^{n} \xi_i$$
$$\text{subject to } y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \beta_0) > 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, \ldots, n. \tag{2.1.11}$$

The two constraints above can be summarized into

$$\xi_i \geq \max\{0, 1 - y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \beta_0)\}$$

This means that the minimum of ξ_i is $\max\{0, 1 - y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \beta_0)\}$. Hence, the minimization statement of (2.1.11) is rephrased as

$$\min_{\boldsymbol{\beta},\beta_0}\left[\sum_{i=1}^{n} \max\{0, 1 - y_i(\boldsymbol{\beta}'\boldsymbol{x}_i + \beta_0)\} + \frac{1}{\gamma}||\beta||^2\right], \tag{2.1.12}$$

where the first term is called *Hinge loss*.

2.1.5 RankBoost

RankBoost (Freund *and others*, 2003) is a well known boosting algorithm for ranking problems. In this subsection we make clear the difference between AUCBoost and RankBoost. In particular we focus on each objective function and show the two boosting methods have different optimal discriminant function.

In general each objective function can be regarded as one of the objective function for ranking (R_U):

$$\mathrm{R}_U(F) = \int\int U(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0 d\boldsymbol{x}_1,$$

where U is a function we choose on our own. If we take a Heaviside function as U, then it becomes AUC and if $U(x) = \exp(-x)$, then it becomes the objective function of RankBoost.

Theorem 2.1.2. *Let U be a convex function with negative derivative U'. Then the function that minimizes R_U is written as:*

$$F = m\left(\frac{g_1}{g_0}\right),$$

where m is a monotonically increasing function.

Proof. For $F_\epsilon = F + \epsilon\,\eta$

$$\begin{aligned}
& \frac{\partial}{\partial\epsilon}\mathrm{R}_U(F_\epsilon)\bigg|_{\epsilon=0} \\
= & \int\int \big(\eta(\boldsymbol{x}_1) - \eta(\boldsymbol{x}_0)\big)U'\big(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)\big)g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\
= & \int\int \eta(\boldsymbol{x})U'\big(F(\boldsymbol{x}) - F(\boldsymbol{y})\big)g_0(\boldsymbol{y})g_1(\boldsymbol{x})d\boldsymbol{y}d\boldsymbol{x} \\
& \qquad - \int\int \eta(\boldsymbol{x})U'\big(F(\boldsymbol{y}) - F(\boldsymbol{x})\big)g_0(\boldsymbol{x})g_1(\boldsymbol{y})d\boldsymbol{x}d\boldsymbol{y} \\
= & \int \eta(\boldsymbol{x})\Big[g_1(\boldsymbol{x})\int U'\big(F(\boldsymbol{x}) - F(\boldsymbol{y})\big)g_0(\boldsymbol{y})d\boldsymbol{y} - g_0(\boldsymbol{x})\int U'\big(F(\boldsymbol{y}) - F(\boldsymbol{x})\big)g_0(\boldsymbol{y})d\boldsymbol{y}\Big]d\boldsymbol{x} \\
= & \ 0.
\end{aligned}$$

Because η is arbitrary, we have

$$\frac{\int U'\big(F(\boldsymbol{y})-F(\boldsymbol{x})\big)g_1(\boldsymbol{y})d\boldsymbol{y}}{\int U'\big(F(\boldsymbol{x})-F(\boldsymbol{y})\big)g_0(\boldsymbol{y})d\boldsymbol{y}}=\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}, \tag{2.1.13}$$

and we define

$$\psi\big(F(\boldsymbol{x})\big)=\frac{\int U'\big(F(\boldsymbol{y})-F(\boldsymbol{x})\big)g_1(\boldsymbol{y})d\boldsymbol{y}}{\int U'\big(F(\boldsymbol{x})-F(\boldsymbol{y})\big)g_0(\boldsymbol{y})d\boldsymbol{y}}.$$

Hence we have

$$\begin{aligned}\frac{\partial\psi\big(F(\boldsymbol{x})\big)}{\partial F(\boldsymbol{x})} &= -\frac{\int U''\big(F(\boldsymbol{y})-F(\boldsymbol{x})\big)g_1(\boldsymbol{y})d\boldsymbol{y}\int U'\big(F(\boldsymbol{x})-F(\boldsymbol{y})\big)g_0(\boldsymbol{y})d\boldsymbol{y}}{\big\{\int U'\big(F(\boldsymbol{x})-F(\boldsymbol{y})\big)g_0(\boldsymbol{y})d\boldsymbol{y}\big\}^2}\\ &\quad -\frac{\int U'\big(F(\boldsymbol{y})-F(\boldsymbol{x})\big)g_1(\boldsymbol{y})d\boldsymbol{y}\int U''\big(F(\boldsymbol{x})-F(\boldsymbol{y})\big)g_0(\boldsymbol{y})d\boldsymbol{y}}{\big\{\int U'\big(F(\boldsymbol{x})-F(\boldsymbol{y})\big)g_0(\boldsymbol{y})d\boldsymbol{y}\big\}^2}\\ &> 0.\end{aligned}$$

So a monotonically increasing function m exists such that

$$F=m\left(\frac{g_1}{g_0}\right).$$

□

Corollary 2.1.1. *The optimal function for RankBoost is written as:*

$$\operatorname*{argmin}_{F\in\mathcal{F}}\mathrm{R}_U(F)=\frac{1}{2}\log\frac{g_1}{g_0}+c,$$

where c is an arbitrary constant and $U(x)=\exp(-x)$.

Proof. From (2.1.13) in Theorem 2.1.2 we have

$$\frac{\int\exp\big(F(\boldsymbol{x})-F(\boldsymbol{y})\big)g_1(\boldsymbol{y})d\boldsymbol{y}}{\int\exp\big(F(\boldsymbol{y})-F(\boldsymbol{x})\big)g_0(\boldsymbol{y})d\boldsymbol{y}}=\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})},$$

and it is equivalent to

$$F(\boldsymbol{x})=\frac{1}{2}\log\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}+\frac{1}{2}\log\frac{\int\exp\big(F(\boldsymbol{y})\big)g_0(\boldsymbol{y})d\boldsymbol{y}}{\int\exp\big(-F(\boldsymbol{y})\big)g_1(\boldsymbol{y})\boldsymbol{y}}.$$

Hence we have

$$F(\boldsymbol{x}) = \frac{1}{2}\log\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})} + c, \tag{2.1.14}$$

where c is an arbitrary constant. □

As a result of Corollary 2.1.1, we see that RankBoost also maximize the area under the ROC curve (AUC), because the optimal discriminant function for RankBoost is a special case of that for AUCBoost in (3.7.6). And it is worth noting that the optimal discriminant function for RankBoost is much similar to that for AdaBoost, because

$$F_{\text{Ada}} = \frac{1}{2}\log\frac{g_1}{g_0} + \frac{1}{2}\log\frac{\pi_1}{\pi_0},$$

where π_0 and π_1 are the prior probability of the population 0 and the population 1, respectively. Hence RankBoost is almost the same as AdaBoost.

2.2 Bayes risk consistency for convex loss functions

The most important property of score function $F(\boldsymbol{x})$ is that a score function optimizing a given objective function must satisfy *Bayes-risk consistency.* We review a theorem proven by (Lugosi and Vayatis, 2004) that shows *Bayes-risk consistency* of convex cost functions under some assumptions.

Consider a class of score functions $F : \mathcal{X} \to [-1, 1]$:

$$\mathcal{F} = \left\{ F(\boldsymbol{x}) = \sum_{i=1}^{N} w_i f_i(\boldsymbol{x}) : N \in \mathbf{N}, w_1, \ldots, w_N \geq 0, \sum_{i=1}^{N} = 1 \right\},$$

which is the convex hull of $\mathcal{C}$: a class of weak classifiers $f(\boldsymbol{x}) \in \{-1, 1\}$'s. Denote the Bayes risk by L^* and define as follows:

$$L^* = \inf_{F} P(\text{sgn}(F(\boldsymbol{X})) \neq Y), \tag{2.2.1}$$

where Y is a class label taking values of $\{-1, 1\}$, and $\operatorname{sgn}(z)$ is a function defined by

$$\operatorname{sgn}(z) = \begin{cases} 1, \text{ if } z > 0 \\ -1, \text{ otherwise.} \end{cases} \tag{2.2.2}$$

Note that the formal definition of sign function is given by

$$\operatorname{sgn}^*(z) = \begin{cases} 1, \text{ if } z > 0 \\ 0, \text{ if } z = 0 \\ -1, \text{ otherwise.} \end{cases} \tag{2.2.3}$$

The loss function L is expressed using the indicator function $I(\cdot)$, as

$$\begin{aligned} L(F) &\equiv P(\operatorname{sgn}(F(\boldsymbol{X})) \neq Y) \\ &= \int\int I(\operatorname{sgn}(F(\boldsymbol{x})) \neq y) p(\boldsymbol{x}, y) d\boldsymbol{x} dy \\ &= \int\int I(\operatorname{sgn}(F(\boldsymbol{x})) \neq y) p(y) p(\boldsymbol{x}|y) d\boldsymbol{x} dy \\ &= \int \Big\{\pi_{-1} I(\operatorname{sgn}(F(\boldsymbol{x})) \neq -1) p_{-1}(\boldsymbol{x}) + \pi_1 I(\operatorname{sgn}(F(\boldsymbol{x})) \neq 1) p_1(\boldsymbol{x})\Big\} d\boldsymbol{x} \\ &= \int \Big\{(1 - \eta(x)) I(\operatorname{sgn}(F(\boldsymbol{x})) = 1) + \eta(\boldsymbol{x}) I(\operatorname{sgn}(F(\boldsymbol{x})) = -1)\Big\} p(\boldsymbol{x}) d\boldsymbol{x} \\ &= E\Big[(1 - \eta(\boldsymbol{x})) I(\operatorname{sgn}(F(\boldsymbol{x})) = 1) + \eta(\boldsymbol{x}) I(\operatorname{sgn}(F(\boldsymbol{x})) = -1)\Big] \end{aligned}$$

where $p_1(\boldsymbol{x}) = p(\boldsymbol{x}|y = 1)$, $p_{-1}(\boldsymbol{x}) = p(\boldsymbol{x}|y = -1)$, $p(\boldsymbol{x}) = \pi_1 p_1(\boldsymbol{x}) + \pi_{-1} p_{-1}(\boldsymbol{x})$ and

$$\eta(\boldsymbol{x}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_1 p_1(\boldsymbol{x})}{\pi_1 p_1(\boldsymbol{x}) + \pi_{-1} p_{-1}(\boldsymbol{x})}. \tag{2.2.4}$$

Hence, we find that the Bayes classifier

$$I(\eta(\boldsymbol{x}) > 1/2) - I(\eta(\boldsymbol{x}) \leq 1/2) \tag{2.2.5}$$

minimizes the loss function:

$$L^*(F) = L(F_B) = E[\min(\eta(\boldsymbol{X}), 1 - \eta(\boldsymbol{X}))].$$

Instead of minimizing $L(F)$ itself, Lugosi and Vayatis (2004) consider an appropriate smooth loss functional to simultaneously avoid overfitting and become computationally feasible in may cases:

$$A(F) = \int\int \phi(-F(\boldsymbol{x})y)p(\boldsymbol{x}, y)d\boldsymbol{x}dy,$$

and the empirical loss

$$A_n(F) = \frac{1}{n}\sum_{i=1}^{n} \phi(-F(\boldsymbol{x})y),$$

where $\phi : [-1, 1] \to \mathbf{R}^+$ is a positive nondecreasing *convex* function such that $\phi(0) = 1$, and the estimator $\hat{F}_n$ minimizes the empirical quantity $A_n(F)$

Assumption 2.2.1. *Let ϕ be a differentiable strictly convex, strictly increasing cost function such that $\phi(0) = 1$,$\lim_{x\to-\infty} = 0$.*

Theorem 2.2.1 (Lugosi and Vayatis (2004))**.** *Assume that the cost function ϕ satisfies Assumption 2.2.1 and that the distribution of $(\boldsymbol{X}, Y)$ and the class $\mathcal{C}$ are such that*

$$\lim_{\lambda\to\infty} \inf_{F\in\lambda\mathcal{F}} A(F) = A^*,$$

where $A^* = \inf A(F)$ over all measurable functions $F : \mathcal{X} \to \mathbf{R}$. Assume that $\mathcal{C}$ has a finite VC dimension.

Let $\lambda_1, \lambda_2, \ldots$ be a sequence of positive numbers satisfying

$$\lambda_n \to \infty and \lambda_n \phi'(\lambda_n)\sqrt{\frac{\ln n}{n}} \to 0, \quad as\ n \to \infty,$$

where $\ln$ *is the logarithm natural and define the estimator $F_n = \frac{1}{\lambda_n}\hat{F}_n \in \mathcal{F}$. Then* $\operatorname{sgn}(F_n(\boldsymbol{x}))$ *is strongly Bayes-risk consistent, that is,*

$$\lim_{n\to\infty} L(\operatorname{sgn}(F_n)) = L^*, \quad almostsurely.$$

Example 2.2.1. *The exponential loss $\phi(z) = \exp(z)$ of AdaBoost satisfies Assumption*

2.2.1, and therefore the Bayes-risk consistency holds. The optimal socre function is

$$F_{\mathrm{Ada}}(\boldsymbol{x}) = \frac{1}{2} \ln \frac{\eta(\boldsymbol{x})}{1-\eta(\boldsymbol{x})}.$$

Example 2.2.2. *Friedman and others (2000) proposed LogitBoost, where* $\phi(\boldsymbol{z}) = \mathrm{logit}(z) = \log_2(1+\exp(z))$. *This case also satisfies Assumption 2.2.1, so the Bayes-risk consistency holds.*

$$F_{\mathrm{Logit}}(\boldsymbol{x}) = \ln \frac{\eta(\boldsymbol{x})}{1-\eta(\boldsymbol{x})}.$$

Chapter 3

A boosting method for maximization of the are under the ROC curve

Abstract

We discuss receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) for binary classification problems in clinical fields. We propose a statistical method for combining multiple feature variables, based on a boosting algorithm for maximization of the AUC. In this iterative procedure, various simple classifiers that consist of the feature variables are combined flexibly into a single strong classifier. We consider a regularization to prevent overfitting to data in the algorithm using a penalty term for non-smoothness. This regularization method not only improves the classification performance but also helps us to get a clearer understanding about how each feature variable is related to the binary outcome variable. We demonstrate the usefulness of score plots constructed componentwise by the boosting method. We describe two simulation studies and a real data analysis in order to illustrate the utility of our method.

Keywords: AUC; Boosting; Classification; ROC curve; Smoothing.

3.1 Introduction

The receiver operating characteristic (ROC) curve has been widely used in medical and biological sciences (Zhou *and others*, 2002; Pepe, 2003), for applications in which the classification performance can be measured by the area under the ROC curve (AUC). This curve has three primary appealing properties. First, it does not assume any specific distributional model, so a method based on the ROC is distribution-free, in contrast to logistic regression analysis or classical linear discriminant analysis under normality assumption. Second, it is independent of the prior probabilities of group membership, so it is able to accommodate case-control studies. Third, the AUC is not influenced by the choice of thresholds that may be changed according to each decision-maker's objective; hence, the AUC expresses the intrinsic accuracy of classification performance. The advantages of the AUC over the odds ratio or relative risk when evaluating the classification performance are discussed by Pepe *and others* (2004).

A procedure for maximizing the AUC using a linear combination of multiple feature variables has been proposed (Pepe and Thompson, 2000) in order to improve on diagnostic accuracy of a single feature variable, and Pepe *and others* (2006) have shown that the AUC-based method can be far superior to logistic regression in certain situations. Ma and Huang (2005) extended this strategy to high-dimensional data by adopting a sigmoid approximation for the AUC. The assumption of linearity gives us easily interpretable results of the analysis, and allows us to get the rough characteristics of each feature variable. However, this strict assumption is often unable to capture informative nonlinear structures in the real world.

Moreover, it has been proved that the optimal combination of feature variables that maximizes the AUC is constructed based on the likelihood ratio (Eguchi and Copas, 2002; McIntosh and Pepe, 2002). This implies that even under a simple setting such as a normality assumption with unequal covariance matrices, the optimal combination is not linear but quadratic. Further details are described in Subsection 4.2.

In this paper, we propose a new statistical method to detect a more essential association between feature variables and a binary outcome variable using a boosting technique, and apply the method to the combination of the feature variables for better classification. A typical one of the boosting methods is AdaBoost (Freund and Schapire, 1997), which is designed to minimize the exponential loss. An AdaBoost-based boosting method for the AUC is presented by Long and Servedio (2007), along with its theoretical justification. The purpose of boosting methods is to construct a strong classifier by combining various weak classifiers. Recently, a variety of loss functions other than the exponential loss have been proposed and discussed in several contexts (Murata *and others*, 2004).

On the other hand, the generalized additive model (GAM) proposed by Hastie and Tibshirani (1986) has wide applications in a variety of research fields. This is mainly because this model can detect the nonlinear effects of feature variables on the objective function flexibly, without sacrificing interpretability:

$$\eta(E(y|\boldsymbol{x})) = F_1(x_1) + \cdots + F_p(x_p),$$

where $\boldsymbol{x} = (x_1, \ldots, x_p)'$, η is a link function and F_k, $k = 1, \ldots, p$, are unspecified functions of x_k. Thus, GAM is also well suited for binary classifications in medical and biological fields, in which the association of the feature vector $\boldsymbol{x}$ with an outcome variable y is of great interest. We consider a model, similar to GAM, that attaches importance to interpretability as well as flexibility, maximizing the AUC for a score function $F(\boldsymbol{x})$ by a boosting algorithm. As a result, we obtain $F(\boldsymbol{x})$ of the form

$$F(\boldsymbol{x}) = F_1(x_1) + \cdots + F_p(x_p),$$

in which we consider score plots of $F_k(x_k)$ against the k-th feature variable x_k. These plots are useful in association studies, for looking at how each feature variable works in the classification and for detecting which feature variable is the most effective one.

This paper is organized as follows. In Section 2, we give a brief review of the ROC curve and discuss the relationship between the AUC and the approximate AUC. In Section 3, we propose AUCBoost, a new boosting method based on the maximization of the AUC. In Section 4 we present two simple simulation studies to investigate the efficiency of AUCBoost, and in Section 5 we demonstrate the application of AUCBoost to a real data set. We close

Section 6 with concluding remarks and ideas for future work.

3.2 Receiver operating characteristic curve

3.2.1 Area under the ROC curve

Let y be a binary class label (y=0, 1), $\boldsymbol{x} \in \mathbf{R}^p$ be a feature vector, and $g_0(\boldsymbol{x})$, $g_1(\boldsymbol{x})$ be probability density functions for each class. We classify a subject with feature vector $\boldsymbol{x}$ into class 1 if a score function $F(\boldsymbol{x})$ is greater than or equal to a threshold value c, and into class 0 otherwise. Then, the false positive rate (FPR) and true positive rate (TPR) are defined as

$$\text{FPR}(c) = \int_{F(\boldsymbol{x}) \geq c} g_0(\boldsymbol{x}) d\boldsymbol{x}, \text{and } \text{TPR}(c) = \int_{F(\boldsymbol{x}) \geq c} g_1(\boldsymbol{x}) d\boldsymbol{x}. \tag{3.2.1}$$

By pairing these probabilities, the ROC curve is given as

$$\text{ROC} = \{(\text{FPR}(c), \text{TPR}(c)) \mid c \in \mathbf{R}\},$$

which is illustrated in Figure 3.1. From (3.2.1), the area under the ROC curve (AUC) is written as

$$\text{AUC}(F) = \int_{\infty}^{-\infty} \text{TPR}(c) d\text{FPR}(c). \tag{3.2.2}$$

The large separation of $g_0(\boldsymbol{x})$ and $g_1(\boldsymbol{x})$ could make the AUC close to 1. However, note that it is also dependent on a score function $F(\boldsymbol{x})$, which we must determine in the analysis of data. Only after employing an adequate $F(\boldsymbol{x})$ for the two probability density functions can we obtain the best value of the AUC. Equation (3.2.2) can be expressed in another manner:

$$\text{AUC}(F) = P(F(\boldsymbol{X}_1) \geq F(\boldsymbol{X}_0)),$$

where $\boldsymbol{X}_0$, $\boldsymbol{X}_1$ are independent p-dimensional random vectors from class 0 and class 1, respectively (Bamber, 1975). The empirical AUC for given observations $\{\boldsymbol{x}_{0i} : i = 1, \ldots, n_0\}$

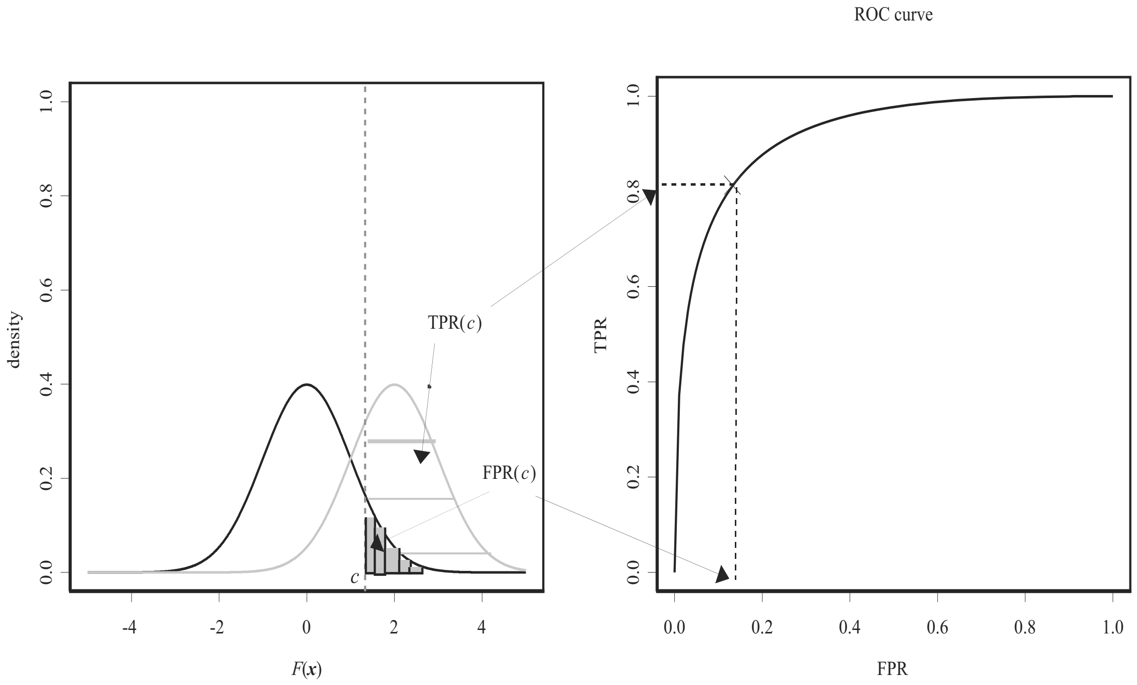


Figure 3.1: The left panel illustrates the definition of FPR and TPR with two probability density functions of $F(\boldsymbol{x})$ for class 0 (black) and 1 (gray), and a threshold c. The right panel is the corresponding ROC curve.

of the class 0 and $\{\boldsymbol{x}_{1j} : j = 1, \ldots, n_1\}$ of the class 1 is given by

$$\overline{\mathrm{AUC}}(F) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathrm{H}(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})), \tag{3.2.3}$$

where $\mathrm{H}(z)$ is the Heaviside function: $\mathrm{H}(z) = 1$ if $z \geq 0$ and 0 otherwise. In the case that $F(\boldsymbol{x})$ is discrete or there are tied values between $F(\boldsymbol{x}_{0i})$ and $F(\boldsymbol{x}_{1j})$, $\mathrm{H}(z)$ is replaced with $\mathrm{H}^*(z)$ that is defined to be 1 if $z > 0, \frac{1}{2}$ if $z = 0$ and 0 if $z < 0$.

3.2.2 Approximate AUC

We would like to obtain an optimal score function in the sense of maximizing the AUC in a class of score functions. It is known that the error rate is minimized by Bayes rule (McLachlan, 2004), which can be expressed using a strictly increasing function of the likelihood ratio. Similarly, the Neyman-Pearson Lemma establishes that the ROC curve for an arbitrary score function is everywhere below the ROC curve for the likelihood ratio (Eguchi and Copas, 2002; McIntosh and Pepe, 2002). That is, the optimal score function that maximizes the AUC is given as

$$F(\boldsymbol{x}) = m\big(\Lambda(\boldsymbol{x})\big), \tag{3.2.4}$$

where $\Lambda(\boldsymbol{x}) = g_1(\boldsymbol{x})/g_0(\boldsymbol{x})$ and m is a strictly increasing function. In this way, we observe that the maximization of the AUC is equivalent to the minimization of the error rate in the sense of Bayes rule.

In practice, the maximization of the empirical AUC presents some difficulties because it consists of a sum of nondifferentiable functions, as seen in equation (3.2.3). This feature prevents us from using gradient-based methods and requires a time-consuming search for the optimal score function (Pepe and Thompson, 2000; Pepe *and others*, 2006). However, such a method becomes impossible to implement as the number of feature variables increases greatly. Therefore, as a means of maximizing the empirical AUC, it has become common to use smooth-function approximations. Eguchi and Copas (2002) used the standard normal distribution function, and Ma and Huang (2005) proposed a sigmoid approximation for this purpose. In this paper, we consider the former approximation:

$$\overline{\mathrm{AUC}}_\sigma(F) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathrm{H}_\sigma(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})),$$

where $\mathrm{H}_\sigma(z) = \Phi\left(z/\sigma\right)$, with Φ being the standard normal distribution function. A smaller scale parameter σ means a better approximation of the Heaviside function $\mathrm{H}(z)$. The choice of the approximation function of $\mathrm{H}(z)$ does not matter so much; the important property is that the first derivative of the approximation function must be symmetric, which is satisfied in both $\mathrm{H}_\sigma(z)$ and the sigmoid function. This property is essential for the proof of Theorem 3.2.1.

Next, we discuss the relationship between the AUC and the approximate AUC. We note that the AUC for a score function $F(\boldsymbol{x})$ has an integral formula given as

$$\mathrm{AUC}(F) = \int\int \mathrm{H}(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_0 d\boldsymbol{x}_1.$$

Similarly, the approximate AUC is given as

$$\mathrm{AUC}_\sigma(F) = \int\int \mathrm{H}_\sigma(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_0 d\boldsymbol{x}_1.$$

Hence, we observe that $\overline{\mathrm{AUC}}_\sigma(F)$ almost surely converges to $\mathrm{AUC}_\sigma(F)$ as n_0 and n_1 both increase to infinity.

Theorem 3.2.1. *Let*

$$\Psi(c) = \mathrm{AUC}_\sigma\Big(F + c\ m\big(\Lambda\big)\Big),$$

where $\Lambda(\boldsymbol{x}) = g_1(\boldsymbol{x})/g_0(\boldsymbol{x})$ *and* m *is a strictly increasing function. Then,* $\Psi(c)$ *is a strictly increasing function of* $c \in \mathbf{R}$*, and*

$$\sup_F \mathrm{AUC}_\sigma(F) = \lim_{c\to\infty} \Psi(c) = \mathrm{AUC}\big(\Lambda\big). \tag{3.2.5}$$

Proof. Let $\zeta(\boldsymbol{x}) = m\big(\Lambda(\boldsymbol{x})\big)$. Then, the first derivative of $\Psi(c)$ with respect to c is given as

$$\int\int \Big(\zeta(\boldsymbol{x}_1) - \zeta(\boldsymbol{x}_0)\Big) \mathrm{H}'_\sigma\Big(F(\boldsymbol{x}_1) + c\ \zeta(\boldsymbol{x}_1) - F(\boldsymbol{x}_0) - c\ \zeta(\boldsymbol{x}_0)\Big) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_0 d\boldsymbol{x}_1,$$

which can be rewritten as

$$\int\int \Big(\zeta(\boldsymbol{x}_0) - \zeta(\boldsymbol{x}_1)\Big) \mathrm{H}'_\sigma\Big(F(\boldsymbol{x}_1) + c\ \zeta(\boldsymbol{x}_1) - F(\boldsymbol{x}_0) - c\ \zeta(\boldsymbol{x}_0)\Big) g_0(\boldsymbol{x}_1) g_1(\boldsymbol{x}_0) d\boldsymbol{x}_1 d\boldsymbol{x}_0,$$

by the exchange of $\boldsymbol{x}_0$ for $\boldsymbol{x}_1$ because of the symmetry: $\mathrm{H}'_\sigma(-z) = \mathrm{H}'_\sigma(z)$. Hence, we conclude that

$$\begin{aligned} 2\frac{\partial}{\partial c}\Psi(c) = \int\int &\Big(\zeta(\boldsymbol{x}_1) - \zeta(\boldsymbol{x}_0)\Big) \mathrm{H}'_\sigma\Big(F(\boldsymbol{x}_1) + c\ \zeta(\boldsymbol{x}_1) - F(\boldsymbol{x}_0) - c\ \zeta(\boldsymbol{x}_0)\Big) \\ &\times g_0(\boldsymbol{x}_0) g_0(\boldsymbol{x}_1) \Big(\Lambda(\boldsymbol{x}_1) - \Lambda(\boldsymbol{x}_0)\Big) d\boldsymbol{x}_0 d\boldsymbol{x}_1, \end{aligned}$$

which is always positive because of the assumption that m is a strictly increasing function. Hence, the function $\Psi(c)$ is strictly increasing.

From the discussion above, it follows that

$$\begin{aligned}
\mathrm{AUC}_\sigma(F) &< \lim_{c\to\infty} \Psi(c) \\
&= \lim_{c\to\infty} \mathrm{AUC}_\sigma\left[c\left\{\frac{F}{c}+\zeta\right\}\right] \\
&= \lim_{c\to\infty} \mathrm{AUC}_{\frac{\sigma}{c}}\left(\frac{F}{c}+\zeta\right) \\
&= \mathrm{AUC}(\zeta) \\
&= \mathrm{AUC}(\Lambda).
\end{aligned}$$

Considering the fact that

$$\lim_{c\to\infty} \Psi(c) \le \sup_F \mathrm{AUC}_\sigma(F),$$

we have

$$\lim_{c\to\infty} \Psi(c) = \sup_F \mathrm{AUC}_\sigma(F),$$

which concludes (3.2.5). □

From Theorem 3.2.1, we observe that

$$\mathrm{AUC}_\sigma(F) < \mathrm{AUC}\big(\Lambda\big),$$

and that no score function $F(\boldsymbol{x})$ can attain the equality above when $\sigma > 0$. Hence, we can perform the supremization of $\mathrm{AUC}_\sigma(F)$ instead of the maximization. This property is not preferable in building an iterative algorithm for maximization of $\mathrm{AUC}_\sigma(F)$; therefore, we propose a regularization scheme for $F(\boldsymbol{x})$ in a subsequent discussion.

3.3 AUCBoost

3.3.1 Objective function

We investigate a classification problem based on a boosting method. The key concept is to construct a powerful score function $F(\boldsymbol{x})$ by combining many various weak classifiers (Hastie *and others*, 2001). Any single weak classifier itself has a very poor ability for classification, whose performance is almost equal to random guessing; however, the combination of a number of them produces a very flexible and strong score function. We aim to construct $F(\boldsymbol{x})$ in such a way based on the AUC.

At first, we prepare a set $\mathcal{F}_k$ for each k-th component of $\boldsymbol{x} \in \mathbf{R}^p$:

$$\mathcal{F}_k = \left\{ f(\boldsymbol{x}) = a\mathrm{H}(x_k - b) + (1-a)/2 \mid a \in \{-1, 1\},\ b \in \mathcal{B}_k \right\},\ k = 1, \ldots, p,$$

where $\mathcal{B}_k$ is a finite discrete set, which is determined by taking every intermediate point of samples or a number of sample quantiles. As seen in the definition, $f(\boldsymbol{x})$ is a simple step function taking one of the two values $\{0, 1\}$. Then, we combine the sets into

$$\mathcal{F} = \bigcup_{k=1}^{p} \mathcal{F}_k, \tag{3.3.1}$$

called the decision stump class, among which we choose weak classifiers to construct $F(\boldsymbol{x})$. The set $\mathcal{F}$ can be modified to include interaction terms that may improve classification performance. However, the interpretation becomes difficult and unclear especially when the number of feature variables is large. Hence, in this paper we focus only on the main effects of feature variables.

In this setting, $F(\boldsymbol{x})$ can be decomposed as the same way as GAM:

$$\begin{aligned} F(\boldsymbol{x}) &= \sum_{f \in \mathcal{F}'_1} \alpha_f f(\boldsymbol{x}) + \cdots + \sum_{f \in \mathcal{F}'_p} \alpha_f f(\boldsymbol{x}) \\ &= F_1(x_1) + \cdots + F_p(x_p), \end{aligned}$$

where $\mathcal{F}'_k$ is a subset of $\mathcal{F}_k$, $k = 1, \ldots, p$, whose elements $f's$ are selected in a boosting algorithm in Subsection 3.2, and α_f means a corresponding coefficient of f. Using these notations, the objective function we propose is given as

$$\overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathrm{H}_\sigma(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) - \lambda \sum_{k=1}^{p} \sum_{x_k \in \mathcal{B}_k} \left\{ F_k^{(2)}(x_k) \right\}^2, \tag{3.3.2}$$

where λ is a smoothing parameter and $F_k^{(2)}(x_k)$ denotes the second-order difference of $F_k(x_k)$: $F_k^{(2)}(x_k) = F_k(x_k^{(-1)}) - 2F_k(x_k) + F_k(x_k^{(+1)})$ with $x_k^{(-1)} < x_k^{(+1)}$. The first term is the approximate empirical AUC based on the standard normal distribution function; the second term gives a penalty for redundant behavior of $F(\boldsymbol{x})$, which focuses on points in $\mathcal{B}_k$ for each k because $F_k(x_k)$ has discontinuities only at the points. Thus, the modeling of $F(\boldsymbol{x})$ is similar to that of GAM. The difference is that the proposed method is based on maximization of the AUC in place of the likelihood, and that we use the second-order difference of $F_k(x_k)$ instead of the second derivative of $F_k(x_k)$ because of its non-smoothness. The iteration method is also different: we maximize the objective function by a boosting method, whereas GAM is implemented by a backfitting algorithm (Hastie et al., 2001). We investigate the difference in detail using numerical simulation data in Section 4.

We note that there is a special relation between the scale parameter σ and the smoothing parameter λ. Equation (3.3.2) can be rewritten as

$$\overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} \mathrm{H}_\sigma(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) - \lambda\sigma^2 \sum_{k=1}^{p}\sum_{x_k \in \mathcal{B}_k} \left\{ \frac{F_k^{(2)}(x_k)}{\sigma} \right\}^2 .$$

Hence, we have

$$\overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \overline{\mathrm{AUC}}_{\sigma',\lambda'}\left(\frac{\sigma'}{\sigma}F\right),$$

if $\lambda\sigma^2 = \lambda'\sigma'^2$. This implies that the maximization of $\overline{\mathrm{AUC}}_{\sigma,\lambda}(F)$ is equivalent to that of $\overline{\mathrm{AUC}}_{1,\lambda\sigma^2}\left(\frac{F}{\sigma}\right)$. Therefore, we have

$$\max_{\sigma,\lambda,F} \overline{\mathrm{AUC}}_{\sigma,\lambda}(F) = \max_{\lambda,F} \overline{\mathrm{AUC}}_{1,\lambda}(F).$$

From this consideration, we can fix $\sigma = 1$ without loss of generality. Henceforth, we discuss

$$\overline{\mathrm{AUC}}_{\lambda}(F) = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} \Phi(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) - \lambda \sum_{k=1}^{p}\sum_{x_k \in \mathcal{B}_k} \left\{ F_k^{(2)}(x_k) \right\}^2,$$

which is rewrite of $\overline{\mathrm{AUC}}_{1,\lambda}(F)$ for notational convenience. This discussion not only leads to a drastic reduction of the computational cost for the implementation of our method, but also has consistency with Theorem 3.2.1. The scale parameter σ, which controls the accuracy of the approximation of the AUC, is not an essential factor in the sense of the supremization of the approximate AUC. On the other hand, the smoothing parameter λ has another important role. As mentioned after Theorem 3.2.1, the approximate AUC has no maximum in itself. The penalty term for smoothness in $\mathrm{AUC}_\lambda(F)$ also guarantees the existence of the maximum of $\mathrm{AUC}_\lambda(F)$, and makes the numerical maximization stable.

Ma and Huang (2005) and Wang *and others* (2007) approximated the empirical AUC by a sigmoid function, and followed a rule of thumb to determine a scale parameter. That is to say, the accuracy of approximation of the empirical AUC is already fixed before running their algorithm. In contrast, we do not impose such a strict condition; we vary only the smoothing parameter λ and select the best value by cross-validation (see Subsection 3.3).

3.3.2 AUCBoost algorithm

Let us give a brief explanation of how the score function $F(\boldsymbol{x})$ is constructed by sequentially selecting $f(\boldsymbol{x})$'s in the set $\mathcal{F}$ defined in (3.3.1). Our approach is based on a boosting learning algorithm to maximize $\overline{\mathrm{AUC}}_\lambda(F)$ in the linear hull of $\mathcal{F}$, with the number of iterations T.

1. Start with a score function $F_0(\boldsymbol{x})$.
2. For $t = 1, \ldots, T$

   a. Find the best weak classifier f_t and calculate the coefficient α_t as

$$\begin{aligned} f_t(\boldsymbol{x}) &= \operatorname*{argmax}_{f\in\mathcal{F}} \frac{\partial}{\partial\alpha}\overline{\mathrm{AUC}}_\lambda(F_{t-1}+\alpha f)\bigg|_{\alpha=0}, \\ \alpha_t &= \operatorname*{argmax}_{\alpha>0} \overline{\mathrm{AUC}}_\lambda(F_{t-1}+\alpha f_t). \end{aligned}$$

   b. Update the score function as

$$F_t(\boldsymbol{x}) = F_{t-1}(\boldsymbol{x}) + \alpha_t f_t(\boldsymbol{x}).$$

3. Finally, output the final score function:

$$F(\boldsymbol{x}) = F_0(\boldsymbol{x}) + \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}).$$

If we have no prior information about the data, we set $F_0(\boldsymbol{x}) = 0$. In step 2.a, we search $\mathcal{F}$ for a $f_t(\boldsymbol{x})$ which maximizes the first derivative of $\overline{\mathrm{AUC}}_\lambda(F)$ at the point $F_{t-1}(\boldsymbol{x}) + \alpha f(\boldsymbol{x})$. This argument is similar to that of Hastie *and others* (2001) and Takenouchi and Eguchi (2004). Next, we calculate the coefficient of $f_t(\boldsymbol{x})$ using the Newton-Raphson method, and add $\alpha_t f_t(\boldsymbol{x})$ to the previous score function. We repeat this process T times and output the final score function. Thus, the resultant score function is an aggregation of $f_t(\boldsymbol{x})$'s with weights α_t's. Further details of this algorithm are as follows.

In step 2.a, we search $\mathcal{F}$ for f_t that satisfies

$$\begin{aligned}
&f_t(\boldsymbol{x}) \\
=&\underset{f\in\mathcal{F}}{\mathrm{argmax}}\ \frac{\partial}{\partial\alpha}\overline{\mathrm{AUC}}_\lambda(F_{t-1}+\alpha f)\bigg|_{\alpha=0} \\
=&\underset{\substack{k\in\{1,\ldots,p\}\\ a\in\{-1,1\}\\ b\in\mathcal{B}_k}}{\mathrm{argmax}}\ \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\phi\big(F_{t-1}(\boldsymbol{x}_{1j}) - F_{t-1}(\boldsymbol{x}_{0i})\big)\big\{a\,\mathrm{H}(x_{1jk}-b) - a\,\mathrm{H}(x_{0ik}-b)\big\} \\
&\quad - 2\lambda\sum_{x_k\in\mathcal{B}_k}\{F_k(x_k^{(-1)}) - 2F_k(x_k) + F_k(x_k^{(+1)})\}\{a\mathrm{H}(x_k^{(-1)}-b) - 2a\mathrm{H}(x_k-b) + a\mathrm{H}(x_k^{(+1)}-b)\},
\end{aligned}$$

where ϕ is the standard normal density function, $F_k(x_k)$ is the k-th component of $F_{t-1}(\boldsymbol{x})$ (a score function of x_k at an iteration number $t-1$), and x_{0ik}, x_{1jk} are the k-th component of $\boldsymbol{x}_{0i}, \boldsymbol{x}_{1j}$, respectively.

Then, the second term in the equation above is calculated into

$$\begin{aligned}
&-2\lambda\Big[\Big\{F_k(b^{(-2)}) - 2F_k(b^{(-1)}) + F_k(b)\Big\}a - \Big\{F_k(b^{(-1)}) - 2F_k(b) + F_k(b^{(+1)})\Big\}a\Big] \\
=&-2\lambda a\Big\{F_k(b^{(-2)}) - 3F_k(b^{(-1)}) + 3F_k(b) - F_k(b^{(+1)})\Big\},
\end{aligned}$$

where an element with a smaller superscript number than that of the minimum element in $\mathcal{B}_k$ is set to the minimum one. Similarly, an element with a larger superscript number than that of the maximum element is set to the maximum one. In regard to the coefficient (α_t) of $f_t(\boldsymbol{x})$, we seek it by the Newton-Raphson method using

$$\begin{aligned}
&\frac{\partial}{\partial\alpha}\overline{\mathrm{AUC}}_\lambda(F_{t-1}+\alpha f_t)\\
=&\frac{1}{n_0n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\phi\Big(F_{t-1}(\boldsymbol{x}_{1j})-F_{t-1}(\boldsymbol{x}_{0i})+\alpha\{f_t(\boldsymbol{x}_{1j})-f_t(\boldsymbol{x}_{0i})\}\Big)\Big(f_t(\boldsymbol{x}_{1j})-f_t(\boldsymbol{x}_{0i})\Big)\\
&\qquad-2\lambda\Big[a_t\Big\{F_{k_t}(b_t^{(-2)})-3F_{k_t}(b_t^{(-1)})+3F_{k_t}(b_t)-F_{k_t}(b_t^{(+1)})\Big\}+2\alpha\Big],
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial^2}{\partial\alpha^2}\overline{\mathrm{AUC}}_\lambda(F_{t-1}+\alpha f_t)\\
=&-\frac{1}{n_0n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\phi\Big(F_{t-1}(\boldsymbol{x}_{1j})-F_{t-1}(\boldsymbol{x}_{0i})+\alpha\{f_t(\boldsymbol{x}_{1j})-f_t(\boldsymbol{x}_{0i})\}\Big)\Big(f_t(\boldsymbol{x}_{1j})-f_t(\boldsymbol{x}_{0i})\Big)^2\\
&\quad\times\Big(F_{t-1}(\boldsymbol{x}_{1j})-F_{t-1}(\boldsymbol{x}_{0i})+\alpha\{f_t(\boldsymbol{x}_{1j})-f_t(\boldsymbol{x}_{0i})\}\Big)-4\lambda, \qquad (3.3.3)
\end{aligned}$$

where

$$f_t(\boldsymbol{x})=a_t\,\mathrm{H}(x_{k_t}-b_t)+(1-a_t)/2.$$

The first term in (3.3.3) is usually negative for an appropriate value of α. However, it happens to be positive in the Newton-Raphson process. Our objective is to obtain α that maximizes $\overline{\mathrm{AUC}}_\lambda(F_{t-1}+\alpha f_t)$, so the sign of (3.3.3) should be always negative. We find that the smoothing parameter λ stabilizes the algorithm of AUCBoost.

3.3.3 Tuning parameter selection

In our method there are two parameters to be determined: a smoothing parameter λ and the iteration number T. We use the following K-fold cross-validation. At first, we partition the whole data set into K subsets of almost equal sizes, and evaluate an objective function such as

$$\mathrm{AUC}_{\mathrm{CV}}(\lambda,T)=\frac{1}{K}\sum_{i=1}^{K}\overline{\mathrm{AUC}}_\lambda^{(i)}(F^{(-i)}),$$

where $F^{(-i)}$ is a score function constructed by AUCBoost using the data set without the i-th subset, and $\overline{\mathrm{AUC}}_{\lambda}^{(i)}$ is the $\overline{\mathrm{AUC}}_{\lambda}$ calculated on the i-th subset alone. We give a typical example of results of $\mathrm{AUC}_{\mathrm{CV}}$ in Figure 3.2, assuming $\boldsymbol{X}_0 \sim \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\boldsymbol{X}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, where $\boldsymbol{\mu}_0 = (0, 0, 0, 0)'$, $\boldsymbol{\mu}_1 = (0, 0.5, 0, 0.5)'$, $\boldsymbol{\Sigma}_0 = \mathrm{diag}(1, 1, 1, 1)$ and $\boldsymbol{\Sigma}_1 = \mathrm{diag}(1, 1, 4, 0.25)$. In this case, the best pair of the parameters seems to be $\lambda = 0.01$ and $T = 200$. The curve with $\lambda = 0.0001$ increases rapidly at the beginning and starts to decline around $T = 20$. The second curve denoted by triangles has a peak around $T = 70$ and shows a moderate tendency to decrease after that point. On the other hand, the best curve with $\lambda = 0.01$ shows that the score function hardly suffers from overfitting to the data. This fact also can be confirmed by observing the corresponding score function $F(\boldsymbol{x})$. The true score function in this setting is a smooth function; however, we observed that the score function with $\lambda = 0.0001$ clearly lacked the smoothness (not shown here). This also indicates overfitting to the data. With an appropriate value of λ and a relatively-large iteration number T, this slow learning process contrasts starkly with the usual regularization technique, i.e., early stopping (Zhang and Yu, 2005). We set the value of K to 10 for simulation studies and 5 for a real data analysis, according to the sample size.

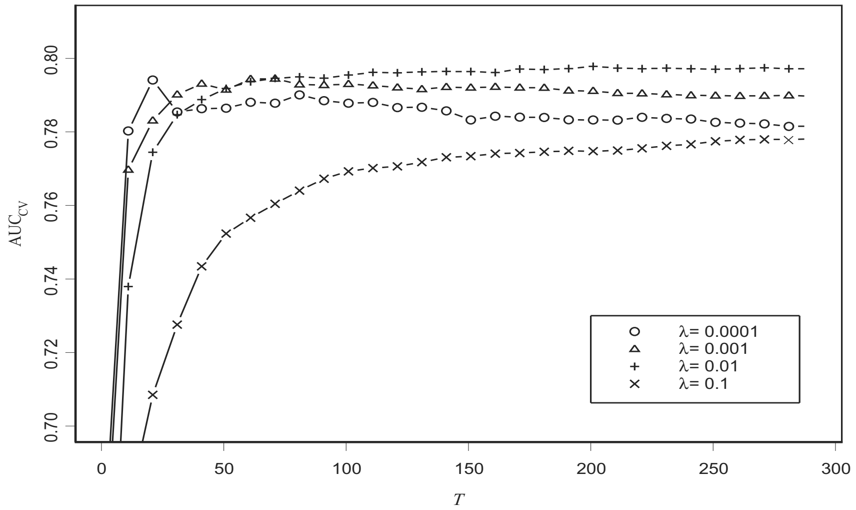


Figure 3.2: Results of $\mathrm{AUC}_{\mathrm{CV}}$ corresponding to different values of λ, as a function of the number of iterations T.

3.3.4 Score plot and score ROC

We discuss the AUCBoost algorithm to select classifiers in the decision stump class $\mathcal{F}$. The choice of the class provides us with useful information regarding the feature variables in a post-analysis of classification. The final score function $F(\boldsymbol{x})$ is decomposed as

$$F(\boldsymbol{x}) \quad = \quad \sum_{k=1}^{p} F_k(x_k).$$

The utility of the plot of $F_k(x_k)$ against x_k (score plot of x_k) is referred to by Friedman *and others* (2000) and Kawakita *and others* (2005). Observing each score plot very carefully, we are able to not only understand how each feature variable x_k influences the classification performance, but also know which feature variable is the most effective and informative one. We discuss this utility more in detail in simulation studies. Another useful way to gauge the efficiency of each feature variable is to draw the ROC curve for $F_k(x_k)$ (score ROC) and calculate the corresponding AUC (score AUC). $F_k(x_k)$ represents the contribution of x_k to the total classification performance; hence, the value of the score AUC shows the utility of x_k. These measurements are more convenient for comparing the utilities of feature variables because we can order them according to their values.

3.4 Simulation studies

3.4.1 Setting

In this section, we present two simulation studies. One is intended to demonstrate that the score function $F(\boldsymbol{x})$ generated by AUCBoost provides a good approximation to the optimal score function, and that score plots are useful for evaluating each feature variable's contribution to $F(\boldsymbol{x})$. The other is designed to show that, in cases where several outliers exist, AUCBoost is much more powerful and robust than other classification methods such as AdaBoost, GAM and the generalized linear model (GLM). The iteration number for AdaBoost is also determined by cross-validation where the objective function is based on the empirical AUC. Cubic splines are used for GAM, and these simulation studies are done using Splus 8.0. Throughout these simulations, the training sample size is set to be 500 (n_0=250, n_1=250) and we evaluated the quality using a test sample of size 200 (n_0=100 n_1=100). Summary statistics are based on 1000 repetitions.

3.4.2 Comparison with the optimal score function

Consider the same situation as that of Subsection 3.3: $\boldsymbol{X}_0 \sim \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\boldsymbol{X}_1 \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, where $\boldsymbol{\mu}_0 = (0, 0, 0, 0)'$, $\boldsymbol{\mu}_1 = (0, 0.5, 0, 0.5)'$, $\boldsymbol{\Sigma}_0 = \mathrm{diag}(1, 1, 1, 1)$ and $\boldsymbol{\Sigma}_1 = \mathrm{diag}(1, 1, 4, 0.25)$. From equation (3.2.4), the optimal score function in this setting is given as

$$F_{\mathrm{N}}(\boldsymbol{x}) = \boldsymbol{x}'(\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1})\boldsymbol{x} + 2(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_0'\boldsymbol{\Sigma}_0^{-1})\boldsymbol{x},$$

which coincides with a linear score function proposed by Su and Liu (1993) if $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$. The score plots constructed by AUCBoost track $F_{\mathrm{N}}(\boldsymbol{x})$ very well as seen in Figure 3.3, where the rug plots at the bottom of each graph depict the data distribution. Clearly, it shows nonlinearity of $F(\boldsymbol{x})$, especially $F_3(x_3)$ and $F_4(x_4)$. From the shape of $F_3(x_3)$ we see that x_3 with class label 0 has a tendency to concentrate around the origin, compared to x_3 with class label 1. On the other hand, in regard to $F_4(x_4)$, we see the opposite tendency of x_4. The flatness of $F_1(x_1)$ means that x_1 is useless for discriminating subjects with class 0 from those with class 1, because weak classifiers for x_1 are rarely chosen, and the weight coefficients are calculated to be very small in the AUCBoost algorithm. Judging from the heights of score plots, x_4 seems to be the most informative one.

Table 3.1 shows the results of the score AUCs and the AUCs calculated by AUCBoost and $\hat{F}_{\mathrm{N}}$, where $\hat{F}_{\mathrm{N}}$ denotes the estimator of F_{N}. As expected, $\hat{F}_{\mathrm{N}}$ achieves superior performance for all AUCs. It is because $\hat{F}_{\mathrm{N}}$ is derived based on the underlying probability distributions; on the other hand, the score function of AUCBoost is constructed by the sample distributions. We also notice that the values of the score AUCs for AUCBoost are in accordance with the heights of the score plots. The utility of x_4 is confirmed again.

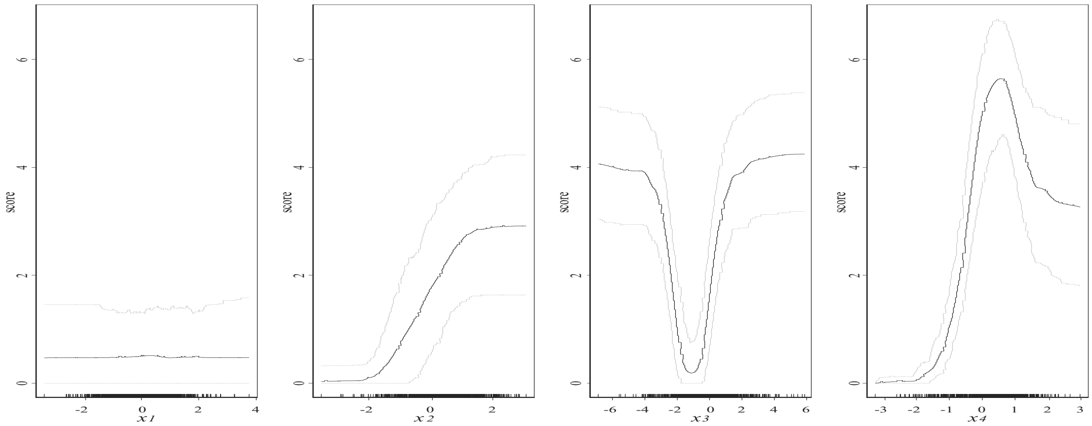


Figure 3.3: Score plots for AUCBoost. The black lines indicate mean score plots and the gray lines indicate the 95 percent pointwise confidence bands.

3.4.3 Comparison with other methods

Next, we relax the conditions of the probability distribution a little, and consider a multivariate t-distribution. This is a more practical setting because it contains several outliers which we often observe in real data. While there are several forms of multivariate t-distribution, we use the most common one. The density function of p-dimensional t-distribution with ν degrees of freedom, mean vector $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\Sigma}^{-1}$, is given as

$$g(\boldsymbol{x}) = \frac{\Gamma(\frac{p+\nu}{2})\sqrt{|\boldsymbol{\Sigma}^{-1}|}}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{p}{2}}}\left[1 + \frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]^{-\frac{p+\nu}{2}}.$$

We use the same parameters as those in the previous subsection: $\boldsymbol{\mu}_0 = (0,0,0,0)'$, $\boldsymbol{\mu}_1 = (0,0.5,0,0.5)'$, $\boldsymbol{\Sigma}_0 = \text{diag}(1,1,1,1)$ and $\boldsymbol{\Sigma}_1 = \text{diag}(1,1,4,0.25)$. To focus on the investigation of the robustness of $F(\boldsymbol{x})$ constructed by AUCBoost, we consider an extreme situation

Table 3.1: *The mean score AUCs and the AUCs with 95 percent confidence bands in parentheses*

	x_1	x_2	x_3	x_4	total
AUCBoost	0.501	0.628	0.700	0.736	0.828
	(0.429,0.579)	(0.548,0.707)	(0.617,0.774)	(0.670,0.799)	(0.772,0.879)
$\hat{F}_{\mathrm{N}}$	0.500	0.638	0.703	0.742	0.840
	(0.425,0.583)	(0.565,0.714)	(0.633,0.779)	(0.668,0.809)	(0.780,0.887)

($\nu = 1$). Figure 3.4 shows score plots and score ROCs of x_3 for AUCBoost, AdaBoost, GAM and GLM. The range of score plots has been adjusted for a better view. Interestingly, the shape of score plots of AUCBoost and AdaBoost are almost the same. This is because both of the boosting methods focus only on points that are useful for the classification. On the other hand, GAM is sensitive to uninformative samples such as outliers, which causes the GAM's performance instability (Kawakita *and others*, 2005). In regard to GLM, it does not capture the useful information about x_3 at all, which is observable from the value of the score AUC (0.496) as well as the shape of the score ROC. The concavity of the shape of ROC is known to be a necessary condition of the optimality (Pepe, 2003). In the last column of Table 3.2, we can see that the corresponding 95 percent confidence band of the AUC for AUCBoost is much narrower than the others. Among all of them, the result for AUCBoost is the most stable with the largest mean AUC value (0.787). The smoothing parameter λ contributes to the stable result of AUCBoost.

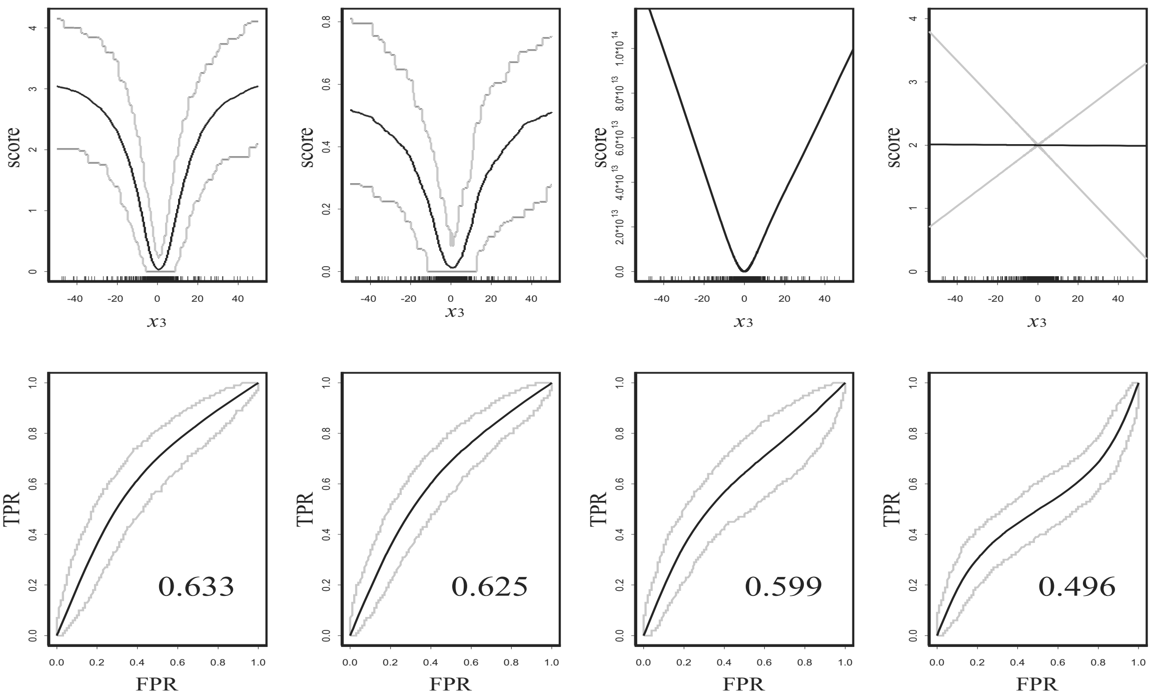


Figure 3.4: Results of score plots (upper panels) and score ROCs (lower panels) for x_3 of AUCBoost, AdaBoost, GAM and GLM. The black lines indicate mean score plots and score ROCs, and the gray lines indicate the 95 percent pointwise confidence bands. The confidence band of the score plot for GAM is omitted, and the minimum values of axises of score plots are set to 0 for a better view.

Table 3.2: *The mean score AUCs and the AUCs with 95 percent confidence bands in parentheses*

	x_1	x_2	x_3	x_4	total
AUCBoost	0.501	0.616	0.633	0.697	0.787
	(0.425,0.578)	(0.541,0.694)	(0.547,0.711)	(0.620,0.764)	(0.723,0.839)
AdaBoost	0.501	0.601	0.625	0.690	0.776
	(0.428,0.579)	(0.522,0.685)	(0.553,0.696)	(0.610,0.761)	(0.705,0.834)
GAM	0.497	0.618	0.599	0.672	0.738
	(0.418,0.583)	(0.543,0.699)	(0.487,0.694)	(0.603,0.748)	(0.661,0.804)
GLM	0.501	0.601	0.496	0.649	0.648
	(0.434,0.572)	(0.388,0.690)	(0.427,0.555)	(0.340,0.744)	(0.533,0.737)

3.5 Application to spinal disease in children data

We apply AUCBoost to a real data set, which can be seen in Statistical Models in S edited by (Chambers and Hastie, 1991). The label is the outcome of corrective spinal surgery of 81 children: whether kyphosis is present or absent. The feature variables are as follows: Age, the age of the child in months; Number, the number of vertebrae in the operation; and Start, the beginning of the range of vertebrae involved in the operation. We used the first 70 samples as training data, and the others as test data. Figure 3.5 shows the score plots for AUCBoost, AdaBoost, GAM and GLM, respectively. We find clear nonlinearity of score plots for Age, except for that of GLM. The peak appears around 100 months. A child of Age 200 is estimated by GLM to have the highest risk of a postoperative deformity; on the other hand, the risk at this age is estimated by GAM to be the lowest. AUCBoost gives results intermediate between these two extremes. The smoothness of score plots for AUCBoost is quite different from that of AdaBoost. This result makes it easy to understand how each feature variable affects the outcome after surgery and to interpret the results of the analysis. It also contributes to preventing the score function $F(\boldsymbol{x})$ from overfitting to the data. The values of the AUCs based on training data are 0.926, 0.997, 0.949 and 0.869 for AUCBoost, AdaBoost, GAM and GLM; however, the values based on test data are 0.777, 0.666, 0.666 and 0.666, respectively.

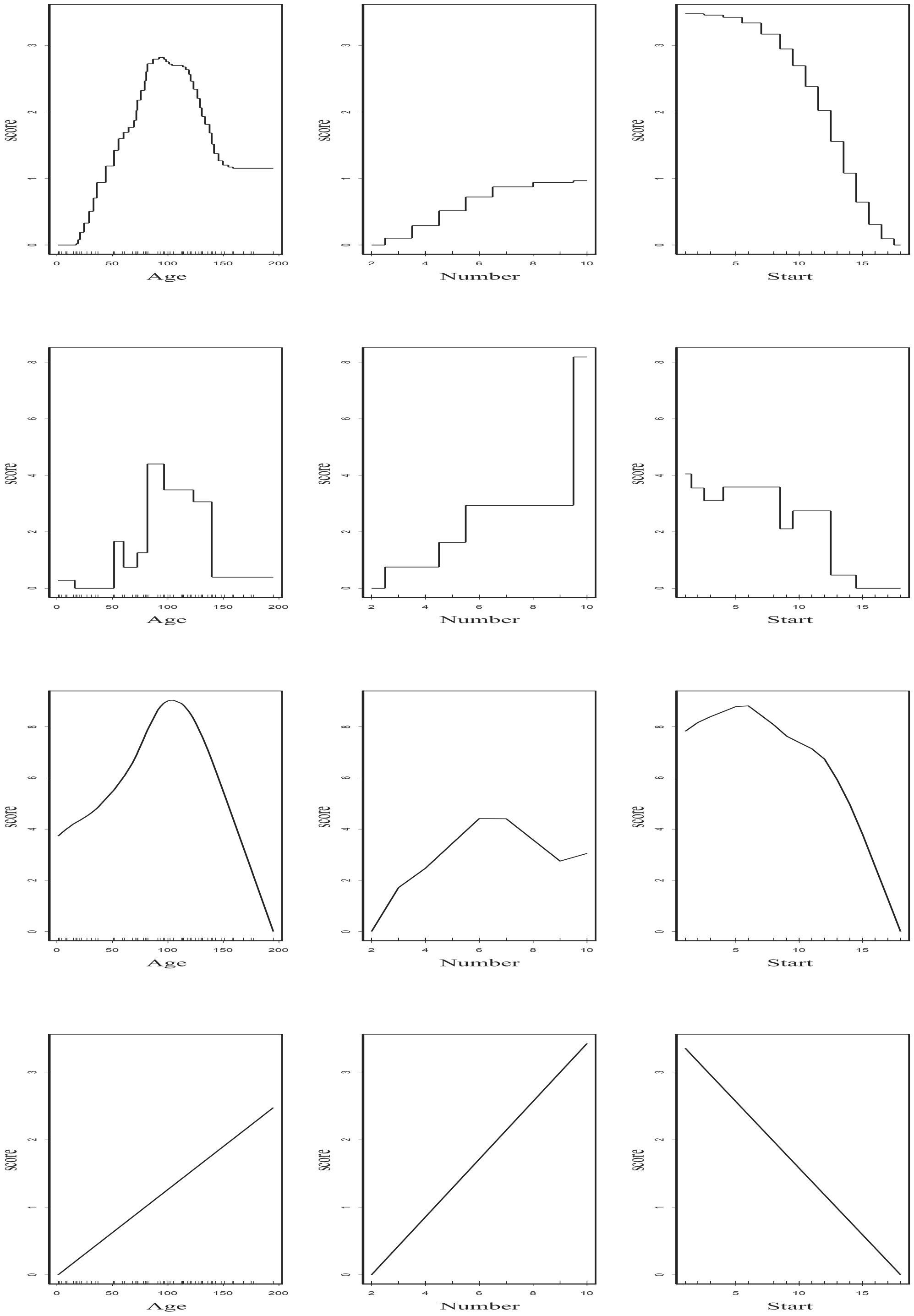


Figure 3.5: Score plots for AUCBoost, AdaBoost, GAM and GLM from top to bottom. The minimum values of each score plot are set to 0 for better view.

3.6 Conclusions and future work

AUCBoost offers a flexible combination of multiple feature variables, which is optimal in the sense of the maximization of the AUC. The smoothing parameter λ in the objective function not only contributes to improvement of the classification performance, but also gives us smoothed score plots, which are very useful in clinical studies. By observing the score plots very carefully, we can understand how each feature variable is associated with a disease or other endpoint, and also evaluate its efficiency by calculating the corresponding AUC.

From the setting of $\mathcal{F}$ which consists of component-based simple classifiers, the score function of AUCBoost has a similar form to that of GAM. However, there are two major differences between them. First, we maximize the AUC instead of the likelihood. Second, we update the score function by sequentially adding weak classifiers, whereas GAM is based on a backfitting algorithm (Hastie *and others*, 2001). The forward stagewise additive modeling gives AUCBoost robustness to distributions of data as seen in Subsection 4.3. Thus, AUCBoost is expected to show stable classification performance in various situations. This property also makes it easy to take discrete or ordered categorical data into consideration, which is difficult or impossible for the backfitting algorithm.

A weak point of AUCBoost is that the selection of the tuning parameter λ and T is time-consuming because we apply a simple cross-validation method. In order to avoid such a computational cost and make it easy to use, a more sophisticated procedure is necessary. Recently, (Ueki and Fueda, 2009) proposed an effective method for determining tuning parameters of maximum penalized likelihood estimator. The idea is based on likelihood, not the AUC; however, it could be modified into AUCBoost and help it reduce its computational costs.

AUCBoost can also be applied to a high-dimensional data analysis, in which variable selection is much more important than in the low-dimensional data analysis we consider in this paper. The AUCBoost algorithm implicitly includes a selection process at each iteration stage, so that informative feature variables are selected as a result after applying AUCBoost. This property is similar to GAMBoost (Tutz and Binder, 2006), which circumvents GAM's restriction to low-dimensional setting. The concept of the partial AUC (pAUC) is also of great interest in the analysis of genetic data. Pepe *and others* (2003) showed the biological utility of the pAUC for ranking informative genes. We will work on developing partial AUCBoost as one of the appealing extensions of AUCBoost.

3.7 Complementary properties of the AUC

Optimal score function for the AUC

In this subsection we derive the optimal discriminant function $F^*(\boldsymbol{x})$ for which the ROC curve lies over any other ROC curves. Suppose p-dimensional random variables $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$ for each population have probability density functions g_0 and g_1, respectively.

At first we fix FPR as

$$\int_{\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}>c} g_0(\boldsymbol{x})d\boldsymbol{x} = \int_{F(\boldsymbol{x})>c'} g_0(\boldsymbol{x})d\boldsymbol{x}, \tag{3.7.1}$$

where $F(\boldsymbol{x})$ is an arbitrary discriminant function. For simplicity we define

$$R = \left\{\boldsymbol{x}\left|\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})} > c\right.\right\}, \quad S = \left\{\boldsymbol{x}\left|F(\boldsymbol{x}) > c'\right.\right\}.$$

Then we can rewrite (3.7.1) as

$$\int_{R\backslash S} g_0(\boldsymbol{x})d\boldsymbol{x} + \int_{R\cap S} g_0(\boldsymbol{x})d\boldsymbol{x} = \int_{S\backslash R} g_0(\boldsymbol{x})d\boldsymbol{x} + \int_{R\cap S} g_0(\boldsymbol{x})d\boldsymbol{x},$$

where $R\backslash S = R\cap(R\cap S)^c$ and R^c is a complement set of R. From (3.7.1) and the definition of R we have

$$\begin{aligned}\int_{R\backslash S} g_1(\boldsymbol{x})d\boldsymbol{x} &\geq c\int_{R\backslash S} g_0(\boldsymbol{x})d\boldsymbol{x} \\ &= c\int_{S\backslash R} g_0(\boldsymbol{x})d\boldsymbol{x} \\ &\geq \int_{S\backslash R} g_1(\boldsymbol{x})d\boldsymbol{x}.\end{aligned} \tag{3.7.2}$$

Finally we have an inequality with respect to TPR as

$$\begin{aligned}\int_R g_1(\boldsymbol{x})d\boldsymbol{x} &= \int_{R\backslash S} g_1(\boldsymbol{x})d\boldsymbol{x} + \int_{R\cap S} g_1(\boldsymbol{x})d\boldsymbol{x} \\ &\geq \int_{S\backslash R} g_1(\boldsymbol{x})d\boldsymbol{x} + \int_{R\cap S} g_1(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_S g_1(\boldsymbol{x})d\boldsymbol{x}.\end{aligned}$$

As a result the optimal discriminant function becomes

$$\begin{aligned}F_{\text{AUC}} &= \text{argmax}_F \ \text{AUC}(F) \\ &= m\left(\frac{g_1}{g_0}\right),\end{aligned} \tag{3.7.3}$$

where m is a monotonically increasing function. This proof is the same as that of Neyman-Pearson fundamental lemma (Neyman and Pearson, 1993), and the fact of (3.7.3) has been implicitly pointed out by Eguchi and Copas (2002) and McIntosh and Pepe (2002). As you see form (3.7.3), the optimal discriminant function could be linear only on the special occasion. For example the each two random variable is normally distributed and the each variance matrix is equal. In the medical research, however, they usually use the linear discriminant function based on the logistic regression. It is very problematic and much of useful information is dismissed.

The derivation of optimal discriminant function above is indirect way, because we have shown the optimality by the fact that TPRs for optimal discriminant function are always above those for the others. But we can also derive the optimal discriminant function by directly maximizing AUC.

Using probability density function g_0 and g_1, AUC can be expressed as

$$\text{AUC}(F) = \int\int \text{H}(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0 d\boldsymbol{x}_1,$$

where H is a Heaviside function. Then we define

$$F_\epsilon = F + \epsilon\,\eta$$

and differentiate $\mathrm{AUC}(F_\epsilon)$ by ϵ as follows:

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\mathrm{AUC}(F_\epsilon)\bigg|_{\epsilon=0} &= \int\!\!\int \big(\eta(\boldsymbol{x}_1)-\eta(\boldsymbol{x}_0)\big)\mathrm{H}'\big(F(\boldsymbol{x}_1)-F(\boldsymbol{x}_0)\big)g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0d\boldsymbol{x}_1\\
&= \int\!\!\int_{F(\boldsymbol{x}_0)=F(\boldsymbol{x}_1)} \big(\eta(\boldsymbol{x}_1)-\eta(\boldsymbol{x}_0)\big)g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0d\boldsymbol{x}_1\\
&= \int\!\!\int_{F(\boldsymbol{x}_1)=F(\boldsymbol{x}_0)} \big(\eta(\boldsymbol{x}_0)-\eta(\boldsymbol{x}_1)\big)g_0(\boldsymbol{x}_1)g_1(\boldsymbol{x}_0)d\boldsymbol{x}_1d\boldsymbol{x}_0\\
&= 0.
\end{aligned}$$

Hence we have

$$\int\!\!\int_{F(\boldsymbol{x}_0)=F(\boldsymbol{x}_1)} (\eta(\boldsymbol{x}_1)-\eta(\boldsymbol{x}_0))(g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)-g_0(\boldsymbol{x}_1)g_1(\boldsymbol{x}_0))d\boldsymbol{x}_0d\boldsymbol{x}_1 = 0.$$

The function $\eta(\boldsymbol{x})$ is arbitrary, so we choose $\eta(\boldsymbol{x}) = \frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}$:

$$\int\!\!\int_{F(\boldsymbol{x}_0)=F(\boldsymbol{x}_1)} g_0(\boldsymbol{x}_0)g_0(\boldsymbol{x}_1)\left\{\frac{g_1(\boldsymbol{x}_1)}{g_0(\boldsymbol{x}_1)} - \frac{g_1(\boldsymbol{x}_0)}{g_0(\boldsymbol{x}_0)}\right\}^2 d\boldsymbol{x}_0d\boldsymbol{x}_1 = 0. \tag{3.7.4}$$

From (3.7.4) it holds that: for ${}^\forall c \in \mathbf{R}$, $c' \in \mathbf{R}^+$ exists such that

$$A_c \subset B_{c'}, \tag{3.7.5}$$

where

$$A_c = \{\boldsymbol{x}\,|F(\boldsymbol{x}) = c\}, \quad B_{c'} = \left\{\boldsymbol{x}\,\middle|\,\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})} = c'\right\}.$$

From the definition above, we have

$$\bigcup_{c\,\in\mathbf{R}} A_c = \bigcup_{c'\,\in\mathbf{R}^+} B_{c'},$$

and

$$A_c \cap A_d = \phi \quad (c \neq d), \quad B_{c'} \cap B_{d'} = \phi \quad (c' \neq d'),$$

so (3.7.5) becomes

$$A_c = B_{c'}.$$

Finally we have

$$F = m\left(\frac{g_1}{g_0}\right), \tag{3.7.6}$$

where m is an arbitrary monotonically increasing function.

Theorem 3.7.1. *The discriminant function which Su and Liu proposed (Su and Liu, 1993) is a linear approximation of $F_n^*(\boldsymbol{x})$ at a special point $\boldsymbol{x}_s$ as below.*

$$\boldsymbol{x}_s = \boldsymbol{\Sigma}_0(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}\boldsymbol{\mu}_0.$$

Proof. Suppose $X_0 \sim \mathrm{N}(\boldsymbol{\mu_0},\boldsymbol{\Sigma_0}),\ X_1 \sim \mathrm{N}(\boldsymbol{\mu_1},\boldsymbol{\Sigma_1})$, then the optimal discriminant function is written as

$$F^*(\boldsymbol{x}) = \boldsymbol{x'}(\boldsymbol{\Sigma_0^{-1}} - \boldsymbol{\Sigma_1^{-1}})\boldsymbol{x} + 2(\boldsymbol{\mu_1'}\boldsymbol{\Sigma_1^{-1}} - \boldsymbol{\mu_0'}\boldsymbol{\Sigma_0^{-1}})\boldsymbol{x} - \boldsymbol{\mu_1'}\boldsymbol{\Sigma_1^{-1}}\boldsymbol{\mu_1} + \boldsymbol{\mu_0'}\boldsymbol{\Sigma_0^{-1}}\boldsymbol{\mu_0}.$$

Here we fix

$$\boldsymbol{x_0} = \boldsymbol{\Sigma_0(\Sigma_0 + \Sigma_1)^{-1}\mu_1} + \boldsymbol{\Sigma_1(\Sigma_0 + \Sigma_1)^{-1}\mu_0},$$

and the derivative at the point is

$$\frac{\partial d_n(\boldsymbol{x_0})}{\partial \boldsymbol{x}} = 2[(\boldsymbol{\Sigma_0^{-1}} - \boldsymbol{\Sigma_1^{-1}})\{\boldsymbol{\Sigma_0(\Sigma_0 + \Sigma_1)^{-1}\mu_1} + \boldsymbol{\Sigma_1(\Sigma_0 + \Sigma_1)^{-1}\mu_0}\} + \boldsymbol{\Sigma_1^{-1}\mu_1} - \boldsymbol{\Sigma_0^{-1}\mu_0}\ ].$$

By the following equation

$$\boldsymbol{(\Sigma_0 + \Sigma_1)^{-1}} = \boldsymbol{\Sigma_0^{-1}} - \boldsymbol{\Sigma_0^{-1}\Sigma_1(\Sigma_0 + \Sigma_1)^{-1}}$$

we have

$$\frac{\partial d_n(\boldsymbol{x_0})}{\partial \boldsymbol{x}} = 4(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})$$

Next we consider $F^*(\boldsymbol{x_0})$. Using the equation

$$(\boldsymbol{\Sigma_0^{-1}} - \boldsymbol{\Sigma_1^{-1}})\boldsymbol{x_0} + \boldsymbol{\Sigma_1^{-1}}\boldsymbol{\mu_1} - \boldsymbol{\Sigma_0^{-1}}\boldsymbol{\mu_0} = 2(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})$$

we have

$$\begin{aligned} d_n(\boldsymbol{x_0}) &= \boldsymbol{x_0'}\{\boldsymbol{2}(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}) + \boldsymbol{\Sigma_1^{-1}}\boldsymbol{\mu_1} - \boldsymbol{\Sigma_0^{-1}}\boldsymbol{\mu_0}\} - \boldsymbol{\mu_1'}\boldsymbol{\Sigma_1^{-1}}\boldsymbol{\mu_1} + \boldsymbol{\mu_0'}\boldsymbol{\Sigma_0^{-1}}\boldsymbol{\mu_0} \\ &= \boldsymbol{2x_0'}(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}) - \boldsymbol{\mu_1'}\boldsymbol{\Sigma_1^{-1}}\boldsymbol{\mu_1} + \boldsymbol{\mu_0'}\boldsymbol{\Sigma_0^{-1}}\boldsymbol{\mu_0} \\ &= -\boldsymbol{2\mu_1'}(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}\boldsymbol{\Sigma_1}(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}) \\ &\qquad -\boldsymbol{2\mu_0'}(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}\boldsymbol{\Sigma_0}(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}) + \boldsymbol{\mu_1'}\boldsymbol{\Sigma_1^{-1}}\boldsymbol{\mu_1} - \boldsymbol{\mu_0'}\boldsymbol{\Sigma_0^{-1}}\boldsymbol{\mu_0} \\ &= (\boldsymbol{\mu_1} - \boldsymbol{\mu_0})'(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}(\boldsymbol{\Sigma_0} - \boldsymbol{\Sigma_1})(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}). \end{aligned}$$

Hence, the tangent line of $F^*(\boldsymbol{x})$ at $\boldsymbol{x_0}$ becomes

$$\begin{aligned} &4(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})'(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}(\boldsymbol{x} - \boldsymbol{x_0}) \\ &\qquad +(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})'(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}(\boldsymbol{\Sigma_0} - \boldsymbol{\Sigma_1})(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}). \end{aligned}$$

This is equivalent to the linear discriminant function $F_l(\boldsymbol{x})$ (See below) proposed by Su and Liu, because ROC curve is invariant to the increasing monotone transformation of discriminant function.

$$F_l(\boldsymbol{x}) = (\boldsymbol{\mu_1} - \boldsymbol{\mu_0})'(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-\boldsymbol{1}}\boldsymbol{x}.$$

□

3.7.1 Convexity of the ROC curve for the optimal score function

The optimal score function is given as

$$F(\boldsymbol{x}) = m\left(\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}\right).$$

Thus, the corresponding FPR and TPR are written as

$$\mathrm{FPR}(c) = \int_{m\left(\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}\right)>c} g_0(\boldsymbol{x})d\boldsymbol{x},\ \mathrm{TPR}(c) = \int_{m\left(\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}\right)>c} g_1(\boldsymbol{x})d\boldsymbol{x}.$$

Then, we have

$$\begin{aligned}\frac{d\mathrm{TPR(c)}}{d\mathrm{FPR(c)}} &= \lim_{\Delta c\to 0} \frac{\int_{c<m\left(\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}\right)<c+\Delta c} g_1(\boldsymbol{x})d\boldsymbol{x}}{\int_{c<m\left(\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}\right)<c+\Delta c} g_0(\boldsymbol{x})d\boldsymbol{x}} \\ &= \lim_{\Delta c\to 0} \frac{\int_{c<m\left(\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}\right)<c+\Delta c} m(c)^{-1} g_0(\boldsymbol{x})d\boldsymbol{x}}{\int_{c<m\left(\frac{g_1(\boldsymbol{x})}{g_0(\boldsymbol{x})}\right)<c+\Delta c} g_0(\boldsymbol{x})d\boldsymbol{x}} \\ &= m^{-1}(c).\end{aligned}$$

Since m is a monotonically increasing function, the inverse function m^{-1} is also monotonically increasing. Hence, the ROC curve for the optimal score function is always convex.

Chapter 4

PSA cut-off nomogram

This section is about the results of the cooperative work with Kent Kanao, Jun Nakashima, Takashi Ohigashi, Eiji Kikuchi, Akira Miyajima, Ken Nakagawa, Mototsugu Oya belonging to a department of urology of Keio University Hospital. The major part of this paper is written by Kent Kanao.

4.1 Introduction

Prostate-specific antigen (PSA) screening for prostate cancer is now widespread but the benefit of PSA screening is still controversial. Recently PSA-based screening has raised concerns that lead to overdetection and overtreatment of some patients (Lin *and others*, 2008). Several studies have shown that the rate of overdetection is increasing especially on elderly men with limited life expectancies (Walter *and others*, 2006; Stangelberger *and others*, 2008). Currently the American Urological Association (AUA) strongly supports that men be informed of the option of active surveillance in lieu of immediate treatment for men who diagnosed with clinically insignificant, since many screen-detected prostate cancers may not need immediate treatment (Greene *and others*, 2009). Therefore, especially for an elderly man who diagnosed with clinically insignificant prostate cancer and has the indication of active surveillance, immediate diagnosis by needle biopsy might be unnecessary.

Now most screening guidelines do not recommend PSA screening in elderly men because of the potential harms of screening (Thompson *and others*, 2007; Luboldt *and others*, 2004).

However, PSA screening is common among the elderly more than age 70, and there is not a little elderly patient with elevated PSA detected by PSA screening or others. AUA recommends the decision to prompt prostate biopsy should take into account multiple factors including free and total PSA, patient age, PSA velocity, PSA density, family history, ethnicity, prior biopsy history and comorbidities (Greene *and others*, 2009). But there is no obvious policy for these elderly men who should prompted prostate biopsy. Therefore it is very important to develop a methodology which can distinguish significant cancer efficiently from the elderly men with elevated PSA using multiple factors and to set an appropriate indication for prostate biopsy. However, the best methods to combine these factors most effectively have not yet been developed. For improved discrimination ability on diagnosis, it is necessary to use multiple factors effectively and get high sensitivity and specificity. To solve this problem, we use AUCBoost, which is the latest boosting algorithm, and we developed PSA cut-off nomogram that avoids overdetection of prostate cancer and decreases unnecessary biopsy in elderly men.

4.2 Methods

From 2004 to 2008, 400 patients over 70 years, with PSA levels of 20.0 ng/ml or less and normal digital rectal examination (DRE) had undergone prostate biopsies in our institute. All cases were diagnosed by systemic needle biopsy (10 cores or more) and graded histologically using the Gleason scoring system. All of the needle biopsy specimens were analyzed by pathologists with special interest in uropathology. The patients who were diagnosed as prostate cancer were divided into clinically significant cancer and insignificant cancer. Clinically significant cancers were defined as having more than two positive cores or Gleason sum of seven or higher and insignificant cancers were the other.

At first the distributions of PSA, F/T ratio, TZ volume and the number of biopsy sessions were estimated. Next receiver operating characteristics (ROC) analysis was performed and the area under the receiver operating characteristics curve (AUC) was used to assess the ability of PSA, F/T ratio, TZ volume and the number of biopsy sessions to discriminate significant cancer.

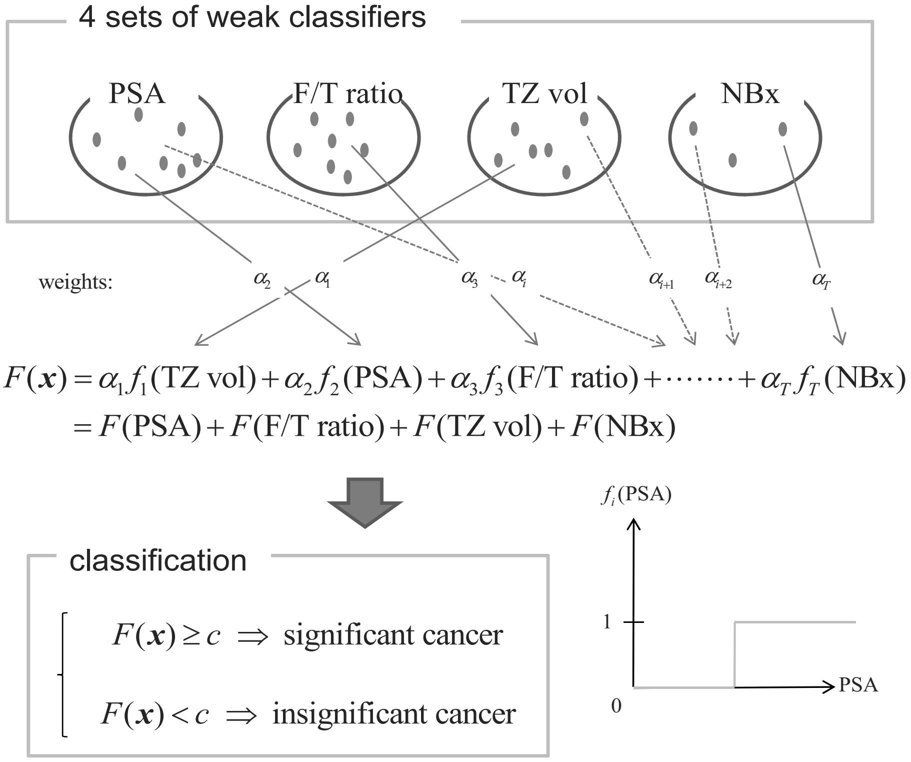

4 sets of weak classifiers
PSA
F/T ratio
TZ vol
NBx
weights:
α_2
α_1
α_3
α_i
α_{i+1}
α_{i+2}
α_T
$F(\boldsymbol{x}) = \alpha_1 f_1(\text{TZ vol}) + \alpha_2 f_2(\text{PSA}) + \alpha_3 f_3(\text{F/T ratio}) + \cdots\cdots + \alpha_T f_T(\text{NBx})$
$= F(\text{PSA}) + F(\text{F/T ratio}) + F(\text{TZ vol}) + F(\text{NBx})$
classification
$F(\boldsymbol{x}) \geq c \Rightarrow$ significant cancer
$F(\boldsymbol{x}) < c \Rightarrow$ insignificant cancer
f_i(PSA)
1
0
PSA

Figure 4.1: The sketch of the AUCBoost algorithm

Furthermore, AUCBoost was used to obtain the most effective combination of these markers. As illustrated in Figure 4.1, AUCBoost combines various weak classifiers $f(\boldsymbol{x})$'s ($\boldsymbol{x}$=PSA, F/T ratio, TZ.vol or Nbx) to produce a score function $F(\boldsymbol{x})$, by which we diagnosis a patient as having significant cancer or not, according to a cut-off value. The conventional and simplest one is $F(\boldsymbol{x}) = f(PSA)$ and the cut-off value is 4 ng/mL but it has low discrimination ability. The weak classifier itself is just a step function; however, the resulting score function has a flexible form so that it can achieve the optimal value of the AUC. The details of the way to calculate the values of the weights α's and to determine the number of the weak classifiers T are described in previous paper (Komori, 2009). Finally we developed nomogram which shows PSA cut-off values with ensuring 95% of sensitivity for significant cancer based on 200 bootstrap resampling repetitions. We used the median to show the results. All tests were carried out with R software version 2.9.0.

4.3 Results

One hundred ninety seven patients were diagnosed with prostate cancer by needle biopsy, 130 were diagnosed with significant and 67 were diagnosed with insignificant cancer. Patient characteristics were shown in Table 4.1. The distribution of PSA, F/T ratio, TZ volume and number of biopsy showed in Figure 4.2. As a result of ROC analysis discriminating significant cancer, the AUC of PSA, F/T ratio, TZ volume and number of biopsy sessions were 0.63, 0.73, 0.79 and 0.60, respectively.(Figure 4.3) Score plots of AUCBoost that depict the relationship between each variable and the type of the cancer (significant of not) were shown in Figure 4.4. The higher value means the stronger tendency to have significant cancer. The height of the score plots indicates the relative importance among the variables; hence, Tz.vol seems to be the most useful for discrimination. The AUC of combination model of PSA, F/T ratio, TZ volume and the number of biopsy by AUC boost was 0.86. PSA cut-off nomogram was developed using AUCBoost to obtain the PSA cut-off value determined by the other three values. (Table 4.5) By use of the nomogram, for example, the PSA cut-off value is 7.4 ng/ml for a 73 y/o man whose TZ volume is 28cc and F/T ratio is 0.20 on first biopsy session. The sensitivity and specificity of the nomogram was 0.95 and 0.45. By use of the nomogram and diagnostic algorithm shown in Figure 4, 122 patients (31%) may avoid prostate biopsy with 5% or less significant cancer overlooked.

4.4 Discussion

In this study we developed PSA cut-off nomogram that avoids overdetection of prostate cancer and decrease unnecessary biopsy in elderly men. Recently the ERSPC trial has demonstrated that PSA-based screening reduced the rate of death from prostate cancer by 20% but indicated that overdetection and overtreatment are probably the most important adverse effects of prostate cancer screening (Schroder *and others*, 2009). Although it is shown that the rate of overdetection is increasing in elderly men, clinically significant cancer with high grade and large volume is also included in such patients. Therefore, it is important to set an appropriate indication for prostate biopsy decreasing overlooked significant cancer as

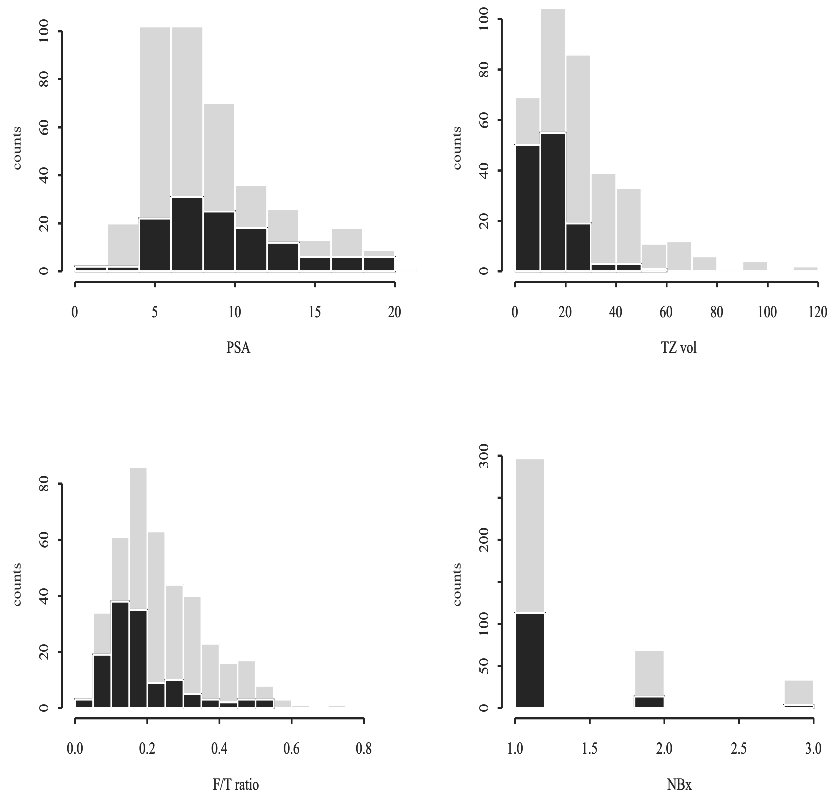


Figure 4.2: The distribution of PSA, F/T ratio, TZ volume and number of biopsy

well as overdetected insignificant cancer. This nomogram was developed for the purpose of decreasing unnecessary biopsy with 5% or less significant cancer overlooked.

In this study clinically significant cancers were defined as having more than two positive cores, Gleason sum of 7 or higher. There is currently no universally accepted definition of clinically significant or insignificant prostate cancer. The gold standard for insignificant disease used is ¡0.5mL of cancer with a Gleason score of 6 or less in the radical prostatectomy specimen (Epstein *and others*, 1994). However, the criteria using post treatment variables and it cannot be used for an informed discussion that might obviate unnecessary or aggressive therapy in certain patients. The most common clinical criterion of low-risk prostate cancer using pretreatment variables is defined as a Gleason score of 6 or less, PSA ¡ 10, and T1c to T2a (Thompson *and others*, 2007). More recently, investigators have shown that the number

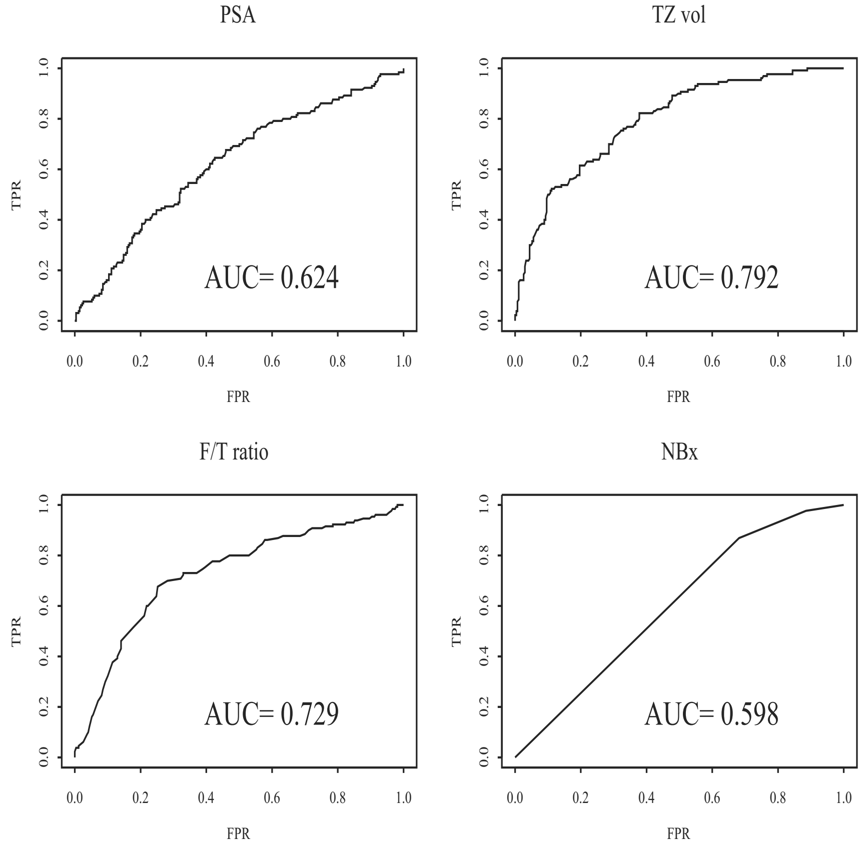


Figure 4.3: The AUC of PSA, F/T ratio, TZ volume and number of biopsy

of biopsies showing cancer may both be helpful in assessing the likelihood of insignificant disease (Antunes *and others*, 2005; Cheng *and others*, 2005; Ochiai *and others*, 2005). In general, active surveillance protocols attempt to identify men with low-risk prostate cancer who are most likely to be safely watched for a period of time and then treated when necessary (Dall'Era *and others*, 2008). Therefore, we define significant cancer using pretreatment variables on the assumption immediate diagnosis by needle biopsy may be unnecessary for elderly men who has the indication of active surveillance.

To develop this nomogram we used FT ratio, TZ volume and number of needle biopsy as well as PSA. In recent years, multiple variables have been taken into account in the risk-estimation for prostate cancer: free and total PSA, patient age, PSA velocity, PSA density, family history, ethnicity, prior biopsy history and many more. A recent studies showed

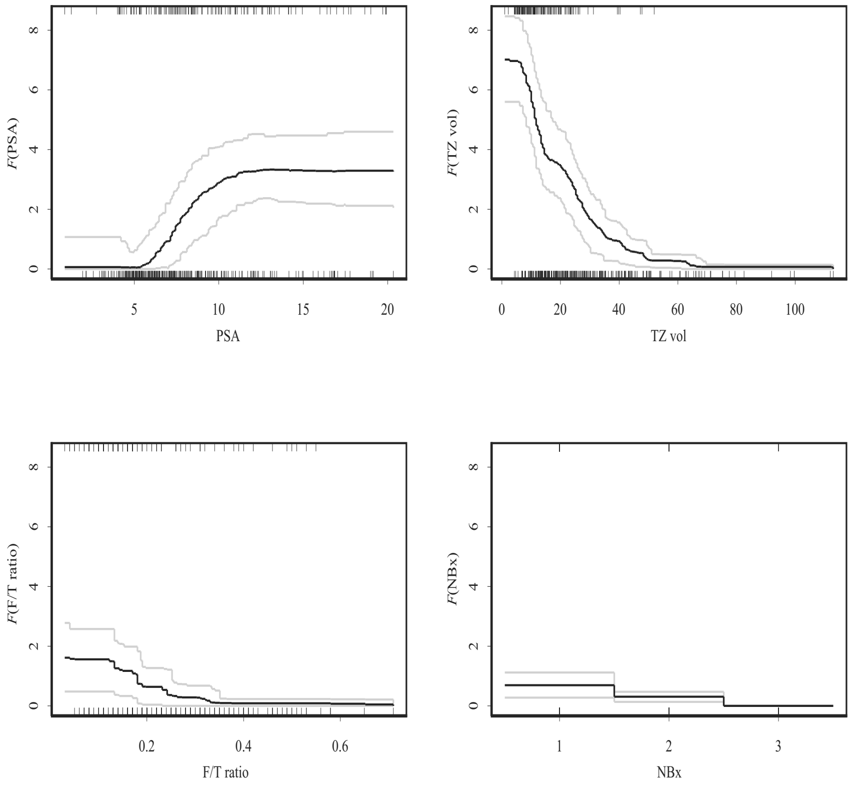


Figure 4.4: Score plots for PSA, F/T ratio, TZ volume and number of biopsy

free/total PSA ratio and PSATZD contribute more effectively as an adjunct to primary prostate screening with total PSA (Catalona *and others*, 1995; Catalona and Slawin, 1998). Additionally, prior biopsy history is also showed to have its predictive power. The predictive power of these variables to discriminate significant cancer is reflected by the AUC. In this study, the AUC of PSA, F/T ratio, TZ volume and number of biopsy sessions were 0.63, 0.73, 0.79 and 0.60, respectively. AUA states that the current policy no longer recommends a single, threshold value of PSA which should prompt prostate biopsy. The decision to proceed to prostate biopsy should be based primarily on PSA and DRE results but should take into account multiple factors. In this study, using AUCBoost these variables contributed to gain high AUC (0.86) in this model. The model might gain higher AUC if it added other variables: family history, PSA kinetics and more, which we did not access in this study.

In this study we used AUCBoost which is the latest boosting algorithm based on a boosting technique which is widely used in the machine learning community. AUCBoost is designed to optimize the AUC in ROC analysis using multiple variables. By use of AUCBoost, most effective combination of these variables can increase 45% of specificity with ensuring 95% of sensitivity for significant cancer. This is considerably higher than previous reports, for instance, Prospective multicenter European trial for patients with PSA levels between 4 and 10 ng/mL showed the specificity of PSA, FT ratio and PSATZD on 95% of sensitivity is 4.2%, 7.7% and 22.3%, respectively.

Nomograms are now considered to be accurate and practical tool for explaining predicted probabilities to patients and several nomograms have already been developed in the fields of urology (Partin and Lamm, 2001; Kattan *and others*, 1998; Kanao *and others*, 2006, 2009). These conventional nomograms were developed to show the prognostic probability not threshold value. In this study, by fixing the sensitivity to 95% and using AUCBoost, we can develop this nomogram which shows cut-off values varying according to the other values of variables. Therefore, this nomogram may be more useful for doing decision making than conventional nomogram. This nomogram take into account multiple variables to help determine the need for prostate biopsy of elderly men, rather than relying on an arbitrary threshold value, and this nomogram may be useful to facilitate discussion of a patient ’ s individualized risk.

In Figure 4.5 we show the diagnostic algorithm of prostate cancer for elderly men using this nomogram. By use of the algorithm 122 patients may avoid prostate biopsy with 5% or less significant cancer overlooked. Of course the determination of prostate biopsy depends on the individual doctor, but this nomogram informs a criterion that does not prompt immediate prostate biopsy.

In this study the sensitivity of this nomogram is fixed to 95%. Theoretically the sensitivity can change into 90% or 80% and it follows that the specificity increases. Now there is no consensus of diagnostic criteria on potential tradeoffs between sensitivity and specificity. Therefore, it become the judgment of the individual doctor how much allows overlook of significant cancer.

In general, the sensitivity and specificity of this nomogram depend on the patient population. Therefore, validation study is need for other population. Before use of this nomogram, it recommends testing ROC analysis and estimate sensitivity and specificity of this nomogram.

This nomogram is new concept nomogram in a point to use AUCBoost and show cut-off value. It is thought that this concept may be useful in the other various clinical fields. Although further validation is necessary to estimate the safety of this nomogram whether mortality increases for patients whose diagnosis are delay by use of this nomogram, it may be accepted a standard diagnostic tool for elderly men with elevated PSA.

4.5 Conclusion

This nomogram may be useful when urologists decide on an indication of prostate biopsy after trans-rectal ultrasonography and F/T ratio test for outpatients who are older than 70 years and have elevated PSA. This nomogram is different from conventional nomogram because it can show a cut-off value not a probability. Therefore, this nomogram may be more useful and practical for doing decision making than conventional nomogram.

Table 4.1: *The patient characteristics*

	(mean ± SD)
Patients	400
significant cancer	130
insignificant cancer	270
Age	74.5±3.8
PSA (ng/ml)	8.33±3.86
F/T ratio	0.24±0.12
TZ vol (cc)	24.9±18.5
Gleason score	
5 or less	54
6	46
7	73
8 or more	74
Number of NBx	
initial	297
2nd	69
3nd	34
Positive core	
1-2	98
3 or more	99

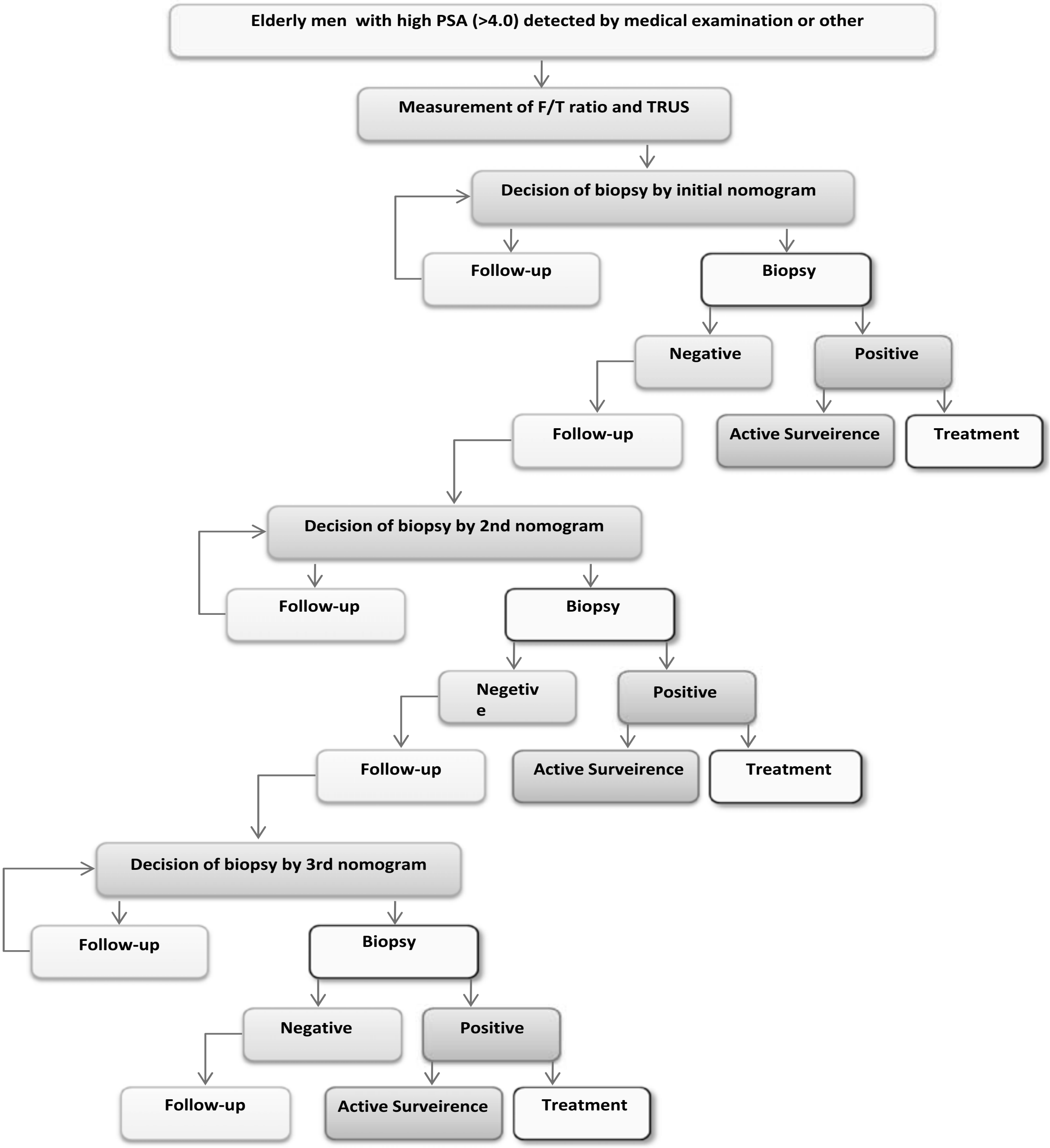

Elderly men with high PSA (>4.0) detected by medical examination or other
Measurement of F/T ratio and TRUS
Decision of biopsy by initial nomogram
Follow-up
Biopsy
Negative
Positive
Follow-up
Active Surveirence
Treatment
Decision of biopsy by 2nd nomogram
Follow-up
Biopsy
Negetiv
e
Positive
Follow-up
Active Surveirence
Treatment
Decision of biopsy by 3rd nomogram
Follow-up
Biopsy
Negative
Positive
Follow-up
Active Surveirence
Treatment

Figure 4.5: Diagnostic algorithm of detection of prostate cancer based on PSA-cut-off nomogram

		TZ vol (cc)										
		0≤	12≤	16≤	20≤	24≤	28≤	32≤	36≤	40≤	44≤	48≤
F/T ratio	0≤	4(4-4)	4(4-4)	4(4-4)	4(4-4)	4(4-6.3)	5.9(4-7.9)	6.8(4-8.8)	7.3(4-9.5)	7.5(4-9.8)	7.9(4-11.4)	8(4-11.8)
	0.04≤	4(4-4)	4(4-4)	4(4-4)	4(4-4.2)	4(4-6.5)	6.3(4-8.1)	7.1(4-8.8)	7.5(4-9.7)	7.8(4-10.4)	8.2(6.3-13.6)	8.3(6.3-13.6)
	0.12≤	4(4-4)	4(4-4)	4(4-4)	4(4-4.6)	4(4-6.5)	6.4(4-8.1)	7.2(4-8.8)	7.6(4-9.8)	7.8(4-10.4)	8.3(6.3-13.6)	8.4(6.3-13.6)
	0.16≤	4(4-4)	4(4-4)	4(4-4.7)	4(4-5.5)	4(4-7.1)	7(4-8.3)	7.6(5.8-9.4)	8.1(6.5-10.7)	8.4(6.7-11)	8.8(7.2-20)	9(7.3-20)
	0.2≤	4(4-4)	4(4-4)	4(4-6.2)	4(4-6.5)	6.5(4-7.5)	7.7(6.3-9.3)	8.4(7.2-10.2)	9(7.5-20)	9.4(7.5-20)	10(7.9-20)	10.2(8.1-20)
	0.24≤	4(4-4)	4(4-4)	4(4-6.6)	5(4-7.1)	6.7(4-8)	7.9(6.4-9.6)	8.6(7.3-11)	9.4(7.5-20)	9.7(7.7-20)	10.6(8.2-20)	11(8.2-20)
	0.28≤	4(4-4)	4(4-4.8)	4(4-6.7)	5.8(4-7.1)	7.1(4-8.3)	8.1(6.9-9.9)	8.8(7.3-12.4)	9.7(7.9-20)	10.2(8-20)	11.1(8.4-20)	12.1(8.4-20)
	0.32≤	4(4-4)	4(4-5.1)	4(4-7.1)	6(4-7.4)	7.2(4-8.6)	8.2(6.9-10.2)	9(7.3-17.3)	9.9(7.9-20)	10.4(8-20)	11.7(8.4-20)	12.6(8.4-20)
	0.36≤	4(4-4)	4(4-5.5)	5.4(4-7.4)	6.3(4-7.5)	7.3(4.9-8.8)	8.4(7.2-10.7)	9.4(7.5-20)	10.2(7.9-20)	10.8(8.1-20)	20(8.5-20)	20(8.7-20)
	0.4≤	4(4-4)	4(4-5.5)	5.4(4-7.4)	6.3(4-7.7)	7.3(4.9-8.8)	8.4(7.2-10.7)	9.4(7.5-20)	10.2(7.9-20)	10.8(8.1-20)	20(8.5-20)	20(8.7-20)

		TZ vol (cc)										
		0≤	12≤	16≤	20≤	24≤	28≤	32≤	36≤	40≤	44≤	48≤
F/T ratio	0≤	4(4-4)	4(4-4)	4(4-4)	4(4-5.8)	4(4-7.3)	6.6(4-8.3)	7.4(4-9.3)	7.7(4-10.6)	8(4-11.1)	8.4(6.4-20)	8.4(6.4-20)
	0.04≤	4(4-4)	4(4-4)	4(4-5.1)	4(4-6)	4(4-7.5)	7.1(4-8.5)	7.5(4.2-9.7)	8.1(4.2-11)	8.4(6.4-11.8)	8.6(6.9-20)	8.8(6.9-20)
	0.12≤	4(4-4)	4(4-4)	4(4-5.5)	4(4-6)	4(4-7.5)	7.1(4-8.5)	7.6(4.2-9.7)	8.1(4.2-11.6)	8.4(6.4-12)	8.7(6.9-20)	8.8(6.9-20)
	0.16≤	4(4-4)	4(4-4)	4(4-6.5)	4(4-6.3)	6.2(4-7.9)	7.4(4-9)	8.1(6.5-10)	8.6(7.3-12.3)	9(7.3-20)	9.6(7.3-20)	9.9(7.6-20)
	0.2≤	4(4-4)	4(4-4.9)	4(4-7.3)	6(4-7.3)	7.1(4-8.4)	8.1(6.9-10)	8.8(7.5-12.1)	9.9(7.9-20)	10.2(8.2-20)	11.4(8.7-20)	13.1(8.7-20)
	0.24≤	4(4-4)	4(4-6.1)	5.5(4-7.5)	6.3(4-7.9)	7.4(4-9)	8.4(7.3-11)	9.4(7.5-20)	10.4(8-20)	11.1(8.4-20)	20(8.9-20)	20(9-20)
	0.28≤	4(4-4)	4(4-6.2)	6.2(4-7.9)	6.6(4-7.9)	7.5(6.3-9)	8.8(7.4-11.4)	9.9(7.7-20)	11(8.4-20)	13.3(8.5-20)	20(9.5-20)	20(9.9-20)
	0.32≤	4(4-4)	4(4-6.5)	6.3(4-8.2)	6.7(4-8.1)	7.5(6.3-9.4)	8.8(7.4-12.6)	9.9(7.7-20)	11.4(8.4-20)	16.6(8.5-20)	20(9.5-20)	20(9.9-20)
	0.36≤	4(4-4)	4(4-6.7)	6.4(4-8.2)	7(4-8.4)	7.8(6.4-9.9)	9(7.4-20)	10.2(7.9-20)	12.1(8.5-20)	20(8.5-20)	20(9.5-20)	20(10-20)
	0.4≤	4(4-4)	4(4-6.8)	6.4(4-8.2)	7(4-8.4)	7.8(6.5-9.9)	9(7.4-20)	10.2(7.9-20)	12.8(8.5-20)	20(8.5-20)	20(9.5-20)	20(10-20)

		TZ vol (cc)										
		0≤	12≤	16≤	20≤	24≤	28≤	32≤	36≤	40≤	44≤	48≤
F/T ratio	0 ≤	4(4-4)	4(4-4)	4(4-6.5)	4(4-6.3)	5.3(4-7.9)	7.1(4-9)	7.7(4-10)	8.2(5.9-13.3)	8.4(6.3-20)	8.8(6.9-20)	9(6.9-20)
	0.04≤	4(4-4)	4(4-4)	4(4-6.7)	4(4-6.7)	6(4-8)	7.4(4-9.1)	8(5.3-10.2)	8.5(6.5-20)	8.7(6.8-20)	9.4(7.2-20)	9.4(7.3-20)
	0.12≤	4(4-4)	4(4-4)	4(4-6.7)	4(4-6.7)	6.1(4-8)	7.4(4-9.3)	8(5.3-10.2)	8.5(6.8-20)	8.8(6.9-20)	9.4(7.3-20)	9.6(7.3-20)
	0.16≤	4(4-4)	4(4-4)	4(4-7.2)	4.8(4-7.3)	6.7(4-8.4)	7.9(6.1-9.9)	8.5(7-11.8)	9.4(7.3-20)	9.8(7.5-20)	10.4(7.7-20)	10.7(8.1-20)
	0.2≤	4(4-4)	4(4-6.5)	5.8(4-7.9)	6.5(4-8.1)	7.4(5.7-9)	8.5(7.3-11.7)	9.6(7.5-20)	10.7(8.2-20)	11.8(8.4-20)	20(9.4-20)	20(9.6-20)
	0.24≤	4(4-4)	4(4-6.9)	6.3(4-8.2)	6.8(4-8.4)	7.6(6.3-9.9)	8.9(7.3-20)	10.2(8.1-20)	12.3(8.4-20)	20(8.8-20)	20(9.9-20)	20(10-20)
	0.28≤	4(4-4)	4(4-7.3)	6.6(4-8.4)	7.2(4-8.5)	8(6.6-9.9)	9.4(7.5-20)	10.6(8.3-20)	20(8.8-20)	20(9.3-20)	20(10.2-20)	20(10.4-20)
	0.32≤	4(4-4)	4(4-7.3)	6.8(4-8.8)	7.3(4-8.8)	8(6.6-10.2)	9.6(7.5-20)	11(8.3-20)	20(8.8-20)	20(9.3-20)	20(10.2-20)	20(10.4-20)
	0.36≤	4(4-4)	4(4-7.5)	7(4-8.8)	7.4(4.9-9)	8.2(7.1-10.7)	9.8(7.5-20)	11.8(8.4-20)	20(9-20)	20(9.5-20)	20(10.4-20)	20(10.4-20)
	0.4≤	4(4-4)	4(4-7.5)	7(4-8.8)	7.4(4.9-9)	8.2(7.1-11)	9.9(7.5-20)	11.9(8.4-20)	20(9-20)	20(9.5-20)	20(10.4-20)	20(10.4-20)

Chapter 5

A boosting method for maximizing the partial area under the ROC curve

Abstract

The receiver operating characteristic (ROC) curve has attracted wide attention for its utility in the medical and biostatistical fields. Given a set of multiple markers obtained from a clinical test or an examination, the area under the ROC curve (AUC) is measured for its ability to discriminate between the controls and cases. Recently, the partial area under the curve (pAUC) has been gaining in popularity, because the pAUC is more suitable for clinical settings in which a high true positive rate is required with a very low false positive rate. Moreover, the pAUC is more sensitive to the effects of markers in clinical evaluation, compared with the AUC, which is often criticized for not properly reflecting these effects. In this context, we have developed a new statistical method that focuses on the pAUC based on a boosting technique. The markers are combined componentially in the boosting algorithm using natural cubic splines or decision stumps (single-level decision trees), according to the types of markers used. We show that the resulting score plots are useful for understanding how each marker is associated with the outcome variable (affected or unaffected). We compare the performance of our boosting method with those of other existing methods, and demonstrate its utility using a real data set.

Keywords: Boosting; Classification; Partial area under the ROC curve; Smoothing.

5.1 Introduction

The receiver operating characteristic (ROC) curve has been widely used in various scientific fields, in situations where the evaluation of discrimination performance is of great concern for the researchers. The area under the ROC curve (AUC) is the most popular metric because it has a simple probabilistic interpretation (Bamber, 1975) and consists of two important rates used to asses classification performance: the true positive rate (TPR) and the false positive rate (FPR). The former is a probability of a affected subject being correctly judged as positive; the latter is that of a unaffected subject being improperly judged as positive. Since the two probabilities characterize different aspects of classification performance, they should be reported separately (Baker, 2003). Hence, the AUC has an advantage over a single measures of performance such as the odds ratio or relative risk (Pepe *and others*, 2004). However, the AUC has been severely criticized for inconsistencies arising between statistical significance derived from the AUC and the corresponding clinical significance when the usefulness of a new marker is evaluated (Cook, 2007). Recently, Pencina *and others* (2008) propose a criterion termed integrated discriminant improvement, and show its advantage over the AUC in the assessment of a new marker. In this context, the partial AUC (pAUC) has been gaining more popularity relative to the AUC in a number of fields (Walter, 2005; Qi *and others*, 2006).

Dodd and Pepe (2003) propose a regression modeling framework based on the pAUC, and apply this framework to investigation of a relationship between multiple markers and the outcome variable. Cai and Dodd (2008) make some modifications to improve the efficiency of the estimation of parameters, and provide graphical tools for the model checking. In regard to classification problems, Pepe and Thompson (2000) propose a method for deriving a linear combination of two markers that optimizes the AUC as well as the pAUC.

However, as recognized by Pepe *and others* (2006), more general approaches are required when the number of markers is quite large; in these cases, marker selection procedure is also indispensable.

In this paper, we propose a new statistical method designed to maximize the pAUC using a boosting technique and the approximate pAUC. The approximation-based method makes it possible to nonlinearly combine more than two markers, based on basis functions of natural cubic splines as well as decision stumps. The resultant score plots for each marker enable us to observe how the markers are associated with the outcome variable in a visually apparent way. Hence, our boosting method attaches importance not only to the classification performance but also to the interpretation of the results, which is essential in clinical and medical fields.

This paper is organized as follows. In Section 2, we give a brief review of the AUC and pAUC, and show a relationship between the pAUC and the approximate pAUC in Theorem 5.2.1. We present a new boosting method, pAUCBoost, in Section 3, and compare it with other existing methods such as SDF (Pepe and Thompson, 2000), AdaBoost (Freund and Schapire, 1997), LogitBoost (Friedman *and others*, 2000) and GAMBoost (Tutz and Binder, 2006) in Section 4. In the next section, we demonstrate the utility of pAUCBoost using a breast cancer data set, in which we use both clinical and genomic data. In Section 6, we summarize and make concluding remarks.

5.2 pAUC and approximate pAUC

5.2.1 Partial area under the ROC curve

Let y denote a class label for cases ($y = 1$) and controls ($y = 0$), and $\boldsymbol{x}$ be a vector of markers as $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$. Given a score function $F(\boldsymbol{x})$ and a threshold c, we judge the subject as positive if $F(\boldsymbol{x}) \geq c$, and as negative if $F(\boldsymbol{x}) < c$. The corresponding false positive rate (FPR) and true positive rate (TPR) are given as

$$\mathrm{FPR}(c) = \int \mathrm{H}(F(\boldsymbol{x}) - c) g_0(\boldsymbol{x}) d\boldsymbol{x}, \ \mathrm{TPR}(c) = \int \mathrm{H}(F(\boldsymbol{x}) - c) g_1(\boldsymbol{x}) d\boldsymbol{x}, \tag{5.2.1}$$

where H is the Heaviside function: $\mathrm{H}(z) = 1$ if $z \geq 0$ and 0 otherwise, and $g_0(\boldsymbol{x})$ and $g_1(\boldsymbol{x})$ are probability density functions for each class. Note that FPR and TPR are also dependent on the score function F; however, for the sake of simplicity, we abbreviate it when the abbreviation does not cause ambiguity.

Then, the ROC curve is defined as a plot of TPR against FPR when the threshold c moves on a real number line:

$$\mathrm{ROC}(F) = \{(\mathrm{FPR}(c), \mathrm{TPR}(c)) |\ c \in \mathbf{R}\}, \tag{5.2.2}$$

and the area under the ROC (AUC) is given as

$$\mathrm{AUC}(F) = \int_{\infty}^{-\infty} \mathrm{TPR}(c) d\mathrm{FPR}(c). \tag{5.2.3}$$

In this setting, we consider a part of the AUC by limiting the value of FPR between α_1 and α_2, which are determined by thresholds c_1 and c_2, respectively:

$$\alpha_1 = \int \mathrm{H}(F(\boldsymbol{x}) - c_1) g_0(\boldsymbol{x}) d\boldsymbol{x},\ \ \alpha_2 = \int \mathrm{H}(F(\boldsymbol{x}) - c_2) g_0(\boldsymbol{x}) d\boldsymbol{x}, \tag{5.2.4}$$

where $0 \leq \alpha_1 < \alpha_2 \leq 1$ $(c_2 < c_1)$. In this paper, we set the values to be 0 and 0.1, respectively. However, it is also worth considering to take $\alpha_1 > 0$ and choose $\alpha_2 - \alpha_1$ to be small enough, so that we essentially maximize TPR for fixed FPR. Then, the pAUC can be divided into a fan-shaped part and a rectangular part:

$$\begin{aligned} \mathrm{pAUC}(F, \alpha_1, \alpha_2) &= \int_{c_1}^{c_2} \mathrm{TPR}(c) d\mathrm{FPR}(c) & (5.2.5) \\ &= \int_{c_1}^{c_2} \int_{c_2 \leq F(\boldsymbol{x}) \leq c_1} \mathrm{H}(F(\boldsymbol{x}) - c) g_1(\boldsymbol{x}) d\boldsymbol{x} d\mathrm{FPR}(c) + \mathrm{TPR}(c_1)(\alpha_2 - \alpha_1) & (5.2.6) \end{aligned}$$

Its probabilistic interpretation is offered by Pepe (2003) as

$$\mathrm{pAUC}(F, \alpha_1, \alpha_2) = P(F(\boldsymbol{X}_1) \geq F(\boldsymbol{X}_0) \mid c_2 \leq F(\boldsymbol{X}_0) \leq c_1). \tag{5.2.7}$$

This means that the observation of $F(\boldsymbol{X}_1)$ is correctly ordered above that of $F(\boldsymbol{X}_0)$, on the condition that the FPR is between α_1 and α_2. Given samples from class 0 $\{\boldsymbol{x}_{0i} : i = 1, 2, \ldots, n_0\}$ and class 1 $\{\boldsymbol{x}_{1j} : j = 1, 2, \ldots, n_1\}$, its empirical form is expressed as

$$\overline{\text{pAUC}}(F, \overline{\alpha}_1, \overline{\alpha}_2) = \frac{1}{n_0 n_1} \sum_{i \in I} \sum_{j=1}^{n_1} \text{H}(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})), \tag{5.2.8}$$

where $\overline{\alpha}_1$ and $\overline{\alpha}_2$ are empirical values that are the closest to α_1 and α_2, respectively; $I = \{i |\ \overline{c}_2 \leq F(\boldsymbol{x}_{0i}) \leq \overline{c}_1\}$, where $\overline{c}_1$ and $\overline{c}_2$ are thresholds determined by $\overline{\alpha}_1$ and $\overline{\alpha}_2$.

5.2.2 Approximate pAUC

As seen in (5.2.8), the empirical pAUC is non-differentiable. Eguchi and Copas (2002) use the standard normal distribution function to approximate the AUC, and applied algorithms in order to maximize the AUC. We extended the method to a new one for maximizing the pAUC, using the approximate pAUC:

$$\text{pAUC}_\sigma(F, \alpha_1, \alpha_2) = \int_{c_1}^{c_2} \int_{c_2 \leq F(\boldsymbol{x}) \leq c_1} \text{H}_\sigma(F(\boldsymbol{x}) - c) g_1(\boldsymbol{x}) d\boldsymbol{x} d\text{FPR}(c) + \text{TPR}(c_1)(\alpha_2 - \alpha_1), \tag{5.2.9}$$

where α_1 and α_2 are defined in (5.2.4), and $\text{H}_\sigma(z)$ is an approximation of $\text{H}(z)$ by the standard normal distribution function, that is, $\text{H}_\sigma(z) = \Phi(z/\sigma)$. Similarly, the corresponding empirical pAUC is defined as

$$\overline{\text{pAUC}}_\sigma(F, \overline{\alpha}_1, \overline{\alpha}_2) = \frac{1}{n_0 n_1} \sum_{i \in I} \Big\{ \sum_{j \in J_{\text{fan}}} \text{H}_\sigma(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) + \sum_{j \in J_{\text{rec}}} \text{H}(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) \Big\}, \tag{5.2.10}$$

where $J_{\text{fan}} = \{j |\ \overline{c}_2 \leq F(\boldsymbol{x}_{1i}) \leq \overline{c}_1\}$ and $J_{\text{rec}} = \{j |\ \overline{c}_1 < F(\boldsymbol{x}_{1i})\}$.

A smaller scale parameter σ implies a better approximation of $\text{H}(z)$. Before discussing a boosting method for the pAUC, we give a theoretical justification of the use of the approximate pAUC in the following theorem.

Theorem 5.2.1. *For a pair of fixed α_1 and α_2, let*

$$\Psi(\gamma) = \text{pAUC}_\sigma\Big(F + \gamma\, m\big(\Lambda\big), \alpha_1, \alpha_2\Big), \tag{5.2.11}$$

where γ is a scalar, $\Lambda(\boldsymbol{x}) = g_1(\boldsymbol{x})/g_0(\boldsymbol{x})$ and m is a strictly increasing function. Then, $\Psi(\gamma)$ is a strictly increasing function of γ, and

$$\sup_F \mathrm{pAUC}_\sigma(F, \alpha_1, \alpha_2) = \lim_{\gamma \to \infty} \Psi(\gamma) = \mathrm{pAUC}\big(\Lambda, \alpha_1, \alpha_2\big). \tag{5.2.12}$$

As proved by Eguchi and Copas (2002) and McIntosh and Pepe (2002), the likelihood ratio $\Lambda(\boldsymbol{x})$ is the optimal score function that maximizes the AUC as well as the pAUC. However, the approximate pAUC does not exactly share the same property as the pAUC, because the $\mathrm{pAUC}_\sigma(F)$ is *not* a concave functional with respect to F. In general the Bayes risk consistency has been well discussed under an assumption of convexity for a variety of loss functions (Lugosi and Vayatis, 2004). Theorem 5.2.1 suggests a weak version of the Bayes risk consistency in the limiting sense.

We also have a following corollary from Theorem 5.2.1.

Corollary 5.2.1. *For any score function of F, let*

$$F_\gamma(\boldsymbol{x}) = F(\boldsymbol{x}) + \gamma\, \eta(\boldsymbol{x}), \tag{5.2.13}$$

where η is a score function, and γ is a scalar. If the FPR of F_γ is fixed to α, then the TPR of F_γ is a increasing function of γ if and only if $\eta = m\big(\Lambda\big)$, where m is a strictly increasing function.

See Appendix 2 for the proof of Corollary 5.2.1. Note that the corollary holds for any α in the range of (0,1); hence, we find that the score function that moves every and all TPR's upward from the original positions, is nothing but the optimal score function derived from likelihood. This fact is not derived from the Neyman-Pearson fundamental lemma (Neyman and Pearson, 1933), from which $m(\Lambda)$ is proved to maximize the AUC. This corollary characterizes another property of the optimal score function.

5.3 pAUCBoost with natural cubic splines

5.3.1 Objective function

We construct a score function $F(\boldsymbol{x})$ in an additive model for the maximization of the pAUC. The set of weak classifiers that we use here consists of component basis functions for representing natural cubic splines, and their standardization factors:

$$\mathcal{F} = \{f(\boldsymbol{x}) = N_{k,l}(x_k)/Z_{k,l} |\ k = 1, 2, \ldots, p,\ l = 1, 2, \ldots, m_k\}. \tag{5.3.1}$$

The basis functions of the natural cubic spline for x_k are defined as

$$N_{k,l}(x_k) = \begin{cases} 1,\ l = 1, \\ x_k,\ l = 2, \\ d_{l-2}(x_k) - d_{m_k-1}(x_k),\ \text{otherwise}, \end{cases} \tag{5.3.2}$$

where

$$d_l(x_k) = \frac{(x_k - \xi_{k,l-2})_+^3 - (x_k - \xi_{k,m_k})_+^3}{\xi_{k,m_k} - \xi_{k,l-2}}, \tag{5.3.3}$$

and z_+ denotes the positive part of z. The standardization factor $Z_{k,l}$ for $N_{k,l}(x_k)$ is given as

$$Z_{k,l} = \begin{cases} 1,\ l = 1, \\ \xi_{k,m_k} - \xi_{k,1},\ l = 2, \\ N_{k,l}(\xi_{k,m_k}) - N_{k,l}(\xi_{k,l-2}),\ \text{otherwise}, \end{cases} \tag{5.3.4}$$

and $\xi_{k,l}$ is one of m_k knots $(\xi_{k,1} < \xi_{k,2} < \ldots < \xi_{k,m_k})$ for x_k. The knots are set to the observed values of x_k or the quantiles depending on the sample size and the number of the components of $\boldsymbol{x}$. We take a moderate number of quantiles for computational cost.

Based on the weak classifiers above, the objective function we propose is given as

$$\begin{aligned}\overline{\mathrm{pAUC}}_{\sigma,\lambda}(F,\overline{\alpha}_1,\overline{\alpha}_2) &= \frac{1}{n_0 n_1}\sum_{i\in I}\Big\{\sum_{j\in J_{\mathrm{fan}}}\mathrm{H}_\sigma(F(\boldsymbol{x}_{1j})-F(\boldsymbol{x}_{0i}))+\sum_{j\in J_{\mathrm{rec}}}\mathrm{H}(F(\boldsymbol{x}_{1j})-F(\boldsymbol{x}_{0i}))\Big\}\\ &\quad -\lambda\sum_{k=1}^{p}\int\big\{F_k''(x_k)\big\}^2dx_k, \qquad (5.3.5)\end{aligned}$$

where $F_k''(x_k)$ is the second derivative of the k-th component of $F(\boldsymbol{x})$, and λ is a smoothing parameter that controls the smoothness of $F(\boldsymbol{x})$. The penalty term prevents $F(\boldsymbol{x})$ from overfitting to the data, as well as ensures the existence of the maximum of the objective function. Hence, it leads to a numerically-stable maximization procedure introduced in the next subsection.

On the other hand, we have

$$\begin{aligned}&\overline{\mathrm{pAUC}}_{\sigma',\lambda'}\Big(\frac{\sigma'}{\sigma}F,\overline{\alpha}_1,\overline{\alpha}_2\Big)\\ =\;& \frac{1}{n_0 n_1}\sum_{i\in I}\left\{\sum_{j\in J_{\mathrm{fan}}}\mathrm{H}_{\sigma'}\Big(\frac{\sigma'}{\sigma}(F(\boldsymbol{x}_{1j})-F(\boldsymbol{x}_{0i}))\Big)+\sum_{j\in J_{\mathrm{rec}}}\mathrm{H}\Big(\frac{\sigma'}{\sigma}(F(\boldsymbol{x}_{1j})-F(\boldsymbol{x}_{0i}))\Big)\right\}\\ &\quad -\lambda'\sum_{k=1}^{p}\int\Big\{\frac{\sigma'}{\sigma}F_k''(x_k)\Big\}^2dx_k \qquad (5.3.6)\\ =\;& \frac{1}{n_0 n_1}\sum_{i\in I}\Big\{\sum_{j\in J_{\mathrm{fan}}}\mathrm{H}_\sigma(F(\boldsymbol{x}_{1j})-F(\boldsymbol{x}_{0i}))+\sum_{j\in J_{\mathrm{rec}}}\mathrm{H}(F(\boldsymbol{x}_{1j})-F(\boldsymbol{x}_{0i}))\Big\}\\ &\quad -\lambda'\frac{\sigma'^2}{\sigma^2}\sum_{k=1}^{p}\int\big\{F_k''(x_k)\big\}^2dx_k. \qquad (5.3.7)\end{aligned}$$

Note that $\overline{c}_1$ and $\overline{c}_2$ that are determined by $\overline{\alpha}_1$ and $\overline{\alpha}_2$ become $\sigma'/\sigma\,\overline{c}_1$ and $\sigma'/\sigma\,\overline{c}_2$ in $\overline{\mathrm{pAUC}}_{\sigma',\lambda'}\Big(\sigma'/\sigma\,F,\overline{\alpha}_1,\overline{\alpha}_2\Big)$; however, the sets of I, J_{fan} and J_{rec} remain unchanged. Hence, we have

$$\overline{\mathrm{pAUC}}_{\sigma,\lambda}(F,\overline{\alpha}_1,\overline{\alpha}_2)=\overline{\mathrm{pAUC}}_{\sigma',\lambda'}\Big(\frac{\sigma'}{\sigma}F,\overline{\alpha}_1,\overline{\alpha}_2\Big), \qquad (5.3.8)$$

if $\lambda\sigma^2=\lambda'\sigma'^2$. This implies that the maximization of $\overline{\mathrm{pAUC}}_{\sigma,\lambda}(F,\overline{\alpha}_1,\overline{\alpha}_2)$ is equivalent to that of $\overline{\text{pAUC}}_{1,\lambda\sigma^2}(F/\sigma, \overline{\alpha}_1, \overline{\alpha}_2)$. Therefore, we have

$$\max_{\sigma,\lambda,F} \overline{\text{pAUC}}_{\sigma,\lambda}(F, \overline{\alpha}_1, \overline{\alpha}_2) = \max_{\lambda,F} \overline{\text{pAUC}}_{1,\lambda}(F, \overline{\alpha}_1, \overline{\alpha}_2), \tag{5.3.9}$$

for any $\sigma > 0$. We remark that the scale parameter σ in the definition of $\overline{\text{pAUC}}_{\sigma,\lambda}$ in (5.3.5) can be fixed to 1 because of equation (5.3.9). Hence, we rewrite the objective function as:

$$\begin{aligned}\overline{\text{pAUC}}_{\lambda}(F, \overline{\alpha}_1, \overline{\alpha}_2) &= \frac{1}{n_0 n_1} \sum_{i \in I} \Big\{ \sum_{j \in J_{\text{fan}}} \Phi(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) + \sum_{j \in J_{\text{rec}}} \text{H}(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i})) \Big\} \\ &\quad -\lambda \sum_{k=1}^{p} \int \big\{ F_k''(x_k) \big\}^2 dx_k, \end{aligned} \tag{5.3.10}$$

without loss of generality, where Φ is the standard normal distribution function.

Note that there exists the maximum value of $\overline{\text{pAUC}}_{\lambda}(F, \overline{\alpha}_1, \overline{\alpha}_2)$ if the penalty term is not zero. The maximum value that is attained by p sets of differential functions can take the larger value by replacing the functions with p sets of natural cubic splines. This can be proved in the same way as the result of generalized additive models (Hastie and Tibshirani, 1990), because the penalty term is the same. Hence, we find that the maximizer of the pAUCBoost objective function is the natural cubic splines.

5.3.2 pAUCBoost algorithm

Here is a boosting algorithm with iteration time T, which is designed to maximize the pAUC.

1. Start with a score function $F_0(\boldsymbol{x}) = 0$ and set each coefficient $\beta_0(f)$ of weak classifiers to be 1 or -1.
2. For $t = 1, ..., T$
   (a) Calculate the values of thresholds $\overline{c}_1$ and $\overline{c}_2$ for each $F_{t-1} + \beta_{t-1}(f)f$.
   (b) Update $\beta_{t-1}(f)$ to $\beta_t(f)$ with a one-step Newton-Raphson iteration.

(c) Find the best weak classifier f_t

$$f_t = \underset{f}{\mathrm{argmax}}\ \overline{\mathrm{pAUC}}_\lambda(F_{t-1} + \beta_t(f)f, \overline{\alpha}_1, \overline{\alpha}_2) \tag{5.3.11}$$

(d) Update the score function as

$$F_t(\boldsymbol{x}) = F_{t-1}(\boldsymbol{x}) + \beta_t(f_t)f_t(\boldsymbol{x}). \tag{5.3.12}$$

3. Finally, output a final score function $F(\boldsymbol{x}) = \sum_{t=1}^{T} \beta_t(f_t)f_t(\boldsymbol{x})$.

The dependency of the $\overline{\mathrm{pAUC}}_\lambda(F_{t-1} + \beta_t(f)f, \overline{\alpha}_1, \overline{\alpha}_2)$ on thresholds $\overline{c}_1$ and $\overline{c}_2$ makes it necessary to pick up the best pair of $(\beta_t(f_t), f_t)$ at the same time in step 2.c. This process is quite different from that of AdaBoost, in which β_t and f_t are determined independently. Because of the dependency and the difficulty of getting the exact solution of $\beta_t(f_t)$, the one-step Newton-Raphson calculation is repeated during the whole boosting process. This is based on a natural assumption that the one-step previous situation in the boosting algorithm is not so different from the current one, because f_t is assumed to be a weak classifier. The pAUCBoost algorithm with natural cubic splines is detailed in Appendix 3.

5.3.3 Tuning procedure

We conduct K-fold cross validation to determine the smoothing parameter λ and the iteration number T. We divide the whole data into K subsets, and calculate the following objective function.

$$\mathrm{pAUC}_{\mathrm{CV}}(\lambda, T) = \frac{1}{K}\sum_{i=1}^{K} \overline{\mathrm{pAUC}}_\lambda^{(i)}(F^{(-i)}, \overline{\alpha}_1, \overline{\alpha}_2), \tag{5.3.13}$$

where $F^{(-i)}$ denotes a score function that is generated by the data without i-th subset; $\overline{\mathrm{pAUC}}_\lambda^{(i)}$ is $\overline{\mathrm{pAUC}}_\lambda$ calculated by the i-th subset only. The optimal parameters are obtained at the maximum value of $\mathrm{pAUC}_{\mathrm{CV}}(\lambda,\ T)$ in a set of grid points $(\lambda,\ T)$. In the case where the values of the $\mathrm{pAUC}_{\mathrm{CV}}(\lambda, T)$ are unstable, we calculate the $\mathrm{pAUC}_{\mathrm{CV}}(\lambda, T)$ 10 times and take the average to determine the optimal parameters. In our subsequent discussion, we set

$K = 10$ and explicitly demonstrate the procedure in the section regarding real data analysis.

5.4 Simulation studies

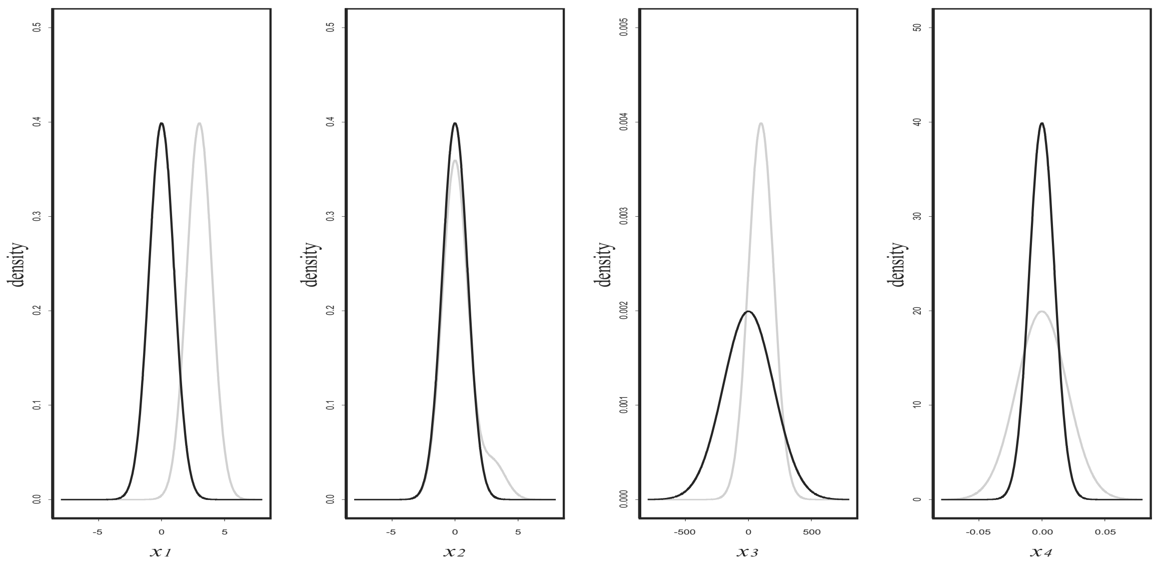


Figure 5.1: Illustration of four different type of sample distributions for class 0 (black) and class 1 (gray).

5.4.1 Setting

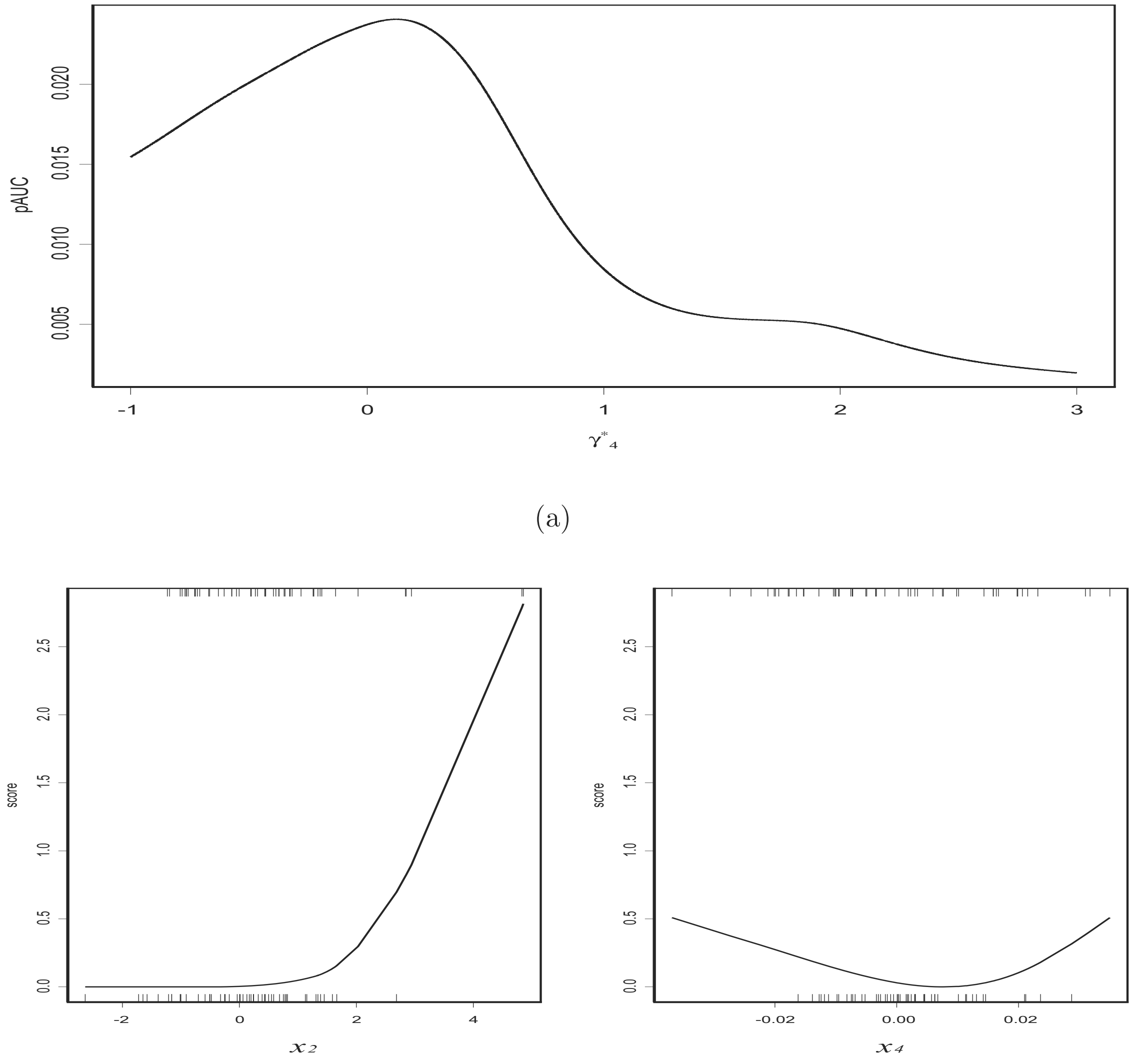
In this section, we compare the performance of pAUCBoost with that of the smooth distribution-free (SDF) method proposed by Pepe and Thompson (2000) in a two-dimensional setting, and with those of other existing boosting methods: AdaBoost, LogitBoost and GAMBoost in a higher-dimensional setting. The simulation setting is similar to that of Pepe *and others* (2003). Suppose that there are four types of sample distributions for each class, $y = 0$ or $y = 1$, as shown in Figure 1. The first panel shows an ideal situation, where we see very little overlap between the two class-conditional distributions. The second situation is of practical interest for disease screening, where FPR must be restricted to be as small as possible, in a case where invasive or costly diagnostic treatments will follow. A small portion of samples from class 0 (controls) is clearly distinguishable from the bulk of samples from class 1 (cases). On the other hand, in the third situation, cases are completely within the range of controls, and therefore not useful for disease screening. The fourth situation is similar to the second one, but some of the samples from cases deviate from controls clearly on both side of the distribution, rather than only on one side. This situation could be worth consideration in a case where high TPR is required with very low FPR in the same way as in the second situation.

In the simulation study, we apply pAUCBoost with $\overline{\alpha}_1 = 0$ and $\overline{\alpha}_2 = 0.1$. The training data set contains 50 controls and 50 cases; the accuracy of the performance is evaluated using a test data set of size 1000 (500 for each class). The results are based on 100 repetitions.

5.4.2 Comparison with SDF

First, we consider the second situation, where we assume normality distributions such as $X_{20} \sim \mathcal{N}(0, 1)$ and $X_{21} \sim \pi\mathcal{N}(0, 1) + (1 - \pi)\mathcal{N}(3, 1)$ with mixing proportion $\pi = 0.9$, and the last situation: $X_{40} \sim \mathcal{N}(0, 1/100)$, $X_{41} \sim \mathcal{N}(0, 4/100)$. The mean value (and the 95 percent confidence interval) of the pAUC based on pAUCBoost is 0.017 (0.012, 0.020); that of SDF is 0.011 (0.005, 0.017). This difference is because SDF assumes linearity of the score function of $F(\boldsymbol{x})$, and the coefficient of x_4, say γ_4, is estimated by SDF to be around 0 as shown in Figure 2 (a), where the coefficient of x_2 is fixed to 1. On the other hand, pAUCBoost captures the nonlinearity of $F(\boldsymbol{x})$ as shown in Figure 2 (b), where the score plot is a component function of $F(\boldsymbol{x})$ for each marker. Although we should first apply a nonlinear transformation of x_4 in this example, it is not practical to examine all marginal distributions and decide the appropriate transformations in general situations.

We have also confirmed that the performance of pAUCBoost is compatible with that of SDF, in a setting where the linearity of the score function is reasonable. We have an average of 0.013 (0.007, 0.017), and 0.013 (0.011, 0.015) for pAUCBoost and the SDF method, respectively, assuming that both of the markers are distributed as $X_{20} \sim \mathcal{N}(0, 1)$ and $X_{21} \sim \pi\mathcal{N}(0, 1) + (1 - \pi)\mathcal{N}(3, 1)$ for controls and cases. Almost the same results are obtained by these quite different statistical methods. SDF uses the estimated values of pAUC to derive a score function; on the other hand, pAUCBoost directly uses the empirical value of the approximate pAUC in its algorithm.



(a)

(b)

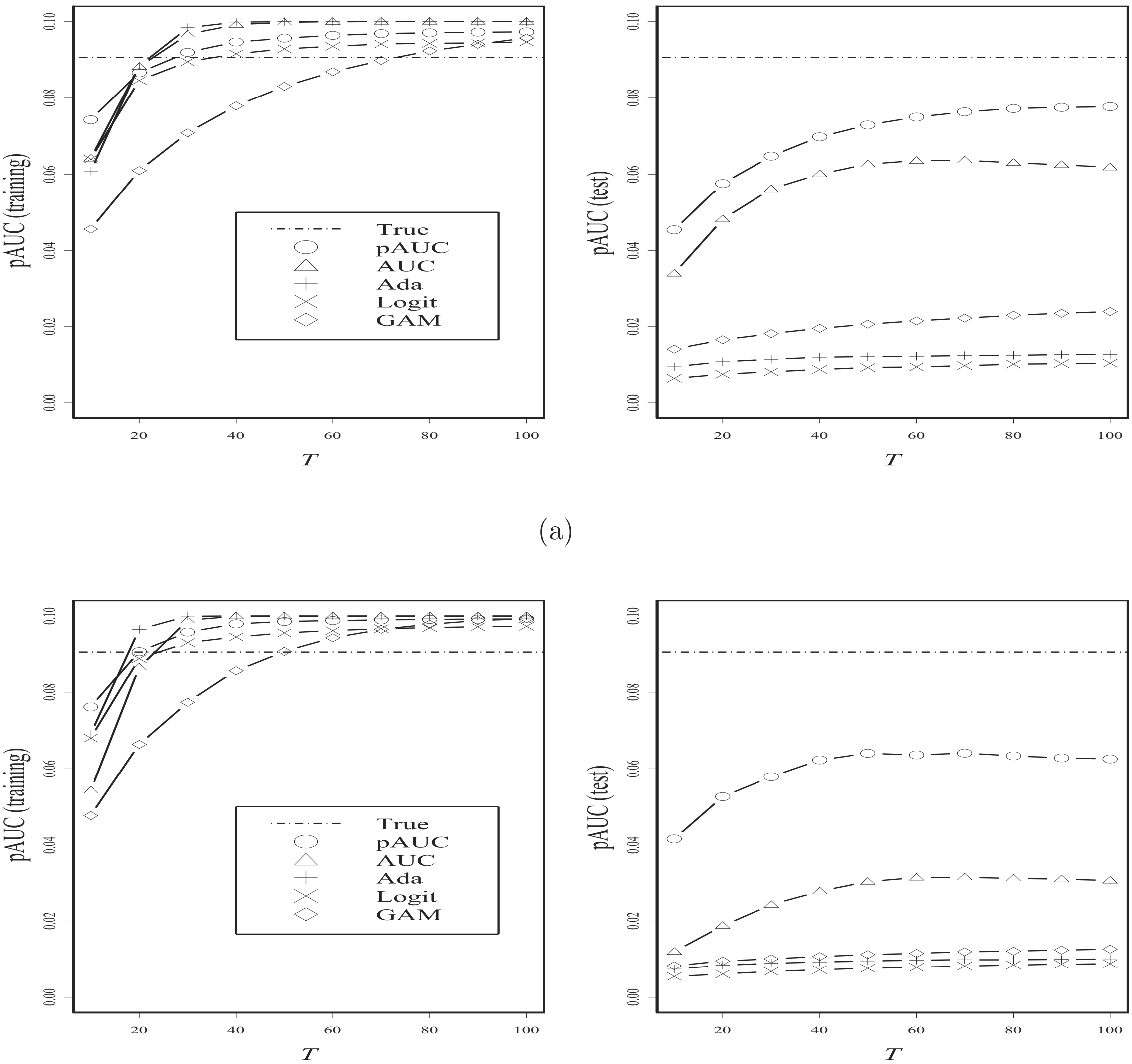
Figure 5.2: (a) Illustration of the estimated value of pAUC by SDF method, where $\gamma_4^* = \gamma_4$ if $-1 \leq \gamma_4 \leq 1$ and $2 - 1/\gamma_4$ otherwise; (b) the resultant score plots by pAUCBoost. The rug plot along the bottom of each graph describes the observation from class 0; the rug plots along the top of each graph describe those that class 0.

5.4.3 Comparison with other boosting methods

Second, we focus on only the most practical situation in disease screening: the second situation in Figure 1. Pepe *and others* (2003) show the utility of the use of the pAUC, in selection of potential genes that are useful for discrimination between normal and cancer tissues. The point is that the value of pAUC reflects the overlap of two distributions of controls and cases, so that we can select genes that are suitable for the purpose of further investigation. For example, some overexpressed genes encourage us to investigate the corresponding protein products. However, the task of how to combine the selected genes for better discrimination is still pending.

Suppose there are 50 independent genes that are informative in the sense of the pAUC, such that $X_{20}^{(i)} \sim \mathcal{N}(0,1)$ and $X_{21}^{(i)} \sim \pi\mathcal{N}(0,1) + (1-\pi)\mathcal{N}(3,1)$, where $\pi = 0.9$ ($i = 1, 2, \ldots, 50$). The assumption of the independence is mainly for simplicity and for making the comparison of pAUCBoost with existing methods clearer. The performance of pAUCBoost in a more realistic situation is demonstrated in Section 6. Figure 3 (a) shows plots of the average of the pAUC against iteration number T for five boosting methods. For all the boosting methods, the values of the pAUC based on the training data reach the upper bound values 0.1 after a number of iterations; however, the values based on the test data show clear differences. The pAUCBoost properly detects the small difference of the two distributions illustrated in the second panel in Figure 1, and shows the best performance. That is, it is the closest to the value of the true pAUC of this setting. On the other hand, AdaBoost, LogitBoost and GAMBoost cannot distinguish the two groups at all. The performance of AUCBoost (pAUCBoost with $\overline{\alpha}_1 = 0$ and $\overline{\alpha}_2 = 1$) is between the two extremes.

Next, we added some noninformative genes to the 50 genes above, i.e., genes that are assumed to be distributed uniformly: $X_{20}^{(i)},\ X_{21}^{(i)} \sim \mathcal{U}(-3,3)$, ($i = 51, \ldots, 100$). The results in the left panel in Figure 3 (b) are the almost the same as those in (a); however, we can find a clear difference between the right panels. The performance of AUCBoost goes down on a large scale, measured by the value of pAUC. This is mainly because of “false discovery”, or selection of noninformative genes by chance. Figure 4 shows the resistance of pAUCBoost to false discovery. The horizontal axis denotes the identification number of genes ($i = 1, \dots, 100$), and the vertical axis is the average number of selected genes, which correspond to the selected spline components, in the five boosting methods during the iteration process with $T = 100$. The total number of genes we use here is 100, so if a boosting method selects informative and noninformative genes equally, the average number is 1 over all genes. On the other hand, if the boosting method selects only the informative genes, the average numbers are expected to be 2 and 0 for informative ($i = 1, \dots, 50$) and noninformative ($i = 51, \dots, 100$) genes, respectively. The boosting methods other than pAUCBoost clearly suffer from false discovery. pAUCBoost has an advantage because it focuses on the essential part of the sample distribution in the sense of the pAUC.



(a)

(b)

Figure 5.3: (a) The results of the pAUC with FPR between 0 and 0.1 for training data (left panel) and test data (right panel) with only informative genes; (b) the results of the pAUC with noninformative genes added.

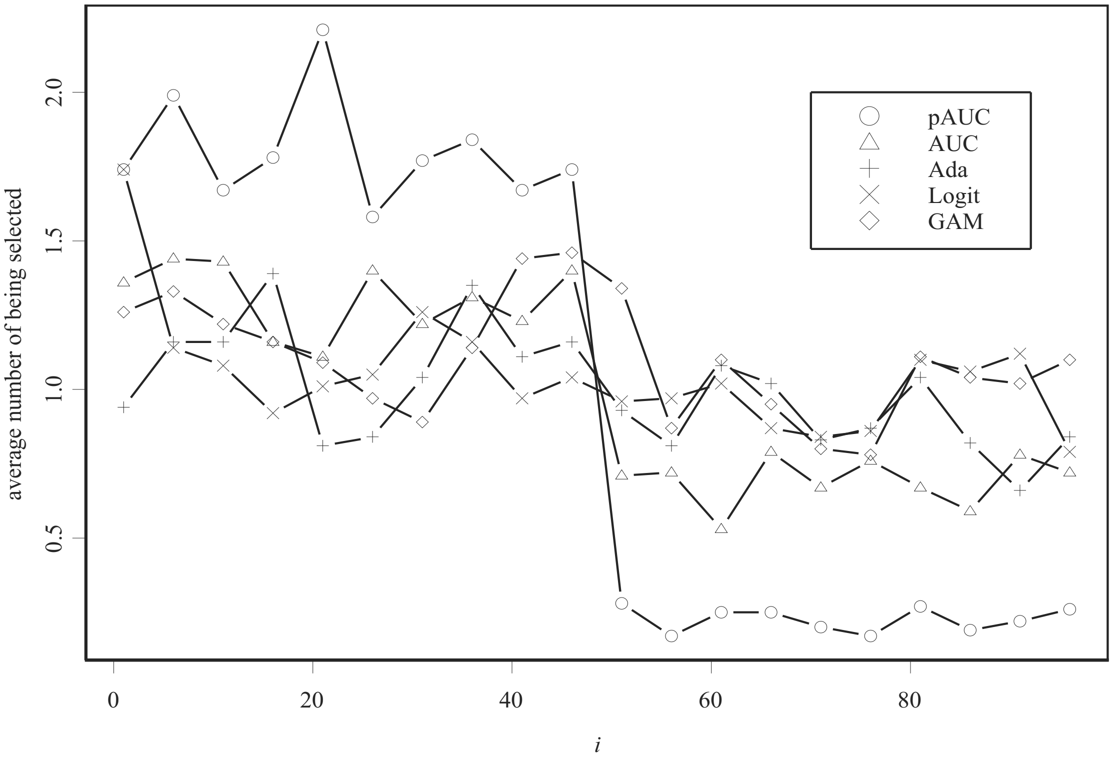


Figure 5.4: Average number of selected genes during the boosting process for five methods. The horizontal axis denotes the identification number of the genes.

5.5 Application of pAUCBoost to breast cancer data

The breast cancer data of van't Veer *and others* (2002) contains not only gene expression profiles but also clinical markers such as Age, age of patients; Size, diameter of breast cancer; Grade, tumour grade; Angi, existence or nonexistence of angioinvasion; ERp, ER expression; PRp, PR expression; and Lymp, existence or nonexistence of lymphocytic infiltrate. First, we apply AUCBoost to these clinical markers and investigate their utility. The weak classifiers we use are natural cubic splines for continuous markers (Age and Size), and decision stumps to discrete markers. The decision stumps for x_k are defined as

$$S_{k,l}(x_k) = \mathrm{H}(x_k - \xi_{k,l}), \tag{5.5.1}$$

where $\xi_{k,l}$ is a knot defined in Section 3. Next, we apply pAUCBoost with $\overline{\alpha}_1 = 0$ and $\overline{\alpha}_2 = 0.1$ to the gene expression data after a pAUC-based filtering process proposed by Pepe *and others* (2003). The training data set and the test data set are the same as those in van't Veer *and others* (2002), that is, 44 patients with good prognosis and 34 patients with distant metastases for training data, and 7 and 12 patients for test data, respectively.

Figure 5 shows the results of the score plot generated by AUCBoost with $\lambda = 0.01$ and $T = 20$, which are determined by a 10-fold cross validation. The Age and Size show almost linear association with the outcome variable; a tendency to develop metastases increases as the value of Grade; patients with negative ER and negative PR are estimated to have high risk of metastases, which are consistent with the result of van't Veer *and others* (2002). We have found that the values of the AUC for training and test data are 0.846 and 0.964, respectively. These results are comparable to those of van't Veer *and others* (2002) that are derived from the gene expression data: 0.882 and 0.869, respectively. This means that clinical markers, not gene expression profiles, also have the ability to discriminate to some extent the patients with good prognosis from those with metastases.

Next, we analyze the gene expression data as follows. The informative genes were selected from the total of 25000 genes according to the criteria that the genes are two-fold regulated and that the significance of regulation $p < 0.01$ in more than 3 patients, which is the same condition as that of van't Veer *and others* (2002). Next, the approximately 5000 filtered genes are ordered based on their values of the pAUC with $\overline{\alpha}_1 = 0$ and $\overline{\alpha}_2 = 0.1$. We assessed the variability using the probability of gene selection proposed by Pepe *and others* (2003). That is

$$P_g(k) = P(\text{gene } g \text{ ranked in the top } k),$$

where k is set to 100 in this analysis. Table 1 shows the results of the top 30 genes based on 1000 bootstrapped samples, along with the values of pAUC and AUC calculated from the

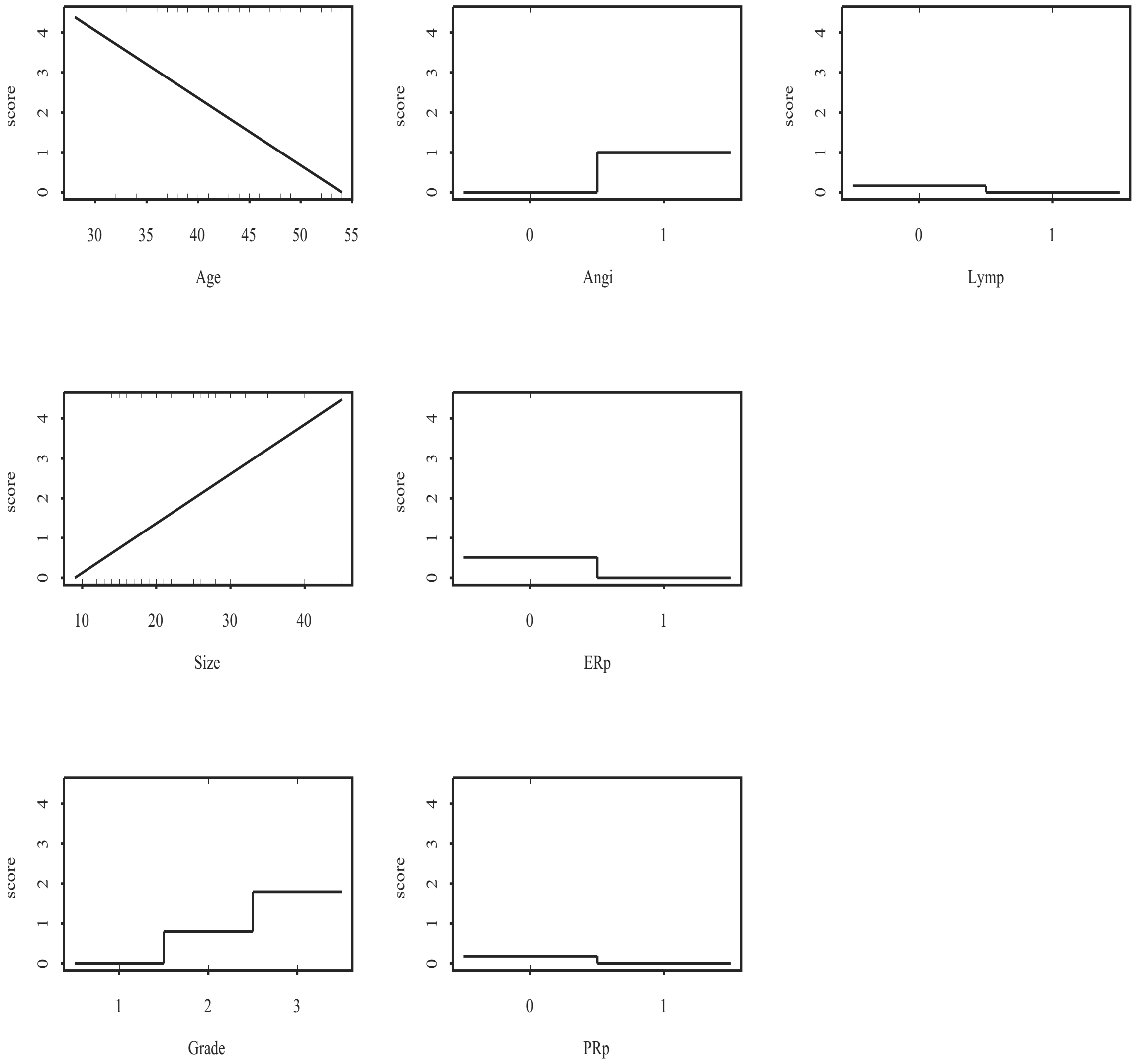


Figure 5.5: Score plots of clinical markers. The rug plot at the bottoms of each score plot shows the observation from patients with good prognosis; the rug plot for patients with distant metastases is described at the top of each score plot.

Table 5.1: *The top 30 genes ranked by the probability of gene selection, and the values of the pAUC and AUC*

No	gene name	$P_g(100)$	pAUC	AUC
1	Contig41613_RC	0.728	0.036	0.666
2	NM_006931	0.728	0.035	0.678
3	Contig40831_RC	0.706	0.037	0.672
4	Contig55574_RC	0.639	0.035	0.654
5	AB023173	0.636	0.034	0.684
6	Contig63649_RC	0.626	0.034	0.749
7	NM_018964	0.586	0.034	0.660
8	AL137615	0.571	0.033	0.655
9	NM_006201	0.541	0.032	0.664
10	NM_001710	0.520	0.032	0.638
11	AA555029_RC	0.519	0.032	0.708
12	NM_020386	0.490	0.030	0.699
13	Contig7558_RC	0.488	0.032	0.659
14	Contig51464_RC	0.482	0.030	0.668
15	NM_014246	0.474	0.032	0.613
16	NM_007359	0.463	0.032	0.696
17	NM_006148	0.450	0.029	0.661
18	NM_004163	0.442	0.029	0.729
19	Contig37562_RC	0.423	0.031	0.630
20	Contig55377_RC	0.416	0.029	0.726
21	Contig47405_RC	0.404	0.029	0.718
22	NM_012261	0.393	0.029	0.721
23	NM_014400	0.379	0.028	0.681
24	Contig44409	0.368	0.029	0.692
25	AL080059	0.364	0.027	0.801
26	Contig60864_RC	0.358	0.029	0.637
27	NM_003748	0.353	0.025	0.793
28	AL080110	0.349	0.026	0.652
29	AL122101	0.343	0.028	0.708
30	NM_018120	0.336	0.026	0.671

original data. As seen in the table, the order of $P_g(100)$ is almost in accordance with that of the pAUC; however, it is quite different from that of the AUC. We picked up significant genes with $P_g(100) > 0.5$, and applied pAUCBoost to the 11 genes. The score plots in Figure 6 describe the association between genes and the outcome variable. Among the 11 genes, Contig41613_RC shows a nonlinear association. That is, the gene expression of the patients with metastases has large variance as shown by the rug plot, compared with that of controls, which has a tendency to take small absolute values and concentrate around the origin. The nonlinearity of the associations can be captured by pAUCBoost in this way. The values of tuning parameter λ and T are determined to be 10^{-6} and 65 by 10-cross validation, as described in the left panel in Figure 7. The right panel shows the pAUC for training (solid) and test (dashed) data, as a function of T with $\lambda = 10^{-6}$. We see that both of the values for training and test data are more than 3 times larger than those of van't Veer *and others* (2002): 0.025 and 0.0084, respectively.

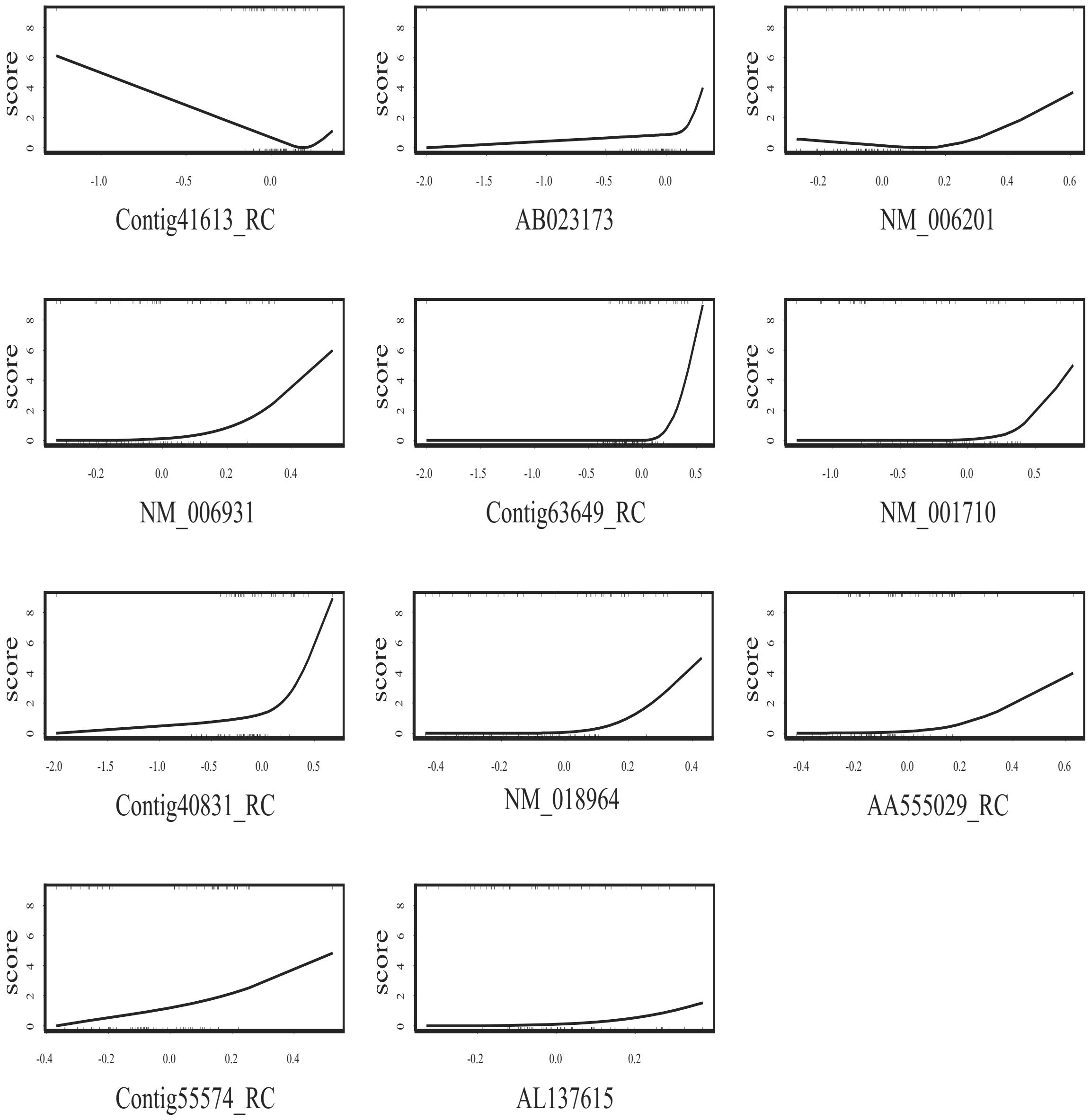


Figure 5.6: Score plots of the selected 11 genes. The rug plot at the bottoms of each score plot shows the observation from patients with good prognosis; the rug plot for patients with distant metastases is described at the top of each score plot.

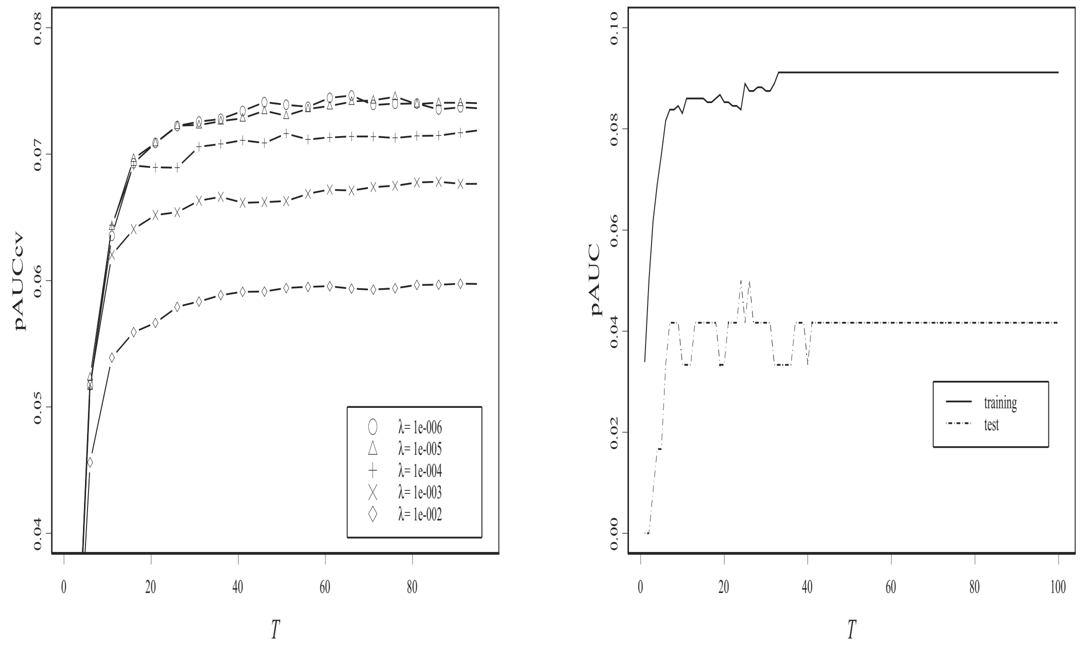


Figure 5.7: The results of 10-fold cross validation with different values of smoothing parameter λ and iteration number T (left panel); the results of the values of pAUC for training data (solid) and test data (dashed) by pAUCBoost, as a function of T with $\lambda = 10^{-6}$ (right panel).

5.6 Conclusions

We have developed the pAUCBoost algorithm, which is designed to maximize the pAUC, based on the approximate pAUC in the additive model. The use of the approximate pAUC is justified by showing a relationship with the non-approximate pAUC in Theorem 5.2.1. The resultant component functions, termed score plots, are useful for understanding the associations between each marker and the outcome variable, as shown in Section 4 and 5. Depending on the types of markers, we employ natural cubic splines, which are the maximizers of the pAUCBoost objective function, as well as decision stumps.

We have also provided a consistent way to analyze gene expression data in the sense of the pAUC, as shown in Section 5. The pAUC is shown to be useful by Pepe *and others* (2003) for selection of informative genes, some of which are overexpressed or underexpressed in cancer tissues. However, how to combine the selected genes and how to discriminate the cancer tissues from normal tissues, have not been addressed. We nonlinearly combined the genes ranked by the pAUC in order to produce a score function, by which the classification of controls and cases is done. Interestingly, we have found 4 genes in common with the 70 genes of van't Veer *and others* (2002): Contig63649_RC, AA555029_RC, Contig40831_RC, NM_006931; 6 genes among the selected 11 genes are related to protein coding. We also applied pAUCBoost to the 70 genes for comparison with the result from the 11 genes. We found that it yielded a poor result, especially about the value of pAUC for test data. Hence, pAUCBoost with FPR restricted to be small should be applied to the genes or markers that have gone through a pAUC-based filtering process beforehand. In the usual analysis setting, in which markers do not have especially high values of the pAUC, AUCBoost is preferable because of its stable performance due to the comprehensive information it can take into the algorithm.

Mainly, there are two types of weak classifiers: smoothing splines and decision stumps. Bühlmann and Yu (2003) proposed to use smoothing splines in the L_2Boost algorithm, and Tutz and Binder (2006) used B-splines in GAMBoost. However, the way of fitting the weak classifiers in pAUCBoost is different from those methods. Our algorithm updates a score function with a basis function of a natural cubic spline for one marker; on the other hand, their algorithms update a score function with *a set* of basis functions for one marker. Hence, our resultant score functions have tendency to have simpler forms, which also leads to simple interpretation of the association between the markers and outcome variable.

In AdaBoost and LogitBoost, decision stumps are used as weak classifiers (Ben-Dor *and others*, 2000; Dettling and Bühlmann, 2003). The advantage of using decision stumps is that we can apply the boosting methods independently of the scale of the marker values. Hence, the decision stump-based method is resistant to outliers, which often occur in real data. However, it easily suffers from false discovery, as clearly shown in Figure 4; this causes poor performance in a setting where non-informative genes are mixed with informative ones.

We have also confirmed that pAUCBoost with decision stumps for weak classifiers shows worse performance than that of pAUCBoost with natural cubic splines. Hence, we have to be much careful about which weak classifiers to be employed; it depends on the types of markers or the purpose of the analysis we are engaged in.

Chapter 6

Ongoing and Future work

Abstract

This paper discusses a boosting method to minimize an index of integrated sensitivity and specificity for density estimation.

6.1 Introduction

Let $\boldsymbol{X}$ and Y be a p-dimensional feature vector and a binary label in context of discriminant analysis. We discuss a new index for assessing performance of a score function $S(\boldsymbol{X})$ defined by

$$L(S) = \frac{\delta(S)}{\sigma(S)}, \tag{6.1.1}$$

where $\delta(S) = \mathrm{E}\{S(\boldsymbol{X})|Y=1\} - \mathrm{E}\{S(\boldsymbol{X})|Y=0\}$ and $\sigma(S) = [\mathrm{E}\{S(\boldsymbol{X}) - \mathrm{E}S(\boldsymbol{X})\}^2]^{1/2}$. The empirical form is given by

$$\overline{L}(S) = \frac{1}{\overline{\sigma}(S)}\Big\{\frac{1}{n_1}\sum_{i=1}^{n} I(y_i=1)S(\boldsymbol{x}_i) - \frac{1}{n_0}\sum_{i=1}^{n} I(y_i=0)S(\boldsymbol{x}_i)\Big\}, \tag{6.1.2}$$

where I is a definition function, $n_y = \sum_{i=1}^{n} I(y_i = y)$ for $y = 0, 1$ and $\overline{\sigma}(S)$ is the sample standard deviation. For $y = 0, 1$, integrated 'one minus specificity' and sensitivity (Pencina *et al*, 2008) are expressed as

$$\int \mathrm{P}(S(\boldsymbol{X}) \geq u | Y = y) du = \mathrm{E}\{S(\boldsymbol{X})|Y = y\}. \tag{6.1.3}$$

On the other hand, we observe a relation to the correlation coefficient $\rho(Y, S(\boldsymbol{X}))$ such that

$$L(S) = \frac{\rho(Y, S(\boldsymbol{X}))}{\sqrt{\pi(1-\pi)}}, \tag{6.1.4}$$

where $\pi = \mathrm{P}(Y = 1)$. We note that for any constants $\alpha > 0$ and β that

$$L(\alpha S + \beta) = L(S). \tag{6.1.5}$$

6.1.1 Bayes consistency

We assume that there exists a score function $S^*(\boldsymbol{X})$ such that

$$L(S^*) = \max_{S \in \mathcal{S}} L(S), \tag{6.1.6}$$

where $\mathcal{S}$ is the space of all score functions. Let $S_\varepsilon = S^* + \varepsilon\, \eta$ for arbitrarily fixed function η. Then, we observe that

$$\frac{\partial}{\partial \varepsilon} L(S_\varepsilon)\Big|_{\varepsilon=0} = 0. \tag{6.1.7}$$

The gradient of $L(S_\varepsilon)$ is given by

$$\frac{\partial}{\partial \varepsilon} L(S_\varepsilon) = \frac{\delta(\eta)}{\sigma(S_\varepsilon)} - \frac{\delta(S_\varepsilon)}{\sigma^3(S_\varepsilon)} \mathrm{E}\big[\{S_\varepsilon(\boldsymbol{X}) - \mu(S_\varepsilon)\}\{\eta(\boldsymbol{X}) - \mu(\eta)\}\big], \tag{6.1.8}$$

where $\mu(S_\varepsilon) = \mathrm{E}S_\varepsilon(\boldsymbol{X})$ and $\mu(\eta) = \mathrm{E}\eta(\boldsymbol{X})$. Using the conditional density of $\boldsymbol{X}$ given $Y = y$: g_y, and the marginal density of $\boldsymbol{X}$: g, it is rewritten as

$$\frac{1}{\sigma(S_\varepsilon)} \int \eta(\boldsymbol{x}) \Big[g_1(\boldsymbol{x}) - g_0(\boldsymbol{x}) - \frac{\delta(S_\varepsilon)}{\sigma^2(S_\varepsilon)} \{S_\varepsilon(\boldsymbol{x}) - \mu(S_\varepsilon)\} g(\boldsymbol{x}) \Big] d\boldsymbol{x}, \tag{6.1.9}$$

where $g = \pi g_1 + (1-\pi)g_0$. Hence, we get

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} L(S_\varepsilon)\Big|_{\varepsilon=0} &= \frac{1}{\sigma(S^*)} \int \eta(\boldsymbol{x}) \\ \times \quad & \Big[g_1(\boldsymbol{x}) - g_0(\boldsymbol{x}) - \frac{\delta(S^*)}{\sigma^2(S^*)} \{S^*(\boldsymbol{x}) - \mu(S^*)\} g(\boldsymbol{x}) \Big] d\boldsymbol{x}. \end{aligned} \tag{6.1.10}$$

We conclude that

$$S^*(\boldsymbol{x}) - \mu(S^*) = \frac{\sigma^2(S^*)}{\delta(S^*)} \frac{g_1(\boldsymbol{x}) - g_0(\boldsymbol{x})}{g(\boldsymbol{x})} \tag{6.1.11}$$

because (6.1.7) must hold for any η. Finally, from (6.1.5) we obtain a simpler expression,

$$\begin{aligned} S^*(\boldsymbol{x}) &= \frac{g_1(\boldsymbol{x}) - g_0(\boldsymbol{x})}{g(\boldsymbol{x})} && (6.1.12) \\ &= \frac{\Lambda(\boldsymbol{x}) - 1}{\pi \Lambda(\boldsymbol{x}) + (1-\pi)}, && (6.1.13) \end{aligned}$$

where $\Lambda(\boldsymbol{x}) = g_1(\boldsymbol{x})/g_0(\boldsymbol{x})$ and we assume $\delta(S^*) > 0$. Note that the existence of S^* is unique except for scale transforms in the sense of (6.1.5). It is clear that S^* is an increasing function of Λ, so we investigate the second derivative in order to confirm the Bayes consistency. It follows from (6.1.8) that

$$\frac{\partial}{\partial \varepsilon} L(S_\varepsilon) = \frac{\sigma(S_\varepsilon)\delta(\eta) - L(S_\varepsilon)\mathrm{cov}(S_\varepsilon, \eta)}{\sigma^2(S_\varepsilon)}. \tag{6.1.14}$$

Then, the second derivative of $L(S_\varepsilon)$ becomes

$$\begin{aligned} & \frac{\partial^2}{\partial \varepsilon^2} L(S_\varepsilon) \\ = \quad & -2\sigma^{-3}(S_\varepsilon)\sigma'(S_\varepsilon)\big\{\sigma(S_\varepsilon)\delta(\eta) - L(S_\varepsilon)\mathrm{cov}(S_\varepsilon, \eta)\big\} \\ & \quad + \sigma^{-2}(S_\varepsilon)\big\{\sigma'(S_\varepsilon)\delta(\eta) - L'(S_\varepsilon)\mathrm{cov}(S_\varepsilon, \eta) - L(S_\varepsilon)\mathrm{cov}'(S_\varepsilon, \eta)\big\}, \end{aligned} \tag{6.1.15}$$

where

$$\begin{aligned}\sigma'(S_\varepsilon) &= \sigma^{-1}(S_\varepsilon)\mathrm{cov}(S_\varepsilon,\eta), &\quad (6.1.16)\\ \mathrm{cov}'(S_\varepsilon,\eta) &= \sigma^2(\eta). &\quad (6.1.17)\end{aligned}$$

Hence, the latter term in (6.1.15) is given

$$\sigma^{-2}(S_\varepsilon)\big\{\sigma^{-1}(S_\varepsilon)\mathrm{cov}(S_\varepsilon,\eta)\delta(\eta) - L'(S_\varepsilon)\mathrm{cov}(S_\varepsilon,\eta) - L(S_\varepsilon)\sigma^2(\eta)\big\}. \quad (6.1.18)$$

As a result, we have

$$\begin{aligned}\frac{\partial^2}{\partial\varepsilon^2}L(S_\varepsilon)\Big|_{\varepsilon=0} &= \sigma^{-2}(S^*)\big\{\sigma^{-1}(S^*)\mathrm{cov}(S^*,\eta)\delta(\eta) - L(S^*)\sigma^2(\eta)\big\}\\ &= \sigma^{-4}(S^*)L(S^*)\big\{\mathrm{cov}^2(S^*,\eta) - \sigma^2(S^*)\sigma^2(\eta)\big\}, &\quad (6.1.19)\end{aligned}$$

because Equation (6.1.14) is zero for S^*. Hence, the second derivative is negative under an assumption of $L(S^*) > 0$.

Through the discussion above, we conclude that the score function S^* given by (6.1.12) attains a maximum of $L(S)$ over $\mathcal{S}$, and the maximum is given as

$$\begin{aligned}\max_{S\in\mathcal{S}} L(S) &= \frac{\delta(S^*)}{\sigma(S^*)}\\ &= \int \frac{\{g_1(\boldsymbol{x}) - g_0(\boldsymbol{x})\}^2}{g(\boldsymbol{x})}d\boldsymbol{x} \Big/ \sqrt{\int \frac{\{g_1(\boldsymbol{x}) - g_0(\boldsymbol{x})\}^2}{g(\boldsymbol{x})}d\boldsymbol{x}}\\ &= \sqrt{\int \frac{\{g_1(\boldsymbol{x}) - g_0(\boldsymbol{x})\}^2}{g(\boldsymbol{x})}d\boldsymbol{x}}. &\quad (6.1.20)\end{aligned}$$

which is invariant with all one-to-one transformations of $\boldsymbol{X}$. Thus we observe the Bayes consistency for the proposed function $L(S)$.

6.1.2 Linear score function

We discuss a simple situation in which a score function is linear: $S_{\boldsymbol{\beta}}(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}$. The corresponding objective function is written by

$$L(S_{\boldsymbol{\beta}}) = \frac{\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}{\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\beta}}}. \tag{6.1.21}$$

where $\boldsymbol{\mu}_y$ is the conditional mean of $\boldsymbol{X}$ given y and $\boldsymbol{V}$ is the variance matrix of $\boldsymbol{X}$; the empirical form is

$$\overline{L}(S_{\boldsymbol{\beta}}) = \frac{\boldsymbol{\beta}^{\mathrm{T}}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_0)}{\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\overline{\boldsymbol{V}}\boldsymbol{\beta}}}. \tag{6.1.22}$$

where $\overline{\boldsymbol{x}}_y$ is the conditional sample mean given y and $\overline{\boldsymbol{V}}$ is the sample variance matrix. We can maximize it by

$$\beta^* = \overline{\boldsymbol{V}}^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_0), \tag{6.1.23}$$

where the maximum is the Mahalanobis distance:

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \overline{L}(S_{\boldsymbol{\beta}}) = \overline{L}(S_{\boldsymbol{\beta}^*}) = \sqrt{(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_0)^{\mathrm{T}}\overline{\boldsymbol{V}}^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_0)}. \tag{6.1.24}$$

Thus, the optimal linear score function is the same as the Fisher linear score function except for the threshold.

6.2 Boosting algorithm for maximization of t-value

We discuss a boosting learning algorithm to maximize the empirical correlation function $\overline{L}(S)$ in a stage-wise manner. Let us prepare a dictionary of classifiers as

$$\mathcal{D} = \{s_\omega : \omega \in \Omega\}. \tag{6.2.1}$$

For example, the dictionary $\mathcal{D}$ is taken by a family of all decision stumps. Henceforth we assume that $\mathcal{D}$ is negation closed, that is, $\forall s \in \mathcal{D}, -s \in \mathcal{D}$.

Boosting algorithm

(i). Take appropriately an initial candidate, $S_1(x)$ in $\mathcal{D}$ as a classifier.

(ii). For any $t = 1, \cdots, M-1$ we get $\alpha_{t+1}, s_{t+1}, S_{t+1}$ from α_t, s_t, S_t as follows:

$$\begin{aligned} (\alpha_{t+1}, s_{t+1}) &= \mathrm{argmax}_{\alpha \in \mathbb{R}, s \in \mathcal{D}} \overline{L}(S_t + \alpha s), & (6.2.2) \\ S_{t+1} &= S_t + \alpha_{t+1} s_{t+1}. & (6.2.3) \end{aligned}$$

Hence we obtain $\hat{S} = S_M$ as the final solution.

We investigate a property associated with $\alpha_{t+1}, s_{t+1}, S_{t+1}$ as defined above to maximize the empirical correlation function $\overline{L}(S)$. Here and hereafter we use a notation $\overline{\delta}$ defined by

$$\overline{\delta}(S) = \frac{1}{n_1} \sum_{i=1}^{n} I(y_i = 1) S(\boldsymbol{x}_i) - \frac{1}{n_0} \sum_{i=1}^{n} I(y_i = 0) S(\boldsymbol{x}_i). \qquad (6.2.4)$$

We get the gradient by

$$\frac{\partial}{\partial \alpha} \overline{L}(S_t + \alpha s) = \frac{\overline{\delta}(s)}{\overline{\sigma}(S_t + \alpha s)} - \overline{L}(S_t + \alpha s) \frac{\overline{\mathrm{cov}}(S_t + \alpha s, s)}{\overline{\sigma}^2(S_t + \alpha s)}, \qquad (6.2.5)$$

where $\overline{\mathrm{cov}}$ denotes the sample covariance, for example,

$$\overline{\mathrm{cov}}(S, T) = \frac{1}{n} \sum_{i=1}^{n} \{(S(\boldsymbol{x}_i) - \overline{S})(T(\boldsymbol{x}_i) - \overline{T})\}. \qquad (6.2.6)$$

We pursue more tractable form for the pair of s_{t+1} and α_{t+1}. For this we first find an exact solution,

$$\alpha_{t+1}(s) = \mathrm{argmax}_{\alpha \in \mathbb{R}} \overline{L}(S_t + \alpha s), \qquad (6.2.7)$$

which is a solution of $(\partial / \partial \alpha) \overline{L}(S_t + \alpha s) = 0$ with respect to α; that is

$$\overline{\delta}(s) = \overline{\delta}(S_t + \alpha s) \frac{\overline{\mathrm{cov}}(S_t + \alpha s, s)}{\overline{\sigma}^2(S_t + \alpha s)}, \qquad (6.2.8)$$

which is

$$\overline{\delta}(s) = \frac{\{\overline{\delta}(S_t) + \alpha\overline{\delta}(s)\}\{\ \overline{\mathrm{cov}}(S_t, s) + \alpha\overline{\sigma}^2(s)\}}{\overline{\sigma}^2(S_t) + 2\alpha\ \overline{\mathrm{cov}}(S_t, s) + \alpha^2\overline{\sigma}^2(s)}, \tag{6.2.9}$$

which, noting that the terms of α^2 cancel, is

$$\begin{aligned} & \overline{\delta}(s)\overline{\sigma}^2(S_t) + 2\alpha\overline{\delta}(s)\ \overline{\mathrm{cov}}(S_t, s) \\ = \quad & \overline{\delta}(S_t)\overline{\mathrm{cov}}(S_t, s) + \alpha\{\overline{\delta}(s)\overline{\mathrm{cov}}(S_t, s) + \overline{\delta}(S_t)\overline{\sigma}^2(s)\}. \end{aligned} \tag{6.2.10}$$

Hence, we obtain the solution,

$$\alpha_{t+1}(s) = \frac{\overline{\delta}(S_t)\overline{\mathrm{cov}}(S_t, s) - \overline{\delta}(s)\overline{\sigma}^2(S_t)}{\overline{\delta}(s)\overline{\mathrm{cov}}(S_t, s) - \overline{\delta}(S_t)\overline{\sigma}^2(s)}. \tag{6.2.11}$$

Secondly, substituting the value into $\overline{L}(S_t + \alpha s)$ in (6.2.2), we get s_{t+1} as

$$s_{t+1} = \mathrm{argmax}_{s\in\mathcal{D}}\overline{L}(S_t + \alpha_{t+1}(s)\ s), \tag{6.2.12}$$

which needs only a light computational burden. Thus we get the updating pair (α_{t+1}, s_{t+1}) by two-stage of the maximization,

$$\overline{L}(S_t + \alpha_{t+1}s_{t+1}) = \max_{s\in\mathcal{D}}\{\max_{\alpha\in\mathbb{R}}\overline{L}(S_t + \alpha s)\}. \tag{6.2.13}$$

We evaluate one step improvement from S_t to $S_{t+1} = S_t + \alpha_{t+1}s_{t+1}$ in the boosting algorithm as follows:

$$\begin{aligned} & \overline{L}(S_t + \alpha_{t+1}s_{t+1})^2 - \overline{L}(S_t)^2 = \frac{\overline{\delta}^2(S_t + \alpha_{t+1}s_{t+1})}{\overline{\sigma}^2(S_t + \alpha_{t+1}s_{t+1})} - \frac{\overline{\delta}^2(S_t)}{\overline{\sigma}^2(S_t)} \\ = \quad & \frac{\overline{\delta}^2(S_t + \alpha_{t+1}s_{t+1})\overline{\sigma}^2(S_t) - \overline{\delta}^2(S_t)\overline{\sigma}^2(S_t + \alpha_{t+1}s_{t+1})}{\overline{\sigma}^2(S_t)\overline{\sigma}^2(S_t + \alpha_{t+1}s_{t+1})} \end{aligned} \tag{6.2.14}$$

of which the numerator is

$$\begin{align}
& \{\overline{\delta}(S_t) + \alpha_{t+1}\overline{\delta}(s)\}^2\overline{\sigma}^2(S_t) - \overline{\delta}(S_t)\{\overline{\sigma}^2(S_t) + 2\alpha_{t+1}\overline{\mathrm{cov}}(S_t, s_{t+1}) + \alpha_{t+1}^2\overline{\sigma}^2(s_{t+1})\} \tag{6.2.15} \\
= \;& 2\overline{\delta}(S_t)\{\overline{\delta}(s_{t+1})\overline{\sigma}^2(S_t) - \overline{\delta}(S_t)\overline{\mathrm{cov}}(S_t, s_{t+1})\}\alpha_{t+1} \nonumber \\
& \qquad\qquad + \{\overline{\delta}^2(s_{t+1})\overline{\sigma}^2(S_t) - \overline{\delta}^2(S_t)\overline{\sigma}^2(s_{t+1})\}\alpha_{t+1}^2 \tag{6.2.16} \\
= \;& 2\overline{\delta}(S_t)\{\overline{\delta}(S_t)\overline{\sigma}^2(s_{t+1}) - \overline{\delta}(s_{t+1})\overline{\mathrm{cov}}(S_t, s_{t+1})\}\alpha_{t+1}^2 \nonumber \\
& \qquad\qquad + \{\overline{\delta}^2(s_{t+1})\overline{\sigma}^2(S_t) - \overline{\delta}^2(S_t)\overline{\sigma}^2(s_{t+1})\}\alpha_{t+1}^2 \tag{6.2.17} \\
= \;& \{\overline{\delta}^2(s_{t+1})\overline{\sigma}^2(S_t) - 2\overline{\delta}(S_t)\overline{\delta}(s_{t+1})\overline{\mathrm{cov}}(S_t, s_{t+1}) + \overline{\delta}^2(S_t)\overline{\sigma}^2(s_{t+1})\}\alpha_{t+1}^2 \nonumber \\
= \;& \overline{\sigma}^2\Big(\overline{\delta}(s_{t+1})S_t - \overline{\delta}(S_t)s_{t+1}\Big)\alpha_{t+1}^2 \nonumber \\
= \;& \overline{\sigma}^2\Big(\overline{\delta}(S_{t+1})S_t - \overline{\delta}(S_t)S_{t+1}\Big). \tag{6.2.18}
\end{align}$$

As a result, we have

$$\overline{L}(S_{t+1})^2 - \overline{L}(S_t)^2 = \overline{\sigma}^2\Big(\overline{L}(S_{t+1})\frac{S_t}{\overline{\sigma}(S_t)} - \overline{L}(S_t)\frac{S_{t+1}}{\overline{\sigma}(S_{t+1})}\Big). \tag{6.2.19}$$

Appendix

A.1 *Proof of Theorem 5.2.1.*

At first, we fix the value of α_1 and α_2 using thresholds $c_{1,F}$ and $c_{2,F}$ as

$$\int \mathrm{H}(F(\boldsymbol{x}_0) - c_{1,F}) g_0(\boldsymbol{x}_0) d\boldsymbol{x} = \alpha_1, \ \int \mathrm{H}(F(\boldsymbol{x}_0) - c_{2,F}) g_0(\boldsymbol{x}_0) d\boldsymbol{x} = \alpha_2, \tag{A.1}$$

where $\alpha_1 < \alpha_2$ $(c_{2,F} < c_{1,F})$. For simplicity, we write

$$\begin{aligned} \mathrm{H}_{F,i}(\boldsymbol{x}) &= \mathrm{H}(F(\boldsymbol{x}) - c_{i,F}), & \text{(A.2)} \\ \mathrm{H}_{i,F}(\boldsymbol{x}) &= \mathrm{H}(c_{i,F} - F(\boldsymbol{x})), \ i = 1, 2. & \text{(A.3)} \end{aligned}$$

Then, the pAUC with FPR being between α_1 and α_2 has an integral formula:

$$\begin{aligned} & \mathrm{pAUC}(F, \alpha_1, \alpha_2) \\ = & \int\int \mathrm{H}_{F,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F}(\boldsymbol{x}_0)\mathrm{H}(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\ = & \int\int \mathrm{H}_{F,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F}(\boldsymbol{x}_0)\mathrm{H}_{F,2}(\boldsymbol{x}_1)\mathrm{H}(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\ = & \int\int \mathrm{H}_{F,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F}(\boldsymbol{x}_0)\mathrm{H}_{F,2}(\boldsymbol{x}_1)\Big\{\mathrm{H}_{1,F}(\boldsymbol{x}_1) + \mathrm{H}_{F,1}(\boldsymbol{x}_1)\Big\}\mathrm{H}(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\ = & \int\int \mathrm{H}_{F,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F}(\boldsymbol{x}_0)\mathrm{H}_{F,2}(\boldsymbol{x}_1)\mathrm{H}_{1,F}(\boldsymbol{x}_1)\mathrm{H}(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\ & + \int\int \mathrm{H}_{F,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F}(\boldsymbol{x}_0)\mathrm{H}_{F,2}(\boldsymbol{x}_1)\mathrm{H}_{F,1}(\boldsymbol{x}_1)\mathrm{H}(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\ = & \int\int \mathrm{H}_{F,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F}(\boldsymbol{x}_0)\mathrm{H}_{F,2}(\boldsymbol{x}_1)\mathrm{H}_{1,F}(\boldsymbol{x}_1)\mathrm{H}(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0)) g_0(\boldsymbol{x}_0) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\ & + \int \mathrm{H}_{F,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F}(\boldsymbol{x}_0) g_0(\boldsymbol{x}_0) d\boldsymbol{x}_0 \int \mathrm{H}_{F,1}(\boldsymbol{x}_1) g_1(\boldsymbol{x}_1) d\boldsymbol{x}_1. \end{aligned} \tag{A.4}$$

Similarly, the approximate pAUC is given as

$$\begin{aligned}
& \mathrm{pAUC}_\sigma(F, \alpha_1, \alpha_2) \\
= & \int\int \mathrm{H}_{F,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F}(\boldsymbol{x}_0)\mathrm{H}_{F,2}(\boldsymbol{x}_1)\mathrm{H}_{1,F}(\boldsymbol{x}_1)\mathrm{H}_\sigma(F(\boldsymbol{x}_1) - F(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\
& + \int \mathrm{H}_{F,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F}(\boldsymbol{x}_0)g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 \int \mathrm{H}_{F,1}(\boldsymbol{x}_1)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_1. \qquad \text{(A.5)}
\end{aligned}$$

In this setting, we will prove Theorem 5.2.1 as follows.

Proof. For simplicity, we define some notations:

$$\zeta(\boldsymbol{x}) = m\big(\Lambda(\boldsymbol{x})\big), \qquad \text{(A.6)}$$

$$F_\gamma(\boldsymbol{x}) = F(\boldsymbol{x}) + \gamma\zeta(\boldsymbol{x}), \qquad \text{(A.7)}$$

$$c_{i,\gamma} = c_{i,F_\gamma}, \ i = 1, 2, \qquad \text{(A.8)}$$

$$c'_{i,\gamma} = \frac{\partial c_{i,\gamma}}{\partial \gamma}, \ i = 1, 2. \qquad \text{(A.9)}$$

Then, the first derivative of $\Psi(\gamma)$ with respect to γ is given as

$$\begin{aligned}
&\frac{\partial}{\partial\gamma}\Psi(\gamma) \\
=\ & \int\int \mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_0)(\zeta(\boldsymbol{x}_0) - c'_{2,\gamma})\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1) \\
&\quad \times\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}_\sigma(F_\gamma(\boldsymbol{x}_1) - F_\gamma(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0d\boldsymbol{x}_1 \\
&+\int\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}'_{1,F_\gamma}(\boldsymbol{x}_0)(c'_{1,\gamma} - \zeta(\boldsymbol{x}_0))\mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1) \\
&\quad \times\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}_\sigma(F_\gamma(\boldsymbol{x}_1) - F_\gamma(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0d\boldsymbol{x}_1 \\
&+\int\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}'(F_\gamma(\boldsymbol{x}_1) - c_{2,\gamma})(\zeta(\boldsymbol{x}_1) - c'_{2,\gamma}) \\
&\quad \times\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}_\sigma(F_\gamma(\boldsymbol{x}_1) - F_\gamma(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0d\boldsymbol{x}_1 \\
&+\int\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1) \\
&\quad \times\mathrm{H}'(c_{1,\gamma} - F_\gamma(\boldsymbol{x}_1))(c'_{1,\gamma} - \zeta(\boldsymbol{x}_1))\mathrm{H}_\sigma(F_\gamma(\boldsymbol{x}_1) - F_\gamma(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0d\boldsymbol{x}_1 \\
&+\int\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1) \\
&\quad \times\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}'_\sigma(F_\gamma(\boldsymbol{x}_1) - F_\gamma(\boldsymbol{x}_0))(\zeta(\boldsymbol{x}_1) - \zeta(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0d\boldsymbol{x}_1 \\
&+\int \Big\{\mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_0)(\zeta(\boldsymbol{x}_0) - c'_{2,\gamma}) - \mathrm{H}'_{F_\gamma,1}(\boldsymbol{x}_0)(\zeta(\boldsymbol{x}_0) - c'_{1,\gamma})\Big\}g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0\mathrm{TPR}(F_\gamma, c_{1,\gamma}) \\
&+\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0\mathrm{TPR}'(F_\gamma, c_{1,\gamma}), \qquad \text{(A.10)}
\end{aligned}$$

where

$$\mathrm{TPR}(F_\gamma, c_{1,\gamma}) = \int \mathrm{H}_{F_\gamma,1}(\boldsymbol{x}_1)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_1. \qquad \text{(A.11)}$$

And the first derivative is rewritten such as

$$
\begin{aligned}
& \frac{\partial}{\partial\gamma}\Psi(\gamma) \\
= \; & \int \mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_0)(\zeta(\boldsymbol{x}_0) - c'_{2,\gamma})g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 \int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}_\sigma(F_\gamma(\boldsymbol{x}_1) - c_{2,\gamma})g_1(\boldsymbol{x}_1)d\boldsymbol{x}_1 \\
& + \int \mathrm{H}'_{1,F_\gamma}(\boldsymbol{x}_0)(c'_{1,\gamma} - \zeta(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 \int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}_\sigma(F_\gamma(\boldsymbol{x}_1) - c_{1,\gamma})g_1(\boldsymbol{x}_1)d\boldsymbol{x}_1 \\
& + \int \mathrm{H}'(F_\gamma(\boldsymbol{x}_1) - c_{2,\gamma})(\zeta(\boldsymbol{x}_1) - c'_{2,\gamma})g_1(\boldsymbol{x}_1)d\boldsymbol{x}_1 \int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_\sigma(c_{2,\gamma} - F_\gamma(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 \\
& + \int \mathrm{H}'(c_{1,\gamma} - F_\gamma(\boldsymbol{x}_1))(c'_{1,\gamma} - \zeta(\boldsymbol{x}_1))g_1(\boldsymbol{x}_1)d\boldsymbol{x}_1 \int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_\sigma(c_{1,\gamma} - F_\gamma(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 \\
& + \int\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1) \\
& \quad \times \mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}'_\sigma(F_\gamma(\boldsymbol{x}_1) - F_\gamma(\boldsymbol{x}_0))(\zeta(\boldsymbol{x}_1) - \zeta(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\
& + \int \Big\{\mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_0)(\zeta(\boldsymbol{x}_0) - c'_{2,\gamma}) - \mathrm{H}'_{F_\gamma,1}(\boldsymbol{x}_0)(\zeta(\boldsymbol{x}_0) - c'_{1,\gamma})\Big\}g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 \mathrm{TPR}(F_\gamma, c_{1,\gamma}) \\
& + \int \mathrm{H}(F_\gamma(\boldsymbol{x}_0) - c_{2,\gamma})\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 \mathrm{TPR}'(F_\gamma, c_{1,\gamma}), \\
= \; & \mathrm{FPR}'(F_\gamma, c_{2,\gamma}) \int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}_\sigma(F_\gamma(\boldsymbol{x}_1) - c_{2,\gamma})g_1(\boldsymbol{x}_1)d\boldsymbol{x}_1 \\
& - \mathrm{FPR}'(F_\gamma, c_{1,\gamma}) \int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}_\sigma(F_\gamma(\boldsymbol{x}_1) - c_{1,\gamma})g_1(\boldsymbol{x}_1)d\boldsymbol{x}_1 \\
& + \mathrm{TPR}'(F_\gamma, c_{2,\gamma}) \int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_\sigma(c_{2,\gamma} - F_\gamma(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 \\
& - \mathrm{TPR}'(F_\gamma, c_{1,\gamma}) \int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_\sigma(c_{1,\gamma} - F_\gamma(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 \\
& + \int\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1) \\
& \quad \times \mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}'_\sigma(F_\gamma(\boldsymbol{x}_1) - F_\gamma(\boldsymbol{x}_0))(\zeta(\boldsymbol{x}_1) - \zeta(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\
& + \Big\{\mathrm{FPR}'(F_\gamma, c_{2,\gamma}) - \mathrm{FPR}'(F_\gamma, c_{1,\gamma})\Big\}\mathrm{TPR}(F_\gamma, c_{1,\gamma}) \\
& + \mathrm{TPR}'(F_\gamma, c_{1,\gamma}) \int \mathrm{H}(F_\gamma(\boldsymbol{x}_0) - c_{2,\gamma})\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0.
\end{aligned}
\tag{A.12}
$$

Since $\mathrm{FPR}(F_\gamma, c_{1,\gamma})$ and $\mathrm{FPR}(F_\gamma, c_{2,\gamma})$ are fixed, we have

$$\begin{aligned}
&\frac{\partial}{\partial\gamma}\Psi(\gamma)\\
= \;& \mathrm{TPR}'(F_\gamma, c_{2,\gamma})\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_\sigma(c_{2,\gamma}-F_\gamma(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0\\
&+\mathrm{TPR}'(F_\gamma, c_{1,\gamma})\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\Big\{1-\mathrm{H}_\sigma(c_{1,\gamma}-F_\gamma(\boldsymbol{x}_0))\Big\}g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0\\
&+\int\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1)\\
&\quad\times\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}'_\sigma(F_\gamma(\boldsymbol{x}_1)-F_\gamma(\boldsymbol{x}_0))(\zeta(\boldsymbol{x}_1)-\zeta(\boldsymbol{x}_0))g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0d\boldsymbol{x}_1. \quad (A.13)
\end{aligned}$$

Next, we investigate the behavior of $\mathrm{TPR}'(F_\gamma, c_{2,\gamma})$. The value of $\mathrm{FPR}(F_\gamma, c_{2,\gamma})$ is fixed, so we have

$$\mathrm{FPR}'(F_\gamma, c_{2,\gamma}) = \int \mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_0)(\zeta(\boldsymbol{x}_0)-c'_{2,\gamma})g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 = 0. \quad (A.14)$$

If $\int \mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_0)g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 = 0$, we have $\int \mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_1)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_1 = 0$; moreover, we have $\mathrm{TPR}'(F_\gamma, c_{2,\gamma}) = 0$. Otherwise, we have

$$c'_{2,\gamma} = \frac{\int \mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_0)\zeta(\boldsymbol{x}_0)g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0}{\int \mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_0)g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0}. \quad (A.15)$$

By substituting it into $\mathrm{TPR}'(F_\gamma, c_{2,\gamma})$, we have

$$\begin{aligned}
&\mathrm{TPR}'(F_\gamma, c_{2,\gamma})\\
=\;& \frac{\int\int K(\boldsymbol{x}_0,\boldsymbol{x}_1)\zeta(\boldsymbol{x}_1)g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0d\boldsymbol{x}_1 - \int\int K(\boldsymbol{x}_0,\boldsymbol{x}_1)\zeta(\boldsymbol{x}_0)g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0d\boldsymbol{x}_1}{\int \mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_0)g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0},
\end{aligned} \quad (A.16)$$

where

$$K(\boldsymbol{x}_0,\boldsymbol{x}_1) = \mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}'_{F_\gamma,2}(\boldsymbol{x}_1) \quad (A.17)$$

Then, the numerator becomes

$$\begin{aligned}
& \int\int K(\boldsymbol{x}_0,\boldsymbol{x}_1)\Big(\zeta(\boldsymbol{x}_1)-\zeta(\boldsymbol{x}_0)\Big)g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\
= & \int\int K(\boldsymbol{x}_1,\boldsymbol{x}_0)\Big(\zeta(\boldsymbol{x}_0)-\zeta(\boldsymbol{x}_1)\Big)g_0(\boldsymbol{x}_1)g_1(\boldsymbol{x}_0)d\boldsymbol{x}_1 d\boldsymbol{x}_0 \\
= & \frac{1}{2}\int\int K(\boldsymbol{x}_0,\boldsymbol{x}_1)\Big(\zeta(\boldsymbol{x}_1)-\zeta(\boldsymbol{x}_0)\Big)\Big(g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)-g_0(\boldsymbol{x}_1)g_1(\boldsymbol{x}_0)\Big)d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\
= & \frac{1}{2}\int\int K(\boldsymbol{x}_0,\boldsymbol{x}_1)\Big(\zeta(\boldsymbol{x}_1)-\zeta(\boldsymbol{x}_0)\Big)\Big(\Lambda(\boldsymbol{x}_1)-\Lambda(\boldsymbol{x}_0)\Big)g_0(\boldsymbol{x}_0)g_0(\boldsymbol{x}_1)d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\
> & \ 0. \qquad \text{(A.18)}
\end{aligned}$$

Hence, we have

$$\mathrm{TPR}'(F_\gamma, c_{i,\gamma}) \geq 0,\ i=1,2, \qquad \text{(A.19)}$$

because we can replace $c_{2,\gamma}$ with $c_{1,\gamma}$, and have the same result.

By looking at the third term in (A.13), we find

$$\mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}'_\sigma(F_\gamma(\boldsymbol{x}_1)-F_\gamma(\boldsymbol{x}_0)) \qquad \text{(A.20)}$$

is invariant to the exchange of $\boldsymbol{x}_0$ for $\boldsymbol{x}_1$ like $K(\boldsymbol{x}_0,\boldsymbol{x}_1)$. Hence by the same argument above, we have

$$\begin{aligned}
& \int\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1) \\
& \quad \times\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}'_\sigma(F_\gamma(\boldsymbol{x}_1)-F_\gamma(\boldsymbol{x}_0))\Big(\zeta(\boldsymbol{x}_1)-\zeta(\boldsymbol{x}_0)\Big)g_0(\boldsymbol{x}_0)g_1(\boldsymbol{x}_1)d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\
= & \frac{1}{2}\int\int \mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_0)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_0)\mathrm{H}_{F_\gamma,2}(\boldsymbol{x}_1)\mathrm{H}_{1,F_\gamma}(\boldsymbol{x}_1)\mathrm{H}'_\sigma(F_\gamma(\boldsymbol{x}_1)-F_\gamma(\boldsymbol{x}_0))\Big(\zeta(\boldsymbol{x}_1)-\zeta(\boldsymbol{x}_0)\Big) \\
& \quad \times\Big(\Lambda(\boldsymbol{x}_1)-\Lambda(\boldsymbol{x}_0)\Big)g_0(\boldsymbol{x}_0)g_0(\boldsymbol{x}_1)d\boldsymbol{x}_0 d\boldsymbol{x}_1 \\
> & \ 0. \qquad \text{(A.21)}
\end{aligned}$$

As a result, we have

$$\frac{\partial}{\partial\gamma}\Psi(\gamma) > 0. \qquad \text{(A.22)}$$

Finally, we have

$$\begin{aligned}
\mathrm{pAUC}_\sigma(F,\alpha_1,\alpha_2) &< \lim_{\gamma\to\infty}\Psi(\gamma)\\
&= \lim_{\gamma\to\infty}\mathrm{pAUC}_\sigma\left[\gamma\left\{\frac{F}{\gamma}+\zeta\right\},\alpha_1,\alpha_2\right]\\
&= \lim_{\gamma\to\infty}\mathrm{pAUC}_{\frac{\sigma}{\gamma}}\left(\frac{F}{\gamma}+\zeta,\alpha_1,\alpha_2\right)\\
&= \mathrm{pAUC}(\zeta,\alpha_1,\alpha_2)\\
&= \mathrm{pAUC}(\Lambda,\alpha_1,\alpha_2). \qquad \text{(A.23)}
\end{aligned}$$

□

A.2 *Proof of Corollary 1*

Proof. Under the condition that $\mathrm{FPR}(F+\gamma\,\eta,c_{F+\gamma\,\eta})=\alpha$, the first derivative of $\mathrm{TPR}(F+\gamma\,\eta,c_{F+\gamma\,\eta})$ regarding to γ is given from (A.18):

$$\begin{aligned}
&\mathrm{TPR}'(F+\gamma\,\eta,c_{F+\gamma\,\eta})\\
=\;&\frac{1}{2}\int\!\!\int K^*(\boldsymbol{x}_0,\boldsymbol{x}_1)\Big(\eta(\boldsymbol{x}_1)-\eta(\boldsymbol{x}_0)\Big)\Big(\Lambda(\boldsymbol{x}_1)-\Lambda(\boldsymbol{x}_0)\Big)g_0(\boldsymbol{x}_0)g_0(\boldsymbol{x}_1)d\boldsymbol{x}_0d\boldsymbol{x}_1\\
&\qquad\Big/\int \mathrm{H}'(F(\boldsymbol{x})+\gamma\,\eta(\boldsymbol{x})-c_{F+\gamma\,\eta})g_0(\boldsymbol{x}_0)d\boldsymbol{x}_0, \qquad \text{(A.24)}
\end{aligned}$$

where

$$K^*(\boldsymbol{x}_0,\boldsymbol{x}_1)=\mathrm{H}'(F(\boldsymbol{x}_0)+\gamma\,\eta(\boldsymbol{x}_0)-c_{F+\gamma\,\eta})\mathrm{H}'(F(\boldsymbol{x}_1)+\gamma\,\eta(\boldsymbol{x}_1)-c_{F+\gamma\,\eta}), \qquad \text{(A.25)}$$

and when the denominator is not zero. Note that if the denominator is zero, we have $\mathrm{TPR}'(F+\gamma\,\eta,c_{F+\gamma\,\eta})=0$ as shown in Appendix 1. The domain of the integration in the numerator is determined by K^*, which is dependent on an arbitrary score function F. Hence, $\mathrm{TPR}'(F+\gamma\,\eta,c_{F+\gamma\,\eta})\geq 0$ only if $\eta=m(\Lambda)$, where m is a strictly increasing function. The sufficiency is confirmed easily. □

A.3 *Details of the pAUCBoost with natural cubic splines.*

For predetermined values of $\overline{\alpha}_1$ and $\overline{\alpha}_2$, calculate the corresponding thresholds $\overline{c}_1$ and $\overline{c}_2$ for all score functions $F_{t-1} + \beta_{t-1}(f)$, where f is in $\mathcal{F}$. The first and second derivative of the objective function, which are used in the Newton-Raphson iteration, are given as:

$$\begin{aligned}
&D_1(\beta_{t-1}(f)) \\
=&\frac{\partial}{\partial \beta_{t-1}(f)} \overline{\mathrm{pAUC}}_\lambda(F_{t-1} + \beta_{t-1}(f)f, \alpha_1, \alpha_2) \\
=&\frac{1}{n_0 n_1} \sum_{i \in I} \sum_{j \in J_{\mathrm{fan}}} \phi\Big(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i}) + \beta_{t-1}(f)\{f(\boldsymbol{x}_{1j}) - f(\boldsymbol{x}_{0i})\}\Big)\Big(f(\boldsymbol{x}_{1j}) - f(\boldsymbol{x}_{0i})\Big) \\
&\qquad - 2\lambda \int \big\{F''_k(x_k) + \beta_{t-1}(f) f''(\boldsymbol{x})\big\} f''_t(\boldsymbol{x}) dx_k,
\end{aligned} \tag{A.26}$$

and

$$\begin{aligned}
&D_2(\beta_{t-1}(f)) \\
=&\frac{\partial^2}{\partial \beta_{t-1}(f)^2} \overline{\mathrm{pAUC}}_\lambda(F_{t-1} + \beta_{t-1}(f)f, \alpha_1, \alpha_2) \\
=&-\frac{1}{n_0 n_1} \sum_{i \in I} \sum_{j \in J_{\mathrm{fan}}} \phi\Big(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i}) + \beta_{t-1}(f)\{f(\boldsymbol{x}_{1j}) - f(\boldsymbol{x}_{0i})\}\Big) \\
&\qquad \times \Big(F(\boldsymbol{x}_{1j}) - F(\boldsymbol{x}_{0i}) + \beta_{t-1}(f)\{f(\boldsymbol{x}_{1j}) - f(\boldsymbol{x}_{0i})\}\Big)\Big(f(\boldsymbol{x}_{1j}) - f(\boldsymbol{x}_{0i})\Big)^2 \\
&\qquad - 2\lambda \int \big\{f''(\boldsymbol{x})\big\}^2 dx_k,
\end{aligned} \tag{A.27}$$

where x_k is the k-th component of a p-dimensional marker vector $\boldsymbol{x}$; $f(\boldsymbol{x})$ is one of $N_{k,l}(x_k)/Z_{k,l}$'s ($l = 1, 2, \ldots, m_k$). The rectangular part of the objective function can be ignored because the first derivatives of TPR's can be approximated to zero by the same argument that the equation of (A.18) becomes zero, when we replace $g_0(\boldsymbol{x}_0)$ and $g_1(\boldsymbol{x}_1)$ with $1/n_0$ and $1/n_1$, respectively. This consideration reduces the amount of calculation by $n_0^* \times n_1^*$, where n_0^* and n_1^* are the cardinalities of I and J_{rec}, respectively. Then, we apply the Newton-Raphson method to get a set of coefficients at iteration time t:

$$\beta_t(f) = \beta_{t-1}(f) - \frac{D_1(\beta_{t-1}(f))}{D_2(\beta_{t-1}(f))}. \tag{A.28}$$

We observed that the value of $\beta_t(f)$ is unstable, especially when the cardinalities of I and J_{rec} are very small. So, we restricted the maximum absolute value to be 1. Using $\beta_t(f)$, the best weak classifier is chosen as

$$f_t = \underset{f}{\operatorname{argmax}}\ \overline{\text{pAUC}}_\lambda(F_{t-1} + \beta_t(f)f, \overline{\alpha}_1, \overline{\alpha}_2). \tag{A.29}$$

We repeat this process T times to get a final score function $F(\boldsymbol{x})$.

The penalty term for x_k consists of a linear combination of $P_{k,l,l'}$'s ($k = 1, \ldots, p; 1 \leq l,\ l' \leq m_k$), where

$$\begin{aligned}
&P_{k,l,l'}\\
&= \int N''_{k,l}(x_k)N''_{k,l'}(x_k)/(Z_{k,l}Z_{k,l'})dx_k\\
&= \begin{cases} 0,\ l \leq 2 \text{ or } l' \leq 2, \\ \frac{6}{(2\xi_{k,m_k}-\xi_{k,m_k-1}-\xi_{k,l-2})(2\xi_{k,m_k}-\xi_{k,m_k-1}-\xi_{k,l'-2})(\xi_{k,m_k}-\xi_{k,l'-2})(\xi_{k,m_k-1}-\xi_{k,l-2})(\xi_{k,m_k-1}-\xi_{k,l'-2})} \\ \times\left\{(\xi_{k,m_k}-\xi_{k,m_k-1})(\xi_{k,m_k-1}-\xi_{k,l-2})+(2\xi_{k,m_k-1}+\xi_{k,l'-2}-3\xi_{k,l-2})(\xi_{k,m_k-1}-\xi_{k,l'-2})\right\},\ \text{otherwise.} \end{cases}
\end{aligned} \tag{A.30}$$

Note that $P_{k,l,l'}$ depends on the scale of x_k. It means the penalty term also depends on the scale. Hence, by adjusting the scale, we make the maximum value of $P_{k,l,l'}$ in accordance for each k so that each x_k is penalized almost equally:

$$\max_{1 \leq l,\ l' \leq m_k} P_{k,l,l'} = P_{k,m_k,m_k} = 1. \tag{A.31}$$

On the other hand, a penalty term for x_k in the case where we use decision stumps as weak classifiers is given as:

$$\sum_{x_k \in \Xi_k} \left\{F_k^{(2)}(x_k)\right\}^2, \tag{A.32}$$

where $F_k^{(2)}(x_k)$ denotes the second-order difference of $F_k(x_k)$: $F_k^{(2)}(x_k) = F_k(x_k^{(-1)}) - 2F_k(x_k) + F_k(x_k^{(+1)})$ for $x_k^{(-1)} < x_k^{(+1)}$, and Ξ_k is a set of $\xi_{k,l}$ $(l = 1, 2, \ldots, m_k)$. In a similar manner to (A.30), the penalty term for x_k is given as a linear combination of $Q_{k,l,l'}$'s:

$$Q_{k,l,l'} = \sum_{x_k \in \Xi_k} S_{k,l}^{(2)}(x_k) S_{k,l'}^{(2)}(x_k) = \begin{cases} 2, \ l = l', \\ -1, \ |l - l'| = 1, \\ 0, \ |l - l'| \geq 2, \end{cases} \tag{A.33}$$

where $S_{k,l}^{(2)}(x_k)$ is a decision stump $S_{k,l}(x_k) = \mathrm{H}(x_k - \xi_{k,l})$. The values of $Q_{k,l,l'}$ are necessary when we use decision stumps together with natural cubic splines as weak classifiers.

Acknowledgements

During my PhD course, I have been supported, encouraged and, warmly guided by many nice professors, colleagues and friends. I am very grateful for all of them; however, there are three important persons to whom I especially have to give heartfelt thanks. I owe to them what I am now and what I am trying to become. The first one is Professor Ritei Shibata who is my pre-supervisor in Keio University and taught me many fundamental things that are necessary for becoming a full-fledged researcher. His severe teaching, training and encouraging have established my basis for current research attitude, toughness and patience. I am very happy that I can be under his tutelage even if it was a very short period.

The second person is medical doctor Kent Kanao in Keio University Hospital. My research has started by his visit to Shibata laboratory to consult with a medical data analysis. At that time, I was in the first grade of master course and was looking for my research topic. MD Kanao kindly accepted my proposal to analyse PSA data that had been collected in Keio University Hospital. We wrote about the results of collaborative work in Chapter 4, which aims to be helpful for medical doctors as well as patients in hospitals. I am looking forward to further cooperative work with MD Kanao and other doctors in Keio University Hospital.

The third person that I appreciate the most is Professor Shinto Eguchi. He accepted my proposal to be a member of Eguchi laboratory when I left Keio University and worried about where I should move on. He gave me a lot of chances to encounter many nice and respectable professors, researchers and his friends regardless of their nationalities. I was able to touch and understand various ways of thinking about statistics, and I was much affected by such nice persons in respect to my research attitude and the view of life as

a researcher. I have been making efforts as much as possible to catch up with all of the excellent persons. During my wonderful PhD course, he had me feeling and experiencing pleasure and excitement of research, which are the most important factor in researcher' life but I did not have before meeting him.

At last, I express my heartfelt thanks to all of my family members. They always helped and supported me not only physically but also mentally. Thank you very much; I will continue to do my best in the future, too.

Bibliography

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician* **52**, 119–126.

Antunes, A. A., Srougi, M. and Dall'Oglio, M. F. (2005). The percentage of positive biopsy cores as a predictor of disease recurrence in patients with prostate cancer treated with radical prostatectomy. *BJU Int* **96**, 1258–1263.

Baker, S. G. (2003). The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *Journal of the National Cancer Institute* **95**, 511–515.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12**, 387–415.

Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association* **101**, 138–156.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. (2000). Tissue classification with gene expression profiles. *Journal of Computation Biology* **7**, 559–583.

Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association* **98**, 324–339.

Cai, T. and Dodd, L. E. (2008). Regression analysis for the partial area under the ROC curve. *Statistica Sinica* **18**, 817–836.

Catalona, W. J. and Slawin, A. W. P. K. M. (1998). Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: a prospective multicenter clinical trial. *JAMA* **279**, 1542–1547.

Catalona, W. J., Smith, D. S. and Wolfert, R. L. (1995). Evaluation of percentage of free serum prostate-specific antigen to improve specificity of prostate cancer screening. *JAMA* **274**, 1214–1220.

Chambers, J. M. and Hastie, T. J. (1991). *Statistical Models in S*, California: Wadsworth and Brooks.

Cheng, L., Poulos, C. K. and Pan, C. X. (2005). Preoperative prediction of small volume cancer (less than 0.5 ml) in radical prostatectomy specimens. *J Urol* **174**, 898–902.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115**, 928–935.

Dall'Era, M. A., Cooperberg, M. R. and Chan, J. M. (2008). Active surveillance for early-stage prostate cancer: review of the current literature. *Cancer* **112**, 1650–1659.

Dettling, M. and Bühlmann, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19**, 1061–1069.

Dodd, L. E. and Pepe, M. S. (2003). Partial AUC estimation and regression. *Biometrics* **59**, 614–623.

Eguchi, S. and Copas, J. (2002). A class of logistic-type discriminant functions. *Biometrika* **89**, 1–22.

Epstein, J. I., Walsh, P. C., Carmichael, M. and Brendler, C. B. (1994). Pathologic and clinical findings to predict tumor extent of nonpalpable (stage T1c) prostate cancer. *JAMA* **271**, 368–374.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* **55**, 119–139.

Freund, Y., Iyer, R., Shapire, R. E. and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research* **4**, 933–969.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics* **28**, 337–407.

Greene, K. L., Albertsen, P. C. and Babian, R. J. (2009). Prostate specific antigen best practice statement. *J Uro* **182**, 2232–2241.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science* **1**, 297–318.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer.

Kanao, K., Mizuno, R., Kikuchi, E., Miyajima, A., Nakagawa, K., Ohigashi, T., Nakashima, J. and Oya, M. (2009). Preoperative prognostic nomogram (probability table) for renal cell carcinoma based on TNM classification. *The Journal of Urology* **181**, 480–485.

Kanao, K., Nakashima, J. and Nakagawa, K. (2006). Preoperative nomograms for predicting stone-free rate after extracorporeal shock wave lithotripsy. *J Urol* **176**, 1453–1456.

Kattan, M. W., Eastham, J. A. and Stapleton, A. M. (1998). A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. *J Natl Cancer Inst* **90**, 766–771.

Kawakita, M., Minami, M., Eguchi, S. and Lennert-Cody, C. E. (2005). An introduction to the predictive technique AdaBoost with a comparison to generalized additive models. *Fisheries Research* **76**, 328–343.

Komori, O. (2009). A boosting method for maximization of the area under the ROC curve. *Annals of the Institute of Statistical Mathematics*, DOI 10.1007/s10463–009–0264–y.

Lin, K., Lipsitz, R. and Janakiraman, T. M. S. (2008). Benefits and harms of prostate-specific antigen screening for prostate cancer: an evidence update for the U.S. Preventive Services Task Force. *Ann Intern Med* **149**, 192–199.

Long, P. M. and Servedio, R. A. (2007). Boosting the area under the ROC curve. *Neural Information Processing Systems.*

Luboldt, H. J., Fornara, P. and Wissbach, L. (2004). Systematic development of a guideline for early detection of prostate cancer: the German way in the evidence gap. *Eur Urol* **46**, 725–730.

Lugosi, B. G. and Vayatis, N. (2004). On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics* **32**, 30–55.

Ma, S. and Huang, J. (2005). Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* **21**, 4356–4362.

McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: Optimality of the risk score. *Biometrics* **58**, 657–664.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*, Hoboken: Wiley & Sons, Inc.

Murata, N., Takenouchi, T., Kanamori, T. and Eguchi, S. (2004). Information geometry of $\mathcal{U}$-boost and Bregman divergence. *Neural Computation* **16**, 1437–1481.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transaction of the Royal Society of London. Series A* **231**, 289–337.

Obuchowski, N. A., Graham, R. J., Baker, M. E. and Powell, K. A. (2001). Ten criteria for effective screening: Their application to multislice CT screening for pulmonary and colorectal cancers. *American Journal of Roentgenology* **176**, 1357–1362.

Ochiai, A., Troncoso, P. and Chen, M. E. (2005). The relationship between tumor volume and the number of positive cores in men undergoing multisite extended biopsy: implication for expectant management. *J Urol* **174**, 2164–2168.

Partin, A. W. and Lamm, L. A. M. D. M. (2001). Contemporary update of prostate cancer staging nomograms (Partin Tables) for the new millennium. *Urology* **58**, 843–848.

Pencina, M. J., D'Agostino Sr., R. B., D'Agostino Jr., R. B. and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**, 157–172.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and prediction*, New York: Oxford University Press.

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123–140.

Pepe, M. S., Cai, T. and Longton, G. (2006). Combining predictors for classification using the area under the Receiver Operating Characteristic curve. *Biometrics* **62**, 221–229.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W. and Newcomb, P. (2004). Limitation of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* **159**, 882–890.

Pepe, M. S., Longton, G., Anderson, G. L. and Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics* **59**, 133–142.

Qi, Y., Joseph, Z. B. and Seetharaman, J. K. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Structure, Function, and Bioinformatics* **63**, 490–500.

Rosset, S., Zhu, J. and Hastie, T. (2004). Boosting as regularized path to maximum margin classifier. *Journal of Machine Learning Research* **5**, 941–973.

Schroder, F. H., Hugosson, J. and Roobol, M. J. (2009). Screening and prostate-cancer mortality in a randomized European study. *N Engl J Med* **360**, 1320–1328.

Stangelberger, A., Waldert, M. and Djavan, B. (2008). Prostate cancer in elderly men. *Rev Urol* **10**, 111–119.

Su, J. Q. and Liu, J. S. (1993). Linear combination of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350–1355.

Takenouchi, T. and Eguchi, S. (2004). Robustifying AdabBoost by adding the naive error rate. *Neural Computation* **16**, 767–787.

Thompson, I., Thrasher, J. B. and Aus, G. (2007). Guideline for the management of clinically localized prostate cancer. *J Urol* **177**, 2106–2131.

Tutz, G. and Binder, H. (2006). Generalized Additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics* **62**, 961–971.

Ueki, M. and Fueda, K. (2009). Optimal tuning parameter estimation in maximum penalized likelihood method. *Annals of the Institute of Statistical Mathematics* in press: DOI 10.1007/s10463-008-0186-0.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

Walter, L. C., Bertenthal, D., Lindquist, K. and Konety, B. R. (2006). PSA screening among elderly men with limited life expectancies. *JAMA* **296**, 2336–2342.

Walter, S. D. (2005). The partial area under the summary ROC curve. *Statistics in Medicine* **24**, 2025–2040.

Wang, Z., Chang, Y. I., Ying, Z., Zhu, L. and Yang, Y. (2007). A parsimonious threshold-independent pretein feature selection method through the area under receiver operating characteristic curve. *Bioinformatics* **23**, 2788–1794.

Woodward, M. (2005). *Epidemiology: Study Design and Data Analysis, second edition*, U.S.A.: Chapman & Hall/CRC.

Zhang, T. and Yu, B. (2005). Boosting with early stopping: convergence and consistency. *The Annals of Statistics* **4**, 1538–1579.

Zhou, X., Obuchowski, N. A. and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*, New York: Wiley & Sons, Inc.