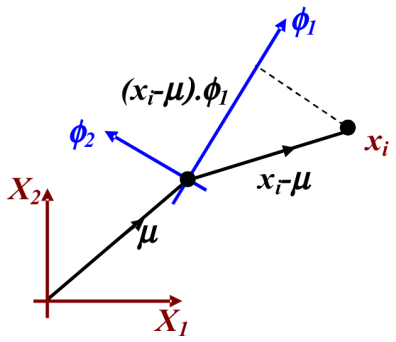
Lecture 15: Linear Discriminant Analysis

In the last lecture we viewed PCA as the process of finding a projection of the covariance matrix. This projection is a transformation of data points from one axis system to another, and is an identical process to axis transformations in graphics. The vector $\boldsymbol{x_i}$ in the original space becomes the vector $\boldsymbol{x_i} - \boldsymbol{\mu}$ in the transformed space. The individual ordinates are the projected lengths of the vector on each of the axes of the transformed space, and this is readily computed using the dot product.



The mean centred data matrix, where each column is one variable, is denoted U and the covariance matrix can be formed from the product:

$$\Sigma = (1/(N-1))U^T U$$

A full projection is defined by a matrix in which each column is a vector defining the direction of one of the new axes:

$$\Phi = [\phi_1, \phi 2, \phi 3 \cdots \phi_m]$$

The projection basis vectors ϕ_i are orthonormal which means that

$$\Phi^T \Phi = I$$

The projection of the data in mean adjusted form can be written:

$$Y = U\Phi$$

Projection of the covariance matrix is

$$\begin{aligned} \Phi^T \Sigma \Phi &= \Phi^T (1/(N-1)) U^T U \Phi \\ &= (1/(N-1)) \Phi^T U^T U \Phi \\ &= (1/(N-1))(U\Phi)^T (U\Phi) \end{aligned}$$

which can be seen to be the covariance matrix of the projected points.

The projections can equally well be found using scatter matrices rather than co-varainace matrices. A scatter matrix is un-normalised, and defined using the mean centered data matrix as before:

$$S = U^T U$$

It is sometimes convenient to write expressions for covariance (or scatter) matrices in a different form, thus:

$$\Sigma = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \overline{x})(x_j - \overline{x})^T$$

Here, the sum is taken over the number N of data points, which are expressed as a vector x_j of dimension n. μ is the mean vector also of dimension n. The implied product is the outer product, not the dot product.

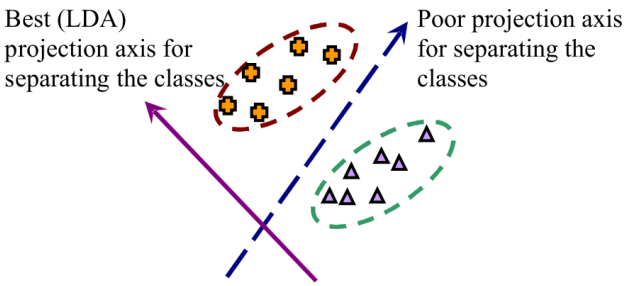
The PCA projection is just one possible projection and it has the property that it diagonalises the covariance matrix. That is it transforms the points such that they are independent of each other. It is found by determining the eigenvectors of Σ:

$$\Phi^T \Sigma \Phi = \Lambda$$

We will now look at a different projection which is the basis of a classification process called linear discriminant analysis LDA.

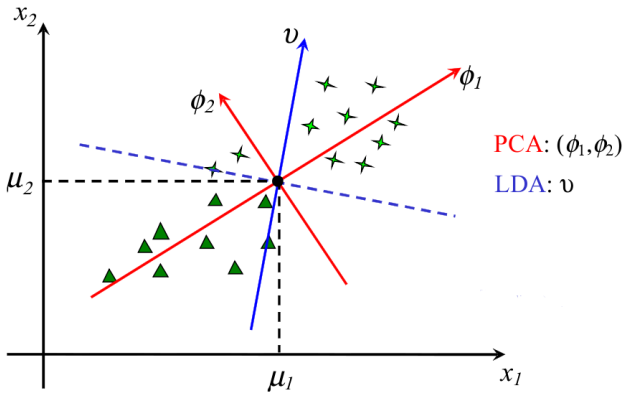
Linear Discriminant Analysis, or simply LDA, is a well-known classification technique that has been used successfully in many statistical pattern recognition problems. It was developed by Ronald Fisher, who was a professor of statistics at University College London, and is sometimes called Fisher Discriminant Analysis (FDA). The primary purpose of LDA is to separate samples of distinct groups. We do this by transforming the data to a different space that is optimal for distinguishing between the classes.

For example, let us suppose we have a simple two variable problem in which there are two classes of objects.



Having made many measurements from examples of the two classes we can find the best Gaussian model for each class. The nominal class boundaries (where the probability of a point belonging to the class falls below a certain limit) are shown by the ellipses. Now, given a new measurement, we want to determine which class it belongs to. One approach might be to compute the Mahalanobis distance between the unknown point and each class mean, and then pick the nearest class. This approach however is computationally very expensive for high dimensional problems since we need to invert the co-variance matrix for each class. For small sample size problems this may not be possible at all since we may not have enough data to make a full rank estimate of the co-variance matrix.

Instead, the approach of the LDA is to project all the data points into new space, normally of lower dimension, which maximises the between-class separability while minimising their within-class variability. In the example above we do this by projecting the points onto the solid axis. Each class will form a single dimensional Gaussian on that axis and the means will be well separated. The variance is also low on that axis, meaning that overall we can classify a new point quickly and easily. In contrast if we were to project the two classes onto the dashed axis, then the resulting distributions would overlap considerably and each class would have high variability, meaning that classification would be very difficult. LDA generalises this process for multiple classes and arbitrarily large numbers of variables. Its main limitation is the implicit assumption that the true covariance matrices of each class are the same.



The above figure illustrates the difference between PCA and LDA. In PCA we take the data as a whole and do not consider any division into classes. The axes are optimal for representing the data as they indicate where the maximum variation actually lies. The LDA axis is optimal for distinguishing between the different classes. In

general the number of axes that can be computed by the LDA method is one less than the number of classes in the problem. In the figure the ϕ_1 axis is less effective for separating the classes than the ν (LDA) axis. The ϕ_2 axis is clearly completely useless for classification.

The first step in the LDA is finding two scatter matrices referred to as the "between class" and "within class" scatter matrices. Suppose in a given problem we have g different classes or (sample groups). Each sample group π_i has a class mean, which we denote $\overline{x_i}$

$$\overline{x_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{i,j}$$

where there are N_i data points in class π_i. We can also define a sample group covariance matrix:

$$\Sigma_i = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (x_{i,j} - \overline{x_i})(x_{i,j} - \overline{x_i})^T$$

and we can define a grand mean for the whole data set:

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{g} N_i \overline{x_i} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{N_i} x_{i,j}$$

The between class scatter matrix is defined as:

$$S_b = \sum_{i=1}^{g} N_i (\overline{x_i} - \overline{x})(\overline{x_i} - \overline{x})^T$$

We could normalise this into a between class covariance matrix by dividing by $N - 1$, but it is not necessary to do so. If all classes were the same size then the $N - i$ could be removed from the equation. The within class matrix is defined as follows:

$$S_w = \sum_{i=1}^{g} (N_i - 1)\Sigma_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \overline{x_i})(x_{i,j} - \overline{x_i})^T$$

The S_w matrix is computed by pooling the estimates of the covariance matrices of each class. Since each Σ_i has rank $N_i - 1$ its rank can be at most $N - g$.

The main objective of LDA is to find a projection matrix Φ_{lda} that maximises the ratio of the determinant of S_b to the determinant of S_w. This can be written:

$$\Phi_{lda} = \arg\max_{\Phi} \frac{|\Phi^T S_b \Phi|}{|\Phi^T S_w \Phi|}$$

This ratio is known as Fishers criterion. To get an intuition of what is means, note that the determinant of the co-variance matrix tells us how much variance a class has. For example, for the co-variance matrix in the PCA (diagonal) projection, the value of the determinant is just the product of the diagonal elements which are the individual variable variances. The determinant has the same value under any ortho-normal projection. So Fishers criterion tries to find the projection that maximises the variance of the class means and minimises the variance of the individual classes.

It has been shown that Φ_{lda} is the solution of the following equation:

$$S_b \Phi - S_w \Phi \Lambda = 0$$

Multiplying by the inverse of S_w we get:

$$\begin{gathered} S_w^{-1} S_b \Phi - S_w^{-1} S_w \Phi \Lambda = 0 \\ S_w^{-1} S_b \Phi - \Phi \Lambda = 0 \\ S_w^{-1} S_b \Phi = \Phi \Lambda \end{gathered}$$

Thus, if S_w is a non-singular matrix, and can be inverted, then the Fisher's criterion is maximised when the projection matrix Φ_{lda} is composed of the eigenvectors of:

$$S_w^{-1} S_b$$

Notice that there will be at most $g - 1$ eigenvectors with non-zero real corresponding eigenvalues. This is because there are only g points to estimate S_b. This again can represent a massive reduction in the dimensionality of the problem. In face recognition for example there may be several thousand variables, but only a few hundred classes.

Once the projection is found all the data points can be transformed to the new axis system along with the class means and co-variances. Allocation of a new point to a class can be done using a distance measure such as the Mahalanobis distance. We will look at methods of classifying in the next lecture. LDA is essentially just a projection method.

The most Discriminant Feature method

The performance of the standard LDA can be seriously degraded if there are only a limited number of total training observations N compared to the dimension of the feature space n.

Since S_w is a function of $N - g$ or fewer linearly independent vectors, its rank is $N - g$ or less. Therefore, S_w is a singular matrix if N is less than $n + g$, or, analogously might be unstable if $N << n$. This is an important problem since, as we saw in the analysis above, it is necessary to invert the S_w matrix to find the LDA basis vectors. To overcome this problem, a two-stage feature extraction technique can be adopted. First the n-dimensional training samples from the original vector space are projected to a lower dimensional space using PCA. Then LDA is applied next to find the best linear discriminant features on that PCA subspace. Thus, the Fishers criterion can be written as:

$$\Phi_{lda} = \arg\max_{\Phi} \frac{|\Phi^T \Phi_{pca}^T S_b \Phi_{pca} \Phi|}{|\Phi^T \Phi_{pca}^T S_w \Phi_{pca} \Phi|}$$

And as before the solution is given by the eigenvectors of:

$$(\Phi_{pca}^T S_w \Phi_{pca})^{-1} (\Phi_{pca}^T S_b \Phi_{pca})$$

This has at most $g - 1$ eigenvectors with non-zero, real corresponding eigenvalues. Let us suppose that there are p principal components. This will be the case if there are at least $p + 1$ independent data points, regardless of how large the number of variables n is. Thus, if we have that $\Phi_{pca}^T S_w \Phi_{pca}$ has dimension $p \times p$, and is estimated from $N - g$ independent observations. Hence providing:

$$g \leq p \leq N - g$$

we can always invert $\Phi_{pca}^T S_w \Phi_{pca}$ and therefore obtain an LDA estimate. The method can always be made to work because, in cases where $p > N - g$ we can just use a smaller number of the eigenvectors.