Chapter 5

Statistical Learning

5.1 Using Statistical Decision Theory

5.1.1 Background and General Method

Suppose the pattern vector, $\mathbf{X}$, is a random variable whose probability distribution for category 1 is different than it is for category 2. (The treatment given here can easily be generalized to R-category problems.) Specifically, suppose we have the two probability distributions (perhaps probability density functions), $p(\mathbf{X} \mid 1)$ and $p(\mathbf{X} \mid 2)$. Given a pattern, $\mathbf{X}$, we want to use statistical techniques to determine its category—that is, to determine from which distribution it was drawn. These techniques are based on the idea of minimizing the expected value of a quantity similar to the error function we used in deriving the weight-changing rules for backprop.

In developing a decision method, it is necessary to know the relative seriousness of the two kinds of mistakes that might be made. (We might decide that a pattern really in category 1 is in category 2, and vice versa.) We describe this information by a *loss function*, $\lambda(i \mid j)$, for $i, j = 1, 2$. $\lambda(i \mid j)$ represents the loss incurred when we decide a pattern is in category i when really it is in category j. We assume here that $\lambda(1 \mid 1)$ and $\lambda(2 \mid 2)$ are both 0. For any given pattern, $\mathbf{X}$, we want to decide its category in such a way that minimizes the expected value of this loss.

Given a pattern, $\mathbf{X}$, if we decide category i, the expected value of the loss will be:

$$L_{\mathbf{X}}(i) = \lambda(i \mid 1)p(1 \mid \mathbf{X}) + \lambda(i \mid 2)p(2 \mid \mathbf{X})$$

where $p(j \mid \mathbf{X})$ is the probability that given a pattern $\mathbf{X}$, its category is j. Our decision rule will be to decide that $\mathbf{X}$ belongs to category 1 if $L_{\mathbf{X}}(1) \leq L_{\mathbf{X}}(2)$, and to decide on category 2 otherwise.

We can use Bayes' Rule to get expressions for $p(j \mid \mathbf{X})$ in terms of $p(\mathbf{X} \mid j)$, which we assume to be known (or estimatible):

$$p(j \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid j)p(j)}{p(\mathbf{X})}$$

where $p(j)$ is the (a priori) probability of category j (one category may be much more probable than the other); and $p(\mathbf{X})$ is the (a priori) probability of pattern $\mathbf{X}$ being the pattern we are asked to classify. Performing the substitutions given by Bayes' Rule, our decision rule becomes:

Decide category 1 iff:

$$\lambda(1 \mid 1)\frac{p(\mathbf{X} \mid 1)p(1)}{p(\mathbf{X})} + \lambda(1 \mid 2)\frac{p(\mathbf{X} \mid 2)p(2)}{p(\mathbf{X})}$$

$$\leq \lambda(2 \mid 1)\frac{p(\mathbf{X} \mid 1)p(1)}{p(\mathbf{X})} + \lambda(2 \mid 2)\frac{p(\mathbf{X} \mid 2)p(2)}{p(\mathbf{X})}$$

Using the fact that $\lambda(i \mid i) = 0$, and noticing that $p(\mathbf{X})$ is common to both expressions, we obtain,

Decide category 1 iff:

$$\lambda(1 \mid 2)p(\mathbf{X} \mid 2)p(2) \leq \lambda(2 \mid 1)p(\mathbf{X} \mid 1)p(1)$$

If $\lambda(1 \mid 2) = \lambda(2 \mid 1)$ and if $p(1) = p(2)$, then the decision becomes particularly simple:

Decide category 1 iff:

$$p(\mathbf{X} \mid 2) \leq p(\mathbf{X} \mid 1)$$

Since $p(\mathbf{X} \mid j)$ is called the *likelihood* of j with respect to $\mathbf{X}$, this simple decision rule implements what is called a *maximum-likelihood* decision.

More generally, if we define $k(i \mid j)$ as $\lambda(i \mid j)p(j)$, then our decision rule is simply,

Decide category1 iff:

$$k(1 \mid 2)p(\mathbf{X} \mid 2) \leq k(2 \mid 1)p(\mathbf{X} \mid 1)$$

In any case, we need to compare the (perhaps weighted) quantities $p(\mathbf{X} \mid i)$ for $i = 1$ and 2. The exact decision rule depends on the the probability distributions assumed. We will treat two interesting distributions.

5.1.2 Gaussian (or Normal) Distributions

The multivariate (n-dimensional) Gaussian distribution is given by the probability density function:

$$p(\mathbf{X}) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} e^{\frac{-(\mathbf{X}-\mathbf{M})^t \mathbf{\Sigma}^{-1} (\mathbf{X}-\mathbf{M})}{2}}$$

where n is the dimension of the column vector $\mathbf{X}$, the column vector $\mathbf{M}$ is called the *mean vector*, $(\mathbf{X}-\mathbf{M})^t$ is the transpose of the vector $(\mathbf{X}-\mathbf{M})$, $\mathbf{\Sigma}$ is the *covariance matrix* of the distribution (an $n \times n$ symmetric, positive definite matrix), $\mathbf{\Sigma}^{-1}$ is the inverse of the covariance matrix, and $|\mathbf{\Sigma}|$ is the determinant of the covariance matrix.

The mean vector, $\mathbf{M}$, with components $(m_1, \ldots, m_n)$, is the expected value of $\mathbf{X}$ (using this distribution); that is, $\mathbf{M} = E[\mathbf{X}]$. The components of the covariance matrix are given by:

$$\sigma_{ij}^2 = E[(x_i - m_i)(x_j - m_j)]$$

In particular, σ_{ii}^2 is called the *variance* of x_i.

Although the formula appears complex, an intuitive idea for Gaussian distributions can be given when $n = 2$. We show a two-dimensional Gaussian distribution in Fig. 5.1. A three-dimensional plot of the distribution is shown at the top of the figure, and contours of equal probability are shown at the bottom. In this case, the covariance matrix, $\mathbf{\Sigma}$, is such that the elliptical contours of equal probability are skewed. If the covariance matrix were diagonal, that is if all off-diagonal terms were 0, then the major axes of the elliptical contours would be aligned with the coordinate axes. In general the principal axes are given by the eigenvectors of $\mathbf{\Sigma}$. In any case, the equi-probability contours are all centered on the mean vector, $\mathbf{M}$, which in our figure happens to be at the origin. In general, the formula in the exponent in the Gaussian distribution is a *positive definite quadratic form* (that is, its value is always positive); thus equi-probability contours are hyper-ellipsoids in n-dimensional space.

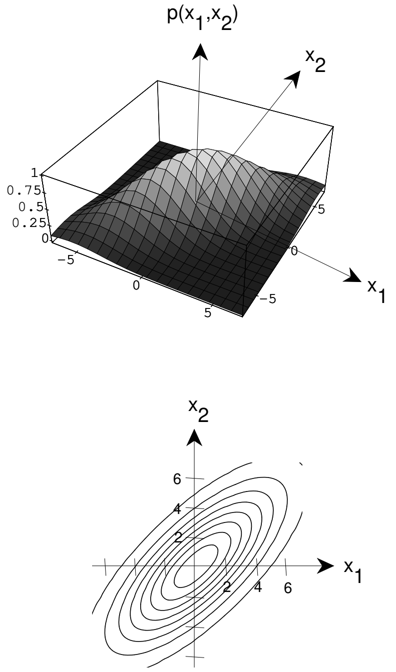

Figure 5.1: The Two-Dimensional Gaussian Distribution

Suppose we now assume that the two classes of pattern vectors that we want to distinguish are each distributed according to a Gaussian distribution but with different means and covariance matrices. That is, one class tends to have patterns clustered around one point in the n-dimensional space, and the other class tends to have patterns clustered around another point. We show a two-dimensional instance of this problem in Fig. 5.2. (In that figure, we have plotted the sum of the two distributions.) What decision rule should we use to separate patterns into the two appropriate categories?

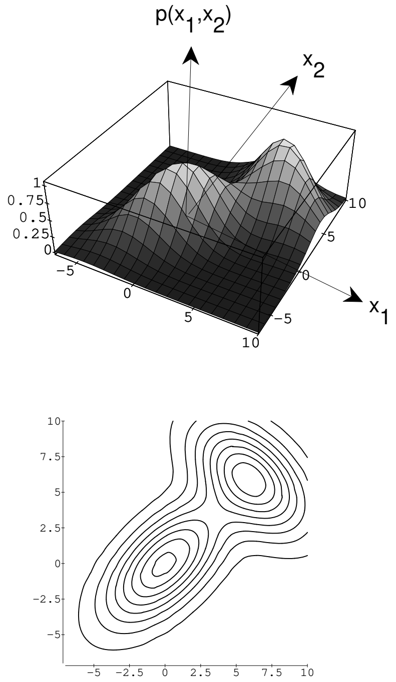

Figure 5.2: The Sum of Two Gaussian Distributions

Substituting the Gaussian distributions into our maximum likelihood formula yields:

Decide category 1 iff:

$$\frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}_2|^{1/2}} e^{-1/2(\mathbf{X}-\mathbf{M}_2)^t \mathbf{\Sigma}_2^{-1} (\mathbf{X}-\mathbf{M}_2)}$$

is less than or equal to

$$\frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}_1|^{1/2}}e^{-1/2(\mathbf{X}-\mathbf{M}_1)^t\mathbf{\Sigma}_1^{-1}(\mathbf{X}-\mathbf{M}_1)}$$

where the category 1 patterns are distributed with mean and covariance $\mathbf{M}_1$ and $\mathbf{\Sigma}_1$, respectively, and the category 2 patterns are distributed with mean and covariance $\mathbf{M}_2$ and $\mathbf{\Sigma}_2$.

The result of the comparison isn't changed if we compare logarithms instead. After some manipulation, our decision rule is then:

Decide category 1 iff:

$$(\mathbf{X}-\mathbf{M}_1)^t\mathbf{\Sigma}_1^{-1}(\mathbf{X}-\mathbf{M}_1) < (\mathbf{X}-\mathbf{M}_2)^t\mathbf{\Sigma}_2^{-1}(\mathbf{X}-\mathbf{M}_2) + B$$

where B, a constant bias term, incorporates the logarithms of the fractions preceding the exponential, *etc.*

When the quadratic forms are multiplied out and represented in terms of the components x_i, the decision rule involves a quadric surface (a hyperquadric) in n-dimensional space. The exact shape and position of this hyperquadric is determined by the means and the covariance matrices. The surface separates the space into two parts, one of which contains points that will be assigned to category 1 and the other contains points that will be assigned to category 2.

It is interesting to look at a special case of this surface. If the covariance matrices for each category are identical and diagonal, with all σ_{ii} equal to each other, then the contours of equal probability for each of the two distributions are hyperspherical. The quadric forms then become $(1/|\mathbf{\Sigma}|)(\mathbf{X}-\mathbf{M}_i)^t(\mathbf{X}-\mathbf{M}_i)$, and the decision rule is:

Decide category 1 iff:

$$(\mathbf{X}-\mathbf{M}_1)^t(\mathbf{X}-\mathbf{M}_1) < (\mathbf{X}-\mathbf{M}_2)^t(\mathbf{X}-\mathbf{M}_2)$$

Multiplying out yields:

$$\mathbf{X}\bullet\mathbf{X} - 2\mathbf{X}\bullet\mathbf{M}_1 + \mathbf{M}_1\bullet\mathbf{M}_1 < \mathbf{X}\bullet\mathbf{X} - 2\mathbf{X}\bullet\mathbf{M}_2 + \mathbf{M}_2\bullet\mathbf{M}_2$$

or finally,

Decide category 1 iff:

$$\mathbf{X}\bullet\mathbf{M}_1 \geq \mathbf{X}\bullet\mathbf{M}_2 + \text{Constant}$$

or

$$\mathbf{X}\bullet(\mathbf{M}_1 - \mathbf{M}_2) \geq \text{Constant}$$

where the constant depends on the lengths of the mean vectors.

We see that the optimal decision surface in this special case is a hyperplane. In fact, the hyperplane is perpendicular to the line joining the two means. The weights in a TLU implementation are equal to the difference in the mean vectors.

If the parameters $(\mathbf{M}_i, \mathbf{\Sigma}_i)$ of the probability distributions of the categories are not known, there are various techniques for estimating them, and then using those estimates in the decision rule. For example, if there are sufficient training patterns, one can use sample means and sample covariance matrices. (Caution: the sample covariance matrix will be singular if the training patterns happen to lie on a subspace of the whole n-dimensional space—as they certainly will, for example, if the number of training patterns is less than n.)

5.1.3 Conditionally Independent Binary Components

Suppose the vector $\mathbf{X}$ is a random variable having binary (0,1) components. We continue to denote the two probability distributions by $p(\mathbf{X} \mid 1)$ and $p(\mathbf{X} \mid 2)$. Further suppose that the components of these vectors are conditionally independent given the category. By conditional independence in this case, we mean that the formulas for the distribution can be expanded as follows:

$$p(\mathbf{X} \mid i) = p(x_1 \mid i)p(x_2 \mid i)\cdots p(x_n \mid i)$$

for $i = 1, 2$

Recall the minimum-average-loss decision rule,

Decide category 1 iff:

$$\lambda(1 \mid 2) p(\mathbf{X} \mid 2) p(2) \leq \lambda(2 \mid 1) p(\mathbf{X} \mid 1) p(1)$$

Assuming conditional independence of the components and that $\lambda(1 \mid 2) = \lambda(2 \mid 1)$, we obtain,

Decide category 1 iff:

$$p(1)p(x_1 \mid 1)p(x_2 \mid 1) \cdots p(x_n \mid 1) \geq p(x_1 \mid 2)p(x_2 \mid 2) \cdots p(x_n \mid 2)p(2)$$

or iff:

$$\frac{p(x_1 \mid 1)p(x_2 \mid 1) \ldots p(x_n \mid 1)}{p(x_1 \mid 2)p(x_2 \mid 2) \ldots p(x_n \mid 2)} \geq \frac{p(2)}{p(1)}$$

or iff:

$$\log \frac{p(x_1 \mid 1)}{p(x_1 \mid 2)} + \log \frac{p(x_2 \mid 1)}{p(x_2 \mid 2)} + \cdots + \log \frac{p(x_n \mid 1)}{p(x_n \mid 2)} + \log \frac{p(1)}{p(2)} \geq 0$$

Let us define values of the components of the distribution for specific values of their arguments, x_i :

$p(x_i = 1 \mid 1) = p_i$
$p(x_i = 0 \mid 1) = 1 - p_i$
$p(x_i = 1 \mid 2) = q_i$
$p(x_i = 0 \mid 2) = 1 - q_i$

Now, we note that since x_i can only assume the values of 1 or 0:

$$\log \frac{p(x_i \mid 1)}{p(x_i \mid 2)} = x_i \log \frac{p_i}{q_i} + (1 - x_i) \log \frac{(1 - p_i)}{(1 - q_i)}$$

$$= x_i \log \frac{p_i(1 - q_i)}{q_i(1 - p_i)} + \log \frac{(1 - p_i)}{(1 - q_i)}$$

Substituting these expressions into our decision rule yields:

Decide category 1 iff:

$$\sum_{i=1}^{n} x_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)} + \sum_{i=1}^{n} \log \frac{(1-p_i)}{(1-q_i)} + \log \frac{p(1)}{p(2)} \geq 0$$

We see that we can achieve this decision with a TLU with weight values as follows:

$$w_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

for $i = 1, \ldots, n$, and

$$w_{n+1} = \log \frac{p(1)}{1-p(1)} + \sum_{i=1}^{n} \log \frac{(1-p_i)}{(1-q_i)}$$

If we do not know the p_i, q_i and $p(1)$, we can use a sample of labeled training patterns to estimate these parameters.

5.2 Learning Belief Networks

To be added.

5.3 Nearest-Neighbor Methods

Another class of methods can be related to the statistical ones. These are called *nearest-neighbor* methods or, sometimes, *memory-based* methods. (A collection of papers on this subject is in [Dasarathy, 1991].) Given a training set Ξ of m labeled patterns, a nearest-neighbor procedure decides that some new pattern, $\mathbf{X}$, belongs to the same category as do its closest neighbors in Ξ. More precisely, a k-nearest-neighbor method assigns a new pattern, $\mathbf{X}$, to that category to which the plurality of its k closest neighbors belong. Using relatively large values of k decreases the chance that the decision will be unduly influenced by a noisy training pattern close to $\mathbf{X}$. But large values of k also reduce the acuity of the method. The k-nearest-neighbor method can be thought of as estimating the values of the probabilities of the classes given $\mathbf{X}$. Of course the denser are the points around $\mathbf{X}$, and the larger the value of k, the better the estimate.

The distance metric used in nearest-neighbor methods (for numerical attributes) can be simple Euclidean distance. That is, the distance between two patterns $(x_{11}, x_{12}, \ldots, x_{1n})$ and $(x_{21}, x_{22}, \ldots, x_{2n})$ is $\sqrt{\sum_{j=1}^{n}(x_{1j} - x_{2j})^2}$. This distance measure is often modified by *scaling* the features so that the spread of attribute values along each dimension is approximately the same. In that case, the distance between the two vectors would be $\sqrt{\sum_{j=1}^{n} a_j^2(x_{1j} - x_{2j})^2}$, where a_j is the scale factor for dimension j.

An example of a nearest-neighbor decision problem is shown in Fig. 5.3. In the figure the class of a training pattern is indicated by the number next to it.

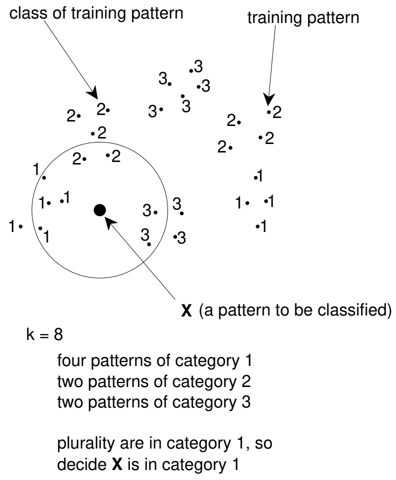

Figure 5.3: An 8-Nearest-Neighbor Decision

See [Baum, 1994] for theoretical analysis of error rate as a function of the number of training patterns for the case in which points are randomly distributed on the surface of a unit sphere and underlying function is linearly separable.

Nearest-neighbor methods are memory intensive because a large number of training patterns must be stored to achieve good generalization. Since memory cost is now reasonably low, the method and its derivatives have seen several practical applications. (See, for example, [Moore, 1992, Moore, *et al.*, 1994]. Also, the distance calculations required to find nearest neighbors can often be efficiently computed by *kd-tree* methods [Friedman, *et al.*, 1977].

A theorem by Cover and Hart [Cover & Hart, 1967] relates the performance of the 1-nearest-neighbor method to the performance of a minimum-probability-of-error classifier. As mentioned earlier, the minimum-probability-of-error classifier would assign a new pattern $\mathbf{X}$ to that category that maximized $p(i)p(\mathbf{X} \mid i)$, where $p(i)$ is the a priori probability of category i, and $p(\mathbf{X} \mid i)$ is the probability (or probability density function) of $\mathbf{X}$ given that $\mathbf{X}$ belongs to category i, for categories $i = 1, \ldots, R$. Suppose the probability of error in classifying patterns of such a minimum-probability-of-error classifier is ε. The Cover-Hart theorem states that under very mild conditions (having to do with the smoothness of probability density functions) the probability of error, ε_{nn}, of a 1-nearest-neighbor classifier is bounded by:

$$\varepsilon \leq \varepsilon_{nn} \leq \varepsilon\left(2 - \varepsilon\frac{R}{R-1}\right) \leq 2\varepsilon$$

where R is the number of categories.

Also see [Aha, 1991].

5.4 Bibliographical and Historical Remarks

To be added.