Classification and *Reductio-ad-Absurdum* Optimality Proofs

Haim Schweitzer (haim@utdallas.edu)
The University of Texas at Dallas

Abstract

Proofs for the optimality of classification in real-world machine learning situations are constructed. The validity of each proof requires reasoning about the probability of certain subsets of feature vectors. It is shown that linear discriminants classify by making the least demanding assumptions on the values of these probabilities. This enables measuring the confidence of classification by linear discriminants. We demonstrate experimentally that when linear discriminants make decisions with high confidence, their performance on real-world data improves significantly, to the point where they beat the best known nonlinear techniques on large portions of the data.

Introduction

In the standard machine learning setup considered here (e.g., (Vapnik 1995; Breiman *et al.* 1984)), a training set T that contains data with known classification is available. The goal is to infer the classification of an instance x that may not be in T. The basic assumption that relates x to T is that there exists a fixed (but usually unknown) probability distribution, from which the data in T as well as x were obtained by random sampling. We limit the discussion to the two-category case where $\{A, B\}$ denote the two possible categories of x.

Typical analysis of classifiers performance measures their average (expected) accuracy. It is usually compared to the performance of the optimal classification rule that minimizes the expected probability of misclassification. This optimal rule is called *the Bayesian classifier*, or *the maximum likelihood classifier* (Duda & Hart 1973; Devijver & Kittler 1982). Without knowledge of the distribution this optimal Bayesian classifier cannot be computed explicitly, except for some special cases (Duda & Hart 1973; Langley, Iba, & Thompson 1992; Haussler, Kearns, & Schapire 1994). Still, it can be expected that the classification obtained by a "good" classifier is identical to the classification obtained by the optimal Bayesian classifier for "many" x values.

This paper investigates assumptions that allow a formal (deductive) proof that the classification of x is optimal. Observe that a formal proof with no assumptions can be given if and only if the optimal classification can be computed. We show that it is possible to give proofs of optimality using assumptions that are milder than a complete knowledge of the distribution. A special case is described, with natural ranking of the assumptions indicating how strong the assumption is. We propose to use this ranking as the degree of confidence in the optimality of the classification. It is shown that classification by linear discriminants is equivalent to choosing the category that can be proved optimal under the least demanding assumption according to this ranking.

In previous work, the design of optimal Bayesian algorithms under various assumptions was investigated. One of the most straight-forward is the Naive-Bayesian (Langley, Iba, & Thompson 1992; Kohavi 1995b), which assumes conditional independence. Other approaches (e.g (Haussler, Kearns, & Schapire 1994)) assume that the data comes from a restricted set of hypotheses. In all these cases the goal is to design an algorithm that is likely to be optimal for any given x when global assumptions hold. The approach that we take here is different. We attempt to determine the optimality for each individual value of x. The motivation is to enable reliable combination of classifiers. For each value of x one may choose a different classifier that is likely to be optimal. That particular classifier need not be optimal for other values of x.

The probabilistic setup

Let O be a set of items. We consider experiments (sampling) that produce elements $o \in O$ according to a fixed probability measure P. A feature of an object is a function $x(o)$ that returns a real value, so that $x(o)$ is a random variable. The vector $x(o) = (x_1(o), \ldots, x_n(o))$ is *the feature vector* of o. We investigate cases where the objects may belong to exactly one of the two categories A and B. The random variable $\alpha(o)$ indicates

Research supported in part by NSF grant IRI-9309135.

the category of the object o, and can take two distinct values: $v_A = -1$ if the object belongs to A, and $v_B = 1$ if the object belongs to B. The expectation of a random variable $x(o)$ is defined as usual to be: $E\{x(o)\} = \int_{o \in O} x(o) d\mathsf{P}\{o\}$.

Let " ' " denote transpose. We use of the following statistics from each individual category:

Statistics from A	Statistics from B
$u_A = E\{x(o) \mid o \in A\}$	$u_B = E\{x(o) \mid o \in B\}$
$R_A = E\{x(o)x(o)' \mid o \in A\}$	$R_B = E\{x(o)x(o)' \mid o \in B\}$
$C_A = R_A - u_A u_A'$	$C_B = R_B - u_B u_B'$
$p_A = \mathsf{P}\{o \in A\}$	$p_B = \mathsf{P}\{o \in B\}$

(1)

The following statistics can be estimated from the joint distribution, or from the priors and the statistics of the individual categories:

$$\begin{aligned} u &= E\{x(o)\} = p_A u_A + p_B u_B \\ R &= E\{x(o)x(o)'\} = p_A R_A + p_B R_B \\ C &= R - uu' \\ k &= E\{\alpha(o)\} = p_A v_A + p_B v_B = p_B - p_A \\ d &= E\{\alpha(o)x(o)\} = p_B u_B - p_A u_A \end{aligned} \tag{2}$$

Proving optimality of classification

The principle of proof by *reductio-ad-absurdum* (e.g., (Copi 1979)) can be put as follows:

> To prove that $p \Rightarrow q$ it is sufficient to derive an "absurd" from p and $\neg q$.

We show that a variation of this principle can be used to prove the optimality of classification in the two-category case. The key observation is identifying the "absurd", which we show to be related to the distinction between optimal classification and correct classification.

Let $\overline{o}$ be an object from A, so that $\alpha(\overline{o}) = v_A$, and A is the correct classification of $\overline{o}$. This, however, does not imply that the *optimal* classification of $\overline{o}$ according to $\overline{x} = x(\overline{o})$ is A. In order for this classification to be optimal (i.e., to minimize the expected probability of error), we must have:

$$\mathsf{P}\{\alpha(o) = v_A \mid x(o) = \overline{x}\} \geq \mathsf{P}\{\alpha(o) = v_B \mid x(o) = \overline{x}\}.$$

Intuitively, we expect the optimal classification and the correct classification to agree on most of the data. Indeed, the optimality implies that there must be an agreement with probability of at least $\frac{1}{2}$ on any portion of the data. (Otherwise always classifying the opposite of the optimal would yield an improvement in accuracy over the optimal.) The following lemma states this fact for particular subsets that we call consistent.

Definition: A subset N of feature vectors is called *consistent* if the optimal classification based on any value $x \in N$ is the same.

Lemma 1: Given $\overline{x} = x(\overline{o})$, let N be a consistent subset that includes $\overline{x}$ and let β be either v_A or v_B. Then the optimal classification of $\overline{o}$ based on $\overline{x}$ is β if and only if:

$$\mathsf{P}\{x(o) \in N \ , \ \alpha(o) = \beta\} \geq \frac{1}{2}\mathsf{P}\{x(o) \in N\}$$

Proof sketch: The optimal classification based on $\overline{x}$ is β if and only if the optimal classification of any $x(o) \in N$ is β. The lemma follows from the fact that the optimal classification must agree with the correct classification with probability of at least $\frac{1}{2}$ on N. □

A direct corollary of Lemma 1 is the following theorem.

Theorem 1: Given $\overline{x} = x(\overline{o})$, let N be a consistent subset that includes $\overline{x}$. Let Q be a real number satisfying:

$$Q \geq 2\mathsf{P}\{x(o) \in N \ , \ \alpha(o) \neq \beta\}$$

The following condition guarantees that the optimal classification of $\overline{o}$ is β:

$$\mathsf{P}\{x(o) \in N\} > Q$$

Proof: These assumptions imply that the condition in Lemma 1 does not hold under the premise that the optimal classification is not β. □

Theorem 1 is the version of *reductio-ad-absurdum* that we use to prove optimality of classification. To prove that the optimal classification based on $\overline{x}$ is β, assume that it is not β, and show a subset consistent with $\overline{x}$ where the agreement between optimal classification and correct classification has probability less than $\frac{1}{2}$.

Large consistent subsets

When the distribution is not known, the probabilities of consistent subsets cannot be computed, so that Theorem 1 cannot be applied. In this section we show that a bound Q, as required by Theorem 1, can sometimes be computed. In these situations the *assumption* that the probability $\mathsf{P}\{x(o) \in N\}$ is larger than Q is sufficient to establish the optimality.

With respect to a feature vector $\overline{x}$, any vector w splits the entire space of feature vectors into the following two half-spaces: $\{x \ ; \ w'(x - \overline{x}) \geq 0\}$, and $\{x \ ; \ w'(x - \overline{x}) \leq 0\}$. Each half-space contains two consistent subsets: the values of x for which the optimal classification is A, and the values of x for which the optimal classification is B. Thus, the vector w partitions the set of feature vectors into four consistent subsets.

Definition: The consistent subset $N(\overline{x}, w, \beta)$ is the set of feature vectors x that should be optimally classified as β, which satisfy: $w'(x - \overline{x}) \geq 0$.

Notice that the union of the following four consistent subsets covers the entire space: $N(\overline{x}, w, v_A)$, $N(\overline{x}, w, v_B)$, $N(\overline{x}, -w, v_A)$, $N(\overline{x}, -w, v_B)$. Therefore, by symmetry a rough estimate to the probability of a consistent subset of this type is $\frac{1}{4}$. We proceed to show how to compute the bound Q for these subsets.

Lemma 2: Let $y(o)$ be a random variable, computed as the following linear function of $x(o)$ and $\alpha(o)$:

$$y(o) = w'x(o) + m\alpha(o) + s,$$

where w is an arbitrary vector, and m, s are arbitrary scalars. Set $V = E\{y^2(o)\}$. For a feature vector $\overline{x}$

and $\beta \in \{v_A, v_B\}$ compute: $\overline{y} = w'\overline{x} + m\beta + s$. Then regardless of the probability measure:

$$\overline{y} > 0 \Rightarrow \mathsf{P}\{x(o) \in N(\overline{x}, w, \beta)\ ,\ \alpha(o) = \beta\} \leq \tfrac{V}{\overline{y}^2}$$
$$\overline{y} < 0 \Rightarrow \mathsf{P}\{x(o) \in N(\overline{x}, -w, \beta)\ ,\ \alpha(o) = \beta\} \leq \tfrac{V}{\overline{y}^2}$$

Proof: When $\overline{y} > 0$, the condition $y(o) \geq \overline{y}$ implies $|y(o)| \geq |\overline{y}|$. Therefore,

$$\begin{aligned}&\mathsf{P}\{x(o) \in N(\overline{x}, w, \beta)\ ,\ \alpha(o) = \beta\}\\ &\leq \mathsf{P}\{w'(x(o) - \overline{x}) \geq 0\ ,\ \alpha(o) = \beta\}\\ &\leq \mathsf{P}\{y(o) \geq \overline{y}\} \leq \mathsf{P}\{|y(o)| \geq |\overline{y}|\} \leq \tfrac{V}{\overline{y}^2}\end{aligned}$$

The last inequality follows from the Chebyshev inequality (see, e.g., (Feller 1970), Page 233). Similarly, when $\overline{y} < 0$ the condition $y(o) \leq \overline{y}$ implies $|y(o)| \geq |\overline{y}|$. Therefore,

$$\begin{aligned}\mathsf{P}\{x(o) \in N(\overline{x}, -w, \beta)\ ,\ \alpha(o) = \beta\} &\leq \mathsf{P}\{y(o) \leq \overline{y}\}\\ &\leq \mathsf{P}\{|y(o)| \geq |\overline{y}|\} \leq \tfrac{V}{\overline{y}^2}\ \square\end{aligned}$$

Observe that V can be explicitly computed in terms of the measurable statistics in (2):

$$V = E\{y^2(o)\} = w'Rw + s^2 + 2sw'u + m^2 + 2m(w'd + sk)$$

Combining Lemma 2 with Theorem 1 gives the following procedure for proving that β is the optimal classification based on $\overline{x}$:

1. Choose values for w, m, s.
2. Compute V and $\overline{y}$ as required by Lemma 2, and set $Q = 2V/\overline{y}^2$.
3. The optimality now follows from the *assumption* that $\mathsf{P}\{x(o) \in N(\overline{x}, \overline{w}, \beta)\} > Q$, where $\overline{w}$ is either w or $-w$, according to the sign of $\overline{y}$.

Optimality proofs and linear classifiers

From the results of the previous section it follows that it is possible to give a proof that A is the optimal classification based on $\overline{x}$, and another proof that B is the optimal classification based on the same $\overline{x}$. Each proof would rely on a different assumption. Since without knowledge of the distribution we are unable to distinguish between the consistent subsets $N(x, w, \beta)$, we propose to use the value of the bound Q as a measure to the "strength" of the assumption. The smaller this value is, the more likely it is that the assumption needed for proving optimality holds. This suggests that the values of s, w, m in Lemma 2 should be chosen to minimize Q. We describe only the minimization with respect to m, and show that the classification reduces to the classic technique of linear discriminants.

The function $f(x) = w'x + s$ is called a linear discriminant (see (Duda & Hart 1973; Devijver & Kittler 1982)) if the classification of $\overline{o}$ based on $\overline{x} = x(o)$ is determined from the sign of $f(\overline{x})$. Simple algebra shows that the variables $\overline{y}$ and V in Lemma 2 can be expressed as:

$$\overline{y} = f(\overline{x}) + m\beta, \quad V = F + m^2 + 2mg,$$

where the constants F, g can be expressed in terms of the statistics in (1), (2) as:

$$\begin{aligned}F &= E\{f^2(x)\} = w'Rw + 2sw'u + s^2,\\ g &= w'd + sk = p_B(w'u_B + s) - p_A(w'u_A + s)\end{aligned} \tag{3}$$

The minimization with respect to m leads to the following result:

Theorem 2: Given $\overline{x} = x(\overline{o})$, compute $f(\overline{x}) = w'\overline{x} + s$ with an arbitrary vector w and an arbitrary scalar s. The following assumption guarantees that the optimal classification of $\overline{o}$ is A:

$$Q_B(\overline{x}) \leq \mathsf{P}\{x(o) \in N(\overline{x}, \overline{w}, B)\},$$

and the following assumption guarantees that the optimal classification of $\overline{o}$ is B:

$$Q_A(\overline{x}) \leq \mathsf{P}\{x(o) \in N(\overline{x}, \overline{w}, A)\},$$

with:

$$Q_A = \frac{2}{1 + 1/G_A} \quad \text{where} \quad G_A = \frac{F - g^2}{(f(\overline{x}) + g)^2}$$

$$Q_B = \frac{2}{1 + 1/G_B} \quad \text{where} \quad G_B = \frac{F - g^2}{(f(\overline{x}) - g)^2}$$

The vector $\overline{w}$ is taken as w if $f(\overline{x}) > 0$, and as $-w$ otherwise.

Proof(sketch): The theorem follows from standard technical minimization showing that:

$$\begin{aligned}Q_A &= \min_m 2(F + m^2 + 2mg)/(f(\overline{x}) - m)^2,\\ Q_B &= \min_m 2(F + m^2 + 2mg)/(f(\overline{x}) + m)^2\end{aligned}$$

and in addition at the minimum the sign of $\overline{y}$ is the same as the sign of $f(\overline{x})$. $\square$

Since the smallest value of Q indicates the most reliable assumption in the *reductio-ad-absurdum* proof, the most likely classification should be determined by choosing *against* the smallest Q. This principle can be summarized as follows:

> **Given $\overline{x} = x(\overline{o})$ compute Q_A, Q_B. If $Q_A < Q_B$, choose B as the most likely classification, ranking the confidence according to how small is Q_A. If $Q_A > Q_B$, choose A as the most likely classification, ranking the confidence according to how small is Q_B.**

We proceed to show that this classification rule reduces to a linear discriminant. When a negative value of $f(x)$ indicates Category A, we can assume without loss of generality that the average value of $f(x)$ on objects from Category A is less than the average value of $f(x)$ on objects from Category B. (Otherwise classify objects as A when $f(x)$ is positive.) This implies that the value of g in (3) is positive.

Corollary (of Theorem 2): If the value of g is positive, the classification obtained by choosing against the smallest among $\{Q_A, Q_B\}$ in Theorem 2 is the same as the classification obtained according to the sign of the linear discriminant $f(x) = w'x + s$.

Proof: $Q_A > Q_B \Leftrightarrow G_A > G_B \Leftrightarrow (f(\overline{x}) + g)^2 < (f(\overline{x}) - g)^2 \Leftrightarrow f(\overline{x}) < 0$ □

This corollary suggests that good choices for w and s can be the values that are used in common linear discriminants. A classic formula for w was proposed in (Fisher 1936). Fisher's formula in terms of the statistics in (1) is:

$$w = (C_A + C_B)^{-1}(u_B - u_A). \quad (4)$$

A modification of Fisher's linear discriminant, sometimes called the Minimum Squared Error (MSE) linear discriminant, is the best linear estimate (in the mean squared error sense) to $\alpha(o)$. The explicit expression in terms of the statistics in (2) is:

$$w = C^{-1}(d - k). \quad (5)$$

These formulas are widely in use, and heavily analyzed. See, e.g., (Duda & Hart 1973; Devijver & Kittler 1982). Once the weight vector w is determined, the standard technique is to compute the value of s so that the projection of the training data by w is optimally separated (Duda & Hart 1973).

Experimental results

It was shown in the previous section that classifying by choosing against the mildest assumption (the smallest value of Q) is identical to the classification obtained by a linear discriminant. In addition, this value of Q tells us how reliable the classification is. In this section we describe experiments on real-world data taken from the UC-Irvine Machine-Learning database. It is shown that, as expected, the accuracy is inversely related to $\min(Q_A, Q_B)$. Classifying only when this value is small (say less than 0.5), improves the accuracy in the classification significantly, to the point where this accuracy beats the best reported results. (Observe, however, that our method does not classify the entire data with this level of accuracy.)

The experiments were conducted as follows. The dataset was divided into a training and a testing portion according to the principle of 10-fold cross-validation (see, e.g., (Kohavi 1995a)). The values of Q_A, Q_B as given in Theorem 2 were computed for each test item, with w, s computed as described below. The accuracy that can be obtained by considering only data with Q values below a fixed threshold was computed, as well as the portion of the data that can be captured with this threshold. This procedure was repeated over 5 runs of 10-fold cross validation. (Thus, the reported results were obtained by repeating a total of 50 training/testing experiments for each dataset.)

Training

The training data is given as a set of m_A feature vectors from Category A and m_B feature vectors from Category B. It is used to compute both the weight vector w and the threshold constant s. The computation of w requires the statistics as given by Equations (1), (2). These statistics were computed by approximating the expectation with averaging. The standard correction for obtaining an unbiased estimate to the covariance matrix (see, e.g., (Duda & Hart 1973)) was used. The specific formulas are:

$$u_A = \frac{1}{m_A}\sum_{x\in A} x \qquad u_B = \frac{1}{m_B}\sum_{x\in B} x$$
$$R_A = \frac{1}{m_A}\sum_{x\in A} xx' \qquad R_B = \frac{1}{m_B}\sum_{x\in B} xx'$$
$$C_A = \frac{m_A}{m_A-1}(R_A - u_A u_A') \qquad C_B = \frac{m_B}{m_B-1}(R_B - u_B u_B')$$

$$u = (m_A u_A + m_B u_B)/(m_A + m_B)$$
$$R = (m_A R_A + m_B R_B)/(m_A + m_B)$$
$$C = \frac{m_A+m_B}{m_A+m_B-1}(R - uu')$$
$$k = (m_B - m_A)/(m_A + m_B)$$
$$d = (m_B u_B - m_A u_A)/(m_A + m_B)$$

The vector w for Fisher's linear discriminant was computed from Equation (4). The MSE weight vector was computed from Equation (5). In each case the value of s was computed by determining the value that best separates the one-dimensional projection of the training data using the following straight-forward approach:

1. The projection $\gamma_i = w'x_i$ is computed for each feature vector in the training set.
2. The samples are sorted in the order of increasing γ.
3. For each value γ_i compute a_i, the number of feature vectors x_j with $\gamma_j \leq \gamma_i$, that are classified as A.
4. The optimal threshold is the value of γ_i for which $2a_i - i$ is maximal. (The value of s is $-\gamma_i$.)

The sorting in Step 2 is a pre-processing for Step 3. To see that the above procedure computes the optimal threshold we observe the following. If γ_i is used as a threshold, the a_i feature vectors with $\gamma_j \leq \gamma_i$ are correctly classified as A. Let $|B|$ be the number of feature vectors that are classified as B. Similar reasoning shows that the number of feature vectors correctly classified as B is $|B| - i + a_i$. Therefore, the total number of correctly classified feature vectors is $2a_i - i + |B|$, and this is maximized when $2a_i - i$ is maximized.

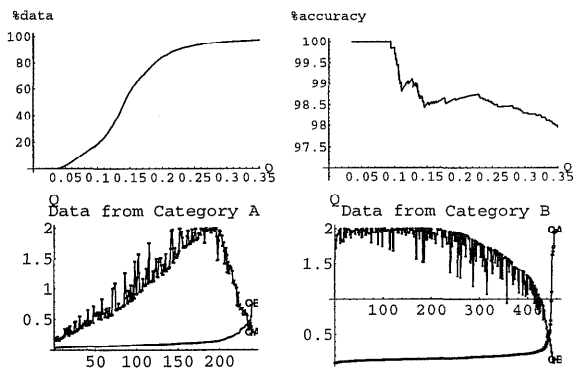
Results

The results of each experiment are described in four plots and two tables. The first plot shows the percentage of feature vectors as a function of a threshold in the value of $Q = \min(Q_A, Q_B)$. The second plot shows percentage accuracy in classifying the feature vectors with the value of Q below the threshold. These numbers were computed from 5 runs of 10-fold cross validation. The third plot shows the values of Q_A, Q_B for each data item from Category A, as computed in the first run of 10-fold Cross-Validation. The data items are arranged in the order of increasing Q_B, and the plotted points are connected with a line. When the classification is correct the Q_B line is below the Q_A line. The fourth plot shows the values of Q_A, Q_B for each data item from Category B, as computed in the

first run of 10-fold Cross-Validation. The data items are arranged in the order of increasing Q_A, and the classification is correct when the Q_A line is below the Q_B line.

Since our experiments show that the MSE outperforms Fisher's linear discriminant, the plots are given only for the MSE.

The Wisconsin breast-cancer dataset: The data contains 699 instances, each defined in terms of 9 numeric attributes. It was originally used in (Wolberg & Mangasarian 1990), and the best reported results achieve accuracy of 97%.



Accuracy with the MSE LDF on the entire data is 96.8%. The largest Q is 0.503. Accuracy and percentage data for 4 thresholds in Q are shown below:

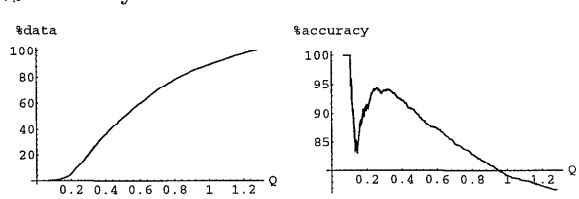
MSE	$Q < .1$	$Q < .25$	$Q < .5$	$Q < 1$
%accuracy	99.8	98.6	96.8	96.8
%data	18.8	93.	100	100

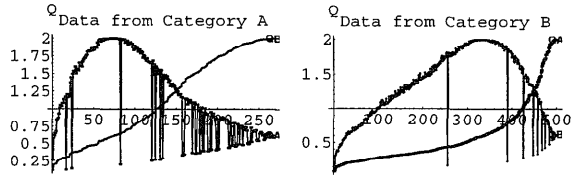
Accuracy with Fisher's LDF on the entire data is 95.7%. The largest Q is 0.791. Accuracy and percentage data for 4 thresholds in Q are shown below:

FISHER	$Q < .1$	$Q < .25$	$Q < .5$	$Q < 1$
%accuracy	99.4	97.6	96.7	95.7
%data	10.4	53.	95	100

The results show that when all the data is considered linear discriminants slightly under-perform the sophisticated classification methods described in (Wolberg & Mangasarian 1990; Breiman 1996). However, the majority of the data satisfies $Q < 0.25$, and for that portion of the data linear discriminants clearly outperform the best reported results.

The Pima Indians Diabetes dataset: This dataset contains 768 instances, each defined in terms of 8 numeric attributes. It was originally used in (Smith *et al.* 1988), and previously reported results achieve around 76% accuracy.





Accuracy with the MSE LDF on the entire data is 76.7%. The largest Q is 1.28. Accuracy and percentage data for 4 thresholds in Q are shown below:

MSE	$Q < .1$	$Q < .25$	$Q < .5$	$Q < 1$
%accuracy	100	94.1	89.4	79.2
%data	0.391	12.3	49	89

Accuracy with Fisher's LDF on the entire data is 72.1%. The largest Q is 1.58. Accuracy and percentage data for 4 thresholds in Q are shown below:

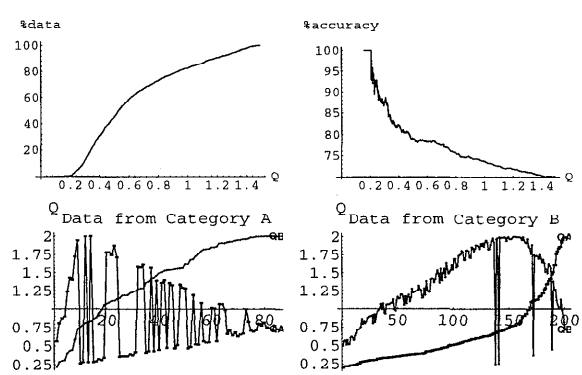
FISHER	$Q < .1$	$Q < .25$	$Q < .5$	$Q < 1$
%accuracy	100	76.5	81.7	77.4
%data	0.026	4.66	34.3	74.7

In this case the values of Q are quite large. Therefore the good performance of the MSE (in comparison to previously reported results) is somewhat surprising. In may indicate that other techniques may be finely tuned to achieve even higher accuracy.

Since there are very few feature vectors satisfying $Q < 0.2$, the behavior of the accuracy graph in this range is very erratic. But it is obvious that linear discriminants easily beat the best reported results for feature vectors satisfying $Q < 0.5$. Notice, however, that this threshold captures less than half of the data.

The Ljubljana breast cancer dataset: This dataset contains 286 instances, each defined in terms of 9 attributes. Previously reported results are in the range of 74-78%.

In our experiments, the non-numeric attributes were converted to numeric attributes in a straightforward way, with the exception of the "breast-quad" attribute. This non-numeric attribute was converted into two numeric attributes, specifying the coordinates of the 2D location. This gives a total of 10 numeric attributes for each instance.



Accuracy with the MSE LDF on the entire data is 70.1%. The largest Q is 1.49. Accuracy and percentage data for 4 thresholds in Q are shown below:

MSE	$Q < .1$	$Q < .25$	$Q < .5$	$Q < 1$
%accuracy	?	91.5	78.9	73.8
%data	0	4.97	45.5	82.8

Accuracy with Fisher's LDF on the entire data is 69.6%. The largest Q is 1.55. Accuracy and percentage data for 4 thresholds in Q are shown below:

FISHER	$Q < .1$	$Q < .25$	$Q < .5$	$Q < 1$
%accuracy	?	89.4	78.8	72.7
%data	0	3.29	48.5	86.4

As in the previous case the values of Q are quite large. Indeed, linear discriminants clearly under-perform the best reported results for this dataset.

Our experiments show that there are no feature vectors satisfying $Q < 0.1$, and the number of feature vectors in the range $Q < 0.25$ is marginal. However, as in the previous case linear discriminants beat the best reported results for feature vectors satisfying $Q < 0.5$. This range captures about half of the data.

From the plots of Q_A, Q_B it is clear that mistakes classifying B items as A occur almost entirely for data items with low confidence. On the other hand, many data items from Category A are classified with relatively high confidence as B. This suggests that these mistakes relate to the data and not to the algorithm. Thus, other algorithms are not expected to perform significantly better with the same data.

Discussion

The performance of the MSE linear discriminant on the entire data is "surprisingly" good. The accuracy clearly decreases as a function of Q, justifying its use as a reliability criterion. Taking only values that are classified with high confidence produces significant improvements in accuracy. With the threshold of $Q < 0.5$ the results of linear discriminants were always better than the best previously reported accuracy on the entire data. Even in complex cases such as the "Diabetes" or the second "Breast-Cancer" dataset, this threshold captures about a half of the data. (It is possible that the performance of other classifiers may be just as good, or even better, on data with low Q values.)

Since our rough estimate of the probability of consistent subsets is 0.25, it is interesting to note that the MSE consistently produces results of above 90% accuracy in this range. This percentage does not go down in the tough cases, but the percentage of data that falls in this range goes down. (In the second Breast-Cancer data this threshold captures less than 5% of the data.)

Observing the values of Q_A, Q_B (the third and fourth plots) shows that the rate of errors is much larger in the right hand side of the plots, since these values correspond to a decision with low confidence. The confidence does not appear to be related to the separation between Q_A and Q_B, but, as predicted, it is related to their minimum.

Concluding remarks

It was shown that *reductio-ad-absurdum* can be used to prove optimality of classification under milder assumptions than a complete knowledge of the distribution. Even though using the ranked assumptions for classification was shown equivalent to linear discriminants, it has the additional advantage of providing a measure of confidence in the classification. This enables combining linear discriminants with other classifiers that may be much more expensive to run, and may require many more examples to achieve generalization.

References

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees.* Wadsworth International Group.

Breiman, L. 1996. Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, UC Berkeley.

Copi, I. M. 1979. *Symbolic Logic.* The Macmillan Company, fifth edition.

Devijver, P., and Kittler, J. 1982. *Pattern Recognition: A Statistical Approach.* London: Prentice Hall.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis.* John Wiley & Sons.

Feller, W. 1970. *An introduction to probability theory and its applications*, volume I. John Wiley & Sons, third edition.

Fisher, R. A. 1936. The use of multiple measurement in taxonomic problems. *Ann. Eugenics* 7:111–132.

Haussler, D.; Kearns, M.; and Schapire, R. 1994. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning* 14:83–113.

Kohavi, R. 1995a. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137–1143. Montreal: Morgan Kaufman.

Kohavi, R. 1995b. *Wrappers for Performance Enhancement and Oblivious Decision Graphs.* PhD dissertation, Stanford, Department of Computer Science.

Langley, P.; Iba, W.; and Thompson, K. 1992. An analysis of Bayesian classifiers. In *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92)*, 223–228. AAAI Press and MIT Press.

Smith, J. W.; Everhart, J. E.; Dickson, W. C.; Knowler, W. C.; and Johannes, R. S. 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Symposium on Computer Applications and Medical Care*, 261–265. IEEE Computer Society Press.

Vapnik, V. 1995. *The nature of Statistical Learning Theory.* Springer-Verlag.

Wolberg, W. H., and Mangasarian, O. 1990. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc. Natl. Acad. Sci. USA* 87:9193–9196.