Fisher Linear Discriminant Analysis

Cheng Li, Bingyu Wang

August 18, 2014

1 What's LDA

Fisher Linear Discriminant Analysis (also called Linear Discriminant Analysis(LDA)) are methods used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates 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 closely related to PCA, for both of them are based on linear, i.e. matrix multiplication, transformations. For the case of PCA, the transformation is baed on minimizing mean square error between original data vectors and data vectors that can be estimated fro the reduced dimensionality data vectors. And the PCA does not take into account any difference in class. But for the case of LDA, the transformation is based on maximizing a ratio of "between-class variance" to "within-class variance" with the goal of reducing data variation in the same class and increasing the separation between classes. Let's see an example of LDA as below(Figure1):

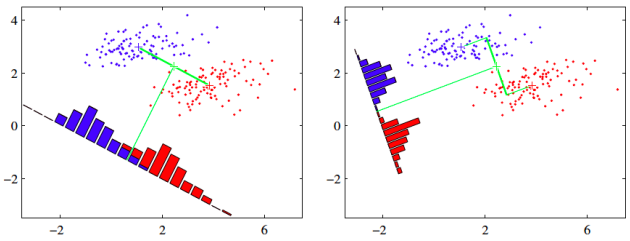

Figure 1: LDA examples

The left plot shows samples from two classes (depicted in red and blue) along with the histograms resulting from projection onto the line joining the class means. Note that there is considerable class overlap in the projected space. The right plot shows the corresponding projection based on the Fisher linear discriminant, showing the greatly improved class separation.

So our job is seeking to obtain a scalar y by projecting the samples X onto a line:

$$y = \theta^T X$$

Then try to find the θ^* to maximize the ratio of "between-class variance" to "within-class variance". Next, we will introduce how to use mathematic way to present this problem.

2 Theory and Model

To figure out the LDA, first we need know how to translate "between-class variance" and "within-class variance" to mathematic language. Then we try to maximize the ratio between these two. To simplify the problem, we start with two classes problem.

2.1 Two Classes Problem

2.1.1 Head the Problem

Assume we have a set of D-dimensional samples $X = \{x^{(1)}, x^{(2)}, ...x^{(m)}\}$, N_1 of which belong to class C_1, and N_2 of which belong to class C_2. We also assume the mean vector of two classes in X-space:

$$u_k = \frac{1}{N_k} \sum_{i \in C_k} x^{(i)} \quad \text{where} \quad k = 1, 2.$$

and in y-space:

$$\hat{u}_k = \frac{1}{N_k} \sum_{i \in C_k} y^{(i)} = \frac{1}{N_k} \sum_{i \in C_k} \theta^T x^{(i)} = \theta^T u_k \quad \text{where} \quad k = 1, 2.$$

One way to define a measure of separation between two classes is to choose the distance between the projected means, which is in y-space, so the **between-class variance** is:

$$\hat{u}_2 - \hat{u}_1 = \theta^T (u_2 - u_1)$$

Also, we can define the **within-class variance** for each class C_k is:

$$\hat{s}_k^2 = \sum_{i \in C_k} (y^{(i)} - \hat{u}_k)^2 \quad \text{where} \quad k = 1, 2.$$

Then, we get the between-class variance and within-class variance, we can define our objective function $J(\theta)$ as:

$$J(\theta) = \frac{(\hat{u}_2 - \hat{u}_1)^2}{\hat{s}_1^2 + \hat{s}_2^2}$$

In fact, if maximizing the objective function J, we are looking for a projection where examples from the class are projected very close to each other and at the same time, the projected means are as farther apart as possible.

2.1.2 Transform the Problem

To find the optimum θ^*, we must express $J(\theta)$ as a function of θ. Before the optimum,we need introduce **scatter** instead of variance.

We define some measures of the scatter as following:

- The scatter in feature space-x: $S_k = \sum_{i \in C_k} (x^{(i)} - u_k)(x^{(i)} - u_k)^T$
- Within-class scatter matrix: $S_W = S_1 + S_2$
- Between-class scather matrix: $S_B = (u_2 - u_1)(u_2 - u_1)^T$

Let's see $J(\theta)$ again:

$$J(\theta) = \frac{(\hat{u}_2 - \hat{u}_1)^2}{\hat{s}_1^2 + \hat{s}_2^2}$$

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

$$\begin{aligned}
\hat{s}_k^2 &= \sum_{i \in C_k} (y^{(i)} - \hat{u}_k)^2 \\
&= \sum_{i \in C_k} (\theta^T x^{(i)} - \theta^T u_k)^2 \\
&= \sum_{i \in C_k} \theta^T (x^{(i)} - u_k)(x^{(i)} - u_k)^T \theta \\
&= \theta^T S_k \theta
\end{aligned}$$

So we can get:

$$\begin{aligned}
\hat{s}_1^2 + \hat{s}_2^2 &= \theta^T S_1 \theta + \theta^T S_2 \theta \\
&= \theta^T S_W \theta
\end{aligned}$$

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

$$\begin{aligned}
(\hat{u}_2 - \hat{u}_1)^2 &= (\theta^T u_2 - \theta^T u_1)^2 \\
&= \theta^T (u_2 - u_1)(u_2 - u_1)^T \theta \\
&= \theta^T S_B \theta
\end{aligned}$$

We can finally express the Fisher criterion in terms of S_W and S_B as:

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

Next, we will maximize this objective function.

2.1.3 Solve the Problem

The easiest way to maximize the object function J is to derive it and set it to zero.

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \left(\frac{\theta^T S_B \theta}{\theta^T S_W \theta} \right) \\
&= (\theta^T S_W \theta) \frac{\partial (\theta^T S_B \theta)}{\partial \theta} - (\theta^T S_B \theta) \frac{\partial (\theta^T S_W \theta)}{\partial \theta} = 0 \\
\implies &= (\theta^T S_W \theta) 2 S_B \theta - (\theta^T S_B \theta) 2 S_W \theta = 0
\end{aligned}$$

Divided by $\theta^T S_W \theta$:

$$
\begin{aligned}
&\implies (\frac{\theta^T S_W \theta}{\theta^T S_W \theta}) S_B \theta - (\frac{\theta^T S_B \theta}{\theta^T S_W \theta}) S_W \theta = 0 \\
&\implies S_B \theta - J S_W \theta = 0 \\
&\implies S_W^{-1} S_B \theta - J\theta = 0 \\
&\implies J\theta = S_W^{-1} S_B \theta \\
&\implies J\theta = S_W^{-1}(u_2 - u_1)(u_2 - u_1)^T \theta \\
&\implies J\theta = S_W^{-1}(u_2 - u_1)(\underbrace{(u_2 - u_1)^T \theta}_{c \in \mathbb{R}}) \\
&\implies J\theta = c S_W^{-1}(u_2 - u_1) \\
&\implies \theta = \frac{c}{J} S_W^{-1}(u_2 - u_1)
\end{aligned}
$$

For now, the problem has been solved and we just want to get the direction of the θ, which is the optimum θ^*:

$$\theta^* \propto S_W^{-1}(u_2 - u_1)$$

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, which is $y = \theta^{*T} X$.

2.2 MultiClasses Problem

Based on two classes problem, we can see that the fisher's LDA generalizes gracefully for multiple classes problem. Assume we still have a set of D-dimensional samples $X = \{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$, and there are totally C classes. Instead of one projection y, mentioned above, we now will seek $(C-1)$ projections $[y_1, y_2, \dots y_{C-1}]$ by means of $(C-1)$ projection vectors θ_i arranged by columns into a projection matrix $\Theta = [\theta_1 | \theta_2 | \dots | \theta_{C-1}]$, where:

$$y_i = \theta_i^T X \implies y = \Theta^T X$$

2.2.1 Derivation

First we will use the scatters in space-x as following:

- Within-class scatter matrix:

$$S_W = \sum_{i=1}^{C} S_i \quad \text{where} \quad S_i = \sum_{i \in C_i} (x^{(i)-u_i})(x^{(i)} - u_i)^T \text{ and } u_i = \frac{1}{N_i} \sum_{i \in C_i} x^{(i)}$$

- Between-class scatter matrix:

$$S_B = \sum_{i=1}^{C} N_i (u_i - u)(u_i - u)^T \quad \text{where} \quad u = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} = \frac{1}{m} \sum_{i=1}^{C} N_i u_i$$

- Total scatter matrix:

$$S_T = S_B + S_W$$

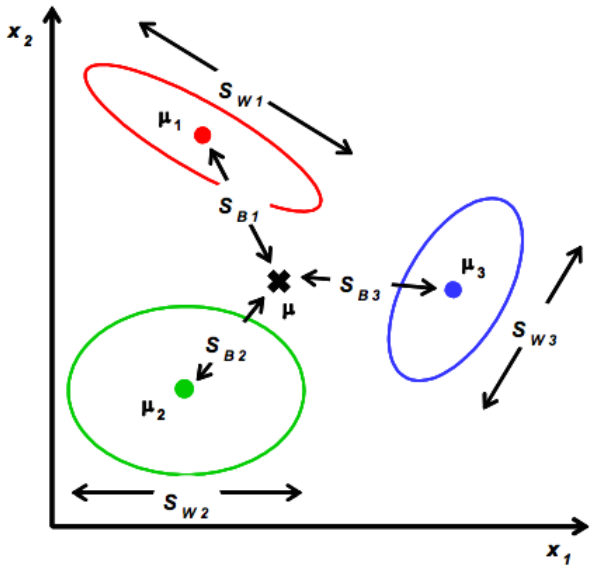

Figure 2: LDA Multi-Class examples

Before moving on, let us see a picture for the multi-class example in Figure2:
Similarly, we define the mean vector and scatter matrices for the projected samples as:

- $\hat{u}_i = \frac{1}{N_i} \sum_{i \in C_i} y^{(i)}$
- $\hat{u} = \frac{1}{N} \sum_{i=1}^{m} y^{(i)}$
- $\hat{S}_W = \sum_{i=1}^{C} \sum_{y \in C_i} (y - \hat{u}_i)(y - \hat{u}_i)^T$
- $\hat{S}_B = \sum_{i=1}^{C} N_i (\hat{u}_i - \hat{u})(\hat{u}_i - \hat{u})^T$

From our derivation for the two-class problem, we can get:

$$\hat{S}_W = \Theta^T S_W \Theta \tag{1}$$

$$\hat{S}_B = \Theta^T S_B \Theta \tag{2}$$

Recall that we are looking for a projection that maximizes the ratio of between-class to within-class scatter. Since the projection is no longer a scalar (it has $C - 1$ dimensions), we use the determinant of the scatter matrices to obtain a scalar objective function:

$$J(W) = \frac{|\hat{S}_B|}{|\hat{S}_W|} = \frac{|\Theta^T S_B \Theta|}{|\Theta^T S_W \Theta|}$$

And now, our job is to seek the projection matrix Θ^*that maximize this ratio. We will not give the derivation process. But we know that the optimal projection matrix Θ^* is the one whose columns are the eigenvectors corresponding to the

largest eigenvalues of the following generalized eigenvalue problem:

$$\begin{aligned}\Theta^* &= [\theta_1^*|\theta_2^*|\dots|\theta_{C-1}^*] \\ &= \text{argmax}\frac{|\Theta^T S_B \Theta|}{|\Theta^T S_W \Theta|} \\ \implies & (S_B - \lambda_i S_W)\theta_i^* = 0\end{aligned}$$

Thus, if S_W is a non-singular matrix, and can be inverted, then the Fisher's criterion is maximized when the projection matrix Θ^* is composed of the eigenvectors of:

$$S_W^{-1} S_B$$

Noticed that, there will be at most $C-1$ eigenvectors with non-zero real corresponding eigenvalues λ_i. This is because S_B is of rank $(C-1)$ or less. So we can see that LDA 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.

3 References

1. L10: Linear discriminants analysis [`http://research.cs.tamu.edu/prism/lectures/pr/pr_l10.pdf`]
2. LDA-linear discriminant analysis [`http://webdancer.is-programmer.com/posts/37867.html`]
3. Lecture 16: Linear Discriminant Analysis [`http://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/IDAPILecture16.pdf`]
4. Nonlinear Dimensionality Reduction Methods for Use with Automatic Speech Recognition By Stephen A. Zahorian and Hongbing Hu