Assessing the performance of classifiers when classes arise from a continuum
(running title: “Assessing classifiers”) [⋆]

Bjørn-Helge Mevik [∗]

Matforsk, Osloveien 1, N-1430 Ås, Norway.

Abstract

The situation where classes arise by dividing the range of a continuous response variable into intervals is discussed. The focus is on assessing the performance of classifiers. Due to the underlying continuum, all misclassifications are not equally grave. The probability of misclassification (pmc) is not optimal in this situation. An alternative performance measure, the squared error rate (sqerr) is proposed. It is related to the mean squared error of regression, and penalises misclassifications according to their severity. Also, because of measurement errors in the response variable, there are misallocated class labels in data sets used for training and testing. Estimates of the pmc and the sqerr are developed for this situation. The estimates are tested and compared on a real data set and in a simulation.

Key words: classification, classes arising from a continuum, measurement error, misallocated class labels, probability of misclassification, squared error rate

1 Introduction

In production industries, the quality of products and raw materials is often characterised by one or more continuous variables, e.g. chemical measurements. Often these variables are expensive, time consuming or otherwise impractical to measure directly. Consequently, one wishes to predict them from other, cheaper or more easily attained measurements.

[⋆] This is a preprint of a paper published in *Computational Statistics and Data Analysis* 46(4) 689–705 © 2003 Elsevier Science B.V. All rights reserved.

[∗] Address: Matforsk, Osloveien 1, N-1430 Ås, Norway. Tel: +47 64970338. Fax: +47 64970333.

Email address: `bjorn-helge.mevik@matforsk.no` (Bjørn-Helge Mevik).

The value of the quality measurement is often used only to allocate the product or raw material into one of several quality classes, for instance “low”, “medium” and “high” values. An example of this is when raw materials have been sorted into categories uniform enough to allow constant process conditions within each category. In this situation, it is natural to use classification to allocate the observations.

Also in situations where the focus is on the continuous quality measurement value, it can be advantageous to group the values into classes. In many applications, for instance in modern spectroscopy, it may be difficult to build a good calibration model for the quality measurement. It has been demonstrated earlier (Næs and Hildrum, 1997) that discriminant analysis can improve on continuous calibration strategies in such situations.

In both of these situations, the data is split into classes corresponding to intervals of a continuous response variable, and one wishes to train a classifier to predict the class of future observations. This has to our knowledge been little described in the literature. Hand et al. (1998) construct the optimal classifier for a special case of this situation, and compare it to linear discriminant analysis. Hand et al. (2001) discuss classifiers for a generalised situation, where classes are defined in terms of other variables. The use of continuous labels for classification is discussed in Water and Duin (1981) and Duin (1982). A complementary situation has been studied by Torgo and Gama (1997), who search for discretisations of the response into intervals in order to use classification methods on regression problems.

The problem has some characteristic features regarding assessing the quality of a classification:

- It is common that the response contains substantial measurement error, which results in uncertain class memberships in the training and test data sets. This leads to uncertain estimates of the quality of a classification.
- For training and testing, one has access to the response measurements in addition to the class labels. This represents information about how certain a class membership is.
- The covariates seldom explain all variability in the true response, so the classification will not be perfect. In the present situation, it can be possible to estimate a lower limit of the misclassification rate of any classifier.
- The classes are ordered, and different misclassifications are not equally grave.

The present paper addresses the problem of assessing the performance of classifiers in when classes arise from a continuum. In Section 2, the problem is described and the pmc is discussed. An alternative performance measure, sqerr, is proposed. The effect of measurement error is discussed in Section 3. Estimates for pmc and sqerr are developed in the following two sections. They are tested on a real data set in Section 6 and in a simulation in Section 7. Finally, some concluding remarks are given in Section 8.

2 Performance measures

The present classification problem can formally be described in the following way. For each object in a population, there is an associated quality characteristic y and a vector $\boldsymbol{x}$ of covariates. The range of y is divided into C adjacent intervals with boundaries $b_0 < b_1 < \cdots < b_C$, where b_0 and b_C might be infinite. The boundaries are considered fixed and given. An object is a member of class j if $y \in [b_{j-1}, b_j)$. One wishes to use the value of $\boldsymbol{x}$ to predict the classes of future objects.

We define the *class membership function* c by $c(t) = j$ such that $t \in [b_{j-1}, b_j)$. Thus the true class of an object is $c(y)$. This will usually be denoted c.

We will think of $(\boldsymbol{x}, y)$ as a random vector following some joint distribution. It is fruitful to write this in the form

$$y = f(\boldsymbol{x}) + \varepsilon, \tag{1}$$

and let $G^{\boldsymbol{x}}$ be the conditional distribution of ε given $\boldsymbol{x}$. The distribution will be assumed to have zero mean for each $\boldsymbol{x}$; thus $\mathsf{E}[y \mid \boldsymbol{x}] = f(\boldsymbol{x})$. The effect of ε is that objects with the same $\boldsymbol{x}$ can belong to different classes, so a perfect classification based on $\boldsymbol{x}$ is usually not possible. The only exception is if, for each $\boldsymbol{x}$, all possible values of y lie in a single interval.

There are several performance measures for assessing classifiers. The *probability of misclassification* (pmc), or *error rate*, is probably the most common, but there are others such as the Brier inaccuracy and different imprecision measures (Hand, 1997). A different strategy is to graphically assess the quality of a classification, for instance by plotting predicted class membership probabilities against each other (Næs and Hildrum, 1997).

2.1 Error rate

The *probability of misclassification* (pmc) or *error rate* of a classifier $\hat{c}$ is defined as the probability that a random future object is misclassified:

$$\mathrm{pmc}(\hat{c}) = \mathrm{Prob}\{\hat{c}(\boldsymbol{x}) \neq c(y)\}. \tag{2}$$

It can be written $\mathsf{E}_{(\boldsymbol{x},y)}\, I\big(\hat{c}(\boldsymbol{x}) \neq c(y)\big)$, where I is a function taking the value 1 if its argument is true, and 0 otherwise.

For a given $\boldsymbol{x}$, the probability that the true class c of an object is different from a given class j is

$$\begin{aligned} \mathrm{pmc}(j;\boldsymbol{x}) &= \mathrm{Prob}\{c \neq j \mid \boldsymbol{x}\} \\ &= \mathrm{Prob}\{y \notin [b_{j-1}, b_j) \mid \boldsymbol{x}\} \\ &= \mathrm{Prob}\Big\{\varepsilon \notin \Big[b_{j-1} - f(\boldsymbol{x}), b_j - f(\boldsymbol{x})\Big) \Bigm| \boldsymbol{x}\Big\} \\ &= 1 - G^{\boldsymbol{x}}\Big(b_j - f(\boldsymbol{x})\Big) + G^{\boldsymbol{x}}\Big(b_{j-1} - f(\boldsymbol{x})\Big), \end{aligned} \tag{3}$$

where $G^{\boldsymbol{x}}$ is the conditional distribution of ε given $\boldsymbol{x}$. The unconditional pmc (2) can be written as

$$\begin{aligned} \mathrm{pmc}(\hat{c}) &= \mathsf{E}_{\boldsymbol{x}}\, \mathrm{Prob}\{\hat{c}(\boldsymbol{x}) \neq c(y) \mid \boldsymbol{x}\} \\ &= \mathsf{E}_{\boldsymbol{x}}\, \mathrm{pmc}(\hat{c}(\boldsymbol{x}); \boldsymbol{x}) \\ &= \mathsf{E}\Big[1 - G^{\boldsymbol{x}}\Big(b_{\hat{c}(\boldsymbol{x})} - f(\boldsymbol{x})\Big) + G^{\boldsymbol{x}}\Big(b_{\hat{c}(\boldsymbol{x})-1} - f(\boldsymbol{x})\Big)\Big]\,. \end{aligned} \tag{4}$$

If we assume that $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ independent of $\boldsymbol{x}$, then $G^{\boldsymbol{x}}(\varepsilon) = \Phi(\varepsilon/\sigma_\varepsilon)$, so (4) becomes

$$\mathrm{pmc}(\hat{c}) = \mathsf{E}\left[1 - \Phi\Big(\frac{b_{\hat{c}(\boldsymbol{x})} - f(\boldsymbol{x})}{\sigma_\varepsilon}\Big) + \Phi\Big(\frac{b_{\hat{c}(\boldsymbol{x})-1} - f(\boldsymbol{x})}{\sigma_\varepsilon}\Big)\right]. \tag{5}$$

For any distribution of $(\boldsymbol{x}, y)$ and partitioning of the range of y, there is a *minimal pmc* ($\mathrm{pmc}_{\mathrm{min}}$), which is usually nonzero. No classifier based on $\boldsymbol{x}$ can have a smaller error rate than this. Let $\mathrm{pmc}_{\mathrm{min}}(\boldsymbol{x}) = \min_j\{\mathrm{pmc}(j;\boldsymbol{x})\}$, i.e., the conditional minimal pmc given $\boldsymbol{x}$. The unconditioned minimal pmc is

$$\begin{aligned} \mathrm{pmc}_{\mathrm{min}} &= \mathsf{E}_{\boldsymbol{x}}\, \mathrm{pmc}_{\mathrm{min}}(\boldsymbol{x}) \\ &= \mathsf{E}_{\boldsymbol{x}} \min_j \Big\{1 - G^{\boldsymbol{x}}\Big(b_j - f(\boldsymbol{x})\Big) + G^{\boldsymbol{x}}\Big(b_{j-1} - f(\boldsymbol{x})\Big)\Big\}, \end{aligned} \tag{6}$$

which becomes

$$\mathsf{E}_{\boldsymbol{x}} \min_j \left\{1 - \Phi\Big(\frac{b_j - f(\boldsymbol{x})}{\sigma_\varepsilon}\Big) + \Phi\Big(\frac{b_{j-1} - f(\boldsymbol{x})}{\sigma_\varepsilon}\Big)\right\} \tag{7}$$

when $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ independent of $\boldsymbol{x}$.

2.2 Squared error rate

When classes are derived by partitioning a continuum, the classes are ordered, and some misclassifications are worse than others. For instance, if there are three classes "good", "acceptable" and "unacceptable", it is worse to classify a "good" object as "unacceptable" than as "acceptable". Even when classifying "good" objects as "acceptable", say, not all errors are equal. If the quality characteristic y of the object is close to the boundary between "good" and "acceptable", the error committed is not as serious as if y is far from the boundary.

By default the error rate does not distinguish between errors; all are considered equal. It can be modified by assigning different penalties to the misclassification between different classes. It can however not distinguish between different misclassifications from for instance "good" into "acceptable".

We propose a new performance measure, called the *squared error rate* (sqerr), that penalises errors according to their severity. It is inspired by the *mean squared error* (MSE) used in regression. The measure is a generalisation of the error rate, in the sense that it is the average of a misclassification penalty, where the penalty depends on not only the class of the object but also its y value.

We define the *squared error penalty function* for a predicted class j when the true quality characteristic value is y, as

$$\operatorname{sqe}(j,y) = \begin{cases} (y-b_{j-1})^2, & y < b_{j-1} \\ 0, & y \in [b_{j-1}, b_j) \\ (y-b_j)^2, & y \geq b_j \end{cases} . \tag{8}$$

That is, the penalty for a correct classification is zero, and the penalty for a misclassification is the squared distance between the y value of the object and the predicted class. The *squared error rate* (sqerr) of a classifier $\hat{c}$ is the expected squared error when classifying future random observations:

$$\operatorname{sqerr}(\hat{c}) = \mathsf{E}_{(\boldsymbol{x},y)} \operatorname{sqe}(\hat{c}(\boldsymbol{x}), y). \tag{9}$$

The proposed measure differentiates misclassifications as described above. Consider for instance a situation with three classes, with boundaries $b_1 = 1.5$ and $b_2 = 2.0$. Suppose an object has $y = 2.5$, thus it belongs to class three. Classifying this object as class two $(1.5 \leq y < 2.0)$ will give a penalty of $(2.5 - 2)^2 = 0.25$, while a prediction of class one $(y < 1.5)$ will give a penalty of $(2.5 - 1.5)^2 = 1$. Similarly, it is worse to classify an object with $y = 6$ as class two than classifying an object with $y = 3$ as class two. The corresponding penalties are 16 and 1, respectively.

Since the penalty is averaged over all objects, not only the misclassified ones, it incorporates information about the error rate as well, so a classifier with small sqerr will usually have both few misclassifications overall and few severe misclassifications. The measure makes it possible to compare classifiers that have approximately the same pmc, and choosing the one with the least severe errors. It could also be argued that a classifier with a larger pmc should be considered better if it has a lower sqerr.

The measure can be interpreted as a squared distance, measuring how far, on the average, the interval of the predicted class is from the true value of y. The reason for using squared distances is twofold. It puts higher penalties on the largest misclassifications, and it makes the measure more similar to the MSE, which is well-known by many statisticians and engineers.

Distances have been used for defining penalties in other situations. Tsokos and Welch (1978) use the squared distance between the means of $\boldsymbol{x}$ for each class as misclassification penalty, and Torgo and Gama (1997) use the absolute distance between the median response values of the intervals as penalty. These were developed for other purposes, and are not directly applicable to the current problem.

The sqerr for a given $\boldsymbol{x}$ and a predicted class j is

$$\begin{aligned}\operatorname{sqerr}(j;\boldsymbol{x}) &= \mathsf{E}_y[\operatorname{sqe}(j,y) \mid \boldsymbol{x}] \\ &= \mathsf{E}_\varepsilon[\operatorname{sqe}(j, f(\boldsymbol{x})+\varepsilon) \mid \boldsymbol{x}] \\ &= \int_{-\infty}^{b_{j-1}-f(\boldsymbol{x})} \Big(f(\boldsymbol{x})+\varepsilon-b_{j-1}\Big)^2 \, dG^{\boldsymbol{x}}(\varepsilon) \\ &\quad + \int_{b_j-f(\boldsymbol{x})}^{\infty} \Big(f(\boldsymbol{x})+\varepsilon-b_j\Big)^2 \, dG^{\boldsymbol{x}}(\varepsilon).\end{aligned} \tag{10}$$

The last equality follows from writing out the definitions of sqe and the expectation explicitly.

If $\varepsilon \sim N(0,\sigma_\varepsilon^2)$ independent of $\boldsymbol{x}$, direct evaluation of the integrals in (10) yields

$$\begin{aligned}&\operatorname{sqerr}(j;\boldsymbol{x}) = \\ &\Big((b_{c-1}-f(\boldsymbol{x}))^2+\sigma_\varepsilon^2\Big)\,\Phi\Big(\frac{b_{c-1}-f(\boldsymbol{x})}{\sigma_\varepsilon}\Big) + (b_{c-1}-f(\boldsymbol{x}))\sigma_\varepsilon^2 g(b_{c-1}-f(\boldsymbol{x})) \\ &+ \Big((b_c-f(\boldsymbol{x}))^2+\sigma_\varepsilon^2\Big)\Big(1-\Phi\Big(\frac{b_c-f(\boldsymbol{x})}{\sigma_\varepsilon}\Big)\Big) - (b_c-f(\boldsymbol{x}))\sigma_\varepsilon^2 g(b_c-f(\boldsymbol{x})),\end{aligned} \tag{11}$$

where g is the probability density function of ε; $g(\varepsilon) = e^{-\varepsilon^2/(2\sigma_\varepsilon^2)}/\Big(\sigma_\varepsilon\sqrt{2\pi}\Big)$.

The sqerr of a classifier (9) can be written $\operatorname{sqerr}(\hat{c}) = \mathsf{E}_{\boldsymbol{x}} \operatorname{sqerr}(\hat{c}(\boldsymbol{x});\boldsymbol{x})$.

Like the pmc, in any given situation, there is a *minimal sqerr*, i.e., the smallest sqerr any classifier can attain:

$$\mathrm{sqerr}_{\min} = \mathsf{E}_{\boldsymbol{x}}\, \mathrm{sqerr}_{\min}(\boldsymbol{x}), \tag{12}$$

where $\mathrm{sqerr}_{\min}(\boldsymbol{x}) = \min_j\{\mathrm{sqerr}(j; \boldsymbol{x})\}$ is the conditional minimal sqerr, given $\boldsymbol{x}$.

3 Measurement error

A problem often encountered in applications is that y cannot be measured without error. The measurement error has the effect that one cannot be certain about the true class of the observations in a data set, especially for observations with y value close to a class boundary. This means that in general, any data set available for training or testing a classifier will contain a fraction of objects with incorrect class labels. The effect of incorrect class labels on discriminant analysis has been studied by several authors (Chhikara and McKeon, 1984; Lachenbruch, 1966, 1974; McLachlan, 1972). The focus in the present paper, however, is on *estimation* of the performance of classifiers.

Letting z be the measurement, we assume the relation $z = y + \delta$, where δ is random. Let H^y be the conditional distribution of δ given y. In general, the distribution might depend on the value of y. Thus the data follows the relation

$$z = f(\boldsymbol{x}) + \varepsilon + \delta. \tag{13}$$

The *apparent class* or *class label* of an object is the class defined by the measured quality characteristic; $c(z)$. A training or test data set of size n will consist of vectors $(\boldsymbol{x}_1, z_1, c(z_1))$, …, $(\boldsymbol{x}_n, z_n, c(z_n))$. The true y values y_i and the associated true classes $c_i = c(y_i)$ are generally not available.

3.1 Estimating the error in a data set

The expected fraction of wrong class labels in a random sample is called the *data error rate* ($\mathrm{err}_{\mathsf{data}}$), i.e.,

$$\mathrm{err}_{\mathsf{data}} = \mathrm{Prob}\{c(z) \neq c\}. \tag{14}$$

For a given y, the probability that the class label $c(z)$ defined by z is different from the true class $c = c(y)$, is

$$\begin{aligned}\mathrm{Prob}\{c(z) \neq c \mid y\} &= \mathrm{Prob}\{z \notin [b_{c-1}, b_c) \mid y\} \\ &= \mathrm{Prob}\{\delta \notin [b_{c-1} - y, b_c - y) \mid y\} \\ &= 1 - H^y(b_c - y) + H^y(b_{c-1} - y),\end{aligned} \tag{15}$$

where H^y is the conditional distribution of δ given y. Thus the data error rate equals

$$\begin{aligned}\mathrm{err}_{\mathrm{data}} &= \mathrm{Prob}\{c(z) \neq c\} \\ &= \mathsf{E}_y \, \mathrm{Prob}\{c(z) \neq c(y) \mid y\} \\ &= \mathsf{E}_y \left[1 - H^y(b_{c(y)} - y) + H^y(b_{c(y)-1} - y)\right].\end{aligned} \tag{16}$$

If we assume that the distribution of δ is $N(0, \sigma_\delta^2)$, and that δ is independent of y, the data error rate can be written

$$\mathsf{E}_y \left[1 - \Phi\Big(\frac{b_{c(y)} - y}{\sigma_\delta}\Big) + \Phi\Big(\frac{b_{c(y)-1} - y}{\sigma_\delta}\Big)\right]. \tag{17}$$

To calculate $\mathrm{err}_{\mathrm{data}}$, both σ_δ and the distribution of y are needed. For many standardised measurements, estimates of σ_δ are available; otherwise σ_δ can be estimated by repeated measurements on objects. The best generally available estimate of y is the measurement z. Thus the distribution of y can be estimated by the empirical distribution of z in a random sample. Given an estimate $\hat{\sigma}_\delta^2$, and a random data set of n observations, the data error rate can be estimated by

$$\widehat{\mathrm{err}}_{\mathrm{data}} = \frac{1}{n} \sum_{i=1}^{n} \left(1 - \Phi\Big(\frac{b_{c(z_i)} - z_i}{\hat{\sigma}_\delta}\Big) + \Phi\Big(\frac{b_{c(z_i)-1} - z_i}{\hat{\sigma}_\delta}\Big)\right). \tag{18}$$

Corresponding to the data error rate, there is a *data squared error rate* ($\mathrm{sqerr}_{\mathrm{data}}$), defined as

$$\mathrm{sqerr}_{\mathrm{data}} = \mathsf{E}_{(y,z)} \, \mathrm{sqe}(c(z), y) = \mathsf{E}_y \Big[\mathsf{E}[\mathrm{sqe}(c(z), y) \mid y]\Big], \tag{19}$$

i.e., the squared error rate of the apparent class labels in a random sample.

For a given response value y, the expected squared error of the apparent class

$c(z)$ is equal to

$$\begin{aligned}
\mathsf{E}[\mathrm{sqe}(c(z), y) \mid y] &= \sum_{j=1}^{C} \mathrm{sqe}(j, y) \operatorname{Prob}\{c(z) = j \mid y\} \\
&= \sum_{j=1}^{C} \mathrm{sqe}(j, y) \big(H^y(b_j - y) - H^y(b_{j-1} - y) \big) \\
&= \sum_{j<c(y)} (y - b_j)^2 \big(H^y(b_j - y) - H^y(b_{j-1} - y) \big) \\
&\quad + \sum_{j>c(y)} (y - b_{j-1})^2 \big(H^y(b_j - y) - H^y(b_{j-1} - y) \big).
\end{aligned} \tag{20}$$

The second equality follows from (15).

If we assume $\delta \sim N(0, \sigma_\delta^2)$, independent of $\boldsymbol{x}$ and y, the data squared error rate can be written

$$\begin{aligned}
\mathsf{E}_y \Bigg[&\sum_{j<c(y)} (y - b_j)^2 \left(\Phi\Big(\frac{b_j - y}{\sigma_\delta}\Big) - \Phi\Big(\frac{b_{j-1} - y}{\sigma_\delta}\Big) \right) \\
&+ \sum_{j>c(y)} (y - b_{j-1})^2 \left(\Phi\Big(\frac{b_j - y}{\sigma_\delta}\Big) - \Phi\Big(\frac{b_{j-1} - y}{\sigma_\delta}\Big) \right) \Bigg].
\end{aligned} \tag{21}$$

Given an estimate of σ_δ^2 and a random set of n observations, this can be estimated by substituting z_i for y, and averaging over the data set:

$$\begin{aligned}
\widehat{\mathrm{sqerr}}_{\mathsf{data}} = \frac{1}{n} \sum_{i=1}^{n} \Bigg(&\sum_{j<c(z_i)} (z_i - b_j)^2 \left(\Phi\Big(\frac{b_j - z_i}{\hat{\sigma}_\delta}\Big) - \Phi\Big(\frac{b_{j-1} - z_i}{\hat{\sigma}_\delta}\Big) \right) \\
&+ \sum_{j>c(z_i)} (z_i - b_{j-1})^2 \left(\Phi\Big(\frac{b_j - z_i}{\hat{\sigma}_\delta}\Big) - \Phi\Big(\frac{b_{j-1} - z_i}{\hat{\sigma}_\delta}\Big) \right) \Bigg).
\end{aligned} \tag{22}$$

4 Estimating the pmc of a classifier

The most widely used estimate of pmc is the fraction of misclassifications when predicting a test set. In the present article, this will be called the *true error count* of the classifier. It is a non-strictly proper measure (Hand, 1997), meaning that no classifier can have lower value of the measure than a classifier that places all objects in the correct class.

Since the test data set in the present situation contains errors among the class labels, the *observed* fraction of misclassifications is in general not equal to the *true* error count. This observed error count will be called the *apparent error count* (erc) of the classifier:

$$\operatorname{erc}(\hat{c}) = \frac{1}{n}\sum_{i=1}^{n} I\Big(\hat{c}(\boldsymbol{x}_i) \neq c(z_i)\Big). \tag{23}$$

The apparent error count is the only generally available error count, and is the combination of true misclassifications and errors in the test data set. It is not a proper measure. A classifier with true $\mathrm{pmc} = 0$, will have an apparent error count equal to the data error of the test set, while a classifier that has an apparent error count of 0, has in reality misclassified the observations with incorrect class label, and thus has a $\mathrm{pmc} > 0$.

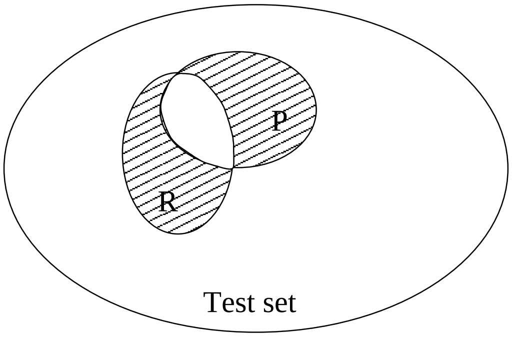

Fig. 1. Illustration of the different errors. The ellipsis R is the set of objects that have incorrect label in the test set. The ellipsis P is the set of objects where the predicted class and the test set label disagree, that is, the apparent errors. The objects in the symmetric difference $P\Delta R$ (the squared area) are misclassified. The objects in the intersection $P \cap R$ might be misclassified if there are more than two classes.

More generally we have the following limits for the true error count:

$$|\operatorname{erc} - d| \leq \text{true error count} \leq \operatorname{erc} + d, \tag{24}$$

where d is the fraction of objects with wrong class label in the data set. The situation is depicted in Figure 1. The lower limit can be attained when all apparently misclassified observations have wrong class labels, or *vice versa*, i.e., $P \subseteq R$ or $R \subseteq P$ in Figure 1. In both cases, the true error count $\geq |\operatorname{erc} - d|$, with the equality always holding if there are only two classes. The upper limit is attained when the set of apparently misclassified observations and the set of test set errors are disjoint, in which case both sets are misclassified by the classifier, and the true error count is $\operatorname{erc} + d$.

The true error count can be anywhere inside this interval of length $2 \times \min\{\operatorname{erc}, d\}$. Thus the accuracy of the apparent error count as an estimate of pmc depends on the fraction of errors in the test set, as well as on the size of estimate itself.

As an attempt to compensate for the uncertainty in the class labels, we introduce a modification of the apparent error count. The idea is to give observations with uncertain class labels small weight when calculating the apparent error count. Given the distribution of the measurement error δ, the conditional probability given y that the class label is correct, $\text{Prob}\{c(z) = c \mid y\}$, can be calculated (see (15)). Assuming $\delta \sim N(0, \sigma_\delta^2)$, independent of $(\boldsymbol{x}, y)$, the *adjusted error count* is

$$\text{erc}_a(\hat{c}) = \sum_{i=1}^{n} w_i I\Big(\hat{c}(\boldsymbol{x}_i) \neq c(z_i)\Big) \Big/ \sum_{i=1}^{n} w_i, \tag{25}$$

where $w_i = \Phi\left((b_{c(z_i)} - z_i)/\hat{\sigma}_\delta\right) - \Phi\left((b_{c(z_i)-1} - z_i)/\hat{\sigma}_\delta\right)$ estimates $\text{Prob}\{c(z_i) = c_i \mid y_i\}$. Multiplying with w_i gives low weight to apparent misclassifications of objects with responses near a border, and dividing by $\sum w_i$ instead of n gives a positive contribution from apparently correct classifications with responses close to a border.

From the formulation of the pmc in (5), we get an alternative estimate:

$$\widehat{\text{pmc}}(\hat{c}) = \frac{1}{n} \sum_{i=1}^{n} \left(1 - \Phi\Big(\frac{b_{\hat{c}(\boldsymbol{x}_i)} - z_i}{\hat{\sigma}_\varepsilon}\Big) + \Phi\Big(\frac{b_{\hat{c}(\boldsymbol{x}_i)-1} - z_i}{\hat{\sigma}_\varepsilon}\Big)\right) \tag{26}$$

This estimate should have lower variance than the error counts, because it is the average of a continuous function, whereas the error counts are (weighted) averages of a discrete function.

4.1 Minimal pmc

When the distribution of ε is known, the minimal pmc (6) can be calculated. Assuming $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ independent of $\boldsymbol{x}$, it can be estimated with

$$\widehat{\text{pmc}}_{\min} = \frac{1}{n} \sum_{i=1}^{n} \min_j \left\{1 - \Phi\Big(\frac{b_j - z_i}{\hat{\sigma}_\varepsilon}\Big) + \Phi\Big(\frac{b_{j-1} - z_i}{\hat{\sigma}_\varepsilon}\Big)\right\} \tag{27}$$

given an estimate of σ_ε and a data set of size n.

5 Estimating the sqerr of a classifier

The definition of the sqerr suggests the estimate

$$\text{sqerc}(\hat{c}) = \frac{1}{n} \sum_{i=1}^{n} \text{sqe}(\hat{c}(\boldsymbol{x}_i), z_i), \tag{28}$$

given a random data set with n observations. This is called the *apparent squared error count.* It estimates the *true squared error count* $\sum \operatorname{sqe}(\hat{c}(\boldsymbol{x}_i), y_i)/n$. As with the apparent error count, the apparent squared error count is not proper. Also, an inequality analogous to (24) holds approximately for the *true* squared error count.

We have that $\mathsf{E}[(b_j - z)^2 \mid y] = (b_j - y)^2 + \sigma_\delta^2$, so $(b_j - z)^2$ overestimates $(b_j - y)^2$. This can be compensated for by subtracting $\hat{\sigma}_\delta^2$ in the estimate. Using this adjustment, we get the following *adjusted squared error count*:

$$\begin{aligned}
&\operatorname{sqerc}_a(\hat{c}) = \\
&\frac{1}{n}\sum_{i=1}^{n}\Big(\big((b_{\hat{c}_i-1} - z_i)^2 - \hat{\sigma}_\delta^2\big)\, I\left(c(z_i) < \hat{c}_i\right) + \big((b_{\hat{c}_i} - z_i)^2 - \hat{\sigma}_\delta^2\big)\, I\left(c(z_i) > \hat{c}_i\right)\Big) \\
&\qquad = \frac{1}{n}\sum_{i=1}^{n}\Big(\operatorname{sqe}(\hat{c}(\boldsymbol{x}_i), z_i) - \hat{\sigma}_\delta^2 I\big(\hat{c}(\boldsymbol{x}_i) \neq c(z_i)\big)\Big) = \operatorname{sqerc} - \hat{\sigma}_\delta^2 \operatorname{erc}, \quad (29)
\end{aligned}$$

namely the apparent squared error count minus $\hat{\sigma}_\delta^2$ times the apparent error count.

From the formulation of $\operatorname{sqerr}(\hat{c})$ as $\mathsf{E}_{\boldsymbol{x}} \operatorname{sqerr}(\hat{c}(\boldsymbol{x}); \boldsymbol{x})$ we get an alternative estimate for $\operatorname{sqerr}(\hat{c})$:

$$\widehat{\operatorname{sqerr}}(\hat{c}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\operatorname{sqerr}}(\hat{c}(\boldsymbol{x}_i); z_i), \qquad (30)$$

where

$$\begin{aligned}
&\widehat{\operatorname{sqerr}}(j; z_i) = \\
&\qquad \big((b_{j-1} - z_i)^2 + \hat{\sigma}_\varepsilon^2\big)\, \Phi\Big(\frac{b_{j-1} - z_i}{\hat{\sigma}_\varepsilon}\Big) + (b_{j-1} - z_i)\hat{\sigma}_\varepsilon^2 g(b_{j-1} - z_i) \\
&\qquad + \big((b_j - z_i)^2 + \hat{\sigma}_\varepsilon^2\big)\Big(1 - \Phi\Big(\frac{b_j - z_i}{\hat{\sigma}_\varepsilon}\Big)\Big) - (b_j - z_i)\hat{\sigma}_\varepsilon^2 g(b_j - z_i). \quad (31)
\end{aligned}$$

5.1 Minimal sqerr

When the distribution of ε is known, the minimal sqerr (12) can be calculated. Given a random sample of size n, a natural estimate of this is

$$\widehat{\operatorname{sqerr}}_{\min} = \frac{1}{n}\sum_{i=1}^{n} \min_j\{\widehat{\operatorname{sqerr}}(j; z_i)\}, \qquad (32)$$

where $\widehat{\operatorname{sqerr}}$ is defined in (31).

6 Example

The estimates were tested on a data set consisting of of near infrared transmission (NIT) and water content measurements of 103 meat samples. The NIT measurements had 100 wavelengths from 850 nm to 1050 nm. The data set is described in more detail in Næs and Isaksson (1992). The data was divided into two classes: 53 samples with at most 65 % water in one class and 50 samples with more than 65 % water in the other.

The water content measurements had little measurement error: The measurements ranged from 50.5 % to 76.9 % and had an estimated standard deviation $\hat{\sigma}_\delta = 0.22$ %. The estimated err_{data} was 0.015, so most samples in the data set can be expected to have the correct class label. The corresponding data sqerr was 3.2×10^{-4}. This ensures that erc and sqerc are almost unbiased. The properties of the estimates under different levels of σ_δ are investigated in Section 7.

A number of classifiers were trained on the first six principal components of the NIT data: linear discriminant analysis (lda), quadratic discriminant analysis (qda), k-nearest neighbour classification with $k = 1$ (1-nn) and $k = 3$ (3-nn) and a classification tree with the gini splitting index (tree) (Breiman et al., 1984). The err and sqerr were estimated, using ten-fold cross-validation. The principal components were recalculated for each cross-validation segment.

Due to the small $\hat{\sigma}_\delta$, the adjusted estimates erc_a and sqerc_a differed very little from the unadjusted estimates (less than 0.01, respectively 0.02) and are therefore not shown. The classifiers were ranged in the following ways (estimates in parentheses):

err: qda (0.18) $<$ lda (0.20) $<$ 3-nn (0.28) $=$ tree (0.28) $<$ 1-nn (0.33)

sqerr: lda (0.55) $<$ qda (1.10) $<$ tree (2.44) $<$ 1-nn (7.74) $<$ 3-nn (11.1)

We see that according to the pmc, qda is slightly better than lda, while lda should be preferred according to the sqerr. Likewise, the order of 1-nn and 3-nn are reversed by the two measures, and tree is considered better than the k-nns by sqerr, but equal to 3-nn by err.

7 Simulation

A simulation was performed to test the different estimates. The simulation was designed to illustrate a three-class situation with a nonlinear relationship between $\boldsymbol{x}$ and y.

Data were generated following the model

$$\begin{gathered} y = x_1 + x_2 + x_2^2 + \varepsilon \quad \text{and} \quad z = y + \delta, \\ \text{where } (x_1, x_2)^t \sim N\left(\mathbf{0}, \begin{pmatrix} 2.0 & 0.4 \\ 0.4 & 1.0 \end{pmatrix}\right), \\ \varepsilon \sim N(0, \sigma_\varepsilon^2) \quad \text{and} \quad \delta \sim N(0, \sigma_\delta^2) \end{gathered} \tag{33}$$

The range of y is the whole real line, and was divided into three intervals at $b_1 = 0$ and $b_2 = 0.6$. Observations belong to class one if $y \in (-\infty, 0)$, class two if $y \in [0, 0.6)$ and class three if $y \in [0.6, \infty)$. The boundaries were chosen to give approximate class frequencies 0.38, 0.12 and 0.50. The apparent class of an objects is analogously defined by the value of the measurement z.

The values of σ_δ and σ_ε were varied according to a 3^2 factorial design, with levels 0.15, 0.3 and 0.9 for σ_ε, and 0, 0.15 and 0.5 for σ_δ. This produces data sets with approximate minimal pmcs of 0.05, 0.10 and 0.20 and data errs of 0, 0.05 and 0.15. The corresponding minimal sqerrs are 0.00075, 0.0060 and 0.13 and data sqerrs 0, 0.0007 and 0.024.

We generated 100 random training data sets of size 100 for each combination of σ_ε and σ_δ. Also, a test set of size 10000 was generated, containing x_1, x_2 and the true response y. For each training data set, the following was estimated:

- The *minimal pmc* and *minimal sqerr*
- The *data err* and *data sqerr*
- The *pmc* and *sqerr* of a number of classifiers trained on the training data set

The pmc and sqerr of each classifier were also calculated by classifying the test data set. Also, the true minimal pmc, minimal sqerr, data err and data sqerr were calculated once for each combination, using the test data set as an approximation to the true distribution.

7.1 Results

The data err was estimated with (18) and the data sqerr was estimated using (22). The biases of the estimates are shown in Figure (2) for the different combinations of σ_ε and σ_δ. Both estimates are unbiased when there is no measurement error, but tend to underestimate when the measurement error increases. The biases were less than 5 % of the corresponding true data err or sqerr, when $\sigma_\delta > 0$.

The average standard deviations for increasing levels of σ_δ were 0, 0.011 and

0.020 for $\widehat{\mathrm{err}}_{\mathrm{data}}$ and 0, 1.4×10^{-4} and 2.0×10^{-3} for $\widehat{\mathrm{sqerr}}_{\mathrm{data}}$. The influence of σ_ε on the standard deviations was negligible.

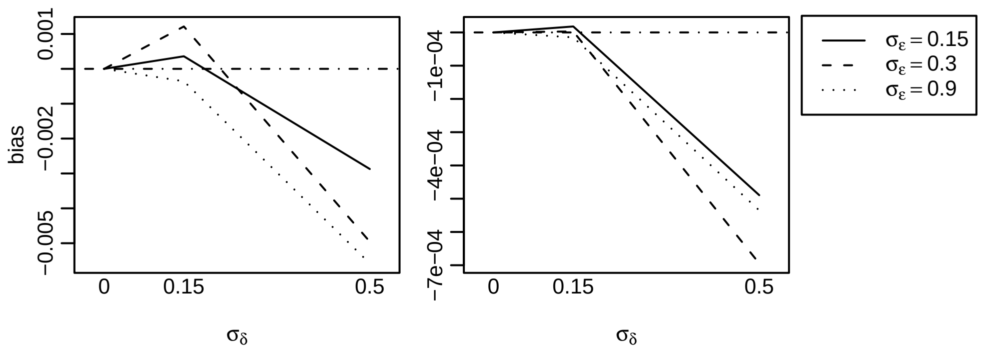

Fig. 2. Bias of the estimated data errors. Left panel: data error rate, right panel: data squared error rate.

The minimal err and sqerr were estimated with (27) and (32), respectively. The biases of these estimates are shown in Figure 3. When $\sigma_\varepsilon = 0.9$ or $\sigma_\delta = 0.5$, the estimates were downward biased with up to 14 % of the corresponding true error. In the other situations, the bias was very small.

The average standard deviations for increasing levels of σ_ε were 0.011, 0.016 and 0.021 for $\widehat{\mathrm{pmc}}_{\mathrm{min}}$, and 2.0×10^{-4}, 1.1×10^{-3} and 1.6×10^{-2} for $\widehat{\mathrm{sqerr}}_{\mathrm{min}}$. The measurement error had only minor effect on the standard deviations.

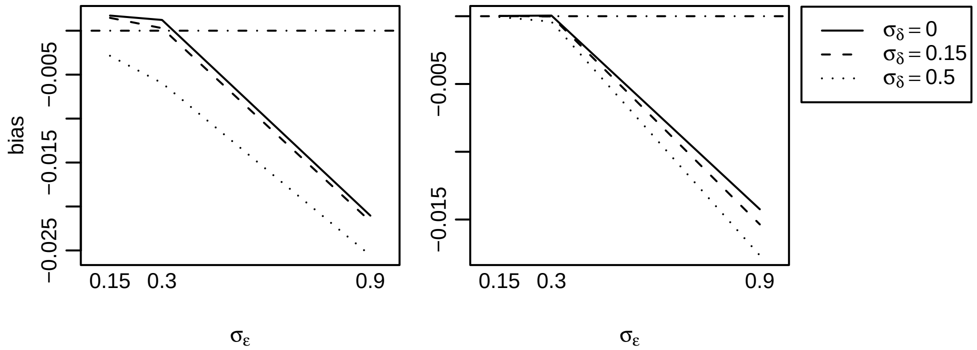

Fig. 3. Bias of the estimated minimal errors. Left panel: minimal error rate, right panel: minimal squared error rate.

Linear discriminant analysis, quadratic discriminant analysis, the k-nearest neighbour rule with $k = 5$ and a regression based classifier were trained on each of the training data sets. The regression based method was as follows: Let $\hat{y}(\boldsymbol{x}) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_2^2$ be the least squares estimate of $f(\boldsymbol{x})$. For each $\boldsymbol{x}$ to be classified, choose the class whose interval contains $\hat{y}(\boldsymbol{x})$; i.e., $c(\hat{y}(\boldsymbol{x}))$.

The pmc was estimated by the apparent error count (23), the alternative pmc estimate (26) and the adjusted error count (25). The sqerr was estimated by the apparent squared error count (28), the alternative sqerr estimate (30), and the adjusted squared error count (29). All estimates were calculated using 20-fold cross validation on the training data set.

The bias and standard deviation of the error rate estimates are summarised in Figures 4 and 5. The $\widehat{\mathrm{pmc}}$ estimate consistently overestimated the pmc, but had lower variance than the other estimates. The apparent error count, erc, also overestimated the pmc, but less than $\widehat{\mathrm{pmc}}$. The adjusted error count, erc_a, underestimated the pmc by about the same amount that erc overestimated by, but had lower variance.

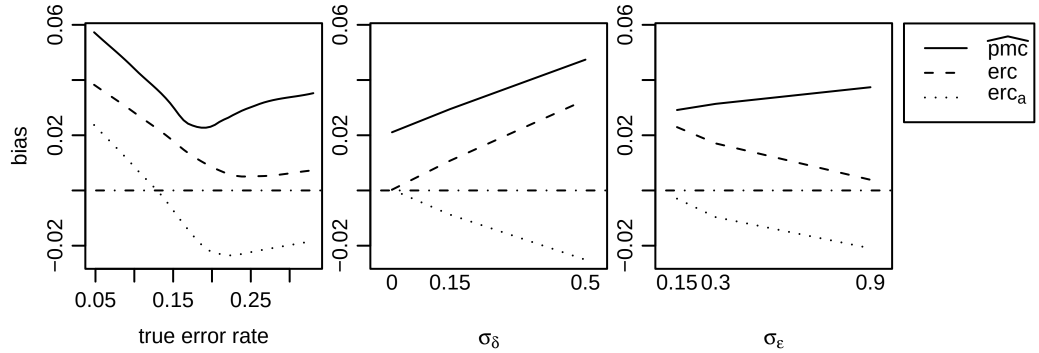

Fig. 4. Bias of the error rate estimates, broken down by true error rate, σ_δ and σ_ε. The left panel shows predicted biases from loess regressions of bias on true error rate. The middle and right panels show average biases for the levels of σ_δ and σ_ε.

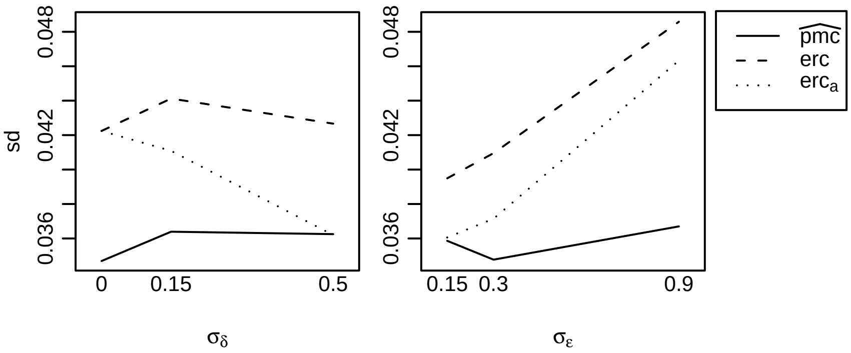

Fig. 5. Average standard deviations of the error rate estimates, for the different levels of σ_ε and σ_δ.

The bias and standard deviation of the sqerr estimates are summarised in Figures 6 and 7. The $\widehat{\mathrm{sqerr}}$ consistently overestimated the sqerr. sqerc was also upward biased, but less than $\widehat{\mathrm{sqerr}}$. On the average, sqerc_a was practically unbiased for all levels of σ_δ or σ_ε. The left panel of Figure 6 seems to imply that sqerc and sqerc_a generally underestimated the sqerr; however the distribution of the true sqerr is quite skewed to the left, and for instance only about 10 % of the true sqerr values were greater than 0.45.

In terms of standard deviation, there is little difference between the estimates. The $\widehat{\mathrm{sqerr}}$ estimate has a slightly higher variance than the other two.

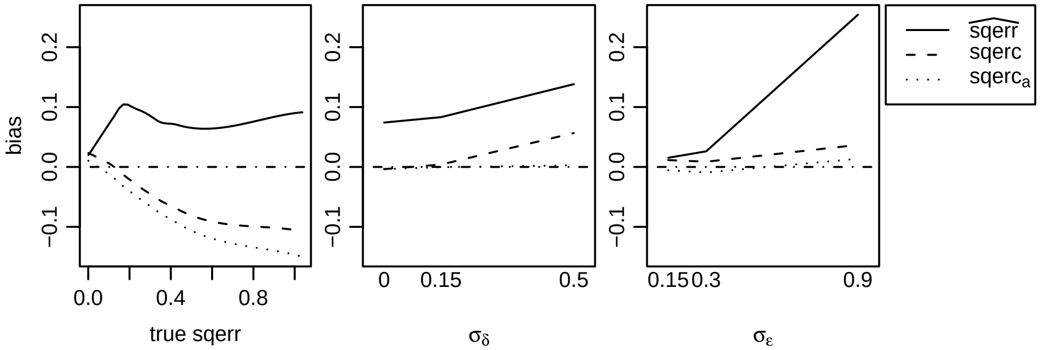

Fig. 6. Bias of the squared error rate estimates, broken down by true squared error rate, σ_δ and σ_ε. The left panel shows predicted biases from loess regressions of bias on true squared error rate. The middle and right panels show average biases for the levels of σ_δ and σ_ε.

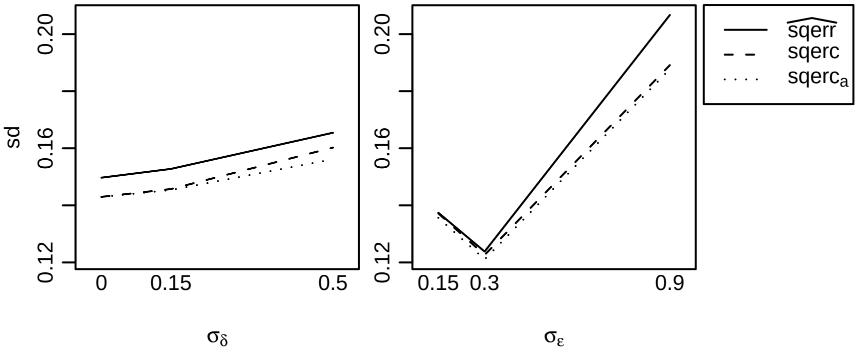

Fig. 7. Average standard deviations of the squared error rate estimates, for the different levels of σ_ε and σ_δ.

8 Discussion

We have proposed a new performance measure for classifiers, the squared error rate (sqerr), specifically designed for the situation where classes arise from a continuum. The measure penalises errors according to their severity. It can be interpreted as a mean distance, and is similar to the MSE of regressions, which is well-known and well-understood by many practitioners. Some people are more accustomed to using RMSE than MSE. For these, it might be preferable to use the square root of the sqerr.

The proposed estimates of data error rate, $\widehat{\text{err}}_{\text{data}}$, and squared data error rate, $\widehat{\text{sqerr}}_{\text{data}}$, performed well in the simulation. When there was 15 % data error rate, they were slightly downward biased, otherwise the bias was negligible. Their standard deviation was between 8 and 22 % of the true values.

The estimates for minimal pmc, $\widehat{\text{pmc}}_{\text{min}}$, and minimal sqerr, $\widehat{\text{sqerr}}_{\text{min}}$, showed similar performance. They underestimated the true values with less than 15 %, and their standard deviation ranged between 10 and 27 % of the true values.

Several estimates for the pmc were proposed. Both the apparent error count, erc, and the adjusted error count, erc_a, was less biased than $\widehat{\mathrm{pmc}}$. They also have the benefit over $\widehat{\mathrm{pmc}}$ that they don't require an estimate of σ_ε, which can be difficult to estimate. Neither of the error counts outperformed the other. They were biased in opposite directions, by about the same amount. The erc_a had a lower standard deviation than the erc, so it might seem slightly better.

Among the estimates for the squared error rate, the adjusted squared error count, sqerc_a, performed best. It was practically unbiased for all levels of σ_δ and σ_ε, but underestimated high values of sqerr. It was also slightly less variable than sqerc, and doesn't need an estimate of σ_ε, as $\widehat{\mathrm{sqerr}}$ does.

There are several techniques for reducing the variance of pmc estimates, for instance by the bootstrap (Efron and Tibshirani, 1993). Similar techniques could be applied to the sqerr estimates, if desirable.

8.1 Estimating σ_ε^2.

Some of the estimates in the present paper require an estimate of $\sigma_\varepsilon^2 = \mathrm{Var}(\varepsilon)$. In this section, it will be assumed that an estimate $\hat{\sigma}_\delta^2$ exists. It can be easier to estimate $\mathrm{Var}(\varepsilon + \delta)$ than $\mathrm{Var}(\varepsilon)$. Assuming that ε and δ are independent, σ_ε^2 can estimated by $\widehat{\mathrm{Var}}(\varepsilon + \delta) - \hat{\sigma}_\delta^2$.

If a data set has been collected according to an experimental design in $\boldsymbol{x}$, and includes response measurements of m separate objects with the same value of $\boldsymbol{x}$, their sample variance is an estimate of $\mathrm{Var}(\varepsilon + \delta \mid \boldsymbol{x})$. This can be averaged over $\boldsymbol{x}$ to give a pooled estimate of $\mathrm{Var}(\varepsilon + \delta)$. Note that the m measurements for each $\boldsymbol{x}$ must be taken on separate objects, not the same object. Otherwise the sample variance estimates σ_δ^2, not $\mathrm{Var}(\varepsilon + \delta)$.

In other situations, estimating $\mathrm{Var}(\varepsilon + \delta)$ is not so straight-forward. One possibility is to fit a model to the data and estimate the variance from the fitted residuals. Due to possible lack of fit or over-fitting, the estimate can be biased. Alternatively, one could use a weighted local estimate of $\mathrm{Var}(\varepsilon + \delta)$ and average this over $\boldsymbol{x}$.

Acknowledgements

The author wishes to thank Prof. Tormod Næs, Matforsk and the University of Oslo, for valuable suggestions and discussions.

References

Breiman, L., Friedman, J., Olshen, R. A., Stone, C. J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA, USA.

Chhikara, R. S., McKeon, J., 1984. Linear discriminant analysis with misallocation in training samples. Journal of the American Statistical Association 79, 899–906.

Duin, R. P. W., 1982. The use of continuous variables for labeling objects. Pattern Recognition Letters 1, 15–20.

Efron, B., Tibshirani, R. J., 1993. An Introduction to the Bootstrap. Vol. 57 of Monographs on Statistics and Applied Probability. Chapman & Hall, New York.

Hand, D. J., 1997. Construction and Assessment of Classification Rules. Wiley Series in Probability and Statistics. John Wiley & Sons, New York.

Hand, D. J., Li, H. G., Adams, N. M., 2001. Supervised classification with structured class definitions. Computational Statistics & Data Analysis 36 (2), 209–225.

Hand, D. J., Oliver, J. J., Lunn, A. D., 1998. Discriminant analysis when the classes arise from a continuum. Pattern Recognition 31 (5), 641–650.

Lachenbruch, P. A., 1966. Discriminant analysis when the initial samples are misclassified. Technometrics 8 (4), 657–662.

Lachenbruch, P. A., 1974. Discriminant analysis when the initial samples are misclassified II: Non-random miscalssification models. Technometrics 16 (3), 419–424.

McLachlan, G. J., 1972. Asymptotic results for discriminant analysis when the initial samples are misclassified. Technometrics 14 (2), 415–422.

Næs, T., Hildrum, K. I., 1997. Comparison of multivariate calibration and discriminant analysis in evaluating NIR spectroscopy for determination of meat tenderness. Applied Spectroscopy 51 (3), 350–357.

Næs, T., Isaksson, T., 1992. Locally weighted regression in diffuse near-infrared transmittance spectroscopy. Applied Spectroscopy 46 (1), 34–43.

Torgo, L., Gama, J., 1997. Search-based class discretization. Lecture Notes in Artificial Intelligence 1224, 266–273.

Tsokos, C. P., Welch, R. L. W., 1978. Bayes discrimination with mean square error loss. Pattern Recognition 10, 113–123.

Water, F. T. B. t., Duin, R. P. W., 1981. Dealing with a priori knowledge by fuzzy labels. Pattern Recognition 14 (1–6), 111–115.