

Note on Fisher Linear Discriminants

Pietro Perona
 California Institute of Technology
 perona@caltech.edu

April 8, 2000

1 Introduction

We have N data points $X_i \in \mathbb{R}^D$. The points belong to G groups. Consider the problem of finding a linear transformation a to one-dimensional space such that the points $X_i a = Z_i \in \mathbb{R}$ are easy to classify. For simplicity we will assume that the N data points X_i have zero mean.

Some notation: given a matrix A indicate with A_i and A^j the i -th row and the j -th column of A , and with A_i^j the i, j -th element of A . Moreover:

$$[i] = g \quad g \text{ is the group of point } i \quad (1)$$

$$X = \begin{bmatrix} X_1 \\ \dots \\ X_N \end{bmatrix} \in \mathbb{R}^{N \times D} \quad \text{the data} \quad (2)$$

$$n = [n_1, \dots, n_G] \quad \text{number of data points in each group} \quad (3)$$

$$N = \text{diag}(n) = G^T G = \begin{bmatrix} n_1, 0, 0, \dots, 0 \\ 0, n_2, 0, \dots, 0 \\ \dots \\ 0, 0, \dots, 0, n_g \end{bmatrix} \quad (4)$$

$$G_i^j = \delta([i], j) \quad i.e. \quad G = \begin{bmatrix} 1, 0, \dots, 0 \\ \dots \\ 0, \dots, 0, 1 \end{bmatrix} \in \mathbb{R}^{N \times G} \quad (5)$$

$$M_g = \frac{1}{n_g} \sum_{[i]=g} X_i \quad \text{Mean of } j\text{-th coordinate in group } g \quad (6)$$

$$M = N^{-1} G^T X = \begin{bmatrix} M_1 \\ \dots \\ M_G \end{bmatrix} \in \mathbb{R}^{G \times D} \quad \text{Matrix collecting the means of each group} \quad (7)$$

$$A = U_A L_A V_A^T \quad \text{the singular value decomposition of } A \quad (8)$$

$$D = I - GN^{-1}G^T \quad (9)$$

$$X_0 = X - GM = DX \quad \text{the data, each referred to group's mean} \quad (10)$$

2 Optimization problem

In order for the points Z_i to be easy to classify one would like to simultaneously maximize the *between-clusters distance* and minimize the *within cluster distance*. These quantities may be defined as:

Between-clusters distance – Consider the means M_g of each group g . One would like to maximize their spread around the overall mean (the origin, since X is zero-mean):

$$B = (GM)^T(GM) = X^T G N^{-1} G^T X \quad (11)$$

Notice that each mean M_g is counted n_g times in order to reflect the frequency of group g .

Within-clusters distance – Consider the spread of the points around each group's center. One would like to minimize:

$$W = X_0^T X_0 = X^T D^T D X \quad (12)$$

In order to optimize both quantities simultaneously Fisher proposed to maximize their ratio with respect to the transformation a :

$$J(a) = \frac{a^T B a}{a^T W a} \quad (13)$$

Taking the derivative with respect to a and equating to zero:

$$DJ(a) = \frac{2Baa^T W a - 2Waa^T B a}{(a^T W a)^2} = 0 \quad (14)$$

$$\text{define } \lambda \doteq \frac{a^T B a}{a^T W a} \quad (15)$$

$$\Rightarrow B a = \lambda W a \quad a^T W a \neq 0 \quad (16)$$

$$(17)$$

Therefore in order to find the value of a we need to solve the *generalized eigenvector problem* $B a = \lambda W a$ subject to $a^T W a \neq 0$.

2.1 Eigenvector problem

Call W^\dagger the generalized inverse of W , i.e. the inverse restricted to the subspace where W is nonsingular. Then:

$$B a = \lambda W a \quad a^T W a \neq 0 \quad (18)$$

$$W^\dagger = V_{X_0} L_{X_0}^{\dagger 2} V_{X_0}^T \quad (19)$$

$$\Rightarrow W^\dagger B a = \lambda a \quad (20)$$

$$\text{define}(U_{WB}, L_{WB}, V_{WB}) \doteq \text{SVD}(W^\dagger B) \quad (21)$$

$$\Rightarrow a = V_{WB}^1 \quad (22)$$

2.2 Alternative approach

An equivalent approach consists in calculating a coordinate transformation $a = Sb$ such that $a^T W a = \|b\|^2$. In this case one may calculate the b that maximizes the numerator, subject to $\|b\| = 1$. One must, however, pay attention to the fact that the solution b must not be in the null space of S .

From the definition of W and B etc.:

$$W = X_0^T X_0 = V_{X_0} L_{X_0}^2 V_{X_0}^T \quad (23)$$

$$S = V_{X_0} L_{X_0}^{-1} \quad (24)$$

$$a = Sb \quad (25)$$

$$b = S^{-1}a = L_{X_0} V_{X_0}^T a \quad (26)$$

$$a^T W a = a^T V_{X_0} L_{X_0}^2 V_{X_0}^T a = b^T b \quad (27)$$

$$a^T B a = b^T S^T B S b \quad (28)$$

$$(U, L, V) \doteq \text{SVD}(S^T B S) \quad (29)$$

$$\Rightarrow b = V^1 \quad (30)$$

$$\Rightarrow a = S V^1 = V_{X_0} L_{X_0}^{-1} V^1 \quad (31)$$

3 Code and References

Check out the `Matlab` function `fisherLD.m` written by Markus Weber. A prize to whoever figures out how the code works.

You will find Fisher linear discriminants discussed in B. Ripley *Pattern recognition and neural networks*, Cambridge University Press, 1996 (SFL library, call n. **qa 76.87 r56 1996**).