Discriminant Adaptive Nearest Neighbor Classification

Trevor Hastie
Department of Statistics, Sequoia Hall
Stanford University
California 94305
trevor@playfair.stanford.edu

Robert Tibshirani
Department of Preventive Medicine and Biostatistics
and Department of Statistics
University of Toronto
tibs@utstat.toronto.edu

Abstract

Nearest neighbor classification expects the class conditional probabilities to be locally constant, and suffers from bias in high dimensions We propose a locally adaptive form of nearest neighbor classification to try to finesse this curse of dimensionality. We use a local linear discriminant analysis to estimate an effective metric for computing neighborhoods. We determine the local decision boundaries from centroid information, and then shrink neighborhoods in directions orthogonal to these local decision boundaries, and elongate them parallel to the boundaries. Thereafter, any neighborhood-based classifier can be employed, using the modified neighborhoods. The posterior probabilities tend to be more homogeneous in the modified neighborhoods. We also propose a method for global dimension reduction, that combines local dimension information. In a number of examples, the methods demonstrate the potential for substantial improvements over nearest neighbour classification.

Introduction

We consider a discrimination problem with J classes and N training observations. The training observations consist of predictor measurements $\mathbf{x} = (x_1, x_2, \ldots x_p)$ on p predictors and the known class memberships. Our goal is to predict the class membership of an observation with predictor vector $\mathbf{x}_0$

Nearest neighbor classification is a simple and appealing approach to this problem. We find the set of K nearest neighbors in the training set to $\mathbf{x}_0$ and then classify $\mathbf{x}_0$ as the most frequent class among the K neighbors. Nearest neighbors is an extremely flexible classification scheme, and does not involve any pre-processing (fitting) of the training data. This can offer both space and speed advantages in very large problems: see Cover (1968), Duda & Hart (1973), or McLachlan (1992) for background material on nearest neighborhood classification.

Cover & Hart (1967) show that the one nearest neighbour rule has asymptotic error rate at most twice the Bayes rate. However in finite samples the curse of dimensionality can severely hurt the nearest neighbor rule. The relative radius of the nearest-neighbor

sphere grows like $r^{1/p}$ where p is the dimension and r the radius for $p = 1$, resulting in severe bias at the target point $\mathbf{x}$. Figure 1 illustrates the situation for a simple example.
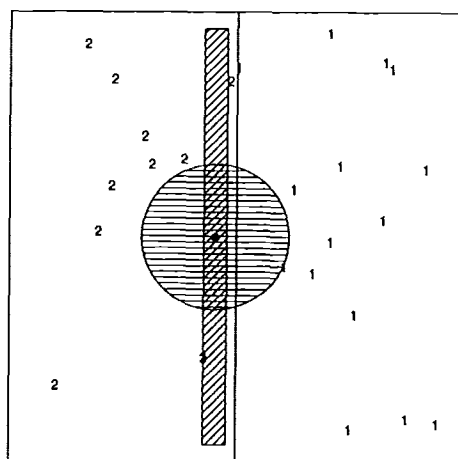

Figure 1: *The vertical strip denotes the NN region using only the X coordinate to find the nearest neighbor for the target point (solid dot). The sphere shows the NN region using both coordinates, and we see in this case it has extended into the class 1 region (and found the wrong class in this instance).*

Our illustration here is based on a 1-NN rule, but the same phenomenon occurs for k-NN rules as well. Nearest neighbor techniques are based on the assumption that locally the class posterior probabilities are constant. While that is clearly true in the vertical strip using only coordinate X, using X and Y this is no longer true.

The techniques outlined in the abstract are designed to overcome these problems. Figure 2 shows an example. There are two classes in two dimensions, one of which almost completely surrounds the other. The left panel shows a nearest neighborhood of size 25 at the target point (shown as origin), which is chosen to be near the class boundary. The right panel shows the same size neighborhood using our discriminant adap-
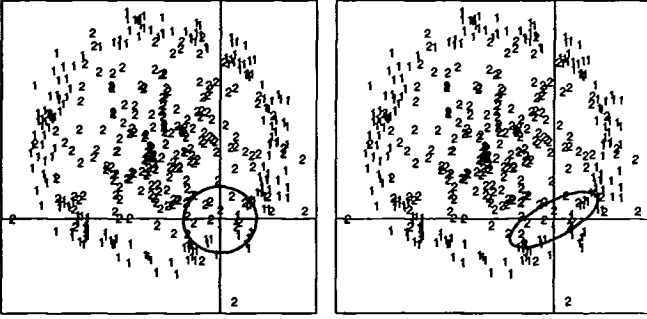
Figure 2: *The left panel shows a spherical neighborhood containing 25 points. The right panel shows the ellipsoidal neighborhood found by the DANN procedure, also containing 25 points. The latter is elongated along the true decision boundary, and flattened orthogonal to it.*

tive nearest neighbour procedure. Notice how the modified neighborhood extends further in the direction parallel to the decision boundary. As we will see in our simulation studies, this new neighborhood can often provide improvement in classification performance.

While the idea of local adaptation of the nearest neighbour metric may seem obvious, we could find few proposals along these lines in the literature. A summary of previous work in given in section .

Discriminant adaptive nearest neighbors

Our proposal is motivated as follows. Consider first a standard linear discriminant (LDA) classification procedure with K classes. Let $\mathbf{B}$ and $\mathbf{W}$ denote the between and within sum of squares matrices. In LDA the data are first sphered with respect to $\mathbf{W}$, then the target point is classified to the class of the closest centroid (with a correction for the class prior membership probabilities). Since only relative distances are relevant, any distances in the complement of the subspace spanned by the sphered centroids can be ignored. This complement corresponds to the null space of $\mathbf{B}$.

We propose to estimate $\mathbf{B}$ and $\mathbf{W}$ locally, and use them to form a local metric that approximately behaves like the LDA metric. One such candidate is

$$\Sigma = \mathbf{W}^{-1}\mathbf{B}\mathbf{W}^{-1}$$
$$= \mathbf{W}^{-1/2}(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2})\mathbf{W}^{-1/2}$$
$$= \mathbf{W}^{-1/2}\mathbf{B}^*\mathbf{W}^{-1/2}. \tag{1}$$

where $\mathbf{B}^*$ is the between sum-of-squares in the sphered space. Consider the action of Σ as a metric for computing distances $(\mathbf{x} - \mathbf{x}_0)^T\Sigma(\mathbf{x} - \mathbf{x}_0)$:

- it first spheres the space using $\mathbf{W}$;

- components of distance in the null space of $\mathbf{B}^*$ are ignored;

- other components are weighted according to the eigenvalues of $\mathbf{B}^*$ when there are more than 2 classes — directions in which the centroids are more spread out are weighted more than those in which they are close

Thus this metric would result in neighborhoods similar to the narrow strip in figure 2: infinitely long in the null space of $\mathbf{B}$, and then deformed appropriately in the centroid subspace according to how they are placed. It is dangerous to allow neighborhoods to extend infinitely in any direction, so we need to limit this stretching. Our proposal is

$$\Sigma = \mathbf{W}^{-1/2}[\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2} + \epsilon\mathbf{I}]\mathbf{W}^{-1/2}$$
$$= \mathbf{W}^{-1/2}[\mathbf{B}^* + \epsilon\mathbf{I}]\mathbf{W}^{-1/2} \tag{2}$$

where ϵ is some small tuning parameter to be determined. The metric shrinks the neighborhood in directions in which the local class centroids differ, with the intention of ending up with a neighborhood in which the class centroids coincide (and hence nearest neighbor classification is appropriate). With this goal in mind one can think of iterating the procedure, and thus successively shrinking in directions in which the class centroids do not coincide.

Here is a summary of the proposal.

Discriminant Adaptive Nearest Neighbor Classifier

0. Initialize the metric $\Sigma = \mathbf{I}$, the identity matrix.

1. Spread out a nearest neighborhood of K_M points around the test point $\mathbf{x}_0$, in the metric Σ.

2. Calculate the weighted within and between sum of squares matrices $\mathbf{W}$ and $\mathbf{B}$ using the points in the neighborhood (see formula (8 below).

3. Define a new metric $\Sigma = \mathbf{W}^{-1/2}[\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2} + \epsilon\mathbf{I}]\mathbf{W}^{-1/2}$.

4. Iterate steps 1, 2, and 3.

5. At completion, use the metric Σ for K-nearest neighbor classification at the test point $\mathbf{x}_0$.

The metric (2) can be given a more formal justification. Suppose we are classifying at a test point $\mathbf{x}_0$ and find a single nearest neighbor $\mathbf{X}$ according to a metric $d(\mathbf{X}, \mathbf{x}_0)$. Let $p(j|\mathbf{x})$ be the true probability of class j at point $\mathbf{x}$.

We consider the *Chi-squared* distance

$$r(\mathbf{X}, \mathbf{x}_0) = \sum_{j=1}^{J} \frac{[p(j|\mathbf{X}) - p(j|\mathbf{x}_0)]^2}{p(j|\mathbf{x}_0)} \tag{3}$$

which measures the distance (appropriately weighted) between the true and estimated posteriors. Small

$r(\mathbf{X}, \mathbf{x}_0)$ implies that the misclassification error rate will be close to the asymptotic error rate for 1NN, which is achieved when $\mathbf{X} = \mathbf{x}_0$ or more generally when $p(j|\mathbf{X}) = p(j|\mathbf{x}_0)$. We show that the first term in our metric (2) approximates $r(\mathbf{X}, \mathbf{x}_0)$.

Assuming that in the neighborhood $\mathbf{x}|j$ has a Gaussian distribution with mean μ_j and covariance Σ, we obtain by a simple first order Taylor approximation

$$p(j|\mathbf{X}) \approx p(j|\mathbf{x}_0) - p(j|\mathbf{x}_0)(\mu_j - \bar{\mu})^T \Sigma^{-1}(\mathbf{X} - \mathbf{x}_0) \quad (4)$$

where $\bar{\mu} = \sum_j p(j|\mathbf{x}_0)\mu_j$. Plugging this into (3) we get

$$r(\mathbf{X}, \mathbf{x}_0) \approx \sum_{j=1}^{J} p(j|\mathbf{x}_0) \left[(\mu_j - \bar{\mu})^T \Sigma^{-1}(\mathbf{X} - \mathbf{x}_0) \right]^2 \quad (5)$$

Thus the approximately best distance metric is $\Sigma^{-1} \sum_j p(j|\mathbf{x}_0)(\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T \Sigma^{-1}$. Estimating Σ by $\mathbf{W}$ and $\sum_j p(j|\mathbf{x}_0)(\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^T$ by $\mathbf{B}$ gives the first term in the metric (2).

By allowing prior uncertainty for the class means μ_j, that is, assume $\mu_j \sim N(\nu_j, \epsilon\mathbf{I})$ in the sphered space, we obtain the second term in the metric (2).

Details of the implementation

Define a weight function at $\mathbf{x}_0$ by

$$k(\mathbf{x}, \mathbf{x}_0; \Sigma, h) = \phi_h(\|\Sigma_0^{1/2}(\mathbf{x} - \mathbf{x}_0)\|). \quad (6)$$

Here Σ_0 is an initial non-negative metric (often I), and ϕ_h is a symmetric real-valued function depending on a parameter h. We use a *tri-cube* function defined over a K-nearest neighborhood $N_K(\mathbf{x}_0)$ of $\mathbf{x}_0$. Formally, we define $d_i = \|\Sigma^{1/2}(\mathbf{x}_i - \mathbf{x}_0)\|$, $h = \max_{i \in N_K(\mathbf{x}_0)} d_i$ and define

$$k(\mathbf{x}_i, \mathbf{x}_0; \Sigma, h) = [1 - (d_i/h)^3]^3 I(|d_i| < h) \quad (7)$$

Let $\mathbf{B}(\mathbf{x}_0; \Sigma_0, h)$ and $\mathbf{W}(\mathbf{x}_0; \Sigma_0, h)$ be the weighted between and within class sum of squares matrices, where the weights assigned to the ith observation are given by $w_i = k(\mathbf{x}_i, \mathbf{x}_0; \Sigma_0, h)$. That is,

$$\mathbf{B}(\mathbf{x}_0; \Sigma_0, h) = \sum_{j=1}^{J} \hat{\pi}_j (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T$$

$$\hat{\pi}_j = \frac{\sum_{y_i=j} w_i}{\sum_{i=1}^{N} w_i} \quad (8)$$

$$\mathbf{W}(\mathbf{x}_0; \Sigma_0, h) = \frac{\sum_{j=1}^{J} \sum_{y_i=j} w_i (\mathbf{x}_i - \bar{\mathbf{x}}_j)(\mathbf{x}_i - \bar{\mathbf{x}}_j)^T}{\sum_{i=1}^{N} w_i} \quad (9)$$

where $\bar{\mathbf{x}}_j$ is the weighted mean of the N_j observations in the jth group. Finally, we let $\mathbf{B}(\mathbf{x}_0; \Sigma_0, h)$ and $\mathbf{B}(\mathbf{x}_0; \Sigma_0, h)$ determine the metric Σ in (2).

Notice that equations (8) and (2) produce a mapping $\Sigma_0 \rightarrow \Sigma$, say $\Sigma = g(\Sigma_0)$. An approach we explore is

to start with $\Sigma_0 = I$ (the identity matrix) and iterate this procedure. The result is a metric Σ for use in a nearest neighbor classification rule at $\mathbf{x}_0$. In our examples we try either a single step of this procedure, or larger number of iterations.

Some remarks about the DANN metric

It is natural to ask whether the mapping $g(\cdot)$ has a fixed point, and if it does, whether an iteration of the form $\Sigma \leftarrow g(\Sigma)$ converges to it. These questions seem difficult to answer in general. To get some insight, it is helpful to consider an equivalent form of the iteration. At each step we take a spherical neighborhood around the test point, estimate the metric Σ, and then transform the predictors via $\mathbf{x}^{new} = \Sigma^{1/2}\mathbf{x}^{old}$. At completion we use a spherical nearest neighbor rule in the final transformed space. It is easy to show that this procedure is equivalent to the one given above. If the metrics estimated in j iterations are $\Sigma_1, \Sigma_2, \dots \Sigma_j$, then the effective metric for the original coordinates is $\Sigma_j^{1/2}\Sigma_{j-1}^{1/2} \cdots \Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2} \cdots \Sigma_{j-1}^{1/2}\Sigma_j^{1/2}$.

Expressed in this way, the fixed points of the iteration satisfy $\mathbf{W}^{-1}\mathbf{B}\mathbf{W}^{-1} + \epsilon\mathbf{W}^{-1} = c\mathbf{I}$. In particular a fixed point occurs when $\mathbf{B}$ is zero and $\mathbf{W}$ is proportional to the identity matrix, in the space of the transformed coordinates.

In practice we find it more effective to estimate only the diagonal elements of $\mathbf{W}$, and assume that the off diagonal elements are zero. This is especially true if the dimension of the predictor space is large, as there will be insufficient data locally to estimate the $O(p^2)$ elements of $\mathbf{W}$. With the diagonal approximation, the two forms of the algorithm are not equivalent: we use the version that transforms the space at each step since a diagonal approximation makes most sense in the transformed coordinates.

If the predictors are spatially or temporally related, we might use a penalized estimate of $\mathbf{W}$ that down-weights components of the covariance that correspond to spatially noisy signals (Hastie, Buja & Tibshirani 1994). A related approach is to pre-filter the predictors using a smooth basis, and then operate in the reduced domain.

In the final neighborhood we perform K nearest neighbor classification. An alternative approach would be to use discriminant analysis to perform the classification, using the locally determined parameters. We are currently investigating this approach.

Choice of tuning parameters

The DANN procedure has a number of adjustable tuning parameters:

K_M : the number of nearest neighbors in the neighborhood $N_{K_M}(\mathbf{x}_0)$ for estimation of the metric;

K : the number of neighbors in the final nearest neighbor rule;

ϵ : the "softening" parameter in the metric.

Test set or cross validation could be used to estimate an optimal values for these parameters. In the examples in the next section we instead use fixed choices. The value of K_m must be reasonably large since the initial neighborhood is used to estimate a covariance: we use $K_M = \max(N/5, 50)$. To ensure consistency one should take K_M to be a vanishing fraction of N, and should also use larger values for higher dimensional problems. A smaller number of neighbors is preferable for the final classification rule to avoid bias: we used $K = 5$, and compared it to standard 5 nearest neighbors. Note that the metric (2) is invariant under non-singular transformations of the predictors, and hence it makes sense to consider fixed values of ϵ. After some experimentation, we found that the value $\epsilon = 1$ works well, and we use this value in the examples below.

Dimension Reduction using Local Discriminant Information

So far our technique has been entirely "memory based", in that we locally adapt a neighborhood about a query point at the time of classification. Here we describe a method for performing a global dimension reduction, by pooling the local dimension information over all points in the training set. In a nutshell we consider subspaces corresponding to eigenvectors of the average local between sum-of-squares matrices.

Consider first how linear discriminant analysis (LDA) works. After sphering the data, it concentrates in the space spanned by the class means $\bar{x}_j$ or a reduced rank space that lies close to these means. If $\bar{x}$ denote the overall mean, this subspace is exactly the principal component hyperplane for the data points $\bar{x}_j - \bar{x}$, weighted by the class proportions.

Our idea to compute the deviations $\bar{x}_j - \bar{x}$ locally in a neighborhood around each of the N training points, and then do an overall principal components analysis for the $N \times J$ deviations. Here are the details. Let $x_j(i)$ be the mean of class j vectors in a neighborhood of the ith training point, and $\bar{x}(i)$ be the overall mean. All means are weighted by the local class membership proportions $\pi_j(i)$, $j = 1, \ldots, J$. Let $\tilde{x}_j(i) = x_j(i) - \bar{x}(i)$, the local centroid deviations. We seek a subspace that gets close in average weighted squared distance to all $N \times J$ of these. Denoting by U ($p \times J$) an orthonormal basis for the $k < p$ dimensional subspace, we minimize the criterion

$$RSS(U) = \sum_{i=1}^{N} \sum_{j=1}^{J} \pi_j(i) \tilde{x}_j^T(i)(I - UU^T)\tilde{x}_j(i),$$

or the total weighted residual sum of squares. It is not hard to show that minimizing $RSS(U)$ amounts to maximizing

$$\text{tr } U^T \left(\sum_{i=1}^{N} B(i) \right) U$$

where the $B(i)$ are the local between sum of squares matrices. This latter problem is solved by finding the largest eigenvectors of the average between sum of squares matrix $\sum_{i=1}^{N} B(i)/N$.
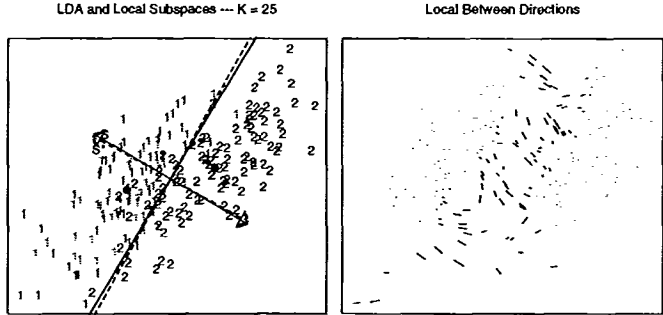

Figure 3: [Left Panel] Two dimensional gaussian data with two classes and correlation 0.65. The solid lines are the LDA decision boundary and its equivalent subspace for classification. The dashed lines were produced by the local procedure described in this section. [Right panel] Each line segment represents the local between information centered at that point.

Figure 3 shows a simple illustrative example. The two classes are Gaussian with substantial within class covariance between the two predictors X_1 and X_2. In the left panel, the solid line is the Gaussian decision boundary that optimally separates the classes. The orthogonal vector labeled S is a one dimensional subspace onto which we can project the data and perform classification. Using the knowledge that the data are Gaussian, it is the leading discriminant direction. The broken lines are the boundaries and equivalent subspace produced by our procedure. In the right panel, each line segment represents the local between information centered at that point. Our procedure uses a principal components analysis of these $N \times J$ line segments to produce the broken line subspace in the left panel.

To allow combination of the local between information in a meaningful way, notice that we have not sphered the data locally before computing the mean deviations. A justification for this is that any local spherical window containing two classes, say, will have approximately a linear decision boundary orthogonal to the vector joining the two means.

Figure 4 shows the eigenvalues of the average between matrix for an instance of a two class, 4 dimensional sphere model with 6 noise dimensions. The decision boundary is a 4 dimensional sphere, although locally linear (full details of this example are given in the next section). For this demonstration we randomly rotated the 10 dimensional data, so that the dimensions to be trimmed are not coordinate directions. The eigenvalues show a distinct change after 4 (the correct
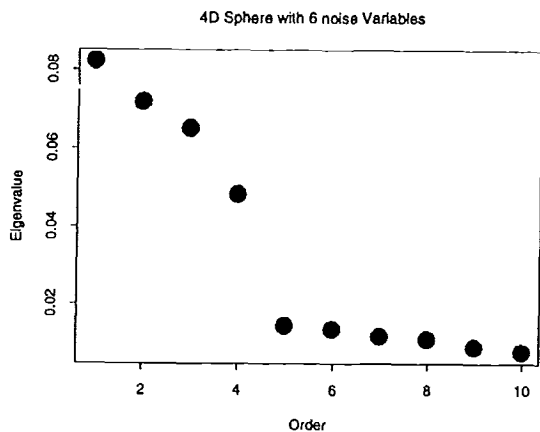
4D Sphere with 6 noise Variables

Figure 4: *The eigenvalues of the average between matrix for the 4D sphere + 6 noise variable problem. Using these first four dimensions followed by our DANN nearest neighbor routine, we get better performance than 5NN in the real 4D subspace.*

dimension), and using our DANN classifier in these four dimensions actually beats ordinary 5NN in the *known* four dimensional sphere subspace.

It is desirable to automate the dimension reduction operation. Since our local information is based on spherical neighborhoods (potentially in high dimensions), we find an iterative approach most successful. We apply this procedure in the full space, and use cross-validated DANN to find the best nested subspace (with a built in bias towards larger subspaces). We then successively repeat these operations in the new subspaces, until no further reduction is deemed suitable by CV. Using DANN in this final subspace is what we have labelled sub-DANN in the boxplots of figures 5.

Examples

The methods

In the following examples we compare several classification approaches:

- LDA—linear discriminant analysis
- reduced LDA—linear discriminant restricted to the (known) relevant subspace, where appropriate.
- 5-NN: 5 nearest neighbor classification
- reduced 5-NN
- DANN— Discriminant adaptive nearest neighbor, one iteration.
- iter-DANN— Discriminant adaptive nearest neighbor, five iterations.
- sub-DANN— Discriminant adaptive nearest neighbor, with automatic subspace reduction. This is described in section .

For all methods, the predictors were first standardized so as to have zero mean and unit variance over the training set, and the test set predictors was standardized by the corresponding training mean and variance. The training and test set sizes were 200 and 500, unless indicated otherwise.

The problems

1. *2 Dimensional Gaussian with 14 Noise.*Two Gaussian classes in two dimensions (X_1, X_2) separated by 2 units in X_1. The predictors have variance $(1,2)$ and correlation 0.75. The additional 14 predictors are independent standard Gaussians.

2. *Unstructured with 8 Noise.* There are 4 classes each with 3 spherical bivariate normal subclasses, having standard deviation 0.25. The means of the 12 subclasses were chosen at random (without replacement) from the integers $[1,2,\cdots 5] \times [1,2,\cdots 5]$. Each training sample had 20 observations per subclass, for a total of 240 observations. The additional 8 predictors are independent standard Gaussians.

3. *4 Dimensional Spheres with 6 Noise.* In this example there are 10 predictors and 2 classes. The last 6 predictors are noise variables, with standard Gaussian distributions, independent of each other and the class membership. The first four predictors in class 1 are independent standard normal, conditioned on the radius being greater than 3, while the first four predictors in class 2 are independent standard normal without the restriction. The first class almost completely surrounds the second class in the four dimensional subspace of the first four predictors. This example was designed to see if DANN could improve upon nearest neighbors in the presence of noise variables.

4. *10 Dimensional Spheres.* As in the previous example there are 10 predictors and 2 classes. Now all 10 predictors in class 1 are independent standard normal, conditioned on the radius being greater than 22.4 and less than 40, while the predictors in class 2 are independent standard normal without the restriction. In this example there are no pure noise variables, the kind that a nearest neighbor subset selection rule might be able weed out. At any given point in the feature space, the class discrimination occurs along only one direction. However this direction changes as we move across the feature space and all variables are important somewhere in the space. The first class almost completely surrounds the second class in the full ten-dimensional space.

Discussion of results

The results for the simulated examples are summarized in Figures 5.

DANN seems to do as well as 5-NN across the board, and offers significant improvements in problems with noise variables. DANN does not do as well as reduced
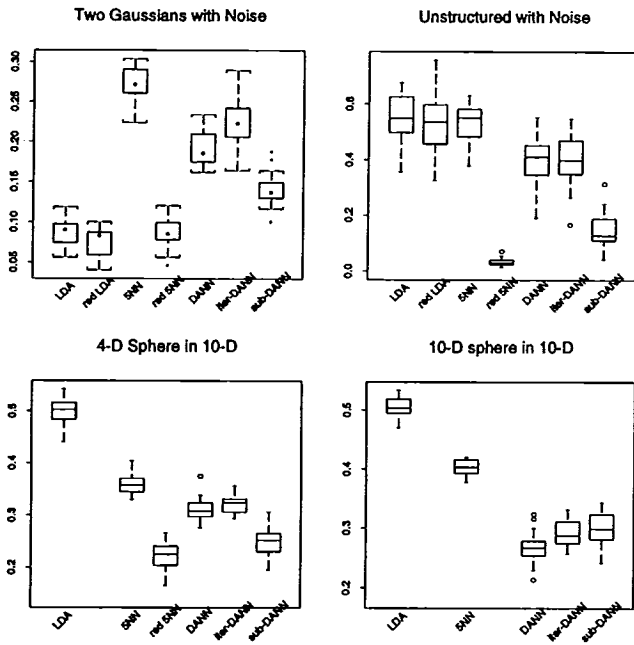
Figure 5: *Boxplots of error rates over 20 simulations.*

Figure 6: *Relative error rates of the methods across the 8 simulated problems. In the top panel the error rate has been divided by the error rate 5-NN, on a simulation by simulation basis. In the bottom panel we have divided by the error rate of LDA.*

nearest neighbors in problems 1 and 3: this is not surprising since in effect we are giving the nearest neighbor rule the information that DANN is trying to infer from the training data. A nearest neighbor method with variable selection might do well in these problems: however this procedure can be foiled by by rotating the relevant subspace away from the coordinate directions.

On the average there seems to be no advantage in carrying out more than one iteration of the DANN procedure. The subspace DANN procedure is the overall winner, producing big gains in problems admitting global dimension reduction.

The top panel of Figure 6 shows error rates relative to 5-NN, accumulated across 8×20 simulated problems (these 4 and another 4 described in Hastie & Tibshirani (1995). The bottom panel shows the rates relative to LDA.

We see that DANN is 20-30% better than 5-NN on the average, and is at most 20% worse. DANN is also better than LDA on the average but can be three times worse (in problem 2).

Image Classification Example

Image scene classification is an important and difficult problem. The example we consider here is classifying satellite images of an area of the earth's surface into land and vegetation type. Other examples include classification of X-rays, such as mammograms, into normal or calcified into normal or cancerous regions.

Figure 7 shows four spectral bands of a section of land, two in the visible spectrum (red and green)
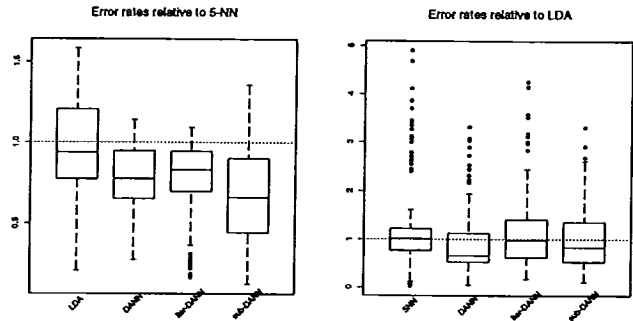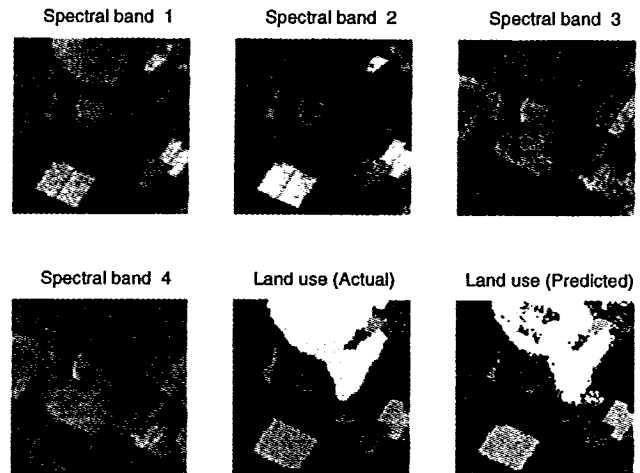

Figure 7: The first four images are the satellite images in the four spectral bands. The fifth image is the known classification, and the final image is the classification map produced by linear discriminant analysis.

A Pixel and its 8 neighbors

Figure 8: The pixel intensities of the 8-neighbors of a pixel (and itself) are used as features for classification

and two in the infra red spectrum. These data are taken from the STATLOG projects archive (Michie, Spigelhalter & Taylor 1994)[1]. The goal is to classify each pixel into one of 7 land types: *red soil, cotton, vegetation stubble, mixture, grey soil, damp grey soil, very damp grey soil.* We extract for each pixel its 8-neighbors, as depicted in figure 8, giving us $(8+1)\times 4 = 36$ features (the pixel intensities) per pixel to be classified. The data come scrambled, with 4435 training pixels and 2000 test pixels, each with their 36 features and the known classification. Included in figure 7 is the true classification, as well as that produced by linear discriminant analysis. Figure 9 shows
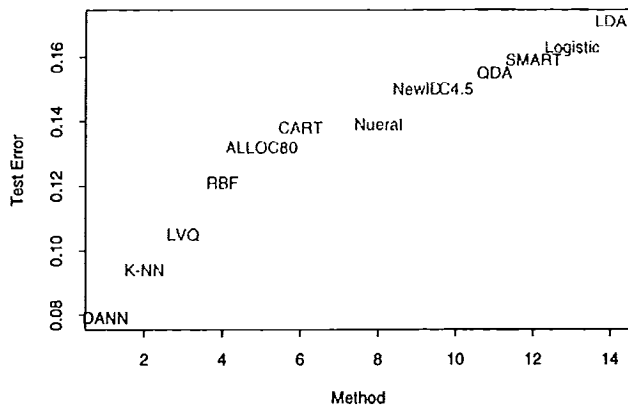
STATLOG results

Figure 9: Misclassification results of a variety of classification procedures on the satellite image test data (taken from Michie et al. (1994)).

the results reported in Michie et al. (1994) for a variety of classifiers; they reported the best result for 5-NN classification. Included in the figure is the result for DANN, which has outperformed 5-NN. We also ran the subspace version of DANN, and figure 10 shows the sequence of of test-error results as a function of subspace size. Again, a low-dimensional subspace actually improves the misclassification error.

Discussion

We have developed an adaptive form of nearest neighbor classification that can offer substantial improvements over standard nearest neighbors method in some problems. We have also proposed a method that uses local discrimination information to estimate a subspace for global dimension reduction.

Short & Fukanaga (1980) proposed a technique close to ours for the two class problem. In our terminology they used our metric with $\mathbf{W} = \mathbf{I}$ and $\epsilon = 0$, with $\mathbf{B}$
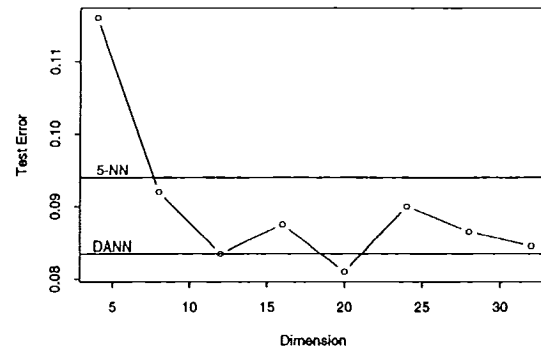
Satellite Image Classification

Figure 10: Misclassification results as a function of subspace size, for the satellite image data

determined locally in a neighborhood of size K_M. In effect this extends the neighborhood infinitely in the null space of the local between class directions, but they restrict this neighborhood to the original K_M observations. This amounts to projecting the local data onto the line joining the two local centroids. In our experiments this approach tended to perform on average 10% worse than our metric, and we did not pursue it further. Short & Fukanaga (1981) extended this to $J > 2$ classes, but here their approach differs even more from ours. They computed a weighted average of the J local centroids from the overall average, and project the data onto it, a one dimensional projection. Even with $\epsilon = 0$ we project the data onto the subspace containing the local centroids, and deform the metric appropriately in that subspace. Myles & Hand (1990) recognized a shortfall of the Short and Fukanaga approach, since the averaging can cause cancellation, and proposed other metrics to avoid this. Although their metrics differ from ours, the Chi-squared motivation for our metric (3) was inspired by the metrics developed in their paper. We have not tested out their proposals, but they report results of experiments with far more modest improvements over standard nearest neighbors than we achieved.

Friedman (1994) proposes a number of techniques for flexible metric nearest neighbor classification. These techniques use a recursive partitioning style strategy to adaptively shrink and shape rectangular neighborhoods around the test point. Friedman also uses derived variables in the process, including discriminant variates. With the latter variables, his procedures have some similarity to the discriminant adaptive nearest neighbor approach.

Other recent work that is somewhat related to this is that of Lowe (1993). He estimates the covariance matrix in a variable kernel classifier using a neural network approach.

There are a number of ways in which this work might

be generalized. In some discrimination problems, it is natural to use specialized distance measures that capture invariances in the feature space. For example Simard, LeCun & Denker (1993), Hastie, Simard & Sackinger (1993), use a transformation-invariant metric to measure distance between digitized images of handwritten numerals in a nearest neighbor rule. The invariances include local transformations of images such as rotation, shear and stroke-thickness. An invariant distance measure might be used in a linear discriminant analysis and hence in the DANN procedure.

Another interesting possibility would be to apply the techniques of this paper to regression problems. In this case the response variable is quantitative rather than a class label. Natural analogues of the local between and within matrices exist, and can be used to shape the neighborhoods for near-neighbor and local polynomial regression techniques. Likewise, the dimension reduction ideas of section can also be applied. There is a strong connection between the latter and the *Sliced Inverse Regression* technique of Duan & Li (1991) for subspace identification. We are currently exploring these directions.

References

Cover, T. (1968), Rates of convergence for nearest neighbor procedures, *in* 'Proc. Hawaii Inter. Conf. on Systems Sciences', Western Periodicals, Honolulu, pp. 413–415.

Cover, T. & Hart, P. (1967), 'Nearest neighbor pattern classification', *Proc. IEEE Trans. Inform. Theory* pp. 21–27.

Duan, N. & Li, K.-C. (1991), 'Slicing regression: a link-free regression method', *Annals of Statistics* pp. 505–530.

Duda, R. O. & Hart, P. E. (1973), *Pattern classification and scene analysis*, Wiley, New York.

Friedman, J. (1994), Flexible metric nearest neighbour classification, Technical report, Stanford University.

Hastie, T. & Tibshirani, R. (1995), Discriminant adaptive nearest neighbor classification, Technical report, Stanford University Statistics Department.

Hastie, T., Buja, A. & Tibshirani, R. (1994), Penalized discriminant analysis, To appear, Annals of Statistics.

Hastie, T., Simard, P. & Sackinger, E. (1993), Learning prototype models for tangent distance, Technical report, AT& Bell Labs.

Lowe, D. G. (1993), Similarity metric learning for a variable kernel classifier, Technical report, Dept. of Comp Sci, Univ. of British Columbia.

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

Michie, D., Spigelhalter, D. & Taylor, C., eds (1994), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood series in Artificial Intelligence, Ellis Horwood.

Myles, J. & Hand, D. J. (1990), 'The multi-class metric problem in nearest neighbour discrimination rules', *Pattern Recognition* 23, 1291–1297.

Short, R. & Fukanaga, K. (1980), A new nearest neighbor distance measure, *in* 'Proc. 5th IEEE Int. Conf. on Pattern Recognition', pp. 81–86.

Short, R. & Fukanaga, K. (1981), 'The optimal distance measure for nearest neighbor classification', *IEEE transactions of Information Theory* IT-27, 622–627.

Simard, P. Y., LeCun, Y. & Denker, J. (1993), Efficient pattern recognition using a new transformation distance, *in* 'Advances in Neural Information Processing Systems', Morgan Kaufman, San Mateo, CA, pp. 50–58.