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...
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: 
\[ E[x(o)] = \sum_{o \in O} x(o)P\{o\} \]

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

<table>
<thead>
<tr>
<th>Statistics from A</th>
<th>Statistics from B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_A = E{x(o)</td>
<td>o \in A}$</td>
</tr>
<tr>
<td>$R_A = E{x(o)x(o')</td>
<td>o \in A}$</td>
</tr>
<tr>
<td>$C_A = R_A - u_Au_A$</td>
<td>$C_B = R_B - u_Bu_B$</td>
</tr>
<tr>
<td>$p_A = P{o \in A}$</td>
<td>$p_B = P{o \in B}$</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>( \mathbf{A} )</th>
<th>( \mathbf{B} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbf{A} )</td>
<td>( \mathbf{B} )</td>
</tr>
</tbody>
</table>

The following condition guarantees that the optimal classification must agree with the correct classification with probability of at least $\frac{1}{2}$ on $N$. 0

**Theorem 1**: Given $x(x)$, let $N$ be a consistent subset that includes $x$. Let $Q$ be a real number satisfying:

\[ Q \geq 2P\{x(o) \in N \; , \; a(o) \neq \beta \} \]

The following condition guarantees that the optimal classification of $x$ is $\beta$:

\[ P\{x(o) \in N\} > Q \]

**Proof sketch**: The optimal classification based on $x$ is $\beta$ if and only if the optimal classification of any $x(o) \in N$ is $\beta$. 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 $N$. 0

A direct corollary of Lemma 1 is the following theorem.

**Theorem 1**: Given $x(x)$, let $N$ be a consistent subset that includes $x$. Let $Q$ be a real number satisfying:

\[ Q \geq 2P\{x(o) \in N \; , \; a(o) \neq \beta \} \]

The following condition guarantees that the optimal classification of $x$ is $\beta$:

\[ P\{x(o) \in N\} > Q \]

**Proof**: These assumptions imply that the condition in Lemma 1 does not hold under the premise that the optimal classification is not $\beta$. 0

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 $x$ is $\beta$, assume that it is not $\beta$, and show a subset consistent with $x$ 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 $P\{x(o) \in N\}$ is larger than $Q$ is sufficient to establish the optimality.

With respect to a feature vector $x$, any vector $w$ splits the entire space of feature vectors into the following two half-spaces: \( \{x \; ; \; w'x \geq 0\} \), and \( \{x \; ; \; w'(x - \bar{x}) \leq 0\} \). 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 $N(\bar{x}, w, \beta)$ is the set of feature vectors $x$ that should be optimally classified as $\beta$, which satisfy: $w'(x - \bar{x}) \geq 0$.

Notice that the union of the following four consistent subsets covers the entire space: $N(\bar{x}, w, v_A)$, $N(\bar{x}, w, v_B)$, $N(\bar{x}, -w, v_A)$, $N(\bar{x}, -w, v_B)$. 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 $\alpha(o)$:

\[ y(o) = w'x(o) + mx(o) + s, \]

where $w$ is an arbitrary vector, and $m, s$ are arbitrary scalars. Set $V = E\{y^2(o)\}$. For a feature vector $x$
and \( \beta \in \{ v_A, v_B \} \) compute: \( \overline{y} = w'x + m\beta + s \). Then regardless of the probability measure:

\[
\overline{y} > 0 \Rightarrow P\{ x(o) \in N(\overline{x}, w, \beta) , \alpha(o) = \beta \} \leq \frac{\overline{y}}{\sqrt{\beta}}
\]

\[
\overline{y} < 0 \Rightarrow P\{ x(o) \in N(\overline{x}, -w, -\beta) , \alpha(o) = \beta \} \leq \frac{\overline{y}}{\sqrt{\beta}}
\]

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

Therefore,

\[
P\{ x(o) \in N(\overline{x}, w, \beta) , \alpha(o) = \beta \} \leq P\{ y(o) \geq \overline{y} \} \leq P\{ |y(o)| \geq |\overline{y}| \} \leq \frac{\overline{y}}{\sqrt{\beta}}
\]

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

Therefore,

\[
P\{ x(o) \in N(\overline{x}, -w, -\beta) , \alpha(o) = \beta \} \leq P\{ y(o) \geq \overline{y} \} \leq P\{ |y(o)| \geq |\overline{y}| \} \leq \frac{\overline{y}}{\sqrt{\beta}} \Box
\]

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

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

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

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:

\[
Q_A = \min_m 2(F + m^2 + 2mg)/(f(\overline{x}) - m)^2,
\]

\[
Q_B = \min_m 2(F + m^2 + 2mg)/(f(\overline{x}) + m)^2
\]

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

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 \( f(x) = w'x + s \).
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:

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

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 $\gamma$.
3. For each value $\gamma_i$ 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 $\gamma_i$ 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 $\gamma_i$ 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 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:

<table>
<thead>
<tr>
<th>MSE</th>
<th>$Q &lt; .1$</th>
<th>$Q &lt; .25$</th>
<th>$Q &lt; .5$</th>
<th>$Q &lt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>%accuracy</td>
<td>99.8</td>
<td>98.6</td>
<td>96.8</td>
<td>96.8</td>
</tr>
<tr>
<td>%data</td>
<td>18.8</td>
<td>93.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>FISHER</th>
<th>$Q &lt; .1$</th>
<th>$Q &lt; .25$</th>
<th>$Q &lt; .5$</th>
<th>$Q &lt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>%accuracy</td>
<td>99.4</td>
<td>97.6</td>
<td>96.7</td>
<td>95.7</td>
</tr>
<tr>
<td>%data</td>
<td>10.4</td>
<td>53.0</td>
<td>95.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

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 70.1%. The largest $Q$ is 1.49. Accuracy and percentage data for 4 thresholds in $Q$ are shown below:
Table: MSE and %accuracy (%data)

<table>
<thead>
<tr>
<th>MSE</th>
<th>$Q &lt; .1$</th>
<th>$Q &lt; .25$</th>
<th>$Q &lt; .5$</th>
<th>$Q &lt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>%accuracy</td>
<td>?</td>
<td>91.5</td>
<td>78.9</td>
<td>73.8</td>
</tr>
<tr>
<td>%data</td>
<td>0</td>
<td>4.97</td>
<td>45.5</td>
<td>82.8</td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th>$Q &lt; .1$</th>
<th>$Q &lt; .25$</th>
<th>$Q &lt; .5$</th>
<th>$Q &lt; 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>%accuracy</td>
<td>?</td>
<td>89.4</td>
<td>78.8</td>
</tr>
<tr>
<td>%data</td>
<td>0</td>
<td>3.29</td>
<td>48.5</td>
</tr>
</tbody>
</table>

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 $Q_A, Q_R$ 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 fails 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**


