Linear Discriminant Analysis

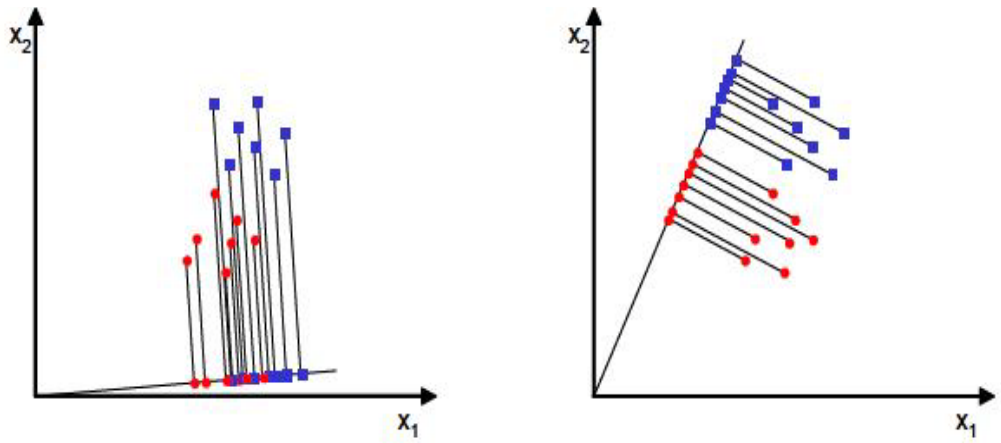
Linear discriminant analysis (LDA) and the related **Fisher's linear discriminant** are methods used in statistics, pattern recognition and machine learning to find a linear combination of features which characterize or separate two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

LDA is also closely related to principal component analysis (PCA) and factor analysis in that both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique.
The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible

Assume we have a set of D-dimensional samples {*x(1, x(2, …, x(N*}, N1 of which belong to class ω1, and N2 to class ω2. We seek to obtain a scalar ***y*** by projecting the samples ***x*** onto a line

$$Y = w^T x$$

Of all the possible lines we would like to select the one that maximizes the separability of the scalars. This is illustrated for the two-dimensional case in the following figures



In order to find a good projection vector, we need to define a measure of separation between the projections

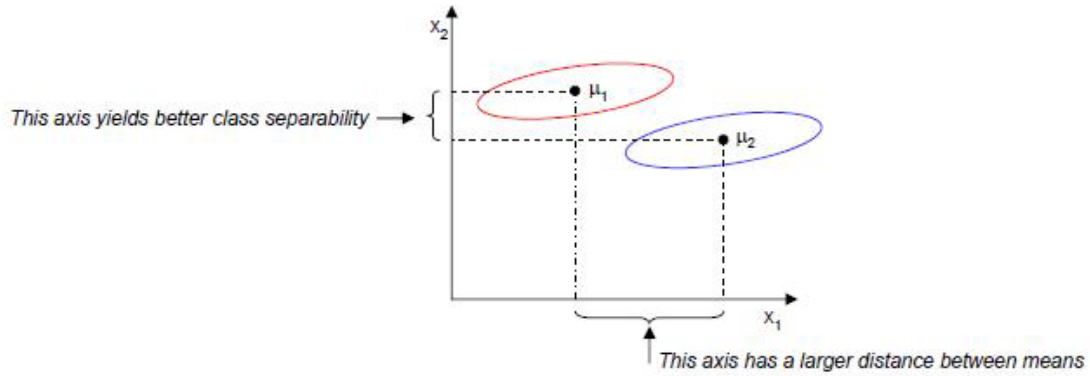
The mean vector of each class in ***x*** and ***y*** feature space is

$$\mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x \quad \text{and} \quad \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i} w^T x = w^T \mu_i$$

We could then choose the distance between the projected means as our objective Function

$$J(w) = \left|\tilde{\mu}_1 - \tilde{\mu}_2\right| = \left|w^T(\mu_1 - \mu_2)\right|$$

However, the distance between the projected means is not a very good measure since it does not take into account the standard deviation within the classes



The solution proposed by Fisher is to maximize a function that represents the difference between the means, normalized by a measure of the within-class scatter

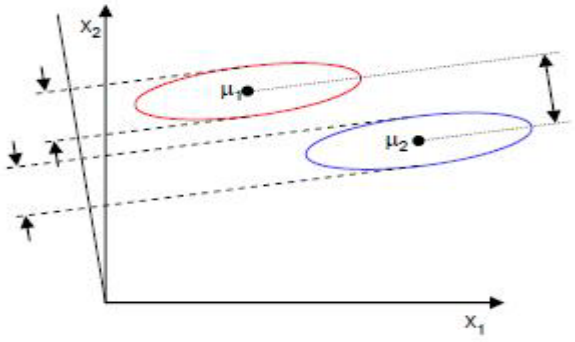
For each class we define the scatter, an equivalent of the variance, as

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2$$

where the quantity $\left(\tilde{s}_1^2 + \tilde{s}_2^2\right)$ is called the within-class scatter of the projected examples
The Fisher linear discriminant is defined as the linear function **wTx** that maximizes the criterion function

$$J(w) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

Therefore, we will be looking for a projection where examples from the same class are projected very close to each other and, at the same time, the projected means are as farther apart as possible



In order to find the optimum projection w*, we need to express J(w) as an explicit function of w

We define a measure of the scatter in multivariate feature space **x**, which are scatter matrices

$$S_i = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$

$$S_1 + S_2 = S_W$$

where SW is called the **within-class scatter matrix**

The scatter of the projection **y** can then be expressed as a function of the scatter matrix in feature space **x**

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2 = \sum_{x \in \omega_i} (w^T x - w^T \mu_i)^2 = \sum_{x \in \omega_i} w^T (x - \mu_i)(x - \mu_i)^T w = w^T S_i w$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^T S_W w$$

Similarly, the difference between the projected means can be expressed in terms of the means in the original feature space

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T \underbrace{(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T}_{S_B} w = w^T S_B w$$

The matrix SB is called the **between-class scatter**. Note that, since SB is the outer product of two vectors, its rank is at most one

We can finally express the Fisher criterion in terms of SW and SB as

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

To find the maximum of J(w) we derive and equate to zero

$$\frac{d}{dw}[J(w)] = \frac{d}{dw}\left[\frac{w^T S_B w}{w^T S_W w}\right] = 0 \Rightarrow$$

$$\Rightarrow [w^T S_W w]\frac{d[w^T S_B w]}{dw} - [w^T S_B w]\frac{d[w^T S_W w]}{dw} = 0 \Rightarrow$$

$$\Rightarrow [w^T S_W w] 2 S_B w - [w^T S_B w] 2 S_W w = 0$$

Dividing by $w^T S_W W$

$$\left[\frac{w^T S_W w}{w^T S_W w}\right] S_B w - \left[\frac{w^T S_B w}{w^T S_W w}\right] S_W w = 0 \Rightarrow$$

$$\Rightarrow S_B w - J S_W w = 0 \Rightarrow$$

$$\Rightarrow S_W^{-1} S_B w - J w = 0$$

- Solving the generalized eigenvalue problem (SW-1SBw=Jw) yields

$$w^* = \underset{w}{\text{argmax}}\left\{\frac{w^T S_B w}{w^T S_W w}\right\} = S_W^{-1}(\mu_1 - \mu_2)$$

This is known as **Fisher's Linear Discriminant** (1936), although it is not a discriminant but rather a specific choice of direction for the projection of the data down to one dimension.