Research Report

Probability Prediction and Classification

Shelby J. Haberman

Research & Development

May 2004
RR-04-19

Probability Prediction and Classification

Shelby J. Haberman

ETS, Princeton, NJ

May 2004

Abstract

Criteria for prediction of multinomial responses are examined in terms of estimation bias. Logarithmic penalty and least squares are quite similar in behavior but quite different from maximum probability. The differences ultimately reflect deficiencies in the behavior of the criterion of maximum probability.

Key words: penalty function, entropy, concentration, misclassification rate, large-sample approximation

Acknowledgements

The author would like to thank Henry Braun, Neil Dorans, and Paul Holland for helpful comments.

Introduction

In statistical applications, it is common to predict a polytomous variable Y by use of one or more continuous or discrete variables. Such applications are encountered in the study of educational testing. A response to a multiple-choice item is a polytomous variable; commonly used holistic scores in grading of essays are polytomous, as are many validity criteria such as whether graduation. Polytomous responses may or may not have obvious associated numerical values. A simple numerical description is not appropriate for the response to a multiple choice question, for there may be a correct response, there may be three different incorrect responses, no response at all may exist, or there may be an invalid response in which more than one choice is marked. On the other hand, a holistic essay score may be a number from 1 to 6, with a higher number indicating a better response. Especially in cases in which no appropriate numerical values correspond to the values of Y, there are basic questions concerning the meaning of a prediction of Y. Given a definition of a prediction of Y, there is then the problem of criteria for evaluation of the prediction. These problems have been treated extensively in the statistical literature (Savage, 1971; Goodman & Kruskal, 1954; Haberman, 1982a; Haberman, 1982b; Gilula & Haberman, 1995b); nonetheless, it is not often appreciated that the appropriate strategy for prediction depends quite strongly on the criterion used to assess the quality of the prediction. In Section 1, prediction criteria based on penalty functions are developed (Haberman, 1982a; Haberman, 1982b), and the criteria based on squared error penalty, logarithmic probability penalty, and misclassification penalty are introduced. In Section 2, samples are used to develop probability predictions (Haberman, 1982a; Haberman, 1982b; Gilula & Haberman, 1995b). In Section 3, some large-sample results are used in a simple case to show that criteria based on misclassification penalty are much different in nature than criteria based on squared error penalty or on logarithmic penalty. This comparison appears to be new. Section 4 examines consequences of the results derived. The most important conclusion is that use of classification error rates is a questionable approach despite its intuitive appeal.

Table 1.

Joint Probabilities of Human and Machine Scores

Machine			Human score				
score	1	2	3	4	5	6	Total
1	0.0080	0.0120	0.0010	0.0010	0.0005	0.0005	0.0230
2	0.0240	0.0440	0.0760	0.0160	0.0010	0.0010	0.1620
3	0.0040	0.0440	0.1000	0.0760	0.0040	0.0010	0.2290
4	0.0010	0.0080	0.0960	0.1760	0.0680	0.0010	0.3500
5	0.0010	0.0040	0.0040	0.0400	0.0800	0.0480	0.1770
6	0.0005	0.0005	0.0010	0.0010	0.0240	0.0320	0.0590
Total	0.0385	0.1125	0.2780	0.3100	0.1775	0.0835	1.0000

1 Penalty Functions

To examine the problem of prediction of polytomous variables, consider a polytomous response random variable Y with values in a finite set $S = \{y_i : 1 \leq i \leq s\}$ with s elements and an explanatory random variable X with values in some space T. For instance, in the case of a holistic essay score from 1 to 6, S is just the set of integers from 1 to 6 and each y_i is the integer i. The variable Y might be the final holistic score obtained from human raters for a randomly selected essay for a specific prompt, and X might be the machine-derived holistic score for the same essay. For illustrative purposes, a joint distribution of X and Y is provided in Table 1. These probabilities are comparable with reported sample data (Feng et al., 2003). In this case, S and T are the same. Let $p_{Y \cdot X}(y|x)$ be the conditional probability that $Y = y$ given that $X = x$ for x in T and y in S, and let $p_Y(y)$, be the marginal probability that $Y = y$. In the essay example, $p_{Y \cdot X}(3|3) = 0.1000/0.2290 = 0.4367$ is the conditional probability of a human score of 3 given a machine score of 3, and $p_Y(3) = 0.2780$ is the marginal probability of a human score of 3. The conditional distribution of Y given X is fully described by the conditional probability vectors $\mathbf{p}_{Y \cdot X}(x)$ with coordinates $p_{Y \cdot X}(y_i|x)$ for integers i from 1 to s. Similarly, the marginal distribution of Y is completely described by the s-dimensional vector $\mathbf{p}_Y$ with coordinates $p_Y(y_i)$ for integers i from 1 to s. In this paper, basic operations on s-dimensional vectors will be common. If $\mathbf{c}$ is an s-dimensional vector with coordinates c_i for $1 \le i \le s$, then the squared Euclidean norm of $\mathbf{c}$ is defined by

$$|\mathbf{c}|_2^2 = \sum_{i=1}^{s} c_i^2,$$

and the maximum norm of $\mathbf{c}$ is defined by

$$|\mathbf{c}|_\infty = \max_{1 \le i \le s} |c_i|.$$

The tie function $t(\mathbf{c})$ is the number of members of the set $A(\mathbf{c})$ of integers i from 1 to s such that $|c_i| = |\mathbf{c}|_\infty$. At times, infinite quantities must be considered. The convention is adopted that $\log 0 = -\infty$ and $0\infty = 0$. The variance of a random variable V is denoted by $\sigma^2(V)$.

To examine sampling, let X_h and Y_h, $1 \le h \le n$, be sampled random variables such that each pair (X_h, Y_h), $1 \le i \le n$, is mutually independent, independent of (X, Y), and distributed as (X, Y). Thus in the essay example, there would be n observed essays. For essay h, X_h would be the machine-derived holistic score, and Y_h would be the human essay score.

The basic problem under study is prediction of the response variable Y by the explanatory variable X. As previously noted, because Y is polytomous and the set of possible values S may not have a useful numerical representation, it is not necessarily appropriate to approximate Y by a single numerical value. On the other hand, Y always has an s-dimensional vector representation $\mathbf{Z}$ with coordinates Z_i for $1 \le i \le s$. Let $\boldsymbol{\delta}_i$ be the s-dimensional vector with coordinates δ_{ij}, $1 \le j \le q$, such that $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ if $i \ne j$. If $Y = y_i$, then $\mathbf{Z} = \boldsymbol{\delta}_i$. Obviously Y determines $\mathbf{Z}$. Given $\mathbf{Z}$, Y is identified as the value y_i such that $Z_i = 1$. For each integer i, the coordinate Z_i is always nonnegative, and the sum $\sum_{i=1}^{s} Z_i = 1$. Let a superscript T denote a transpose. Then in the case of holistic scoring, $\mathbf{Z}$ is the six-dimensional vector $(0, 0, 1, 0, 0, 0)^T$ if the holistic score Y is 3. The unconditional expected value of $\mathbf{Z}$ is $\mathbf{p}_Y$, and $\mathbf{p}_{Y\cdot X}(x)$ is the conditional expected value of $\mathbf{Z}$ given $X = x$. Naturally, $p_Y(y)$ and $p_{Y\cdot X}(y|x)$ are both nonnegative, and

$$\sum_{i=1}^{s} p_Y(y_i) = \sum_{i=1}^{s} p_{Y\cdot X}(y_i|x) = 1.$$

Thus the observation $\mathbf{Z}$ and the probability vectors $\mathbf{p}_Y$ and $\mathbf{p}_{Y\cdot X}(x)$ are all members of the unit simplex Q of s-dimensional vectors $\mathbf{a}$ with nonnegative coordinates a_i and sum $\sum_{i=1}^{s} a_i = 1$, and Z is always a vertex of Q. In this paper, probability predictors are studied. Here a probability predictor $\mathbf{q}(X)$ is a random vector such that $\mathbf{q}$ is a function from the space T of values of X to the unit simplex Q (Savage, 1971; Haberman, 1982a; Haberman, 1982b). In this fashion, $\mathbf{p}_{Y\cdot X}(X)$ is a probability predictor, as is the constant predictor $\mathbf{p}_{YC}$ equal to $\mathbf{p}_Y$ for any possible value of X. One may regard a probability predictor as an approximation of $\mathbf{Z}$. Because the original observation Y is a one-to-one function of the vector $\mathbf{Z}$, a probability predictor also provides a type of prediction of Y.

To study probability prediction, accuracy of prediction must be considered. Several common approaches exist that can be described within a common framework (Savage, 1971; Haberman, 1982a; Haberman, 1982b) based on a nonnegative and possibly infinite penalty function L designed to measure the discrepancy between the observed vector Z and the probability predictor $\mathbf{q}(X)$. For any value y of Y and any member $\mathbf{a}$ of the unit simplex, L assumes a value $L(y, \mathbf{a})$. The observed penalty is $L(y, \mathbf{q}(x))$ if $Y = y$ and $X = x$. In this report, the three penalty functions considered are the squared error penalty function L_C defined by

$$L_C(y_i, \mathbf{a}) = |\boldsymbol{\delta}_i - \mathbf{a}|_2^2,$$

with the logarithmic penalty function L_H defined by

$$L_H(y_i, \mathbf{a}) = -\log a_i,$$

and the misclassification penalty function L_M defined by

$$L_M(y_i, \mathbf{a}) = \begin{cases} 1, & i \notin A(\mathbf{a}), \\ 1 - 1/t(\mathbf{a}), & i \in A(\mathbf{a}). \end{cases}$$

Thus the squared error function $L_C(Y, \mathbf{a})$ is the squared Euclidean distance between the observed vector $\mathbf{Z}$ and the probability prediction $\mathbf{a}$. If $Y = y_i$, then the logarithmic probability penalty $L_H(Y, \mathbf{a}) = -\log a_i$ is minus the logarithm of the probability a_i predicted for the observed value y_i of Y. The misclassification rate penalty $L_M(Y, \mathbf{a})$ is based on the idea of classification of Y as the value y_i with the highest probability a_i. In this fashion, $L_M(Y, \mathbf{a})$ is 1 if the observed value y_i of Y has assigned probability a_i less than the maximum probability $|\mathbf{a}|_\infty$ assigned to a value of Y, so that Y is not classified correctly by the classification rule. If the assigned probability a_i for $Y = y_i$ is larger than any other probability a_j assigned to $Y = y_j \neq y_i$, then $L_M(Y, \mathbf{a})$ is 0, for Y is classified correctly by the classification rule. The case of $t(\mathbf{a}) > 1$ and i in $A(\mathbf{a})$ is a bit more complicated. In this instance, classification is ambiguous, so that Y is randomly classified with probability $1/t(\mathbf{a})$ as y_j if j is in $A(\mathbf{a})$. For $Y = y_i$, the probability of incorrect classification is then $1 - 1/t(\mathbf{a})$.

As an example, consider the case of human essay scoring. Let the actual score be 3, and let a probability predictor assign probability 0.1 to scores 1 and 6, probability 0.15 to scores 2 and 5, and probability 0.25 to scores 3 and 4. Here the maximum predicted probability 0.25 is assigned to both scores 3 and 4. The squared error penalty is

$$(0 - 0.1)^2 + (0 - 0.15)^2 + (1 - 0.25)^2 + (0 - 0.25)^2 + (0 - 0.15)^2 + (0 - 0.1)^2 = 0.69,$$

the logarithmic penalty is

$$-\log(0.25) = 1.386,$$

and the misclassification rate penalty is

$$1 - 1/2 = 0.5.$$

The penalty function L satisfies the regularity conditions that the penalty $L(Y, \mathbf{a})$ is finite if the probability a_i assigned to the outcome y_i is positive and if $Y = y_i$. It is also assumed that the penalty $L(Y, \mathbf{a})$ is 0 if, and only if, the probability a_i assigned to the outcome y_i is 1 and $Y = y_i$, so that $\mathbf{a} = \mathbf{Z}$. These requirements hold if L is L_C, L_H, or L_M. In addition, L_C never exceeds 2, and L_M never exceeds 1.

The fundamental assumption is that the smallest expected penalty from use of a constant prediction function is observed if the prediction function is $\mathbf{p}_{YC}$. For members $\mathbf{a}$ and $\mathbf{b}$ of the unit simplex, let

$$D^*(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{s} a_i L(y_i, \mathbf{b})$$

be the expected penalty $E(L(Y, \mathbf{q}(X)))$ if $\mathbf{p} = \mathbf{a}$ and if $\mathbf{q}(x)$ is $\mathbf{b}$ for all x in T. Obviously the expected penalty $D^*(\mathbf{a}, \mathbf{b})$ is nonnegative. If b_i is positive whenever a_i is positive, then $D*(\mathbf{a}, \mathbf{b})$ is finite. Unless $\mathbf{a} = \mathbf{b} = \boldsymbol{\delta}_i$ for some possible value y_i of the dependent variable Y, $D^*(\mathbf{a}, \mathbf{b})$ is positive. For a given $\mathbf{a}$, it is assumed that the expected penalty $D(\mathbf{a}, \mathbf{b})$ is smallest if $\mathbf{b} = \mathbf{a}$, so that

$$D(\mathbf{a}) = D^*(\mathbf{a}, \mathbf{a}) \leq D^*(\mathbf{a}, \mathbf{b}),$$

and $D(\mathbf{a})$ is nonnegative and finite. These requirements hold if the penalty function L is L_C, L_H, or L_M (Haberman, 1982a; Haberman, 1982b). Rather remarkably, these requirements provide a justification for use of logarithmic probability. If the number s of possible values of Y is at least 3, then a penalty function L that satisfies all regularity conditions and satisfies the condition that $L(y_i, \mathbf{a})$ is determined by a_i for each i from 1 to s must be equal to dL_H for some positive real d (Savage, 1971).

The function D is used to define basic measures of dispersion and association (Haberman, 1982a; Haberman, 1982b). The unconditional dispersion measure J_Y is defined to be the expected penalty $D(\mathbf{p}_Y) = E(L(Y, \mathbf{p}_{YC}))$ from probability prediction of Y by the constant predictor $\mathbf{p}_{YC}$. This measure is nonnegative and finite, and J_Y is 0 if, and only if, $p_Y(y) = 1$ for some possible value y of Y, so that $Y = y$ and $\mathbf{Z} = \mathbf{p}_Y$ with probability 1. In this latter case, Y is said to be essentially constant.

The dispersion measures associated with squared error penalty and logarithmic penalty are commonly used. In the case of the squared error function L_C, $D(\mathbf{p})$ is the Gini concentration

$$C_Y = 1 - \sum_{y \in S} p_Y(y)]^2$$

(Gini, 1912). It should be noted that C_Y is simply the probability that the sample variables Y_1 and Y_2 satisfy $Y_1 \neq Y_2$. Thus in the case of essay scoring, C_Y is the probability that two randomly chosen essays have the same human holistic score. In the case of log penalty, $D(\mathbf{p}_Y)$ is the Shannon entropy

$$H_Y = -\sum_{y \in S} p_Y(y) \log p_Y(y)$$

(Shannon, 1948). For misclassification rate, $D(\mathbf{p}_Y)$ is the minimum classification rate error

$$M_Y = 1 - \max_{y \in S} p_Y(y)$$

obtained if Y is classified as having a constant value y without regard to X. Because

$$\sum_{y \in S} [p_Y(y)]^2 \leq \left[\max_{y \in S} p_Y(y)\right] \sum_{y \in S} p_Y(y) = \max_{y \in S} p_Y(y),$$

with equality only if some $p_Y(y)$ is 1, the minimum classification rate error M_Y never exceeds the concentration C_Y, $C_Y \leq 1 - s^{-1}$, and $M_Y < C_Y$ if Y is not essentially constant (Gilula & Haberman, 1995a; Goodman & Kruskal, 1954). Because the logarithm of a positive real number d never exceeds $d - 1$, and $\log(d)$ only equals $d - 1$ if $d = 1$,

$$H_Y \geq \sum_{y \in S} p_Y(y)[1 - p_Y(y)] = C_Y,$$

with equality only if Y is essentially constant. In addition, $H_Y \leq \log(s)$ (Gilula & Haberman, 1995a). In the case of the probabilities for essay scores in Table 1, the best strategy for classification is to classify all essays by the score 4. The misclassification rate M_Y is then $1 - 0.3100 = 0.6900$. As expected, the concentration $C_Y = 0.7740$ exceeds $M_Y = 0.6900$ and C_Y is less than $1 - 6^{-1} = 0.8333$. In addition, the entropy $H_Y = 1.6043$ exceeds the concentration C_Y and is less than $\log(6) = 1.7918$.

The conditional dispersion measure $J_{Y \cdot X}(x)$ of Y given $X = x$ is the dispersion $D(\mathbf{p}_{Y \cdot X}(x))$. The conditional dispersion measure $J_{Y \cdot X}$ is the expected penalty

$$E(J_{Y \cdot X}(X)) = E(L(Y, \mathbf{p}_{Y \cdot X}(X)))$$

from use of the conditional probability prediction $\mathbf{p}_{Y \cdot X}(X)$ for Y. This measure is the smallest possible expected penalty $E(L(Y, \mathbf{q}(X)))$ for prediction of Y by a probability predictor $\mathbf{q}(X)$. Because $\mathbf{p}_{YC}$ is a probability predictor, $0 \leq J_{Y \cdot X} \leq J_Y$. The conditional dispersion $J_{Y \cdot X}$ is 0 if, and only if, for some function c from T to S, $Y = c(X)$ with probability 1. Thus Y may be said to be essentially determined by X. At the other extreme, the conditional and unconditional dispersions J_Y and $J_{Y \cdot X}$ are the same if X and Y are independent, so that $\mathbf{p}_{Y \cdot X}(X) = \mathbf{p}_Y$ with probability 1.

If the squared error penalty function is used, then the conditional dispersion $J_{Y\cdot X}(x)$ given $X = x$ is the conditional concentration

$$C_{Y\cdot X}(x) = 1 - \sum_{y\in S}[p_{Y\cdot X}(y|x)]^2$$

of Y given $X = x$, and the conditional concentration of Y given X is

$$C_{Y\cdot X} = E(C_{Y\cdot X}(X)) = 1 - \sum_{y\in S} E([p(y|X)]^2).$$

Thus $C_{Y\cdot X}$ is the conditional probability that $Y_1 \neq Y_2$ given that $X_1 = X_2$. If the logarithmic penalty is employed, then $J_{Y\cdot X}(x)$ is the conditional entropy

$$C_{Y\cdot X}(x) = -\sum_{y\in S} p_{Y\cdot X}(y|x)\log p_{Y\cdot X}(y|x)$$

of Y given $X = x$, and

$$H_{Y\cdot X} = E(H_{Y\cdot X}(X)) = -\sum_{y\in S} E(p_{Y\cdot X}(y|X)\log p_{Y\cdot X}(y|X))$$

is the conditional entropy of Y given X. For misclassification rate, $J_{Y\cdot X}(x)$ is the minimum classification error rate

$$M_{Y\cdot X}(x) = 1 - \max_{y\in S} p_{Y\cdot X}(y|x)$$

for Y given $X = x$ and $J_{Y\cdot X}$ is the minimum classification error rate

$$M_{Y\cdot X} = E(M_{Y\cdot X}(X)) = 1 - E(\max_{y\in S} p_{Y\cdot X}(y|X))$$

from classification of Y by use of a function of X. As in the unconditional case, $M_{Y\cdot X}(x) \leq C_{Y\cdot X}(x) \leq H_{Y\cdot X}(x)$, and $M_{Y\cdot X} \leq C_{Y\cdot X} \leq H_{Y\cdot X}$. For example, in the case of essay scoring, $C_{Y\cdot X} = 0.6266$ exceeds $M_{Y\cdot X} = 0.524$ and is less than $H_{Y\cdot X} = 1.1851$.

If the expected penalty function D is strictly concave, then the conditional dispersion measure $J_{Y\cdot X}$ is equal to the unconditional dispersion measure J_Y only if X and Y are independent. In the case of squared error penalty and logarithmic penalty, the function D is strictly concave. Thus $C_{Y\cdot X} = C_Y$ or $H_{Y\cdot X} = H_Y$ implies independence of the dependent variable Y and the independent variable X. For misclassification rate, D is not strictly concave, and the conditional classification error rate $M_{Y\cdot X}$ may equal the unconditional classification error rate M_Y without independence of X and Y. For an extreme case, let Y and X have possible values 1 and 2, so that s is 2 and S and T are the integers 1 and 2. Let the probability $p_X(x)$ that $X = x$ be 0.5 for x equal 1 or 2, let $p_{Y\cdot X}(1|1) = 1$, $p_{Y\cdot X}(2|1) = 0$, and $p_{Y\cdot X}(1|2) = p_{Y\cdot X}(2|2) = 0.5$, so that $p_Y(1) = 0.75$ and $p_Y(2) = 0.25$. Obviously X and Y are dependent; however, $M_{Y\cdot X} = M_Y = 0.25$. In contrast, $C_{Y\cdot X} = 0.25$ is less than $C_Y = 0.375$, and $H_{Y\cdot X} = 0.3466$ is less than 0.5623.

The dispersion measures described in this section are commonly used to construct analogues of the coefficient of determination of regression analysis to describe the strength of the relationship between Y and X. This practice is particularly well known if T is finite, so that X is polytomous (Goodman & Kruskal, 1954). If $J_Y > 0$, then

$$\rho_{Y\cdot X} = 1 - \frac{J_{Y\cdot X}}{J_Y}$$

measures the proportional reduction in loss from use of X as a predictor of Y. Given the inequality constraints on the conditional dispersion $J_{Y\cdot X}$ and the unconditional dispersion J_Y, it follows that $0 \leq \rho_{Y\cdot X} \leq 1$, with $\rho_{Y\cdot X} = 1$ if, and only if, Y is essentially determined by X. If X and Y are independent, then $\rho_{Y\cdot X}$ is 0. If D is strictly concave, then $\rho_{Y\cdot X}$ is only 0 if X and Y are independent.

The Goodman and Kruskal λ coefficient $\lambda_{Y\cdot X}$ is $1 - M_{Y\cdot X}/M_Y$, and the Goodman and Kruskal τ coefficient $\tau_{Y\cdot X}$ is $1 - C_{Y\cdot X}/C_Y$ (Goodman & Kruskal, 1954). The Theil uncertainty coefficient $U_{Y\cdot X}$ is $1 - H_{Y\cdot X}/H_Y$ (Theil, 1971). As evident from the relationships between independence and equality of conditional and unconditional dispersion measures, $\tau_{Y\cdot X}$ and $U_{Y\cdot X}$ are only 0 if X and Y are independent. In contrast, $\lambda_{Y\cdot X}$ may be 0 for dependent X and Y. In the example used previously to illustrate the possibility of dependence with equal conditional and unconditional classification error rates, $\lambda_{Y\cdot X} = 0$, $\tau_{Y\cdot X} = 0.3515$, and $U_{Y\cdot X} = 0.3837$. Note that $\tau_{Y\cdot X}$ and $U_{Y\cdot X}$ are relatively similar, and they both suggest that an appreciable reduction in error is achieved by use of X in prediction of Y. On the other hand, $\lambda_{Y\cdot X}$ suggests that X has no value as a predictor of Y. In the example of essay scoring, differences in results are less dramatic, for squared error penalty leads to $\tau_{Y\cdot X} = 1 - 0.6266/0.7740 = 0.1904$, logarithmic probability penalty leads to $U_{Y\cdot X} = 1 - 1.1851/1.6043 = 0.2613$, and misclassification error leads to $\lambda_{Y\cdot X} = 1 - 0.5240/0.6900 = 0.2406$. In all cases, the indication is that the machine-generated essay score permits a modest improvement in probability prediction relative to the prediction achievable without the machine-generated score.

2 Sampling and Probability Prediction

A probability prediction of Y given X may be developed by use of sample observations. For some simple examples, for y in S and x in T, let $f_Y(y)$ be the number of integers h with $Y_h = y$, let $f_{YX}(y, x)$ be the number of integers h with $Y_h = y$ and $X_h = x$, and let $f_X(x)$ be the number of integers h with $X_h = x$. One might consider the probability predictor $\hat{\mathbf{p}}_{YC}$ with coordinate i equal to $\hat{p}_Y(y_i) = n^{-1}f_Y(y_i)$, the fraction of the integers h with $Y_h = y_i$. For a slightly more complex case, consider $\hat{\mathbf{p}}_{Y\cdot X}$, where $\hat{p}_{Y\cdot X}(x)$ is $\hat{p}_{YC}$ if $f_X(x) = 0$ and $\hat{p}_{Y\cdot X}(x)$ has coordinate i equal to the relative frequency

$$\hat{p}_{Y\cdot X}(y_i|x) = f_{YX}(y_i, x)/f_X(x)$$

if $f_X(x) > 0$. The functions $\hat{p}_{Y\cdot X}$ and $\hat{p}_{YC}$ have the disadvantage that they can provide probability predictors that have coordinates equal to 0. As a consequence, alternatives of interest are $\hat{\mathbf{p}}_{YC\boldsymbol{\alpha}}$ and $\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}}$, where $\boldsymbol{\alpha}$ is an s-dimensional vectors with positive coordinates α_i with sum $\alpha_+ = \sum_{i=1}^{s} \alpha_i$, $\hat{\mathbf{p}}_{YC\boldsymbol{\alpha}}(x)$ has coordinate i equal to

$$\hat{p}_{YC\boldsymbol{\alpha}}(y_i|x) = \frac{f_Y(y_i) + \alpha_i}{n + \alpha_+}$$

for any x in T and $\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}}(x)$ has coordinate i equal to

$$\hat{p}_{Y\cdot X\boldsymbol{\alpha}}(y_i|x) = \begin{cases} \frac{f_Y(y_i)+\alpha_i}{n+\alpha_+}, & f_X(x) = 0, \\ \frac{f(y_i,x)+\alpha_i}{f_X(x)+\alpha_+}, & f_X(x) > 0. \end{cases}$$

It is possible to consider $\alpha_i = 0.5$. This choice is consistent with common estimation procedures for logarithms of ratios of probabilities (Anscombe, 1956).

In general, a function $\hat{\mathbf{q}}$ of the observations X_h and Y_h, $1 \le h \le n$, is considered. For any given value of the X_h and Y_h, $\hat{\mathbf{q}}$ is a probability predictor with value $\hat{\mathbf{q}}(x)$ at x in T, and $\hat{\mathbf{q}}(x)$ has coordinate i equal to $\hat{q}(y_i|x)$. The function $\hat{\mathbf{q}}(X)$ equal to $\hat{\mathbf{q}}(x)$ if $X = x$ is assumed to be an s-dimensional random vector. The function $\hat{\mathbf{q}}$ may be termed a sample probability predictor. It is easily seen that $\hat{\mathbf{p}}_{YC}$, $\hat{\mathbf{p}}_{YC\boldsymbol{\alpha}}$, $\hat{\mathbf{p}}_{Y\cdot X}$, and $\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}}$ are all sample probability predictors. Many more complex sample probability predictors may be constructed by use of log-linear models (Gilula & Haberman, 1995b).

To assess the value of the sample probability predictor $\hat{\mathbf{q}}$, the expected penalty is evaluated. The penalty under study is a random variable $L(Y, \hat{\mathbf{q}}(X))$ that depends on the observations X and Y under study and on the sampled variables X_h and Y_h. To find the expected penalty efficiently requires several arguments involving conditional expectations. Given $X = x$ and given the observed X_h and Y_h, the conditional expected value of the penalty $L(Y, \hat{\mathbf{q}}(X))$ from probability prediction of Y from $\hat{q}$ is the random variable

$$D^*(\mathbf{p}_{Y \cdot X}(x), \hat{\mathbf{q}}(x)) \geq J_{Y \cdot X}(x).$$

Let

$$F(\hat{\mathbf{q}}|x) = D^*(\mathbf{p}_{Y \cdot X}(x), \hat{\mathbf{q}}(x)) - J_{Y \cdot X}(x) \geq 0$$

denote the conditional excess expected penalty given $X = x$ and the X_h and Y_h. Then

$$F(\hat{\mathbf{q}}|x) = \sum_{y \in S} p_{Y \cdot X}(y|x)[L(y, \hat{\mathbf{q}}(x)) - L(y, \mathbf{p}_{Y \cdot X}(x))]$$

(Haberman, 1982a). Given $X = x$, the conditional expected excess penalty is

$$\Delta(\hat{\mathbf{q}}|x) = E(F(\hat{\mathbf{q}}|x)),$$

and the expected excess penalty is

$$B(\hat{\mathbf{q}}) = E(\Delta(\hat{\mathbf{q}}|X)).$$

The expected penalty is then

$$I(\hat{\mathbf{q}}) = B(\hat{\mathbf{q}}) + J_{Y \cdot X}.$$

In the case of squared error, $F(\hat{\mathbf{q}}|x)$ is

$$F_C(\hat{\mathbf{q}}|x) = \sum_{y \in S} [\hat{q}(y|x) - p_{Y \cdot X}(y|x)]^2,$$

and $\Delta(\hat{\mathbf{q}}|x)$ is

$$\Delta_C(\hat{\mathbf{q}}|x) = \sum_{y \in S} \{\sigma^2(\hat{q}(y|x)) + [E(\hat{q}(y|x) - p_{Y \cdot X}(y|x)]^2\}$$

(Haberman, 1982a). The expected excess penalty is then

$$B_C(\hat{\mathbf{q}}) = E(\Delta_C(\hat{\mathbf{q}}|X)),$$

and the expected penalty is

$$I_C(\hat{\mathbf{q}}) = B_C(\hat{\mathbf{q}}) + C_{Y\cdot X}.$$

In the case of the logarithmic probability penalty, $F(\hat{\mathbf{q}}|x)$ is

$$F_H(\hat{\mathbf{q}}|x) = \sum_{y\in S} p_{Y\cdot X}(y|x) \log\left[\frac{p_{Y\cdot X}(y|x)}{\hat{q}(y|x)}\right]$$

(Haberman, 1982a), so that $F_H(\hat{\mathbf{q}}|x)$ is ∞ if $\hat{q}(y|x)$ is 0 for some y such that $p_{Y\cdot X}(y|x) > 0$. The expected excess penalty $B(\hat{\mathbf{q}})$ is then

$$B_H(\hat{\mathbf{q}}) = E(\Delta_H(\hat{\mathbf{q}}|X)),$$

and the expected penalty is

$$I_H(\hat{\mathbf{q}}) = B_H(\hat{\mathbf{q}}) + H_{Y\cdot X}.$$

Both $B_H(\hat{\mathbf{q}})$ and $I_H(\hat{\mathbf{q}})$ may be infinite.

For misclassification rate, $F(\hat{\mathbf{q}}|x)$ becomes $F_M(\hat{\mathbf{q}}|x)$, where $F_M(\hat{\mathbf{q}}|x)$ is the difference between $|\mathbf{p}_{Y\cdot X}(x)|_\infty$ and the average of the $p_{Y\cdot X}(y|x)$ for y in S such that $\hat{q}(y|x) = |\mathbf{q}(x)|_\infty$. The conditional expected excess penalty given $X = x$ becomes

$$\Delta_M(\hat{\mathbf{q}}|x) = E(F_M(\hat{\mathbf{q}}|x)),$$

the expected excess penalty becomes

$$B_M(\hat{\mathbf{q}}) = E(\Delta_M(\hat{\mathbf{q}}|X)),$$

and the expected penalty is

$$I_M(\hat{\mathbf{q}}) = B_M(\hat{\mathbf{q}}) + M_{Y\cdot X}.$$

3 Penalty Criteria in Large Samples

The large-sample behavior of the expected excess penalty is very different for misclassification rate penalty than for squared error penalty or for logarithmic probability penalty. As evident in Section 2, because the logarithmic probability penalty can be infinite, some differences exist for results for logarithmic penalty functions and squared error penalty functions. In the cases of squared error penalty and logarithmic probability penalty, large-sample properties have been explored previously in the case of log-linear models (Haberman, 1982a; Gilula & Haberman, 1995b); however, a more precise and simpler discussion of differences is provided by examination of the case of $\hat{\mathbf{p}}_{Y \cdot X}$ and $\hat{\mathbf{p}}_{Y \cdot X\boldsymbol{\alpha}}$ for T, the range of X, a finite set with v elements, and for the probability $p_X(x)$ that $X = x$ positive for each x in T. To facilitate comparison of results for different choices of X, it is helpful to use the minimum expected value

$$g = n \min_{x \in T} p_X(x)$$

of the counts $f_X(x)$. In the case of squared error penalty, $B_C(\hat{\mathbf{p}}_{Y \cdot X})$ and $B_C(\hat{\mathbf{p}}_{Y \cdot X\boldsymbol{\alpha}})$ are both of order g^{-1}. In the case of logarithmic probability penalty, $B_H(\hat{\mathbf{p}}_{Y \cdot X})$ is infinite, and $B_H(\hat{\mathbf{p}}_{Y \cdot X\boldsymbol{\alpha}})$ is of order g^{-1}. In the case of misclassification rate penalty, $B_M(\hat{\mathbf{p}}_{Y \cdot X})$ and $B_M(\hat{\mathbf{p}}_{Y \cdot X\boldsymbol{\alpha}})$ are equal and are typically of order $\exp(-\beta g)$ for some real $\beta > 0$.

To verify these claims, a few basic results concerning the distribution of the frequency counts $f_X(x)$ should be noted. The probability that $f_X(x) = 0$ is $r(x) = [1 - p_X(x)]^n$. Note that $\log(d) < d - 1$ if d is a positive real number other than 1. It follows that $r(x)$ is equal to

$$\exp\{n \log[1 - p_X(x)]\} < \exp[-np_X(x)] \le \exp(-g),$$

so that convergence of this probability to 0 is exponentially fast for each x in T. It is also helpful to note that

$$m(x) = E(1/f_X(x)|f_X(x) > 0)$$

is bounded below by

$$u_1(x) = \frac{1 - k(x)}{(n+1)p_X(x)}$$

for

$$k(x) = \frac{np_X(x)r(x)}{1 - r(x)}$$

and bounded above by

$$u_1(x) + 3u_2(x)$$

for

$$u_2(x) = \frac{u_1(x) - k(x)/2}{(n+2)p_X(x)}$$

(Stephan, 1945). Thus $np_X(x)m(x)$ differs from 1 by a term of order g^{-1}.

Given these preliminaries, each penalty function requires separate attention.

3.1 Squared Error

In the case of squared error penalty, both $B_C(\hat{\mathbf{p}}_{Y\cdot X})$ and $B_C(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$ are well approximated by

$$n^{-1}\sum_{x\in T} C_{Y\cdot X}(x) \le g^{-1}C_{Y\cdot X}.$$

Thus the estimated excess penalty is of order g^{-1}. Although the details of verification of this claim are a bit complicated, the basic principles are readily summarized.

The simplest case to consider is $\hat{\mathbf{p}}_{Y\cdot X}$. In this instance, given that $f_X(x) > 0$, the conditional expectation of $\hat{p}_{Y\cdot X}(y|x)$ is $p_{Y\cdot X}(y|x)$. Given that $f_X(x) = 0$, the conditional expectation of $\hat{p}_{Y\cdot X}(y|x)$ is

$$P(Y = y|X \ne x) = \frac{p_Y(y) - p_Y(y|X = x)p_X(x)}{1 - p_X(x)}.$$

It follows that

$$E(\hat{p}_{Y\cdot X}(y|x)) = [1 - r(x)]p_{Y\cdot X}(y|x) + r(x)P(Y = y|X \ne x)$$

differs from $p_{Y\cdot X}(y|x)$ by a term of order $\exp(-g)$. In like fashion, the conditional variance of $\hat{p}_{Y\cdot X}(y|x)$ given $f_X(x) > 0$ is

$$p_{Y\cdot X}(y|x)[1 - p_{Y\cdot X}(y|x)]/f_X(x),$$

and the conditional variance of $\hat{p}_{Y\cdot X}(y|x)$ given $f_X(0) = 0$ is

$$P(Y = y|X \ne x)[1 - P(Y = y|X \ne x)]/n.$$

It follows that

$$np_X(x)\sigma^2(\hat{p}_{Y\cdot X}(y|x)) - p_{Y\cdot X}(y|x)[1 - p_{Y\cdot X}(y|x)]$$

is of order g^{-1}. Because

$$\sum_{y\in S} p_{Y\cdot X}(y|x)[1 - p_{Y\cdot X}(y|x)] = C_{Y\cdot X}(x),$$

it follows that

$$B_C(\hat{\mathbf{p}}_{Y\cdot X}) - n^{-1}\sum_{x\in T} C_{Y\cdot X}(x)$$

is of order g^{-2}. Because $C_{Y\cdot X}(x)$ cannot exceed $1 - s^{-1}$ and $g \leq n/v$, $B_C(\hat{\mathbf{p}}_{Y\cdot X})$ is of order g^{-1}. Slight changes in arguments can be used to show that

$$B_C(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}}) - n^{-1}\sum_{x\in T} C_{Y\cdot X}(x)$$

is also of order g^{-2}. Note that $C_{Y\cdot X}(x)$ is positive if distinct y and y' exist such that the conditional probabilities $p_{Y\cdot X}(y|x)$ and $p_{Y\cdot X}(y'|x)$ are both positive.

For the probabilities in Table 1, if $n = 1,000$, then g is 23, and the expected excess squared error penalty is about 0.00378. For $n = 10,000$, g is 230, and the expected excess is about 0.000378.

3.2 Logarithmic Probability

In the case of logarithmic probability, $\hat{p}_{Y\cdot X}(y|x)$ is 0 with positive probability for a case with $p_{Y\cdot X}(y|x) = 0$ unless some function c on T exists for which $Y = c(X)$. In this trivial case, the expected excess penalty $B_H(\hat{\mathbf{p}}_{Y\cdot X})$ is 0. Otherwise, $B_H(\hat{\mathbf{p}}_{Y\cdot X})$ is infinite, although it should be noted that, for any positive real d, the expected minimum of d and $F_H(\hat{\mathbf{p}}_{Y\cdot X})$ approaches $v(s-1)/n$ whenever each conditional probability $p_{Y\cdot X}(y|x)$ is positive (Gilula & Haberman, 1995b).

More interesting results are available if $\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}}$ is considered. It is simplest to confine attention to the case in which $p_{Y\cdot X}(y|x)$ is always positive. To facilitate comparison of results for different choices of X, the condition may be used that a positive real d exists such that $p_{Y\cdot X}(y|x) \geq d$ for all y and x. The basic result obtained is quite simple, for the expected excess penalty $B_H(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$ differs from $v(s-1)/n$ by a term of order g^{-2}. This result is predictable given current literature (Gilula & Haberman, 1995b). It should be noted that, for sufficiently large sample sizes, $B_H(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$ is at least s times larger than $B_C(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$.

The argument required to prove results for the logarithmic penalty is rather similar to that used for squared error; however, a bit more effort is required because $\log \hat{p}_{Y\cdot X\boldsymbol{\alpha}}(y_i|x)$ is a nonlinear function of $\hat{p}_{Y\cdot X}(y_i|x)$ that may be as small as $\log(\alpha_i/(n+\alpha_+))$. Details are not considered here; however, it is worth noting that two basic principles are involved. For $\hat{p}_{Y\cdot X}(y|x) > 0$, the logarithms $\log[\hat{p}_{Y\cdot X\boldsymbol{\alpha}}(y|x)/p_{Y\cdot X}(y|x)]$ are approximated by

$$\frac{\hat{p}_{Y\cdot X\boldsymbol{\alpha}}(y|x) - p_{Y\cdot X}(y|x)}{p_{Y\cdot X}(y|x)}$$

by use of the elementary expansion

$$\log(b/a) = \frac{b-a}{c}$$

for b and a real and positive and c a real number between b and a. To place limits on the probability that $|\hat{p}_{Y\cdot X\boldsymbol{\alpha}}(y|x) - p_{Y\cdot X}(y|x)|$ exceeds some small quantity δ, large deviation theory is used (Bahadur & Ranga Rao, 1960). Let a and b be positive real numbers less than 1, and let

$$d = a\log\left(\frac{a}{b}\right) + (1-a)\log\left(\frac{1-a}{1-b}\right).$$

Let f be a binomial random variable with sample size $k > 0$ and probability b. If $a > b$, then the probability that $f/k > a$ does not exceed $\exp(-kd)$. If $a < b$, then the probability that $f/k < a$ does not exceed $\exp(-kd)$.

For $n = 1,000$ and for probabilities defined as in Table 1, the expected excess logarithmic probability penalty is about 0.03. If n is 10,000, then the expected excess penalty is reduced to 0.003. As expected, these values are somewhat larger than the corresponding ones for squared error penalty.

3.3 Misclassification Penalty

For misclassification penalty, $B_M(\hat{\mathbf{p}}_{Y\cdot X}) = B_M(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$ because $\hat{p}_{Y\cdot X}(y|x) > \hat{p}_{Y\cdot X}(y'|x)$ if, and only if, $\hat{p}_{Y\cdot X\boldsymbol{\alpha}}(y|x) > \hat{p}_{Y\cdot X\boldsymbol{\alpha}}(y'|x)$ for y and y' in S and x in T. In addition, $B_M(\hat{\mathbf{p}}_{Y\cdot X})$ is trivially 0 if Y satisfies the equiprobability condition that $p_{Y\cdot X}(y|x) = s^{-1}$ for all y and x. In other cases, large-deviation theory may be applied to obtain an upper bound on $B_M(\hat{\mathbf{p}}_{Y\cdot X})$. One finds that $B_M(\hat{\mathbf{p}}_{Y\cdot X})$ is of order $\exp(-\beta g)$ for some $\beta > 0$. As will be evident from examination of Table 1, this exponential rate of convergence to 0 does not necessarily imply that $B_M(\hat{\mathbf{p}}_{Y\cdot X})$ is negligible even for relatively large samples.

To examine the required large-deviation theory, let $p(y, x)$ be the probability that $X = x$ and $Y = y$, so that $p(y, x) = p_X(x)p_{Y\cdot X}(y|x)$. Let $p_{Y\cdot X}(y|x) > p_{Y\cdot X}(y'|x)$, and let

$$\gamma(y, y'|x) = \{[p_{Y\cdot X}(y|x)]^{1/2} - [p_{Y\cdot X}(y'|x)]^{1/2}\}^2.$$

For real nonnegative a and b, $a > b$,

$$(a^{1/2} - b^{1/2})(a^{1/2} + b^{1/2}) = a - b.$$

It follows that

$$\gamma(y, y'|x) > \frac{1}{4}[p_{Y\cdot X}(y|x) - p_{Y\cdot X}(y'|x)]^2.$$

If

$$m(y, y'|x) = 1 - p(y, x) - p(y', x) + 2[p(y, x)p(y', x)]^{1/2},$$

$$v(y, y'|x) = p(y, x) + p(y', x) - [p(y, x) - p(y', x)]^2,$$

and

$$\zeta(y, y'|x) = \begin{cases} 2[1 - p(y', x)/p(y, x)]^{-1}, & s = 2, \\ \{1 - [p(y', x)/p(y, x)]^{1/2}\}^{-1}, & s > 1, \end{cases}$$

then

$$m(y, y'|x) = 1 - p_X(x)\gamma(y, y'|x) < 1,$$

the probability $\xi(y, y'|x)$ that $f_{YX}(y', x) \geq f_{YX}(y, x)$ does not exceed $[m(y, y'|x)]^n$, and $\xi(y, y'|x)$ is well approximated by

$$\xi_0(y, y'|x) = \frac{[m(y, y'|x)]^n}{[2\pi n v(y, y'|x)]^{1/2}\zeta(y, y'|x)}$$

in the sense that

$$\frac{\xi(y, y'|x) - \xi_0(y, y'|x)}{\xi(y, y'|x)}$$

is of order g^{-1} (Bahadur & Ranga Rao, 1960). Use of the inequality $d - 1 < \log(d)$ for positive real $d \neq 1$ implies that

$$[m(y, y'|x)]^n < \exp[-np_X(x)\gamma(y, y'|x)].$$

In addition, for any x in T,

$$F_M(\hat{\mathbf{p}}_{Y\cdot X}|x) \leq \sum_{y\in S}[|\mathbf{p}_{Y\cdot X}|_\infty - p_{Y\cdot X}(y|x)]u(y|x),$$

where $u(y|x)$ is 1 if $\hat{p}_{Y \cdot X}(y|x) \geq \hat{p}_{Y \cdot X}(y_i|x)$ for each i in $A(\mathbf{p}_{Y \cdot X}(x))$ and $u(y|x)$ is 0 otherwise. Let $p_{Y \cdot X}(c(x)|x) = |\mathbf{p}_{Y \cdot X}(x)|_\infty$. It follows that

$$B_M(\hat{\mathbf{p}}_{Y \cdot X}) \leq \sum_{x \in T} p_X(x) \sum_{y \in S} [|\mathbf{p}_{Y \cdot X}(x)|_\infty - p_{Y \cdot X}(y|x)][r(x) + \xi(c(x), y|x)]]\}.$$

Thus a real $\tau > 0$ and $\beta > 0$ exists such that $B_M(\hat{\mathbf{p}}_{Y \cdot X})$ is less than $\tau \exp(-\beta g)$ for g sufficiently large. The size of β is at least one quarter the square of the smallest difference $p_{Y \cdot X}(c(x)|x) - p_{Y \cdot X}(y_i|x)$ for i not in $A(\mathbf{p}_{Y \cdot X}(x)$ and x in T, and τ can be selected not to exceed $2(s-1)$. The bound is quite generous, as is evident from the more accurate approximation to $\xi(y, y'|x)$.

For the probabilities in Table 1, if $n = 1,000$, then use of upper bounds shows that the expected excess penalty does not exceed 0.0129, but the refined approximation yields 0.00434. For $n = 10,000$, the upper bound is 0.0000959, and the refined approximation is 0.0000140. Note that the value for $n = 1,000$ is quite comparable to that for squared error, but the expected excess for $n = 10,000$ is very small. In general, the exponential rate of the convergence to 0 of $B_M(\hat{\mathbf{p}}_{Y \cdot X})$ implies, at least for a large enough sample size, that a much smaller expected excess penalty is achieved for misclassification penalty than is achieved in the case of squared error or logarithmic probability penalty.

3.4 Comparison of Expected Penalties

For an additional simple illustration of the implications of the large-sample properties of the sample probability predictors under study, consider $s = 2$, and define a uniformly distributed random variable W with range $(1/4, 3/4)$ such that the conditional probability that $Y = 1$ given that $W = w$ is w. Let v be a positive integer, and let X be the largest integer not greater than $2v(W - 1/4)$, so that $p^X(x) = v^{-1}$ for integers x from 0 to $v - 1$, and

$$p^{Y \cdot X}(1|x) = \frac{1}{4} + \frac{2x+1}{4v}.$$

Straightforward calculations show that, in the case of squared error, the condition concentration of Y given W is

$$C_{Y \cdot W} = 2 \int_{1/4}^{3/4} 2w(1-w)dw = \frac{11}{24},$$

the conditional concentration of Y given X is

$$C_{Y\cdot X} = \frac{11}{24} + \frac{1}{24v^2},$$

and the expected excess penalties $B_C(\hat{\mathbf{p}}_{Y\cdot X})$ and $B_C(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$ are well approximated by $vC_{Y\cdot X}/n$. As n approaches ∞ and v/n approaches 0, the expected penalties $I_C(\hat{\mathbf{p}}_{Y\cdot X})$ and $I_C(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$ are well approximated by

$$\left(\frac{11}{24} + \frac{1}{24v^2}\right)(1 + v/n).$$

At this point, there is a tradeoff to consider. More categories v in the definition of X leads to a smaller conditional dispersion $C_{Y\cdot X}$ but a larger approximation for the estimated excess penalties $B_C(\hat{\mathbf{p}}_{Y\cdot X})$ and $B_C(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$. For n large, the optimal situation has v approximately equal to $(3n/11)^{1/3}$, so that the expected penalties $I_C(\hat{\mathbf{p}}_{Y\cdot X})$ and $I_C(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$ are well approximated by

$$C_{Y\cdot W} + \frac{1}{6}\left(\frac{11}{3n}\right)^{2/3}.$$

Note that for sufficiently large n, the expected penalties from use of probability predictors from sample data becomes increasingly close to the expected penalty achieved through prediction of Y by W under the condition that $\mathbf{p}_{Y\cdot W}$ is known. The difference in expected penalties is of order $n^{-2/3}$. As an illustration of results, consider $n = 1,000$. In this case, v, which must be an integer, may be taken as 6, and $I_C(\hat{\mathbf{p}}_{Y\cdot X})$ and $I_C(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$ exceed $C_{Y\cdot W}$ by about 0.0039.

In like manner, for the logarithmic penalty, the conditional entropy of Y given W is

$$\begin{aligned} H_{Y\cdot W} &= -2\int_{1/4}^{3/4} [w\log(w) + (1-w)\log(1-w)]dw \\ &= \frac{1}{2} + 2\log(2) - \frac{9}{8}\log(3) \\ &= 0.650, \end{aligned}$$

and the conditional entropy $H_{Y\cdot X}$ of Y given X is well approximated by

$$H_{Y\cdot W} + \frac{1}{48v^2}\int_{1/4}^{3/4}[w(1-w)]^{-1}dw = H_{Y\cdot W} + \frac{1}{24v^2}\log(3).$$

It follows that the expected penalty $I_H(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}}))$ is well approximated by

$$H_{Y\cdot W} + \frac{1}{24v^2}\log(3) + \frac{v}{2n}.$$

To reduce expected penalty from use of $\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}}$, the optimal choice of v for large n has v approximately equal to $(4^{-1}n\log 3)^{1/3}$, so that $I_H(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$ is well approximated by

$$H_{Y\cdot W} + \frac{2}{3}\left(\frac{\log 3}{4n^2}\right)^{1/3}.$$

For sufficiently large n, the expected penalty from use of sample data becomes increasingly close to the expected penalty achieved through prediction of Y by W under the condition that $\mathbf{p}_{Y\cdot W}$ is known. As in the case of concentration, the difference in expected penalties is of order $n^{-2/3}$. For $n = 1,000$, the optimal choice of v is 6, and the expected penalty for the sample predictor exceeds the expected penalty for $\mathbf{p}_{Y\cdot X}$ by 0.0043. Thus results for squared error and logarithmic penalty are quite similar.

The situation is very different for misclassification penalty. Here the conditional misclassification rate $M_{Y\cdot W}$ of Y given W is easily seen to be $(1/2+1/4)/2 = 3/8$, and the conditional misclassification rate $M_{Y\cdot X}$ of Y given X is $3/8$ for v even and $3/8+1/(8v^2)$ for v odd. In terms of the expected penalty $I_M(\hat{\mathbf{p}}_{Y\cdot X}) = I_M(\hat{\mathbf{p}}_{Y\cdot X\boldsymbol{\alpha}})$, the optimal choice of v is 2. In this case, the expected penalty is very close to $M_{Y\cdot W} + (1/4)\xi(2,1|x)$, and $\xi(2,1|1)$ is bounded above by 0.9841^n. It follows that the expected excess misclassification penalty is less than 10^{-7} if $n = 1,000$, a figure drastically smaller than the corresponding values for squared error penalty or logarithmic probability penalty.

4 Conclusions

The large-sample properties associated with squared error penalty, logarithmic probability penalty, and misclassification penalty indicate that misclassification penalty exhibits very different behavior than do the other penalties. One might think that the asymptotic results imply a superiority of misclassification penalty on the grounds that, for a sufficiently large sample size, if each conditional probability $p_{Y\cdot X}(y|x)$ is positive, then the expected excess penalty is smaller for misclassification penalty than for the other choices of penalty functions. In reality, the apparent advantage of misclassification penalty reflects a very serious flaw in the criterion. The misclassification rate is very insensitive to variations in predicted probabilities unless two or more predicted probabilities are nearly the same. For example, in the example of prediction of a dichotomous response, the predictor X for

Table 2.

Joint Probabilities of Human and Machine Scores

Machine	Human score						
score	1	2	3	4	5	6	Total
1	0.0110	0.0120	0.0000	0.0000	0.0000	0.0000	0.0230
2	0.0100	0.0760	0.0760	0.0000	0.0000	0.0000	0.1620
3	0.0000	0.1000	0.1000	0.0290	0.0000	0.0000	0.2290
4	0.0000	0.0000	0.1740	0.1760	0.0000	0.0000	0.3500
5	0.0000	0.0000	0.0000	0.0800	0.0800	0.0170	0.1770
6	0.0000	0.0000	0.0000	0.0000	0.0270	0.0320	0.0590
Total	0.0210	0.1880	0.3500	0.2850	0.1070	0.0490	1.0000

$v = 2$ is as effective as the original variable W in terms of misclassification rate despite the substantial variability of the conditional probability $p_Y(y|w)$ as a function of w. On the other hand, the differences $C_{Y \cdot X} - C_{Y \cdot W} = 1/96$ and $H_{Y \cdot X} - H_{Y \cdot W} = 0.011$ are fairly substantial.

The example of essay scoring exhibits a similar issue. The same conditional misclassification rate is achieved if Table 1 is modified to yield Table 2. Table 1 and Table 2 are quite different. Other measures reflect the change. The conditional concentration $C_{Y \cdot X}$ is changed from 0.6266 to 0.5457, and the conditional entropy $H_{Y \cdot X}$ is changed from 1.1851 to 0.8873. The latter two measures reflect the decreased dispersion in Table 2 relative to Table 1.

In practice, attempts to use misclassification penalty rather than more sensitive penalty functions are likely to obscure actual improvements in prediction. For example, progress in the machine scoring of essays can be expected to be obscured as long as criteria based on misclassification rates are employed.

References

Anscombe, F. (1956). On estimating binomial response relations. *Biometrika*, *43*, 461–464.

Bahadur, R. R., & Ranga Rao, R. (1960). On deviations of the sample mean. *The Annals of Mathematical Statistics*, *31*, 1015–1027.

Feng, X., Dorans, N. J., Patsula, L. N., & Kaplan, B. (2003). *Improving the statistical aspects of e-rater: Exploring alternative feature reduction and combination rules* (ETS RR-03-15). Princeton, NJ: ETS.

Gilula, Z., & Haberman, S. J. (1995a). Dispersion of categorical variables and penalty functions: Derivation, estimation, and comparability. *Journal of the American Statistical Association*, *90*, 1447–1452.

Gilula, Z., & Haberman, S. J. (1995b). Prediction functions for categorical panel data. *The Annals of Statistics*, *23*, 1130–1142.

Gini, C. (1912). *Variabilità e mutabilità: Contributo allo studio delle distribuzioni e delle relazioni statische.* Bologna, Italy: Cuppini.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross-classifications. *Journal of the American Statistical Association*, *49*, 732–764.

Haberman, S. J. (1982a). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, *77*, 568–580.

Haberman, S. J. (1982b). Measures of association. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (vol. 1, pp. 130–137). New York: John Wiley.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*, 783–801.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Stephan, F. F. (1945). The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate. *The Annals of Mathematical Statistics*, *16*, 50–61.

Theil, H. (1971). On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, *76*, 103–154.