PATTERN CLASSIFICATION AND SCENE ANALYSIS

RICHARD O. DUDA
PETER E. HART

Stanford Research Institute,
Menlo Park, California

A WILEY-INTERSCIENCE PUBLICATION

JOHN WILEY & SONS
New York • Chichester • Brisbane • Toronto • Singapore

By substituting 0 or 1 for x_i and $x_{j(i)}$, the reader can verify that

$$P(x_i \mid x_{j(i)}) = [p_i^{x_i}(1 - p_i)^{1-x_i}]^{x_{j(i)}}[q_i^{x_i}(1 - q_i)^{1-x_i}]^{1-x_{j(i)}} \quad (62)$$

where

$$p_i = P(x_i = 1 \mid x_{j(i)} = 1) \quad (63)$$

and

$$q_i = P(x_i = 1 \mid x_{j(i)} = 0). \quad (64)$$

By letting $q_1 = P(x_1 = 1)$, substituting Eq. (62) in Eq. (61), taking the logarithm, and collecting terms, we obtain the *Chow expansion:*

$$\log P(x) = \sum_{i=1}^{d} \log(1 - q_i) + \sum_{i=1}^{d} x_i \log \frac{q_i}{1 - q_i} + \sum_{i=2}^{d} x_{j(i)} \log \frac{1 - p_i}{1 - q_i} + \sum_{i=2}^{d} x_i x_{j(i)} \log \frac{p_i(1 - q_i)}{(1 - p_i)q_i}. \quad (65)$$

Similar results for higher-order dependence can be obtained in an obvious way.

A few observations about these results are in order. First, we note that if the variables are indeed independent, $p_i = q_i$ and the last two sums in the expansion disappear, leaving the familiar expansion for the independent case. When dependence exists, we obtain additional linear and quadratic terms. Of course, the linear terms can be combined, so that the expansion effectively contains a constant, d linear terms, and $d - 1$ quadratic terms.

Comparing this with the second-order Rademacher-Walsh or Bahadur-Lazarsfeld expansions, either of which requires $d(d - 1)/2$ quadratic terms, we see that the savings can be appreciable. Of course, the savings can only be realized if we know the *dependence tree*, the function $j(i)$ which exhibits the limited dependence of one variable on preceding variables. If the dependence tree cannot be inferred from the physical significance of the variables, it may be necessary to compute all of the correlation coefficients merely to find the significant ones. However, even in this case it should be pointed out that one might prefer to use the Chow expansion because the resulting approximate probabilities are always nonnegative and sum to one.

4.10 FISHER'S LINEAR DISCRIMINANT

One of the recurring problems encountered in applying statistical techniques to pattern recognition problems is what Bellman calls the curse of dimensionality. Procedures that are analytically or computationally manageable in low-dimensional spaces can become completely impractical in a space of 50 or 100 dimensions. Thus, various techniques have been developed for

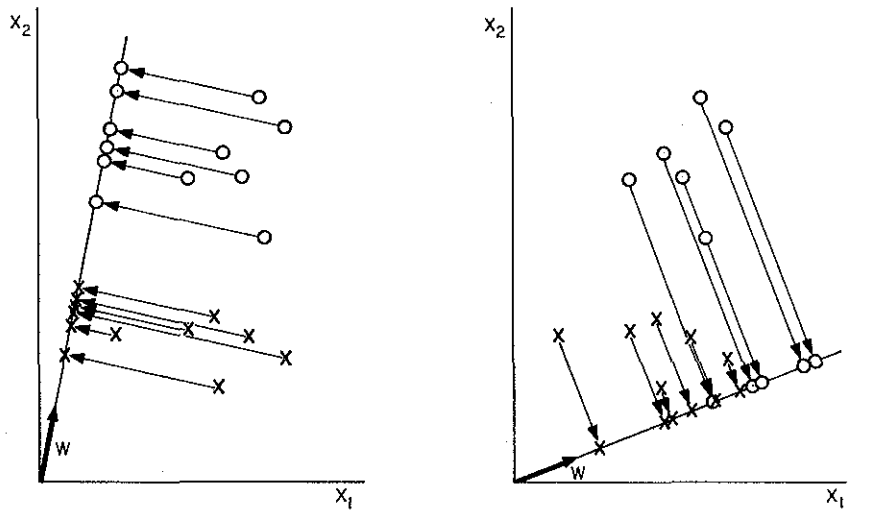

FIGURE 4.6. Projection of samples onto a line.

reducing the dimensionality of the feature space in the hope of obtaining a more manageable problem.

We can reduce the dimensionality from d dimensions to one dimension if we merely project the d-dimensional data onto a line. Of course, even if the samples formed well-separated, compact clusters in d-space, projection on an arbitrary line will usually produce a confused mixture of samples from all of the classes. However, by moving the line around, we might be able to find an orientation for which the projected samples are well separated. This is exactly the goal of classical discriminant analysis.

Suppose that we have a set of n d-dimensional samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$, n_1 in the subset $\mathscr{X}_1$ labelled ω_1 and n_2 in the subset $\mathscr{X}_2$ labelled ω_2. If we form a linear combination of the components of $\mathbf{x}$, we obtain the scalar

$$y = \mathbf{w}^t\mathbf{x} \quad (66)$$

and a corresponding set of n samples $y_1, \ldots, y_n$ divided into the subsets $\mathscr{Y}_1$ and $\mathscr{Y}_2$. Geometrically, if $\|\mathbf{w}\| = 1$, each y_i is the projection of the corresponding $\mathbf{x}_i$ onto a line in the direction of $\mathbf{w}$. Actually, the magnitude of $\mathbf{w}$ is of no real significance, since it merely scales y. The direction of $\mathbf{w}$ is important, however. If we imagine that the samples labelled ω_1 fall more or less in one cluster while those labelled ω_2 fall in another, we want the projections falling on the line to be well separated, not thoroughly intermingled. Figure 4.6 illustrates the effect of choosing two different values for $\mathbf{w}$ for a two-dimensional example.

A measure of the separation between the projected points is the difference of the sample means. If $\mathbf{m}_i$ is the d-dimensional sample mean given by

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{X}_i} \mathbf{x}, \tag{67}$$

then the sample mean for the projected points is given by

$$\begin{aligned} \tilde{m}_i &= \frac{1}{n_i} \sum_{y \in \mathcal{Y}_i} y \\ &= \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{X}_i} \mathbf{w}^t \mathbf{x} = \mathbf{w}^t \mathbf{m}_i. \end{aligned} \tag{68}$$

It follows that $|\tilde{m}_1 - \tilde{m}_2| = |\mathbf{w}^t(\mathbf{m}_1 - \mathbf{m}_2)|$, and that we can make this difference as large as we wish merely by scaling $\mathbf{w}$. Of course, to obtain good separation of the projected data we really want the difference between the means to be large relative to some measure of the standard deviations for each class. Rather than forming sample variances, we define the *scatter* for projected samples labelled ω_i by

$$\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2. \tag{69}$$

Thus, $(1/n)(\tilde{s}_1^2 + \tilde{s}_2^2)$ is an estimate of the variance of the pooled data, and $\tilde{s}_1^2 + \tilde{s}_2^2$ is called the total *within-class scatter* of the projected samples. The *Fisher linear discriminant* is then defined as that linear function* $\mathbf{w}^t\mathbf{x}$ for which the *criterion function*

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \tag{70}$$

is maximum.

To obtain J as an explicit function of $\mathbf{w}$, we define the *scatter matrices* S_i and S_W by

$$S_i = \sum_{\mathbf{x} \in \mathcal{X}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \tag{71}$$

and

$$S_W = S_1 + S_2. \tag{72}$$

Then

$$\begin{aligned} \tilde{s}_i^2 &= \sum_{\mathbf{x} \in \mathcal{X}_i} (\mathbf{w}^t\mathbf{x} - \mathbf{w}^t\mathbf{m}_i)^2 \\ &= \sum_{\mathbf{x} \in \mathcal{X}_i} \mathbf{w}^t(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t\mathbf{w} \\ &= \mathbf{w}^t S_i \mathbf{w}, \end{aligned} \tag{73}$$

* It should be noted that we are now using the term "discriminant function" to mean any function of $\mathbf{x}$ that is helpful in solving the decision problem; we do not insist that the resulting discriminant function be used directly to define the classifier. Because $y = \mathbf{w}^t\mathbf{x}$ is a sum of random variables, it is common to make reference to the central limit theorem and to assume that $p(y \mid \omega_i)$ is a normal density, thereby simplifying the problem of obtaining a classifier. When this assumption is not justified, one can still afford to use fairly elaborate methods to estimate $p(y \mid \omega_i)$ and derive an "optimal" classifier.

so that

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^t S_W \mathbf{w}. \tag{74}$$

Similarly,

$$\begin{aligned} (\tilde{m}_1 - \tilde{m}_2)^2 &= (\mathbf{w}^t\mathbf{m}_1 - \mathbf{w}^t\mathbf{m}_2)^2 \\ &= \mathbf{w}^t(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{w} \\ &= \mathbf{w}^t S_B \mathbf{w}, \end{aligned} \tag{75}$$

where

$$S_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t. \tag{76}$$

The matrix S_W is called the *within-class scatter matrix*. It is proportional to the sample covariance matrix for the pooled d-dimensional data. It is symmetric and positive semidefinite, and is usually nonsingular if $n > d$. S_B is called the *between-class scatter matrix*. It is also symmetric and positive semidefinite, but because it is the outer product of two vectors, its rank is at most one. In particular, for any $\mathbf{w}$, $S_B\mathbf{w}$ is in the direction of $\mathbf{m}_1 - \mathbf{m}_2$, and S_B is quite singular.

In terms of S_B and S_W, the criterion function J can be written as

$$J(\mathbf{w}) = \frac{\mathbf{w}^t S_B \mathbf{w}}{\mathbf{w}^t S_W \mathbf{w}}. \tag{77}$$

This expression is well known in mathematical physics as the generalized Rayleigh quotient. It is easy to show that a vector $\mathbf{w}$ that maximizes J must satisfy

$$S_B\mathbf{w} = \lambda S_W \mathbf{w} \tag{78}$$

which is a generalized eigenvalue problem. If S_W is nonsingular we can obtain a conventional eigenvalue problem by writing

$$S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w}. \tag{79}$$

In our particular case, it is unnecessary to solve for the eigenvalues and eigenvectors of $S_W^{-1}S_B$ due to the fact that $S_B\mathbf{w}$ is always in the direction of $\mathbf{m}_1 - \mathbf{m}_2$. Since the scale factor for $\mathbf{w}$ is immaterial, we can immediately write the solution

$$\mathbf{w} = S_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \tag{80}$$

Thus, we have obtained Fisher's linear discriminant, the linear function with the maximum ratio of between-class scatter to within-class scatter. The problem has been converted from a d-dimensional problem to a hopefully more manageable one-dimensional problem. This mapping is many-to-one, and in theory can not possibly reduce the minimum achievable error rate. In

neral, one is willing to sacrifice some of the theoretically attainable per-rmance for the advantages of working in one dimension. When the con-ional densities $p(\mathbf{x} \mid \omega_i)$ are multivariate normal with equal covariance atrices Σ, one need not even sacrifice any performance. In that case we call that the optimal decision boundary has the equation

$$\mathbf{w}^t\mathbf{x} + w_0 = 0$$

here

$$\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

d where w_0 is a constant involving $\mathbf{w}$ and the prior probabilities. If we use mple means and the sample covariance matrix to estimate $\boldsymbol{\mu}_i$ and Σ, we tain a vector in the same direction as the $\mathbf{w}$ of Eq. (80) that maximizes J. aus, for the normal, equal-covariance case, the optimal decision rule is erely to decide ω_1 if Fisher's linear discriminant exceeds some threshold, d to decide ω_2 otherwise.

11 MULTIPLE DISCRIMINANT ANALYSIS

or the c-class problem, the natural generalization of Fisher's linear dis-iminant involves $c - 1$ discriminant functions. Thus, the projection is from d-dimensional space to a $(c - 1)$-dimensional space, and it is tacitly sumed that $d \geq c$. The generalization for the within-class scatter matrix is vious:

$$S_W = \sum_{i=1}^{c} S_i \tag{81}$$

here, as before,

$$S_i = \sum_{\mathbf{x}\in\mathcal{X}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \tag{82}$$

nd

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x}\in\mathcal{X}_i} \mathbf{x}. \tag{83}$$

The proper generalization for S_B is not quite so obvious. Suppose that we efine a *total mean vector* $\mathbf{m}$ and a *total scatter matrix* S_T by

$$\mathbf{m} = \frac{1}{n}\sum_{\mathbf{x}} \mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i\mathbf{m}_i \tag{84}$$

nd

$$S_T = \sum_{\mathbf{x}} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^t. \tag{85}$$

Then it follows that

$$\begin{aligned} S_T &= \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{X}_i} (\mathbf{x} - \mathbf{m}_i + \mathbf{m}_i - \mathbf{m})(\mathbf{x} - \mathbf{m}_i + \mathbf{m}_i - \mathbf{m})^t, \\ &= \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{X}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t + \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{X}_i} (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t \\ &= S_W + \sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t. \end{aligned}$$

It is natural to define this second term as the between-class scatter matrix, so that the total scatter is the sum of the within-class scatter and the between-class scatter:

$$S_B = \sum_{i=1}^{c} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^t \tag{86}$$

and

$$S_T = S_W + S_B. \tag{87}$$

If we check the two-class case, we find that the resulting between-class scatter matrix is n_1n_2/n times our previous definition. We could redefine S_B for the two-class case to obtain complete consistency, but we shall recall Emerson's remark that a foolish consistency is the hobgoblin of little minds and proceed.

The projection from a d-dimensional space to a $(c - 1)$-dimensional space is accomplished by $c - 1$ discriminant functions

$$y_i = \mathbf{w}_i^t\mathbf{x} \qquad i = 1, \ldots, c - 1. \tag{88}$$

If the y_i are viewed as components of a vector $\mathbf{y}$ and the weight vectors $\mathbf{w}_i$ are viewed as the columns of a d-by-$(c - 1)$ matrix W, then the projection can be written as a single matrix equation

$$\mathbf{y} = W^t\mathbf{x}. \tag{89}$$

The samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ project to a corresponding set of samples $\mathbf{y}_1, \ldots, \mathbf{y}_n$ which can be described by their own mean vectors and scatter matrices. Thus, if we define

$$\tilde{\mathbf{m}}_i = \frac{1}{n_i}\sum_{\mathbf{y}\in\mathcal{Y}_i} \mathbf{y} \tag{90}$$

$$\tilde{\mathbf{m}} = \frac{1}{n}\sum_{i=1}^{c} n_i\tilde{\mathbf{m}}_i \tag{91}$$

$$\tilde{S}_W = \sum_{i=1}^{c} \sum_{\mathbf{y}\in\mathcal{Y}_i} (\mathbf{y} - \tilde{\mathbf{m}}_i)(\mathbf{y} - \tilde{\mathbf{m}}_i)^t \tag{92}$$

and

$$\tilde{S}_B = \sum_{i=1}^{c} n_i(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})^t \tag{93}$$