

# Recent Developments in Discriminant Analysis from an Information Geometric Point of View

Shinto Eguchi

Institute of Statistical Mathematics, Tokyo

and

John Copas

Department of Statistics, University of Warwick

## Abstract

This paper concerns the problem of classification based on training data. A framework of information geometry is given to elucidate the characteristics of discriminant functions including logistic discrimination and AdaBoost. We discuss a class of loss functions from a unified viewpoint.

*Keywords:* AdaBoost; Bayes rule; Information geometry; Kullback-Leibler divergence; LogitBoost; Logistic regression; U boost.

## 1 INTRODUCTION

There have been many recent advances in the methodology of classification amongst the communities of statistics, neural computation, machine learning, and artificial intelligence. In particular, the idea of the decision tree has influenced researchers involved in classification problems in several fields. See Breiman *et al.* (1984). A number of interesting proposals in this area have emerged from computational learning theory. Schapire (1990) presented a learning theoretic discussion of several proposed algorithms. The idea of combining weak learners has proved attractive and stimulating amongst the statistical community. See, for example, Freund (1995) for the concept of *boosting by majority*. Freidman *et al.* (2000) give an interesting interpretation of the boosting method from a statistical point of view. By developing a framework for the information geometry of recently proposed methods of classification, our aim in this paper is to give a more general geometric interpretation of discriminant analysis.

We consider a Utopia for statisticians where we can get a true underlying distribution  $g(x)$  rather than an empirical data set. In practice of course we have to estimate  $g$ , or test a hypothesis about  $g$  based on a given data set. The

maximum likelihood method has been applied to data in various context beyond the specialties originally envisaged by Fisher (1922). In our Utopia maximum likelihood is implemented by

$$\hat{f}(g) = \arg \max_{f \in \mathcal{M}} E_g \log f(X).$$

Here  $\mathcal{M}$  is a statistical model, which is usually parameterized by finite number of parameters. We observe that

$$E_g \log g(X) \geq E_g \log f(X)$$

with equality if and only if  $f = g$ . Thus the Kullback-Leibler divergence is deeply connected with the maximum likelihood method, of which geometrical discussion will be given in the following section.

We focus on a standard situation in a binary classification, where an input vector  $X$  and binary output  $Y$  have a joint distribution  $q(x, y) = \pi_y g_y(x)$  for  $x \in R^p$ ,  $y \in \{-1, +1\}$ . The corresponding posterior distribution,  $q(y|x)$ , is

$$q(y|x) = \frac{q(x, y)}{\sum_{y'} q(x, y')}.$$

The Bayes rule of allocation for any given  $x$  is

$$\begin{aligned} y(x) &= \arg \max_{y \in \{-1, +1\}} \log q(y|x) \\ &= \operatorname{sgn} \left( \log \frac{q(x, +1)}{q(x, -1)} \right), \end{aligned}$$

where  $\operatorname{sgn}(a)$  denotes the sign of  $a$  in the usual sense.

In practice we have to implement this using a model  $p \in \mathcal{M}^*$ , where

$$\mathcal{M}^* = \{p(x, y) = \pi_y f_y(x) : \pi_{-1} + \pi_{+1} = 1, f_{-1} \in \mathcal{M}, f_{+1} \in \mathcal{M}\}.$$

The universal applicability of maximum likelihood gives

$$\hat{p} = \hat{p}(q) = \arg \max_{p \in \mathcal{M}^*} E_q \log p(X, Y). \tag{1}$$

The corresponding classifier is just the plug-in procedure

$$\hat{y}(x) = \operatorname{sgn} \left( \log \frac{\hat{p}(x, +1)}{\hat{p}(x, -1)} \right).$$

Basically, this is an abstract version of Fisher discriminant analysis.

Alternatively, Day and Kerridge (1967) discussed a conditional likelihood approach, and proposed

$$\hat{p} = \hat{p}(q) = \arg \max_{p \in \mathcal{M}^*} E_q \log p(Y|X). \quad (2)$$

One of the aims of our paper is to give a discussion of the superiority of this conditional method (2) over the Fisher discriminant method (1).

We discuss loss functions for classifiers or discriminant functions which have the favorable property that the Bayes rule attains minimum risk. The exponential loss driving AdaBoost satisfies this property, in addition to the more familiar logistic loss. See Vapnik (1998), Friedman *et al.* (2000) and Schapire and Singer (2000). We present a class of loss functions for which boosting algorithms can be presented in a unified way. The idea of additive logistic model in Friedman *et al.* (2000) helps us to discuss boosting in a rather general setting.

## 2 BRIEF REVIEW OF INFORMATION GEOMETRY

Let us give a brief review of information geometry in statistics. See Amari (1985) and Nagaoka and Amari (2000). Differential geometric methods have been developed to offer comprehensive insights into statistical inference. Given a data space  $\mathcal{X}$ , a subset of the Euclidean space, the object for geometry is a space  $\mathcal{P}$  of probability distributions over  $\mathcal{X}$ . For simplicity we only discuss the regular case in which all probability measures or distributions are absolutely continuous with respect to a fixed measure  $\mu$ , so that we can identify each distribution as a density.

The Kullback-Leibler divergence over  $\mathcal{P}$ ,

$$\text{KL}(g, f) = E_g \left\{ \log \frac{g(X)}{f(X)} \right\}, \quad (3)$$

plays a fundamental role on the theory of statistical inference. Here  $E_g$  denotes mathematical expectation with respect to the distribution  $g$ . The pair of affine connections on this space, called the exponential and mixture connections, help us to grasp the essential asymmetry of  $\text{KL}(g, f)$  in  $f$  and  $g$ . There is thus a contrast with Riemannian space in which the Riemannian metric satisfies symmetry, and

for which the metric connection is unique. The exponential connection is defined by the geodesic

$$g_\theta(x) = c_\theta \{f(x)\}^\theta \{g(x)\}^{1-\theta}$$

connecting distributions  $f$  and  $g$  in  $\mathcal{P}$ , where  $c_\theta$  is the normalizing constant

$$c_\theta = \left( \int \{f(x)\}^\theta \{g(x)\}^{1-\theta} d\mu(x) \right)^{-1}.$$

This can be rewritten as

$$g_\theta(x) = g(x) \exp\{\theta t(x) - \psi(\theta)\}, \quad (4)$$

where  $t(x) = \log f(x)/g(x)$  and  $\psi(\theta) = \log c(\theta)$ . We can thus associate the exponential connection with the exponential family with parameter  $\theta$  and statistic  $t(x)$ . Here  $\theta$  is called the natural parameter and  $t(x)$  is called the canonical statistic. It can easily be extended to a geodesical plane of finite dimension, and the discussion of the infinite dimensional case leads to the mathematical formulation of infinite dimensional exponential family, as in Pistone (1995). The exponential connection leads to the exponential curvature tensor for a manifold embedded in the space  $\mathcal{P}$ . It is equivalent to the idea of statistical curvature introduced by Efron (1975).

The mixture connection is defined by

$$h_\eta(x) = \eta f(x) + (1 - \eta)h(x). \quad (5)$$

It is related with the maximum likelihood method as follows. Let

$$\mathcal{M} = \{f(x, \omega) : \omega \in \Omega\}$$

be a statistical model, where  $x = (x_1, \dots, x_p)$  and  $\omega = (\omega_1, \dots, \omega_d)$ . Suppose that we observe data  $\{x_i : i = 1, \dots, n\}$  from a distribution  $g(x)$ , which is not necessarily assumed to be in  $\mathcal{M}$ . See White (1982) for a general discussion. Then the average of the log-likelihood function on the statistical model  $\mathcal{M}$  is

$$\ell_n(f) = \frac{1}{n} \sum_{i=1}^n \log f(x_i), \quad f \in \mathcal{M} \quad (6)$$

which has the population version

$$\ell(f, g) = E_g\{\log f(X)\}.$$

Here we view the log-likelihood as a function on  $\mathcal{M}$  rather than on  $\Omega$ . We thus interpret the estimator to be a functional mapping the empirical distribution of the data into  $\mathcal{M}$ . The maximum likelihood estimator is equivalent to minimizing the Kullback-Leibler divergence in this population or abstract version (Huber, 1985),

$$\hat{f}(g) = \arg \min_{f^* \in \mathcal{M}} \text{KL}(g, f^*),$$

since  $\text{KL}(g, f)$  is equivalent to  $\ell(f, g)$  up to a constant. In fact  $\ell_n(f)$  converges almost surely to the expected log likelihood function  $\ell(f, g)$  as  $n \rightarrow \infty$ .

The inverse image of  $\hat{f}(g)$  is

$$\mathcal{A}(f) = \{g : \text{KL}(g, f) = \min_{f^* \in \mathcal{M}} \text{KL}(g, f^*)\},$$

which is called the ancillary leaf associated with the maximum likelihood method. By definition  $f \in \mathcal{A}(f)$ , which means that the estimator is Fisher-consistent, or the map  $\hat{f}$  is idempotent. We observe that the codimension of  $\mathcal{A}(f)$  is the same as the dimension of the statistical model  $\mathcal{M}$ . The tubular neighborhood around the model  $\mathcal{M}$  is expressed by

$$\bigcup_{f \in \mathcal{M}} \mathcal{A}(f).$$

Note that the leaf  $\mathcal{A}(f)$  intersects with  $\mathcal{M}$  at  $f$ , again implying the consistency of the maximum likelihood estimator. The leaf is geodesical in the sense of mixture connection, that is,

$$h_\eta \in \mathcal{A}(f) \quad (\forall \eta, 0 \leq \eta \leq 1)$$

where the mixture geodesic  $h_\eta$  as in (5) and  $h$  is in  $\mathcal{A}(f)$ . In accordance with this, we see that the mixture connection characterizes the maximum likelihood method.

Now we review a dualistic structure associated with the exponential and mixture connections, in which the Kullback-Leibler divergence plays a key role. Let us assume that  $f$ ,  $g$  and  $h$  satisfy

$$\text{KL}(h, g) = \text{KL}(h, f) + \text{KL}(f, g).$$

Then we find that, for any  $\theta$  and  $\eta$

$$\text{KL}(h_\eta, g) = \text{KL}(h_\eta, f) + \text{KL}(f, g_\theta),$$

where  $p_\theta$  and  $q_\eta$  are defined at (4) and (5), respectively. This dualistic structure leads to a new understanding about exponential families and maximum likelihood. Let  $\mathcal{N}$  be an exponential family and  $\mathcal{M}$  be a statistical model embedded in  $\mathcal{N}$ . Assume that  $\mathcal{M}$  is exponentially geodesical, meaning that any exponential geodesic (4) is also in  $\mathcal{M}$ . Now suppose we sample data from a statistical model within  $\mathcal{M}$ , and summarize these data by the sufficient summary  $\hat{g}$ . Then

$$\text{KL}(\hat{g}, f) = \text{KL}(\hat{g}, \hat{f}) + \text{KL}(\hat{f}, f) \quad (7)$$

for any  $f \in \mathcal{M}$ , where  $\hat{f} = \arg \min_{f^* \in \mathcal{M}} \text{KL}(\hat{g}, f^*)$ . Thus the Pythagorean theorem (7) leads to a dualistic Euclidean geometry. This can be thought of as a generalisation of the standard Pythagorean identity in the statistics of linear regression models, in which least squares has a simple geometric interpretation over Euclidean space. In this analogy, we could think of  $\hat{g}$  as an observed set of responses,  $f$  the set of expected values under some “true” regression model, and  $\hat{f}$  the corresponding set of fitted values. Equation (7) is like the standard least squares identity, in which “true mean square” equals “residual mean square” plus “regression mean square”. In this way we can interpret (7) as the analysis of variance for general applications of maximum likelihood.

### 3 CLASSICAL DISCRIMINANT ANALYSIS

We now consider an information geometric interpretation of discriminant analysis and pattern recognition. We are particularly interested in recent developments which have appeared in the literatures of statistics, computational learning theory, and neural networks, including the bagging method, the arching method, the boosting algorithm, and support vector machines. Our idea is to focus on a loss function for any given allocation rule. Let  $X$  and  $Y$  be input and output variables, where in our context  $X = (X_1, \dots, X_p)$  is an explanatory vector and  $Y$  a group label for  $X$ , whose joint distribution is

$$q(y, x) = \pi_y g_y(x)$$

for  $x \in R^p$  and  $y = 1, \dots, G$ , where  $\pi_y$  is the probability of the label  $y$  and  $g_y(x)$  is the conditional distribution of  $X$  given  $Y = y$ . For identification we assume that these distributions  $\{g_y(x) : y = 1, \dots, G\}$  are distinct.

### 3.1 Optimality of the Bayes rule

Let  $F(x, y)$  be a discriminant function that leads to the classification

$$\hat{y}_F(x) = \arg \max_{1 \leq y \leq g} F(x, y)$$

for any given feature vector  $x$ . For a measure to assess the performance of the discriminant function  $F(x, y)$ , we could take the error rate, or the misclassification probability,

$$\text{err}(F) = \text{P}(\hat{y}_F(X) \neq Y).$$

The Bayes rule defines the discriminant function by the posterior distribution

$$F^*(x, y) = q(y|x).$$

A fundamental result is that the Bayes rule yields the minimization of the error rate amongst the class of all discriminant functions, that is,

$$F^* = \arg \min_F \text{err}(F).$$

The Kullback-Leibler divergence between  $q(x, y) = \pi_y g_y(x)$  and  $p(x, y) = \pi'_y f_y(x)$  is

$$\text{KL}(q, p) = \sum_{y=1}^G \pi_y \text{KL}(g_y, f_y) + \sum_{y=1}^G \pi_y \log \frac{\pi_y}{\pi'_y}.$$

By Bayes theorem, the conditional distribution of  $Y = y$  given  $X = x$  is

$$p(y|x) = \frac{\pi_y g_y(x)}{p_{\text{mar}}(x)},$$

where  $p_{\text{mar}}$  is the marginal distribution of  $X$  given by

$$p_{\text{mar}}(x) = \sum_{y'=1}^G \pi_{y'} g_{y'}(x).$$

The Kullback-Leibler divergence between two joint distributions  $q$  and  $p$  can now be decomposed into

$$\text{KL}(q, p) = E_q \left\{ \log \frac{q(Y|X)}{p(Y|X)} \right\} + \text{KL}(q_{\text{mar}}, p_{\text{mar}}), \quad (8)$$

where  $p_{\text{mar}}$  and  $q_{\text{mar}}$  are the marginal distributions of  $X$  induced from  $p$  and  $q$  respectively.

### 3.2 Full and conditional likelihood

Our model is

$$\mathcal{M}^* = \{p(x, y) = \pi_y f_y(x) : \sum_{y=1}^G \pi_y = 1, f_y \in \mathcal{M} (y = 1, \dots, G)\}.$$

Thus the dimension of  $\mathcal{M}^*$  is  $(d + 1)G - 1$ , where  $d$  is the dimension of  $\mathcal{M}$ .

Given a training data set  $\{(x_i, y_i) : i = 1, \dots, n\}$ , the maximum likelihood method can be applied directly to the log-likelihood function

$$\ell_n(p) = \frac{1}{n} \sum_{i=1}^n \sum_{y=1}^G z_y(y_i) \log f_y(x_i) + \frac{n_y}{n} \sum_{y=1}^G \log \pi_y,$$

where  $z_a(b) = 1$  if  $a = b$  and  $= 0$  otherwise, and  $n_y$  is the count of the number of observations with  $y_i = y$ . Let

$$\hat{p} = \arg \max_{p \in \mathcal{M}^*} \ell_n(p)$$

which gives the discriminant function

$$\hat{F}(x, y) = \log \hat{p}(y|x), \tag{9}$$

where  $\hat{p}(y|x) = \hat{p}(x, y) / \sum_{y'} \hat{p}(x, y')$ . Thus maximum likelihood estimation reduces to  $\hat{F}$  using the plug-in-rule.

An alternative approach is implemented by logistic regression. The conditional log-likelihood is

$$\ell^{(C)}(p) = \frac{1}{n} \sum_{i=1}^n \sum_{y=1}^G z_y(y_i) \log p(y|x)$$

whose maximizer  $\hat{p}^{(C)}$  directly leads to the estimated  $\hat{F}^{(C)}(x, y) = \hat{p}^{(C)}(y|x)$  without any reduction as in (9).

Recall that the general theory of maximum likelihood estimation is equivalent to the minimization of the Kullback-Leibler divergence from  $q(x, y)$  to  $p(x, y)$  in the abstract sense discussed in the Introduction. The population version  $\hat{p}^{(C)}$  is equivalent to the minimization of  $-E_q\{\log p(Y|X)\}$ , or

$$E_q \left\{ \log \frac{q(Y|X)}{p(Y|X)} \right\}, \tag{10}$$

which is the first term of the right side of (8). An important characteristic of the conditional estimator  $\hat{F}^{(C)}(x, y)$  is that it is robust to misspecification of the marginal distribution. See Day and Kerridge (1967) and Efron (1975) for the original discussion of robustness in this sense.

## 4 NEW DEVELOPMENTS

### 4.1 Binary classification

For notational convenience we now confine ourselves to the case of binary classification  $g = 2$ , letting  $y = -1, +1$  as discussed in Introduction. We can easily extend this to a multi-class classification problem by using some additional notation. We now restrict to a class of classification rules, the class  $\mathcal{F}$  of all discriminant functions  $F(x)$ . Note that because we are now in the binary case we are able to simplify the notation for discriminant functions from  $F(x, y)$  to  $F(x)$  by defining, for any given  $F(x)$  in  $\mathcal{F}$ , the prediction of  $Y$  as

$$\hat{y}_F(x) = \text{sgn}(F(x)).$$

The error rate is now expressed by

$$\text{err}(F) = \text{P}(YF(X) < 0).$$

Defining the log-likelihood ratio

$$\Lambda(x) = \log \frac{\pi_{+1}p_{+1}(x)}{\pi_{-1}p_{-1}(x)}, \quad (11)$$

leads to the predictor  $\hat{y}_\Lambda$ . The Bayes rule is given by the discriminant  $\Lambda$ , and so attains the minimum of  $\text{err}$  in  $\mathcal{F}$  as discussed above in the multi-class case.

### 4.2 $U$ loss function

To generalize our discussion, we now consider loss functions with a natural optimality property in terms of the error rate. A loss function  $L$  over  $\mathcal{F}$  is said to be Bayes-optimal if

$$\Lambda = \arg \min_{F \in \mathcal{F}} L(F). \quad (12)$$

We introduce the loss function  $L_U$  by

$$L_U(F) = E\{U(YF(X))\},$$

calling the  $U$  loss function, where  $U$  a generic function. We first give some examples of  $U$  and then prove a theorem for establishing Bayes-optimality.

EXAMPLES

(i) Error rate: Let  $U(z) = 1$  if  $z < 0$  and 0 otherwise. Then the resulting loss function  $L_U$  is the error rate.

(ii) Logistic regression: the generic function

$$U(z) = -\log \frac{\exp(z)}{1 + \exp(z)}$$

gives minus the logistic regression loss

$$E \left\{ \log \frac{\exp(YF(X))}{1 + \exp(YF(X))} \right\}$$

(iii) Exponential loss: Let  $U(z) = \exp(-\frac{1}{2}z)$ . In this case the resulting loss function is

$$E \left[ \exp \left\{ \frac{1}{2} Y F(X) \right\} \right]$$

which leads to the AdaBoost algorithm. Note that the usual definition does not have the factor  $\frac{1}{2}$ , but we use this version to facilitate the comparison with other loss functions in our unified discussion. We will see that the functional optimization of the exponential loss function leads to the Hellinger distance, which is another motivation for modifying the usual definition by this factor  $\frac{1}{2}$ .

(iv) Squared error: The generic function  $U(z) = \frac{1}{2}\{1 + \exp(z)\}^{-2}$  presents the squared error:

$$E \left[ \left\{ \frac{1 + Y}{2} - \frac{\exp(F(X))}{1 + \exp(F(X))} \right\}^2 \right]$$

since

$$\left\{ \frac{1 + y}{2} - \frac{\exp(f)}{1 + \exp(f)} \right\}^2 = \frac{1}{\{1 + \exp(yf)\}^2}$$

for  $y \in \{-1, +1\}$ .

The following theorem allows us to show that all these loss functions satisfy Bayes optimality (12).

**THEOREM 1.**

*Let  $U$  be a decreasing function such that*

$$U'(-z) = \exp(z)U'(z). \tag{13}$$

Then the loss function  $L_U(F) = E_p\{U(YF(X))\}$  satisfies

$$\Lambda = \arg \min_{F \in \mathcal{F}} L_U(F). \quad (14)$$

PROOF. We get

$$\begin{aligned} E\{U(YF(X)) - U(Y\Lambda(X)) | X = x\} &= P(Y = +1 | X = x)\{U(F(x)) - U(\Lambda(x))\} \\ &\quad + P(Y = -1 | X = x)\{U(-F(x)) - U(-\Lambda(x))\}, \end{aligned}$$

which is

$$P(Y = +1 | X = x) \left[ \int_{\Lambda(x)}^{F(x)} U'(z) dz - \exp\{-\Lambda(x)\} \int_{\Lambda(x)}^{F(x)} U'(-z) dz \right].$$

Hence it follows from the assumption on  $U$  that

$$\begin{aligned} &E\left\{\left(U(YF(X)) - U(Y\Lambda(X))\right) | X = x\right\} \\ &= P(Y = +1 | X = x) \int_{\Lambda(x)}^{F(x)} (1 - \exp\{z - \Lambda(x)\}) U'(z) dz \geq 0. \end{aligned}$$

This completes the proof.

We observe that the generic functions defined in (i)-(iv) above all satisfy (13), so the corresponding loss functions are all Bayes optimal. However, the naive square error  $E\{Y - F(X)\}^2$  is not Bayes optimal, as the generic function  $(z - 1)^2$  violates the condition (13).

Any loss function satisfying Bayes optimality naturally leads us to a divergence over  $\mathcal{F}$ :

$$D(\Lambda, F) = E\{U(YF(X)) - U(Y\Lambda(X))\}.$$

By definition this divergence  $D$  is nonnegative and equals zero only if  $F = \Lambda$ .

Friedman, *et al.* (2000) propose the LogitBoost based on the loss function (ii) in contrast with the AdaBoost algorithm associated with (iii). The two loss functions are apparently different, however we can smoothly connect them by the one-parameter family

$$U_\alpha(z) = \int^z \frac{dz}{\exp\left(\frac{1+\alpha}{2} z\right) + \exp\left(\frac{1-\alpha}{2} z\right)}.$$

In fact  $U_\alpha$  satisfies (13), and  $U_1$  and  $U_0$  are equal to the functions defined in (ii) and (iii) respectively.

We can also observe another proof given by a straightforward extension of the proof for the special case of the exponential loss (iii) given by Friedman *et al.* (2000). In fact

$$E(U(YF(X)) - U(Y\Lambda(X))|X = x) = \\ P(Y = +1|X = x)U(F(x)) + P(Y = -1|X = x)U(-F(x)),$$

so differentiation yields

$$\frac{\partial}{\partial F(x)} E \{U(YF(X))|X = x\} = \\ U'(F(x)) \{P(Y = +1|X = x) - \exp(F(x))P(Y = -1|X = x)\} \quad (15)$$

because of the assumption (13). Hence (15) vanishes only when  $F = \Lambda$ , and so the conditional expectation of  $U(YF(X))$  given  $X = x$  has a minimum at  $F = \Lambda$ . This follows by noting that if  $\Lambda_\epsilon(x) = (1 - \epsilon)\Lambda(x) + \epsilon F(x)$  then

$$\frac{\partial^2}{\partial \epsilon^2} E \{U(Y\Lambda_\epsilon(X))|X = x\} \Big|_{\epsilon=0} = U'(\Lambda(x))P(Y = +1|x)\{\Lambda(x) - F(x)\}^2 > 0$$

from the assumption on  $U$ .

Eguchi and Copas (2001) give another interesting proof from the Neyman-Pearson lemma. Here, we consider the problem of testing hypothesis about the underlying distribution  $f$ , testing the null hypothesis H:  $f = g_{+1}$  against the alternative hypothesis A:  $f = g_{-1}$ . The Neyman-Pearson lemma asserts that the log-likelihood ratio test is the most powerful. Let

$$\delta(z) = \pi_{+1}\{P_{+1}(\Lambda(X) > z) - P_{+1}(F(X) > z)\} \\ - \exp(z)\pi_{-1}\{P_{-1}(\Lambda(X) > z) - P_{-1}(F(X) > z)\}.$$

The Neyman-Pearson lemma also implies that  $\delta(z) \geq 0$  for all  $z$ . After some algebra, we arrive at the integration formula

$$L_U(F) - L_U(\Lambda) = - \int_{-\infty}^{\infty} U'(z)\delta(z)dz.$$

We mention that another familiar measure of performance of a discriminant function, the Area under the ROC Curve, corresponds to a  $U$  function outside

our class. This loss function is

$$L(F) = - \int_{-\infty}^{\infty} G_{-1}(z) dG_{+1}(z)$$

which is minus one times the area under the ROC curve. See Eguchi and Copas (2001) for an application to medical screening. This suggests the following enlargement to our class of loss functions, by defining

$$U^*(z) = \int^z \exp(t) U'(t) dt$$

and

$$L_U(F) = E \left[ \frac{y+1}{2} U(F(X)) + \frac{y-1}{2} U^*(F(X)) \right].$$

An example is the one-parameter family of generic functions

$$U_\beta(z) = \frac{1 - \exp(\beta z)}{\beta} \quad \text{and} \quad U_\beta^*(z) = \frac{1 - \exp((1-\beta)z)}{1-\beta}.$$

The limits of  $U_\beta$  are

$$\lim_{\beta \rightarrow 0} U_\beta(z) = z \quad \text{and} \quad \lim_{\beta \rightarrow 0} U_\beta^*(z) = \exp(z) - 1$$

and

$$\lim_{\beta \rightarrow 1} U_\beta(z) = 1 - \exp(-z) \quad \text{and} \quad \lim_{\beta \rightarrow 1} U_\beta^*(z) = z.$$

### 4.3 Optimized loss functions

We have introduced a class of Bayes optimal loss functions on the space  $\mathcal{F}$  of discriminant functions through Section 2.4. We now investigate the explicit form of the loss function  $L_U(F)$  when the discriminant function  $F$  is the log-likelihood  $\Lambda$ . In general we get

$$\begin{aligned} L_U(\Lambda) &= \int \pi_{+1} p_{+1}(x) U \left( \log \frac{\pi_{+1} p_{+1}(x)}{\pi_{-1} p_{-1}(x)} \right) d\mu(x) \\ &\quad + \int \pi_{-1} p_{-1}(x) U^* \left( \log \frac{\pi_{+1} p_{+1}(x)}{\pi_{-1} p_{-1}(x)} \right) d\mu(x). \end{aligned}$$

For our four examples (i)-(iv) above, we have the following expressions:

EXAMPLES (continued)

(i) Error rate is

$$\begin{aligned} L_U(\Lambda) &= \int \pi_{+1} p_{+1}(x) 1_{[\pi_{+1} p_{+1}(x) \leq \pi_{-1} p_{-1}(x)]} d\mu(x) \\ &\quad + \int \pi_{-1} p_{-1}(x) 1_{[\pi_{+1} p_{+1}(x) > \pi_{-1} p_{-1}(x)]} d\mu(x). \end{aligned}$$

(ii) Logistic loss is

$$L_U(\Lambda) = -E\{\log p(Y|X)\}.$$

As discussed at the end of Section 3.2, the minimization of the logistic loss is equivalent to maximizing the conditional likelihood.

(iii) Exponential loss is

$$L_U(\Lambda) = \sqrt{\pi_{+1}\pi_{-1}} \int \sqrt{p_{+1}(x)p_{-1}(x)} d\mu(x).$$

The squared Helinger distance is

$$H^2(\pi_{+1}p_{+1}, \pi_{-1}p_{-1}) = \int \left\{ \sqrt{\pi_{+1}p_{+1}(x)} - \sqrt{\pi_{-1}p_{-1}(x)} \right\}^2 d\mu(x),$$

which is connected with the exponential loss by  $H^2 = 2 - 2L_U$ .

(iv) Squared error is

$$L_U(\Lambda) = \int \frac{\pi_{+1}p_{+1}(x)\pi_{-1}p_{-1}(x)}{\pi_{+1}p_{+1}(x) + \pi_{-1}p_{-1}(x)} d\mu(x).$$

If we consider the  $\chi$ -square divergence

$$\chi^2(\pi_{+1}p_{+1}, \pi_{-1}p_{-1}) = \int \frac{\{\pi_{+1}p_{+1}(x) - \pi_{-1}p_{-1}(x)\}^2}{\pi_{+1}p_{+1}(x) + \pi_{-1}p_{-1}(x)} d\mu(x),$$

then  $\chi^2 = 1 - 4L_U$ .

## 4.4 $U$ Boost

In this section we discuss the implementation of the classification rule based on a given loss function, and using a given set of training data. Our approach is to model the discriminant function in the total space  $\mathcal{F}$ . First, we consider a classical setting, in which our model  $\mathcal{M}^*$  embedded in  $\mathcal{F}$  is given by the linear form

$$\mathcal{M}^* = \left\{ F(x, \beta) = \sum_{j=1}^p \beta_j t_j(x) : \beta \in B \right\}. \quad (16)$$

This model can easily be generalized by introducing some flexibility into the functions  $t_j(x)$ , possibly with their own shape parameters, including for example sigmoidal functions, wavelet basis functions or radial basis functions. For all such models we use the notation  $F(\cdot, \beta)$  for the general parametric form of the discriminant function.

We begin with a key equation:

$$yU'(yF(x)) = \frac{y^* - p(x)}{A(F(x))} \quad (17)$$

where

$$y^* = \frac{y+1}{2}, \quad p(F) = \frac{\exp\{F\}}{1 + \exp\{F\}}, \quad A_U(F) = \{U'(F)(1 + \exp\{F\})\}^{-1}.$$

We note that if  $U$  is the generic function of logistic loss as defined at (ii) in Section 3.3, then  $A_U(F) = 1$ .

Given training data  $\{(x_i, y_i) : i = 1, \dots, N\}$  the empirical version of the loss function  $L_U(F)$  is given by

$$\bar{L}_U(F) = \frac{1}{2N} \sum_{i=1}^N \{(1 + y_i)U(F(x_i)) + (1 - y_i)U^*(F(x_i))\}.$$

For the linear model (16) or a general model  $F(\cdot, \beta)$ , the gradient vector of the empirical loss is, noting equation (17),

$$\sum_{i=1}^N \frac{y_i^* - p(F(x_i, \beta))}{A_U(F(x_i, \beta))} \frac{\partial}{\partial \beta} F(x_i, \beta) = 0,$$

which reduces to a weighted logistic estimating equation with the weight function  $\{A_U(F(x, \beta))\}^{-1}$ .

Now let us consider a form of additive generalized model:

$$F(x) = \sum_{j=1}^M f_j(x).$$

Here we evolve  $F(x)$  into  $F(x) + f(x)$  using the loss function  $L_U(F)$ . A simple calculation gives

$$\begin{aligned} s_U(x) &= \frac{\partial}{\partial f(x)} E[U(Y\{F(X) + f(X)\} | X = x)] \Big|_{f(x)=0} \\ &= E\{YU'(YF(X)) | X = x\} \end{aligned}$$

and

$$\begin{aligned} H_U(x) &= \frac{\partial^2}{\partial f(x)^2} E[U(Y\{F(X) + f(X)\} | X = x)] \Big|_{f(x)=0} \\ &= E\{U''(YF(X)) | X = x\} \end{aligned}$$

From these the Newton update is given by

$$F(x) - \frac{E \{YU'(YF(X))|X = x\}}{E \{U''(YF(X))|X = x\}}.$$

In this way a straightforward extension of LogitBoost in Friedman *et al.* (2000) is proposed in the following:

1. Start with weights  $w_i = 1/n, i = 1, \dots, n, F(x) = 0$ .
2. Repeat for  $m = 1, \dots, M$ :
  - (a) Fit the regression function  $f_m(x)$  by weighted least-squares of  $y_i$  on  $x_i$  with weights  $w_i$ ,
  - (b) Update  $F(x) \leftarrow F(x) + f_m(x)$ ,
  - (c) Update  $w_i \leftarrow w_i U''(-y_i f_m(x_i))$  and renormalize.
3. Output the classifier  $\text{sgn}[F(x)] = \text{sgn} \left[ \sum_{m=1}^M f_m(x) \right]$ .

In practice this can also be seen as an extension of Gentle AdaBoost (Friedman *et al.*, 2000). However, this formulation is quite formal, so to study the applicability of the method we would have to investigate its properties in a more concrete setting, including worst case evaluations and benchmark tests for the particular form of the  $U$  function being assumed.

## References

- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont.
- Day, N. E. and Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics* **23**, 313-323.
- Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis is. *J. American Statistical Association*. **70**, 892-898.
- Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics* **3**, 1189-1217.
- Eguchi, S. and Copas, J. B. (2000). A class of logistic-type discriminant functions. In revision for *Biometrika*.

- Fisher (1922). On the mathematical foundations of the theoretical statistics. *Phi. Trans. R. Soc., A*, **222**, 309-368.
- Freund, Y. (2001). Boosting a weak learning algorithm by majority. *Information and Computation*, **121**(2), 256-285.
- Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics*, **28**, 337-407.
- Huber, P. J. (1985). Projection pursuit (with discussion). *Annals of Statistics* **13**, 435- 475.
- Pistone, G. and Sempi, C. (1995). An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Annals of Statistics* **23**, 1543-1561.
- Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, **5**, 197-227.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-26.