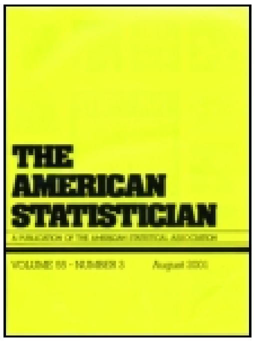

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Local Versus Global Models for Classification Problems

David J Hand & Veronica Vinciotti

To cite this article: David J Hand & Veronica Vinciotti (2003) Local Versus Global Models for Classification Problems, The American Statistician, 57:2, 124-131, DOI: 10.1198/0003130031423

To link to this article: https://doi.org/10.1198/0003130031423

Published online: 01 Jan 2012.

Local Versus Global Models for Classification Problems: Fitting Models Where it Matters

David J. HAND and Veronica VINCIOTTI

It is generally argued that predictive or decision making steps in statistics are separate from the model building or inferential steps. In many problems, however, predictive accuracy matters more in some parts of the data space than in others, and it is appropriate to aim for greater model effectiveness in those regions. If the relevant parts of the space depend on the use to which the model is to be put, then the best model will depend also on this intended use. We illustrate using examples from supervised classification.

KEY WORDS: Inference; Decision making; Logistic discrimination; Model fitting; Prediction.

1. INTRODUCTION

The traditional approach to prediction and decision analysis in statistics and operations research separates the model building (inferential) aspects from the predictive aspects. Lindley (1965, p. 66–67), for example, says:

> . . . the inference problem is basic to any decision problem, because the latter can only be solved when the knowledge of [the probability model] . . . is completely specified . . . The person making the inference need not have any particular decision problem in mind. The scientist in his laboratory does not consider the decisions that may subsequently have to be made concerning his discoveries. His task is to describe accurately what is known about the parameters in question.

Similar remarks were made by Cox (1958), Nelder (1994, p. 226), and others. These authors noted that two distinct processes are involved, and Lindley, in particular, argued that it is not necessary to consider the use to which the model will be put when carrying out the model-building inference. In this article, we argue that this is often not true, and that, in general, it is necessary to take the prospective use of the model into account when building it.

The two-stage strategy, of inference followed by prediction or decision, involves the implicit assumption that the loss functions involved at the decision stage will not affect the model being built: that different models are not best suited to different loss functions. Unfortunately, this is generally only true if the model is correct; that is, if the model has the same form as the underlying data generating process. In such situations, the models are generally unbiased, although, as we discuss in the following, the variance of the estimate can be minimized by appropriate choice of estimator. However, the truth is that the model is hardly ever correct, and is sometimes ill-specified. There are almost always aspects of the relationships between the predictor variables and the response which are not reflected in the assumed model form. This means that the fitted model is simply that model which is "closest," in some sense, to the underlying data generating model (random variation aside). It follows that, by varying the metric used to define "closeness," one can obtain different models, each good fits in their own way. Clearly, under these circumstances, one would like to use that metric which reflects the loss function to be used in the prediction problem, so obtaining a model which is good for the particular problem being addressed.

The inference problem can be viewed as a decision problem in which the loss function has not been clearly specified, precisely because various different kinds of decision may have to be made in the future. If the loss function is not or cannot be specified, it is sensible to aim to build a model which is reasonable under a wide range of potential loss functions. It is therefore common to adopt a simple measure of how well the postulated distributions fit the underlying distributions, as far as we can tell from the data, often using metrics based on likelihood or penalized likelihood. In many problems, however, the intended use of the model means that the loss function can be specified, or at least can be narrowed down to a class of such functions. When this is the case, use of loss functions such as likelihood—if this measure is not in the given class—can lead to suboptimal, even poor, results. Or, at least, it can lead to results that are worse than those obtained using the loss function derived from the intended use to which the model will be put.

We illustrate these ideas below, in the context of supervised classification rules, and present an adjusted *local* logistic discriminant approach which takes the objectives into account.

2. GLOBAL VERSUS LOCAL GOODNESS-OF-FIT MEASURES

At the model-fitting step, traditional statistical approaches measure the distance between the fitted model and the underlying distributions using global measures of proximity—with penalized variants of likelihood being particularly popular. By "global" here we mean that all aspects of the data and the distributions contribute to the estimate of goodness of fit. Sometimes different parts of the data may be weighted differently, so that a more efficient estimator results. For example, the standard least squares regression estimator, arising from the normal-based likelihood theory, minimizes the criterion $R_1 = \sum(y_i - \hat{y}_i)^2$ (with $\hat{y}_i$ an estimate of y_i, the value of the response variable for case i) and thus applies equal weights to the likelihood contributions from each data point. However, if there is reason to suspect heteroscedasticity in the response, then weighted least squares is adopted, $R_1 = \sum w_i(y_i - \hat{y}_i)^2$, with larger weights being applied to observations which are thought to have smaller variance.

David J. Hand is Professor, and Veronica Vinciotti is a Research Fellow, Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex, UB8 3PH United Kingdom (E-mail addresses: d.j.hand@ic.ac.uk *and* Veronica.Vinciotti@brunel.ac.uk). The work of Veronica Vinciotti on this project was partially supported by a grant from the Institute of Actuaries and Fair, Isaac. We are grateful to Martin Crowder and two anonymous referees for comments on an earlier version of this article.

DOI: 10.1198/0003130031423

This is perfectly reasonable: observations with lower variance contribute more information. Note, however, the key point that the rationale for adopting such a weighting is solely one of efficiency of the estimate, and that no account is taken of potential future applications for the predictions. That is, the choice of weights ignores the fact that different points may be of differing degree of importance in the context of the problem.

This strategy is all very well, but for many predictive problems all parts of the distributions are not of equal importance: the information conveyed by some of the observations is less important than the information conveyed by others. If the model is properly specified, this is not an issue since under these circumstances a good fit in some region will not induce a poor fit in another region. However, when the model is not properly specified, then a good fit in some region may well detract from quality of fit in another: improving the fit of a misspecified model in some part of the space by forcing a close fit between the model and the underlying distributions in that region may well degrade the fit elsewhere. As we illustrate below, for some predictive problems, only parts of the space of predictor variables are of interest, so that it makes sense to focus attention in those regions, improving the fit there, even if fit elsewhere is degraded. This can be achieved through differential weighting. Of course, adopting arbitrary weightings will probably mean that the estimator is less efficient than it might be. However, we might hope that the improvement in fit in the relevant part of the space more than compensates for the increase in variance. Note that, in general, one will not be able to obtain an exact fit to the "truth" in the relevant part of the space, even if one focuses attention in this region. Only in those (presumably rare) cases where the model is properly specified locally will this be possible. In general, one will be seeking that model which minimizes some measure of discrepancy between the model and the "truth" locally.

3. SUPERVISED CLASSIFICATION

We illustrate these ideas using supervised classification methods (Hand 1981; McLachlan 1992; Ripley 1996; Hand 1997; Webb 1999).

In a supervised classification problem, we are provided with a *training* or *design* set of data, comprising descriptive measurement vectors of a sample of n cases along with identifying labels for each of these n cases indicating to which of several possible classes each of them belongs. The aim is to use these data to construct a rule which will permit one to predict the class of a new case, based solely on its descriptive measurement vector. For convenience we will restrict the following discussion to the two-class case, but the arguments are generally valid.

A large number of methods for tackling such problems have been devised. Of particular relevance in the present context is that such methods have been developed by several distinct intellectual communities, including statistics, but also including machine learning, data mining, pattern recognition, and other disciplines. These different communities have approached the problem from rather different perspectives, with the result that different kinds of tools have been developed. The statistical community has tended to think in terms of probabilities of class membership, $P(c \mid \mathbf{x})$, where c is the class label and $\mathbf{x}$ is the vector of measurements. Since, for simplicity, we are restricting ourselves to the two-class case, c is 0 or 1. A classification will be made by comparing $P(0 \mid \mathbf{x})$ with some threshold, t: assign an object with measurement vector $\mathbf{x}$ to class 0 if $P(0 \mid \mathbf{x}) > t$. Other groups have focused attention on estimating the *decision surface*, the surface at which $P(0 \mid \mathbf{x}) = t$, or the equivalent $f(P(0 \mid \mathbf{x})) = f(t)$, where f is some monotonic increasing transformation. A new case will be classified as belonging to class 0 if $f(P(0 \mid \mathbf{x})) > f(t)$ and to class 1 otherwise. Of course, the fact is that in all such problems only the decision surface is really relevant. Provided $P(0 \mid \mathbf{x}) > t$, we do not care by how much $P(0 \mid \mathbf{x})$ exceeds t, the same classification will result. This means that attention directed on accurate estimation of $P(0 \mid \mathbf{x})$ in regions far from $P(0 \mid \mathbf{x}) = t$ may be misdirected. Friedman (1997) demonstrated this in an elegant article showing that bias in the estimate of $P(0 \mid \mathbf{x})$ did not necessarily adversely affect classification performance and could, in fact, improve it.

Classification performance is often measured by *misclassification rate*—the proportion of new objects that a supervised classification rule assigns to the wrong classes. Using misclassification rate is equivalent to adopting the threshold $t = 1/2$ in the above and, as discussed in detail by Hand (1997), is based on the implicit assumption that the cost of misclassifying a class 0 point to class 1 is the same as the converse. In general, if the cost of misclassifying a class i point is k_i, then the classification threshold which minimizes the overall loss is given by $t = k_1/(k_0 + k_1)$.

To illustrate these ideas, let us consider basic methods from each of the statistics and machine learning communities. Both of the methods we consider produce linear decision surfaces.

Logistic discriminant analysis is a familiar statistical tool for supervised classification. Such a model takes $P(0 \mid \mathbf{x})$ as having the form $P(0 \mid \mathbf{x}) = \exp(\beta'\mathbf{x})/[1 + \exp(\beta'\mathbf{x})]$, with binomial variation about this level. If $P(0 \mid \mathbf{x})$ is compared with a threshold t, the decision surface is $\beta'\mathbf{x} = \log(t/1 - t) = T$, a linear form, and the decision is made by comparing $\beta'\mathbf{x}$ with T. The maximum likelihood estimates of the coefficients β are generally found by an iteratively weighted least squares scheme, which uses the criterion

$$R_2 = \sum \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i(1 - \hat{y}_i)},$$

so allowing for the heteroscedasticity arising from the binomial distribution. We see from this that the weights, $1/\hat{y}_i(1 - \hat{y}_i)$ take their smallest values where $\hat{y} \approx 0.5$ and their largest values where $\hat{y} \approx 0$ or 1. This is fine for reasons of efficiency, but may not be so useful for the predictive problem presented. For example, if misclassification rate is of prime interest, then a classification threshold $t = 0.5$ will be adopted, so that regions of the $\mathbf{x}$ space closest to $P(0 \mid \mathbf{x}) = 0.5$ are of special concern and, as we noted earlier, regions with very different values of $P(0 \mid \mathbf{x})$ are of limited concern. If the model is properly specified, then the contours of $P(0 \mid \mathbf{x})$ are all parallel so that accurate estimation of contours where $P(0 \mid \mathbf{x})$ is far from 0.5 provides information on the $P(0 \mid \mathbf{x}) = 0.5$ contour. However, suppose the model is not properly specified. In particular, suppose that the contours of $P(0 \mid \mathbf{x})$ are straight lines which are not parallel (possible if the data have finite support in $\mathbf{x}$ space—we give an example below). Then adopting a model in which the contours have parallel orientation is likely to lead to a biased orientation for most contours. In particular, the estimate of the $P(0 \mid \mathbf{x}) = 0.5$ contour may well be biased. More generally, the estimate of any contour $P(0 \mid \mathbf{x}) = t$ is likely to be biased. We can describe this situation by saying that the standard logistic discriminant approach dissipates its accuracy of fit across the entire $\mathbf{x}$ space, instead of focusing it where it matters. This can become particularly important when the classes are of very different sizes. For example, in banking, about 0.1% of transactions are thought to be fraudulent; in credit scoring, in some circumstances, less than 5% of cases are regarded as bad risks; and in screening for rare medical conditions sometimes the proportion of people in the disease class can be very small indeed. In such unbalanced situations, the cost associated with misclassifying the smaller class is generally much higher than the reverse, so that the classification threshold t will normally be far from 0.5 (or else one will often end up assigning everyone to the larger class, which defeats the aim). This means that the contour of interest for classification purposes is very different from the 0.5 contour, so that aggregating over (differently shaped and oriented) contours from elsewhere in the $\mathbf{x}$ space can be very detrimental.

Turning to machine learning, an early tool for supervised classification was the perceptron. This (also) adopted a linear form $g(\mathbf{x}) = \beta'\mathbf{x}$ for the decision surface. However, the key difference between this model and logistic discriminant analysis lies in the criterion optimized. In particular, perceptron algorithms focus on misclassified design set points, not on general quality of model fit. Various estimation procedures have been developed. If, for the ith design set vector $\mathbf{x}_i$, $i = 1, \ldots, n$, we define vectors $\mathbf{y}_i = (1 - 2c_i)\mathbf{x}_i$ (where $c_i = 0, 1$ is the true class of the ith object in the design set), then a good solution will make $\beta'\mathbf{y}_i$ positive for as many design set samples as possible. Perceptron algorithms then seek the β vector which minimizes

$$C_1 = \sum_{y_i \in Y} (-\beta'\mathbf{y}_i),$$

where Y is the set of design set points which are misclassified. This can be minimized by steepest descent, with updating step $\beta_{k+1} = \beta_k + \rho_k \sum_{y_i \in Y} \mathbf{y}_i$. If the design set classes are linearly separable, this procedure is guaranteed to converge to a solution. More generally, if ρ_k decreases with each step the effect of misclassified points diminishes over time, while still guaranteeing finding a solution if the classes are linearly separable. More sophisticated versions introduce a *margin*, requiring not only that $\beta'\mathbf{y}_i > 0$ for all design set points if possible (i.e., that they are correctly classified), but that $\beta'\mathbf{y}_i > b$, and including in the set Y any points which do not satisfy this. This modification forces the decision surface away from design set points and so improves generalization ability. From a statistical perspective, however, it is rather ad hoc.

The criterion underlying the perceptron approach is a *local* criterion in the sense that, if the decision surface is altered slightly, only those design set points close to the decision surface can contribute to the change in the criterion value. That is, only such close points can have their predicted classification changed by such a small adjustment. The predicted classification of points far from the decision surface will remain unaltered by a small adjustment, so their contribution to the criterion value (which will be either 0 or $1/n$) will remain unchanged. This is unlike the criterion of likelihood, where an adjustment to the model changes the contributions made by each of the design set points to the value of the likelihood.

4. LOCAL LOGISTIC DISCRIMINATION

This section presents a modification to the standard statistical approach of logistic discrimination which properly acknowledges the fact that the prediction phase does influence our choice of model. We begin by defining a measure of goodness of fit between the observed data and the postulated model form which places most emphasis in those regions where accuracy most matters. In particular, in the context of supervised classification problems, we wish to make sure that our models are most accurate in the vicinity of the true decision surface. The parameters in a logistic discriminant model are traditionally estimated by maximizing the likelihood

$$L = \prod_{i=1}^{n} \hat{P}(c_i \mid \mathbf{x}_i), \tag{1}$$

where c_i is again the true class of the ith object in the design set, and $\hat{P}(c_i \mid \mathbf{x}_i)$ is the value of $P(c_i \mid \mathbf{x}_i)$ estimated from the model. Now, if we knew where the true decision surface was, then a natural way to try to make our models accurate in its vicinity would be to weight the points close to the surface more heavily than those further away. For example, we could use a modified likelihood function

$$L_w = \prod_{i=1}^{n} \hat{P}\left(c_i \mid \mathbf{x}_i\right)^{w_i} \tag{2}$$

with weights w_i decreasing with increasing distance from the decision surface. If the decision surface is that set of points in $\mathbf{x}$ space for which $P(0 \mid \mathbf{x}) = t$, then a natural metric for measuring the "distance from the decision surface" for any point $\mathbf{x}$ would be the difference between t and $P(0 \mid \mathbf{x})$. The difficulty is, of course, that we do not know $P(0 \mid \mathbf{x})$ so we cannot measure the difference between t and $P(0 \mid \mathbf{x})$. This suggests the use of an iterative procedure, in which distance is measured using an initial estimate of $P(0 \mid \mathbf{x})$, $\hat{P}_0(0 \mid \mathbf{x})$, which can be compared with t to yield weights in L_w. This L_w will yield new estimates of $P(0 \mid \mathbf{x})$, $\hat{P}_1(0 \mid \mathbf{x})$, and so on. The initial estimate could be based on any classification method: we experimented with both standard unweighted logistic discrimination and with nearest neighbor methods, using the former for the analyses described below.

In adopting such an iterative procedure, it is still necessary to decide how rapidly the weight should decay with distance between t and $\hat{P}_s(0 \mid \mathbf{x})$. We have experimented with four approaches. In each case, (because we had large datasets available) we used the standard "design set, validation set, test set" paradigm, but any other methods such as leave-one-out, cross-validation, or bootstrap methods could be used.

Method 1. This method uses only one iterative step. Define the weight function as

$$w_i(t) = \exp\left\{-\lambda\left(\hat{P}(0 \mid \mathbf{x}_i) - t\right)^2\right\} \tag{3}$$

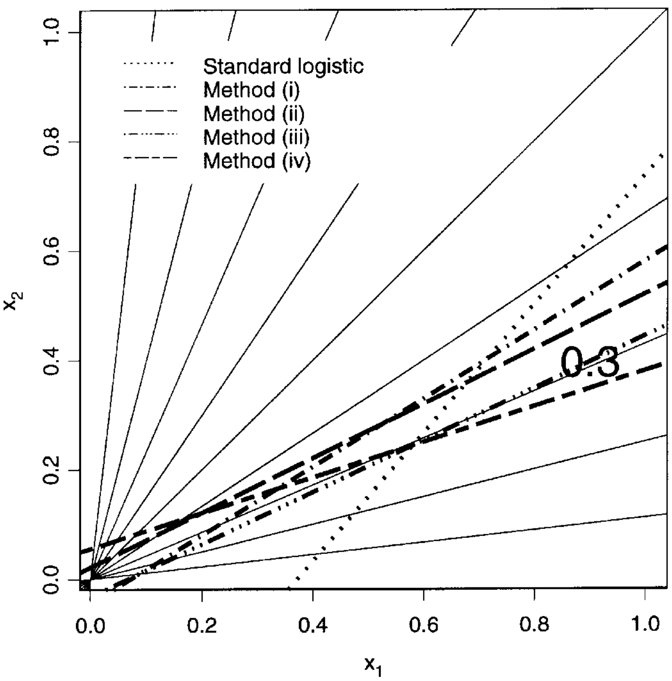

Figure 1. Estimates of the $P(0 \mid \mathbf{x}) = 0.3$ contour by standard logistic discrimination and weighted logistic discrimination.

with $\lambda \geq 0$ a tuning parameter. When $\lambda = 0$ the method reduces to standard logistic discrimination, weighting all design points equally. When $\lambda > 0$ more weight is placed on points near to t. Begin by obtaining initial estimates of $P(0 \mid \mathbf{x})$ using standard logistic discrimination based on the entire design set. From these estimates, find the value of λ that minimizes the overall cost-weighted misclassification loss on the validation set. Finally, using the test set, evaluate the expected future cost-weighted misclassification loss for the λ which yields the minimum. Other weight functions could be used, but we suspect that the resulting performance will be more sensitive to the choice of "bandwidth" λ than to the shape of w.

Method 2. This method also uses just one iterative step, also beginning with standard logistic regression to give initial estimates, $\hat{P}_0(0 \mid \mathbf{x})$, of $P(0 \mid \mathbf{x})$. However, the weight function $w_i(t)$ is now defined as taking the value 1 if the point $\mathbf{x}_i$ is among those m points for which $(\hat{P}_0(0 \mid \mathbf{x}) - t)^2$ is smallest and taking the value 0 otherwise. The choice of m is made using the validation set, and future cost-weighted misclassification loss calculated for this m from the test set.

Method 3. This uses the same weight function as Method 1, but proceeds through multiple steps. Beginning with $\lambda = 0$ (standard logistic discrimination), λ is gradually incremented, each time calculating the cost-weighted loss on the validation set. We chose to stop when the validation set loss had remained constant for 50 values of λ, and took the smallest of these 50 values of λ as our estimate, but other methods could be used. Again we estimated future performance using the test set.

Method 4. This uses the same weight function as Method 2, but proceeds through multiple steps, as in Method 3. The method begins with $m = n$, so that all of the data points are used and the initial model is again the standard logistic model. Then m is gradually reduced (either one point at a time, or in larger groups if the dataset is large). For each value of m an estimate of the cost-weighted loss is obtained from the validation set, stopping when no change has been observed over 50 iterations. Future performance is then estimated from the test set.

For comparison purposes, we have also included standard ("global") logistic discrimination in the example described below.

5. EXAMPLES

5.1 Simulated Data: Logistic Regression

A simple illustration of the situation we have been describing is given in Figure 1, which shows, in the unbroken lines, contours of $P(0 \mid \mathbf{x})$ for a two-dimensional $\mathbf{x}$ space. These contours are linear, but not parallel. Singularity at the lower left corner is avoided by taking $h(\mathbf{x})$, the distribution of $\mathbf{x}$, to have value 0 in the vicinity of $\mathbf{x} = \mathbf{0}$. Logistic discrimination applied to data generated with these contours will yield estimated contours with slope which depends on the distribution $h(\mathbf{x})$. Regions where $h(\mathbf{x})$ is most dense will have most influence on the estimated orientation of the contours. To illustrate, we took a sample of 4,000 points, uniformly distributed across the unit square apart from a small region near $\mathbf{x} = \mathbf{0}$, assigning them label 0 with probability $x_2/(x_1 + x_2)$ and label 1 otherwise, so that they were drawn from distributions with the indicated contours of $P(0 \mid \mathbf{x})$. The symmetry of this arrangement means that the contours estimated by logistic regression will (apart from sampling fluctuation) be parallel to the 0.5 contour—the diagonal of the square. This implies that contours farther away will be poorly estimated. We took the 0.3 contour as an example, indicated by the label in the figure. The estimate of this contour obtained by standard logistic discrimination is shown by a dotted line—it is seen to be almost parallel to the 0.5 contour. The contours estimated by the procedure described above are shown by various style of broken lines. They all show a much better fit to the true contour.

5.2 Earnings Data

These data were obtained from the UCI Machine Learning repository (Merz and Murphy 1996). They consist of 30,162 observations, on 15 variables. The task is the one of predicting whether a given adult earns more than $50,000 per year, based on information such as age, race, sex, marital-status, education and bank account state. It is not possible to present contour plots for this dataset so, instead, in Figure 2, we show (the vertical axis) the differences in overall cost-weighted losses between the standard logistic discrimination and local logistic discrimination methods as the relative size of the costs k_0 and k_1 varies. In particular, as described in Section 3, with these misclassification costs, the overall loss due to misclassifying future cases will be minimized if points are classified into class 0 when $P(0 \mid \mathbf{x}) > k_1/(k_0 + k_1)$ and into class 1 otherwise. The ratio $t = k_1/(k_0 + k_1)$ thus defines the contour of interest, and values of this ratio from 0.1 to 0.9 in steps of 0.1 are plotted as the horizontal axis in Figure 2. In this figure, a value above 0 on the vertical axis means that the overall loss of the standard logistic approach exceeds that of the local logistic approach. We see that

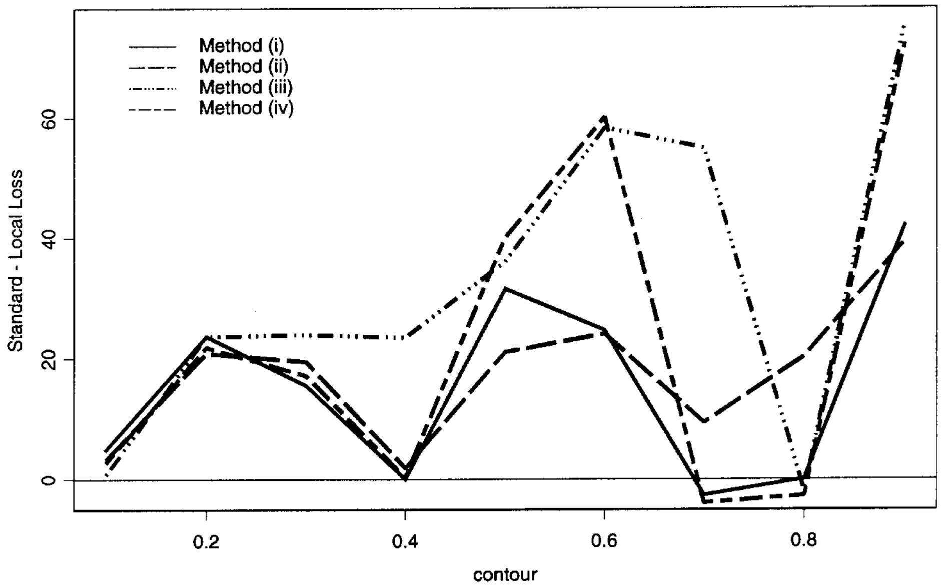

Figure 2. Loss due to standard model minus loss due to local models, as the relative misclassifbation costs vary: earnings data.

the local model is almost always superior to the standard model, and, when it is not superior it is only just inferior.

5.3 Unsecured Personal Loan Data

These data were supplied by a major UK bank. They describe unsecured personal loans for 21,618 customers who applied for a two-year loan between January 1st, 1995, and December 30th, 1996. Customers were assigned to one of two classes according to whether or not they missed payments in more than three months. Sixteen variables from the original application for the loan were used to build the models. These include age, time at current address, time with bank, time with current employer, loan amount, occupation, previous credit reference agency searches, credit protection, cheque guarantee card, and loan purpose.

In problems such as this, models linear in the application variables are strongly preferred because of their interpretative simplicity. There is often a legal requirement that a simple explanation for a rejection should be available to customers who are not granted a loan, and this has typically been in terms of the weights accorded to different variables. Logistic discrimination is by far the most popular method, and is very widely used. However, this merely means that a linear decision surface is required. Logistic discrimination also enforces the condition that the estimated contours are parallel (that all of the decision surfaces in which we might be interested are oriented in the same direction). There is no reason for this assumption, and it seems unlikely to hold in practice. The local model allows it to be relaxed.

The results of fitting the models to these data are shown in Figure 3, using the same style of plot as Figure 2. Once again, it can be seen that the local models almost always outperform standard logistic discrimination. Note also that, with the class definitions used above, 11% of the customers lie in one class and 89% in the other. It would be unusual in such cases, where the classes'

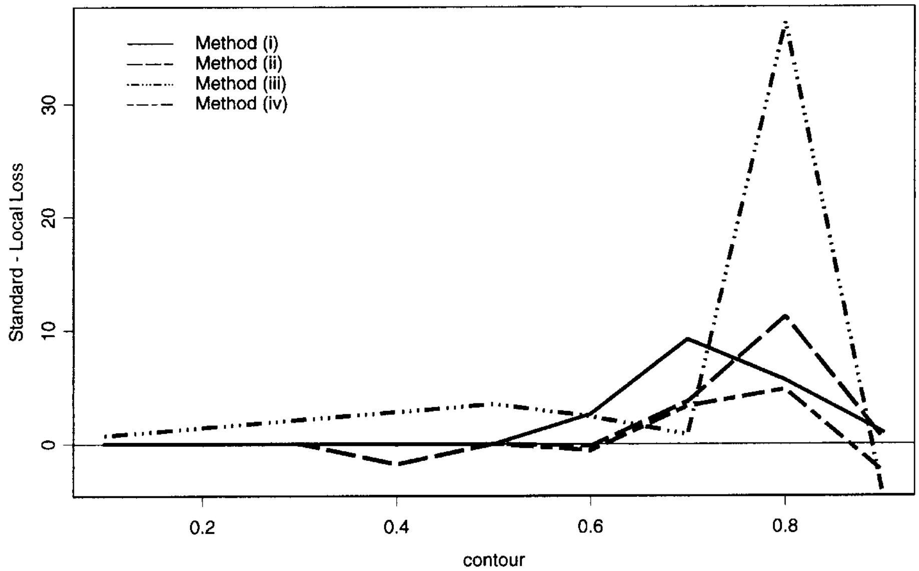

Figure 3. Loss due to standard model minus loss due to local models, as the relative misclassifbation costs vary: loan data.

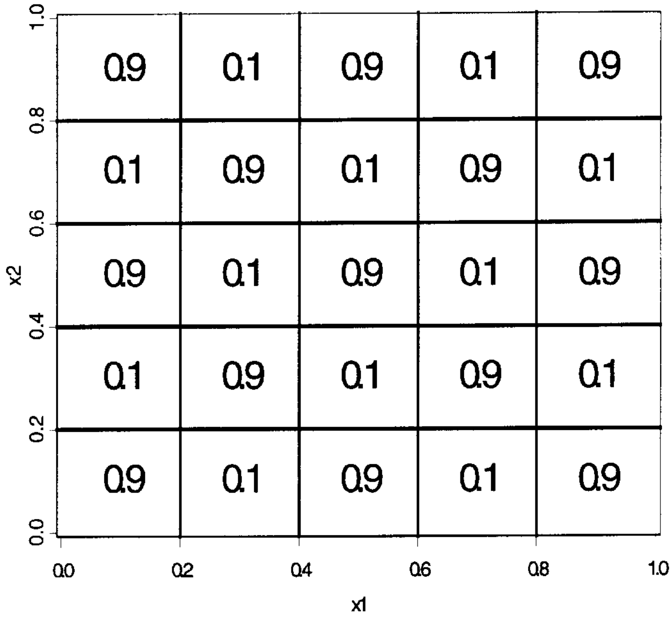

*Figure 4. Values of P(0 | **x**) in a bivariate example.*

sizes are very unbalanced, to use equal misclassification costs (i.e., a threshold $t = 0.5$). Such a choice could easily lead to all customers being classified into the larger class. Usually a value of t is chosen which corresponds to weighting misclassified customers from the smaller class more heavily than misclassified customers from the larger class. That is, the contour which is of interest is far from the $P(0 \mid \mathrm{x}) = 0.5$ contour. In this particular example, large values of t would be appropriate, and we see that for such values the local method is substantially better than the global method.

5.4 Simulated Data: Nearest Neighbor

Although we have chosen to illustrate the ideas using logistic models, the behavior described in this article is not confined to classic parametric models, but can also arise in models traditionally regarded as nonparametric. For example, nearest neighbor models (e.g., Webb 1999, sec. 3.3) are popular "nonparametric" classification tools. To classify a new point **x**, one examines the classes of the m design set points which are nearest to **x**. The proportion of these points which belong to class 0, m_0/m, serves as an estimate of $P(0 \mid \mathrm{x})$, and is compared with the threshold t. However, it is necessary to choose a value for m. Current strategies for choosing m ignore the value of the threshold t which is to be used in the classification rule, and generally choose a value to apply throughout the data space. (There are more sophisticated methods which allow m to vary over the space, but these are rarely used.) The problem is illustrated by the (very artificial) distribution in Figure 4. This shows a checkerboard pattern in which $P(0 \mid \mathrm{x})$ takes the values 0.9 and 0.1 in alternate squares.

Table 1. Confusion Matrices and Resulting Losses When Different Numbers of Nearest Neighbors are Used With the Contour

	True class 1		*True class 0*
	*(a) True P(0 | **x**)*		
Pred class 1	1743		197
Pred class 0	197		1863
Total loss		197	
	*(b) 3-nearest neighbor estimate of P(0 | **x**)*		
	True class 1		*True class 0*
Pred class 1	1513		474
Pred class 0	427		1586
Total loss		450.5	
	*(c) 51-nearest neighbor estimate of P(0 | **x**)*		
	True class 1		*True class 0*
Pred class 1	776		892
Pred class 0	1164		1168
Total loss		1028	

We consider two extreme situations of possible interest: misclassification costs $k_0 = k_1 = 0.5$, leading to $t = 0.5$, and misclassification costs $k_0 = 0.05$ and $k_1 = 0.95$, leading to $t = 0.95$. In the first case it is important to distinguish between the alternating squares, since minimum loss will be obtained by classifying new points which fall in the squares with $P(0 \mid \mathrm{x}) = 0.9$ as class 0 and points which fall in the other squares as class 1. In this case a fairly small value of m might be expected to do well. In contrast, in the second case, all regions of the large square have true probability less than the threshold. Small values of m are likely to lead to occasional local regions which have estimated probability above t. The larger the value of m in this second case, the less likely this is to occur, and hence the lower the overall loss is likely to be.

To test these qualitative arguments, we generated a test dataset of 4,000 cases, uniformly distributed across the unit square, and assigned each point to class 0 with a probability of 0.1 or 0.9 according to its position in the square. Table 1 shows the results when the costs $k_0 = k_1 = 0.5$ are used, so that $t = 0.5$. Panel (a) shows the number of points predicted to belong to each class when each point is assigned to a class by comparing its known true probability of belonging to class 0 with t. Since the known true probabilities are used, this is an estimate of the Bayes optimal cost. Panel (b) shows the same thing when each point is assigned to a class by comparing the $m = 3$ nearest neighbor estimate of $P(0 \mid \mathrm{x})$ with t. Panel (c) shows the same thing when each point is assigned to a class by comparing the $m = 51$ nearest neighbor estimate of $P(0 \mid \mathrm{x})$ with t. The figure beneath each panel shows the total cost-weighted loss of the misclassifications using each method. We see that, as predicted, in this $t = 0.5$ case, substantially lower loss results when $m = 3$ is used than when $m = 51$ is used.

In contrast, Table 2 shows the results obtained when $k_0 = 0.05$ and $k_1 = 0.95$ (so that $t = 0.95$). Now, also as predicted, lower loss results when the larger value of m is used. (In fact, the loss is the same as that produced by the Bayes solution.)

6. DISCUSSION

If parameters of a model are estimated by some global measure of goodness of fit such as likelihood, then the problems described in this article can arise. If the family of models is misspecified or "incorrect," in that it does not include the form of the "true" data generating process, a situation which we might typically assume to be the case, then the global measure of goodness of fit is, in effect, averaging the fit over the entire data space. If, however, the application means that interest lies in a local region of the space, then models which are superior can be constructed by adjusting the goodness of fit measure to concentrate in the critical regions. This argument is true for both frequentist

Table 2. Confusion Matrices and Resulting Losses When Different Numbers of Nearest Neighbors are Used With the Contour

(a) True $P(0 \mid \mathbf{x})$

	True class 1	True class 0
Pred class 1	1940	2060
Pred class 0	0	0
Total loss	103	

(b) 3-nearest neighbor estimate of $P(0 \mid \mathbf{x})$

	True class 1	True class 0
Pred class 1	1814	1173
Pred class 0	126	197
Total loss	178.35	

(c) 51-nearest neighbor estimate of $P(0 \mid \mathbf{x})$

	True class 1	True class 0
Pred class 1	1940	2060
Pred class 0	0	0
Total loss	103	

and Bayesian modeling strategies, since both use the likelihood, which is an aggregate measure of goodness of fit.

Note that often a misspecified model is deliberately adopted because it has other merits in the context of the problem. For example, in many problems simplicity and interpretability are advantages which can outweigh accuracy. Thus, in many credit scoring problems in retail banking (e.g., Thomas, Edelman, and Crook 2002, p. 88) it is important that the classification rule should be easily interpretable because it is a common (often legal) requirement that one should be able to explain the basis for rejecting an applicant for a loan (Hand 2001; Hand and Adams 2000). Because of this, linear functions of applicant characteristics are the most popular kind of model. This means that modern sophisticated and highly nonlinear tools such as neural networks, multivariate adaptive regression splines, and support vector machines are not suitable for this application. In fact, logistic discriminant analysis, based on a linear combination of the applicant characteristics, is particularly widely used. However, as we have noted above, logistic discriminant analysis assumes that the probability contours $P(0 \mid \mathbf{x})$ are parallel. The model thus provides a common orientation of these estimated contours which is an average, in some sense, over all of the contours. This is perhaps unwise since we are in fact interested in only one of the contours. Local logistic discriminant analysis yields an estimate (still linear, as required) of that single contour of interest.

In general, in any domain in which aspects beyond straightforward generalizability are also important (interpretability, in the credit scoring example), then one may deliberately choose a model which cannot fit all the possible idiosyncrasies of the unknown underlying distribution. In such cases, if only particular parts of the model are crucial, and this will depend on the reasons for fitting the model, then one may do better by focusing effort on only part of the data space. Of course, in order to take advantage of this, one needs to be confident that the region focused on is really the region of concern. In many problems, there is ambiguity or uncertainty about exactly which region is of interest. For example, in supervised classification problems it is often very difficult to give precise values for the relative misclassification costs (see, e.g., Kelly, Hand, and Adams 1998; Adams and Hand 1999, 2000). In such problems it would be unwise to concentrate effort on too narrow a part of the space.

An idea related to that discussed above is that of boosting. This was originally developed by the computational learning theory community (e.g., Schapire 1990; Freund 1995; Freund and Schapire 1997) but has since been properly formalized by statisticians (Friedman, Hastie, and Tibshirani 2000; Hastie, Tibshirani, and Friedman 2001). Although boosting also focuses attention in the critical region, it does this by combining multiple classification rules, to form an additive model, rather than simply choosing that single rule which is best suited for the contour in question. The component rules are combined in a weighted sum, and since each component is generally a nonlinear transformation of the predictor space $\mathbf{x}$ (each is a map to $\{0, 1\}$, the set of class labels) the resulting weighted sum is a nonlinear transformation of $\mathbf{x}$.

Tibshirani and Hastie (1987) also discussed local likelihood estimation, but their aim and model differ in two main ways from ours, both hinging on how one defines proximity to the contour of interest. First, their aim was to construct models which provide good fits to the data independently of any possible restrictions on model form. In particular, whereas, for reasons of interpretability, in most of the illustrations above we have constrained our models to be simple linear functions of $\mathbf{x}$, they fit flexible nonlinear models which may be markedly nonlinear. Indeed, the possibility of nonlinearity is one of the motivations underlying their work. Second, the local neighborhood they adopted as being relevant to the estimate at a particular value of $\mathbf{x}$ is defined in terms of an interval in $\mathbf{x}$. For example, if an estimate is required at a point $\mathbf{x}^*$, they base their local model on the nearest few design set points to $\mathbf{x}^*$. In contrast, our "local neighborhood" is defined in terms of distance from the unknown contour $P(0 \mid \mathbf{x})$. Yet others who have also explored local likelihood methods (e.g., Copas 1995; Loader 1996; Hjort and Jones 1996; Hall and Tao 2002) did not restrict their model forms, and so used definitions of neighborhood which are different from ours.

[Received January 2002. Revised November 2002.]

REFERENCES

Adams, N. M., and Hand, D. J. (1999), "Comparing Classifiers When the Misallocation Costs are Uncertain," *Pattern Recognition*, 32, 1139–1147.

——— (2000), "Improving the Practice of Classifier Performance Assessment," *Neural Computation*, 12, 305–311.

Copas, J. B. (1995), "Local Likelihood Based on Kernel Censoring," *Journal of the Royal Statistical Society*, Series B, 57, 221–235.

Cox, D. R. (1958), "Some Problems Connected With Statistical Inference," *Annals of Mathematical Statistics*, 29, 357–372.

Freund, Y. (1995), "Boosting a Weak Learning Algorithm by Majority," *Information and Computation*, 121, 256–285.

Freund, Y., and Schapire, R. E. (1997), "A Decision-Theoretic Generalization of Online Learning and an Application to Boosting," *Journal of Computer and System Sciences*, 55, 119–139.

Friedman, J. H. (1997), "On Bias, Variance, 0/1-loss, and the Curse of Dimensionality," *Data Mining and Knowledge Discovery*, 1, 55–77.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2000), "Additive Logistic Regression: A Statistical Review of Boosting," *The Annals of Statistics*, 28, 337–374.

Hall, P., and Tao, T. (2002), "Relative Efficiencies of Kernel and Local Likelihood Density Estimators," *Journal of the Royal Statistical Society*, Series B, 64, 537–547.

Hand, D. J. (1981), *Discrimination and Classification*, Chichester: Wiley.

——— (1997), *Construction and Assessment of Classification Rules*, Chichester: Wiley.

——— (2001), "Modelling Consumer Credit Risk," *IMA Journal of Management Mathematics*, 12, 139–155.

Hand, D. J., and Adams, N. M. (2000), "Defining Attributes for Scorecard Construction," *Journal of Applied Statistics*, 27, 527–540.

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

Hjort, N. L., and Jones, M. C. (1996), "Locally Parametric Nonparametric Density Estimation," *The Annals of Statistics*, 24, 1619–1647.

Kelly, M. G., Hand, D. J., and Adams, N. M. (1998), "Defining the Goals to Optimise Data Mining Performance," in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, eds. R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro, Menlo Park: AAAI Press, pp. 234–238.

Lindley, D. V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2: Inference*, Cambridge: Cambridge University Press.

Loader, C. R. (1996), "Local Likelihood Density Estimation," *The Annals of Statistics*, 24, 1602–1618.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

Merz, C. J., and Murphy, P. M. (1996), "UCI Repository of Machine Learning Database," available at http://www.ic.uci.edu/mlearn/MLRepository.html.

Nelder, J. A. (1994), "The Statistics of Linear Models: Back to Basics," *Statistics and Computing*, 4, 221–234.

Ripley B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press.

Schapire R. E. (1990), "The Strength of Weak Learnability," *Machine Learning*, 5, 197–227.

Thomas, L. C., Edelman, D. B., and Crook, J. N. (2002), *Credit Scoring and its Applications*, Philadelphia: SIAM.

Tibshirani, R., and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567.

Webb, A. R. (1999), *Statistical Pattern Recognition*, London: Arnold.