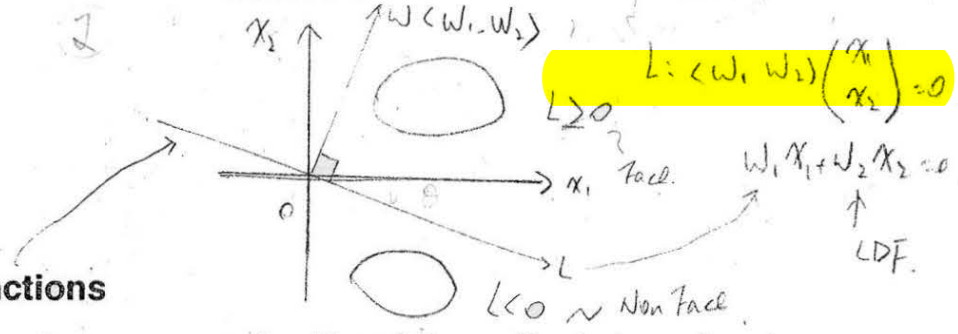
Linear Discriminant Functions and the Discrete and Binary Feature Cases

4

Art is the imposing of a pattern on experience, and our aesthetic enjoyment in recognition of the pattern.

Dialogues of Alfred North Whitehead [1953], Chapter 29
Alfred North Whitehead 1861–1947



4.1 INTRODUCTION

Linear Discriminant Functions

In this section we generalize the concept and utility of *linear discriminant functions (LDFs)*. Although LDFs are a fundamental concept in StatPR, they are also useful to introduce certain types of neural network structures and training formulations. Many StatPR models and associated classification strategies lead to discriminant functions that are of the form:

$$g_i(\underline{x}) = \underline{w}_i^T \underline{x} + w_{i0} \quad i = 1, 2, \ldots c \tag{4-1}$$

The decision regions are then separated by hyperplanes (Appendix 6). Conceptually, a d-dimensional feature vector, $\underline{x}$, is reduced to a single dimension or number that is then used for classification.[1] Throughout this chapter we consider the $c = 2$ class case, which may be extended by considering $c > 2$ classes pairwise.

Figure 1 summarizes the two major approaches considered in this chapter. The first involves projection of the d-dimensional data onto an appropriate line, thereby reducing the feature data to a single measurement. We show several approaches for

[1] This is one version of a problem in *Multiple Discriminant Analysis* where feature data are projected from R^d into R^q, where typically $q \ll d$.

determining the parameters of this line in order to obtain 'good' discrimination (classification) ability. The second approach involves determination of a suitable separating hyperplane in R^d. Both iterative and 'batch' solutions for the determination of this plane from the sample data are considered. In either case, the solution structure remains the same irrespective of whether the underlying statistical characterizations and the Bayesian approach lead to such a solution; that is, we are forcing these structures on the solution.

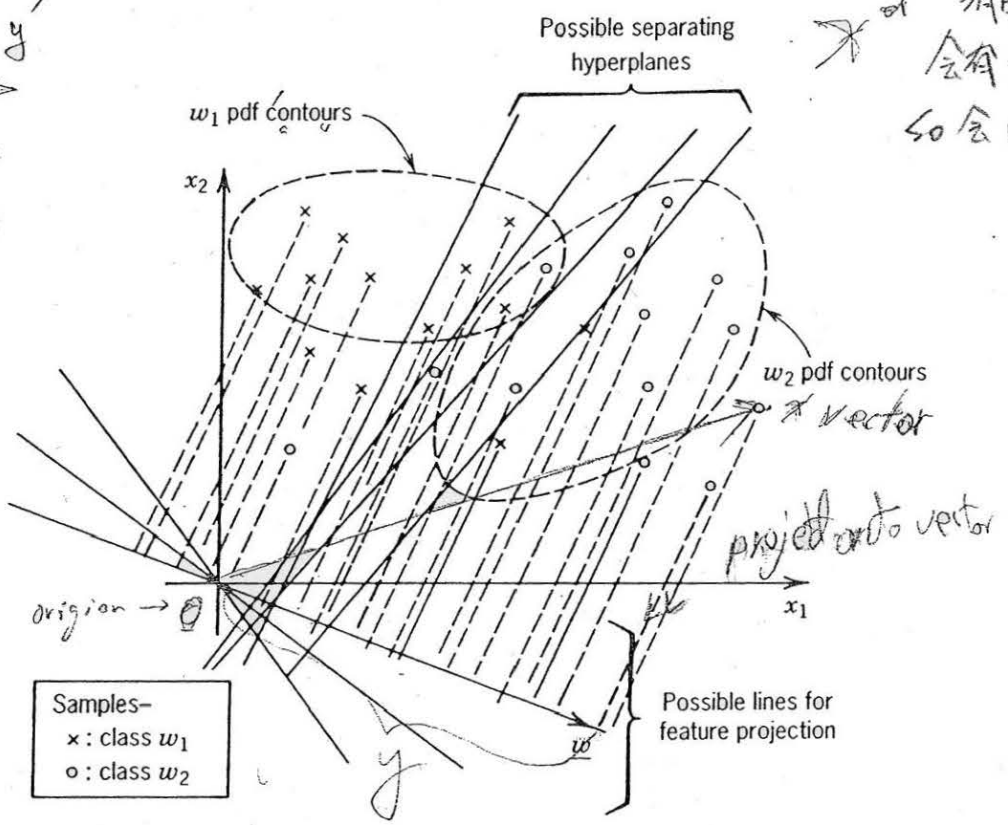


Figure 1: The $d = 2$ example of chapter objectives.

Fisher's Linear Discriminant

The Fisher approach [Fisher 1936] is based on *projection* of d-dimensional data onto a *line*. The hope is that these projections onto a line will be well *separated by class*. Thus, the line is oriented to maximize this class separation. The real utility of this type of an approach is when the size of the feature vector is very large, for example $d = 50,\ 100,$ or larger. We begin with a $c = 2$ class example. The given training set

$$H = \{\underline{x}_1, \underline{x}_2, \ldots \underline{x}_n\} = \{H_1,\ H_2\} \qquad (4-2)$$

is partitioned into $n_1 \leq n$ training vectors in subset H_1, corresponding to class w_1, and $n_2 \leq n$ training vectors in set H_2, corresponding to class w_2, where $n_1 + n_2 = n$. The feature vector projections are formed via

$$y_i = \underline{w}^T \underline{x}_i =< \underline{w}, \underline{x}_i > \quad i = 1, 2, \ldots n \qquad (4-3)$$

If we further constrain $\| \underline{w} \| = 1$, each y_i is the projection of $\underline{x}_i$ onto a line *in the direction of* $\underline{w}$. Note that this line always goes through the origin in R^d. The problem is to choose the direction of $\underline{w}$, given H, such that y_i from H_1 and y_i from H_2 fall into (ideally) distinct clusters along the line, denoted Y_1 and Y_2. Ideally, if $y_i \in Y_1$, then $\underline{x}_i \in H_1$ and if $y_i \in Y_2$, then $\underline{x}_i \in H_2$.

Measures of Projected Data Class Separation. One *measure of separation* of the projections is the *difference of the means of the projections*. For example, $|\mu_{Y_1} - \mu_{Y_2}|^2$ is such a measure, where, using (4-3)

$$\mu_{Y_i} = E\{y_i | \underline{x}_i \in w_i\} = E\{\underline{w}^T \underline{x}_i | \underline{x}_i \in w_i\} \qquad (4-4)$$

This measure may be shown to be related to the H_1, H_2 sample means through $\underline{w}$:

$$\underline{m}_i = \frac{1}{n_i} \sum_{\underline{x}_i \in H_i} \underline{x}_i \qquad (4-5a)$$

Thus, the *projection mean* for each class is a scalar, given by

$$\bar{m}_i = \frac{1}{n_i} \sum_{\underline{x}_i \in H_i} \underline{w}^T \underline{x}_i = \frac{1}{n_i} \sum_{y_i \in \underline{y}_i} y_i$$

$$= \underline{w}^T \frac{1}{n_i} \sum_{\underline{x}_i \in H_i} \underline{x}_i = \underline{w}^T \underline{m}_i \qquad (4-5b)$$

where $\underline{m}_i$ is the sample mean of the vectors in H_i. The difference of the projection means using sample data is therefore

$$|\bar{m}_1 - \bar{m}_2| = |\underline{w}^T (\underline{m}_1 - \underline{m}_2)| \qquad (4-5c)$$

The difference between the means of the projected data alone is insufficient for a good classifier, as is shown in Figure 2. Although we want well-separated (class) projections, *they should not be intermingled*. To achieve this, we need to consider variances of y_i in Y_i *relative* to the means. Therefore, a better class separation measure is the ratio (difference of means)/(variance of within-class data). For example, a reasonable criterion in the $c = 2$ case is

$$J(\underline{w}) = \frac{(\mu_{Y_1} - \mu_{Y_2})^2}{\sigma_{Y_1}^2 + \sigma_{Y_2}^2} \qquad (4-6a)$$

or, in the case of sample data,

$$J(\underline{w}) = \frac{(\bar{m}_1 - \bar{m}_2)^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} \qquad (4-6b)$$

where $\hat{\sigma}_i^2$ is a measure of within-class scatter of the projected data. Instead of using variances in (4-6b), we could define the within-class scatter of the projection data as

$$\bar{s}_i^2 = \sum_{y \in Y_i} (y - \bar{m}_i)^2 \qquad (4-7a)$$

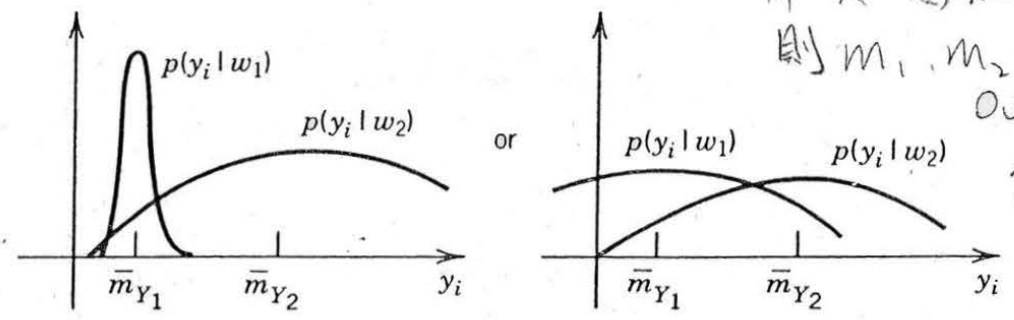


Figure 2: The importance of defining good separation measures for projected data.

Recall that $\hat{\sigma}^2 = [1/(n-1)] \sum_{i=1}^{n} (x_i - m_i)^2$ is an unbiased variance estimator. Therefore, for n_i samples in H_i (from w_i)

$$\bar{s}_i^2 \approx (n_i - 1)\hat{\sigma}_i^2 \qquad (4-7b).$$

Fisher showed that a reasonable measure of projected data separability is the criterion function that is a scaled version of (4-6b):

$$J(\underline{w}) = \frac{(\bar{m}_1 - \bar{m}_2)^2}{\bar{s}_1^2 + \bar{s}_2^2} \qquad (4-8)$$

The value of $\underline{w}$ that maximizes (4-6a), (4-6b), or (4-8) is used in a linear discriminant function (of the form $\underline{w}^T \underline{x}$) to yield *Fisher's linear discriminant*. To proceed further, it is necessary to rewrite J in (4-8) as an *explicit function* of $\underline{w}$. Notice $\underline{w}$ influences *both* the numerator and the denominator of (4-8), since both m_i and s_i are functions of y_i. We therefore have created an optimization problem, that is, to determine the direction of $\underline{w}$ such that the criterion of (4-8) is maximum.

Solution for Criteria of (4-6a) (Exact Means and Covariances Known). Denoting the mean and covariance of the d-dimensional vectors in class w_i as $\underline{\mu}_i$ and Σ_i respectively, it is straightforward to show

$$\mu_{Y_i} = \underline{w}^T \underline{\mu}_i \qquad i = 1, 2 \qquad (4-9a)$$

and

$$\sigma_{Y_i}^2 = \underline{w}^T \Sigma_i \underline{w} \qquad i = 1, 2 \qquad (4-9b)$$

A maximum of (4-6a) is found by setting $\partial J / \partial \underline{w} = \underline{0}$. This requires

$$\frac{\partial J}{\partial \underline{w}} = \frac{\partial J}{\partial \mu_{Y_1}} \frac{\partial \mu_{Y_1}}{\partial \underline{w}} + \frac{\partial J}{\partial \mu_{Y_2}} \frac{\partial \mu_{Y_2}}{\partial \underline{w}} + \frac{\partial J}{\partial \sigma_{Y_1}^2} \frac{\partial \sigma_{Y_1}^2}{\partial \underline{w}} + \frac{\partial J}{\partial \sigma_{Y_2}^2} \frac{\partial \sigma_{Y_2}^2}{\partial \underline{w}} = \underline{0} \qquad (4-9c)$$

Equation 4-9c together with (4-9a) and (4-9b) yields the constraint on the optimal $\underline{w}$, denoted $\hat{\underline{w}}$:

$$\frac{2(\mu_{Y_1} - \mu_{Y_2})}{\sigma_{Y_1}^2 + \sigma_{Y_2}^2} (\underline{\mu}_1 - \underline{\mu}_2) - \frac{2(\mu_{Y_1} - \mu_{Y_2})^2}{(\sigma_{Y_1}^2 + \sigma_{Y_2}^2)^2} (2\Sigma_1 + 2\Sigma_2)\hat{\underline{w}} = \underline{0} \qquad (4-9d)$$

or

$$\hat{w} = \frac{1}{2}k(\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2) \qquad (4-9e)$$

where

$$k = \frac{\sigma_{Y_1}^2 + \sigma_{Y_2}^2}{\mu_{Y_1} - \mu_{Y_2}} \qquad (4-9f)$$

is merely a scale term that affects only $\| w \|$. Equation 4-9e is intuitively pleasing and bears a striking similarity to Bayesian results in the Gaussian case (Chapter 2). Therefore, by estimating Σ_i and μ_i from H_i, the direction of $\hat{w}$ may be determined.

Solution for Criterion of (4-8), Based on Sample Data. An alternative procedure is now shown. Defining a *scatter matrix* S_i as

$$S_i = \sum_{x \in H_i} (x - m_i)(x - m_i)^T \qquad i = 1, 2 \qquad (4-10a)$$

and

$$S_W = S_1 + S_2 \qquad (4-10b)$$

the denominator of (4-8) may be formulated as

$$J(\hat{w}) = \tilde{s}_1^2 + \tilde{s}_2^2 = w^T S_W w \qquad (4-10c)$$

Similarly, the numerator of (4-8), using (4-5c), may be rewritten in terms of the sample means as

$$(\tilde{m}_1 - \tilde{m}_2)^2 = w^T (m_1 - m_2)(m_1 - m_2)^T w = w^T S_B w \qquad (4-10d)$$

where S_B is the *between-class scatter matrix*. Since S_B is the outer product of a vector with itself, it has rank one. Therefore, (4-8) becomes

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \qquad (4-11a)$$

where the sample data in H_1 and H_2 are used to determine S_W and S_B. Forming $\partial J / \partial w = 0$ leads to

$$S_W \hat{w} (\hat{w}^T S_B \hat{w})(\hat{w}^T S_W \hat{w})^{-1} = S_B \hat{w} \qquad (4-11b)$$

which yields a *generalized eigenvector problem*, where the scalar term $(\hat{w}^T S_B \hat{w})(\hat{w}^T S_W \hat{w})^{-1} = \lambda$. Thus, we seek a solution to

$$\lambda S_W \hat{w} = S_B \hat{w} \qquad (4-11c)$$

If S_W^{-1} exists, a simple solution for the *direction* of $\hat{w}$ is

$$\hat{w} = (S_W^{-1} S_B)\hat{w} \qquad (4-11d)$$

and a solution for $\underline{w}$ may be found by solving for an e-vector of $(S_W^{-1}S_B)$. An alternative solution is based on the fact that $S_B\hat{\underline{w}}$, in (4-11b), has direction $\underline{m}_1 - \underline{m}_2$, since $(\underline{m}_1 - \underline{m}_2)(\underline{m}_1 - \underline{m}_2)^T\hat{\underline{w}} = (\underline{m}_1 - \underline{m}_2)k$. Therefore,

$$\hat{\underline{w}} = S_W^{-1}(\underline{m}_1 - \underline{m}_2) \qquad (4-12)$$

Figure 3 shows an example of the Fisher approach in the $c = 2$ class case.

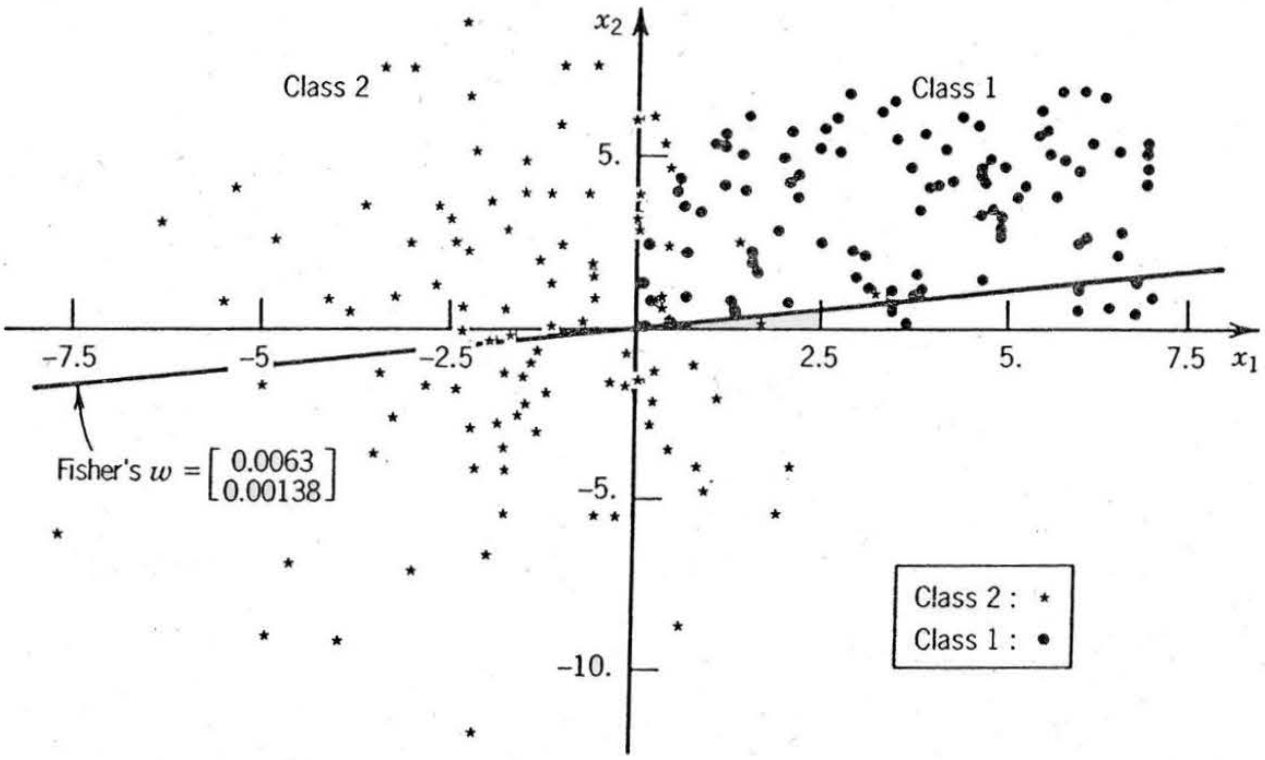


Figure 3: An example of Fisher's linear discriminant function. The line onto which feature data are projected is shown. (The line passes through the origin.)

4.2 DISCRETE AND BINARY CLASSIFICATION PROBLEMS

Classification Procedures for Discrete Feature Data

Previous formulations assumed continuous-valued features. When features are modeled as realizations of discrete random variables, density functions are replaced by probabilities (Appendix 2). Therefore, Bayes rule in the case of a discrete feature vector becomes

$$P(w_j|\underline{x}) = \frac{P(\underline{x}|w_j)P(w_j)}{P(\underline{x})} \qquad (4-13)$$

where

$$P(\underline{x}) = \sum_{i=1}^{c} P(\underline{x}|w_j)P(w_i) \qquad (4-14)$$