Data Pre-Processing-VI

(Feature Extraction- LDA)

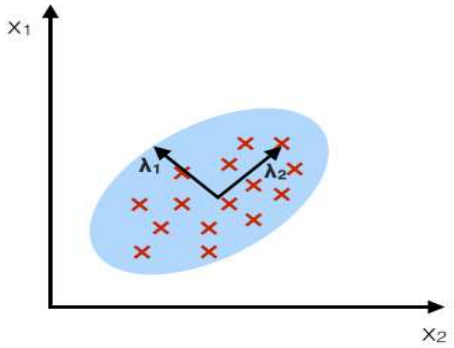
Dr. JASMEET SINGH
ASSISTANT PROFESSOR, CSED
TIET, PATIALA

Linear Discriminant Analysis (LDA)

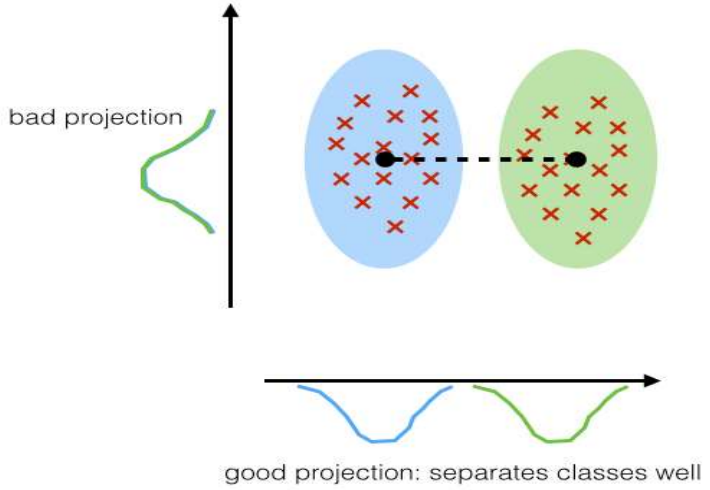
- Both Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are linear transformation techniques that are commonly used for dimensionality reduction.
- PCA can be described as an **"unsupervised"** algorithm, since it "ignores" class labels and its goal is to find **the directions (the so-called principal components) that maximize the variance in a dataset.**
- In contrast to PCA, **LDA is "supervised" and computes the directions ("linear discriminants") that will represent the axes that that maximize the separation between multiple classes.**
- Unlike **PCA that calculates eigenvalues and eigenvectors of the covariance matrix** of the data, **LDA calculates eigen value and eigen vectors using the within-class and between class scatter (variance) matrix**.

Linear Discriminant Analysis (LDA)

PCA:
component axes that maximize the variance

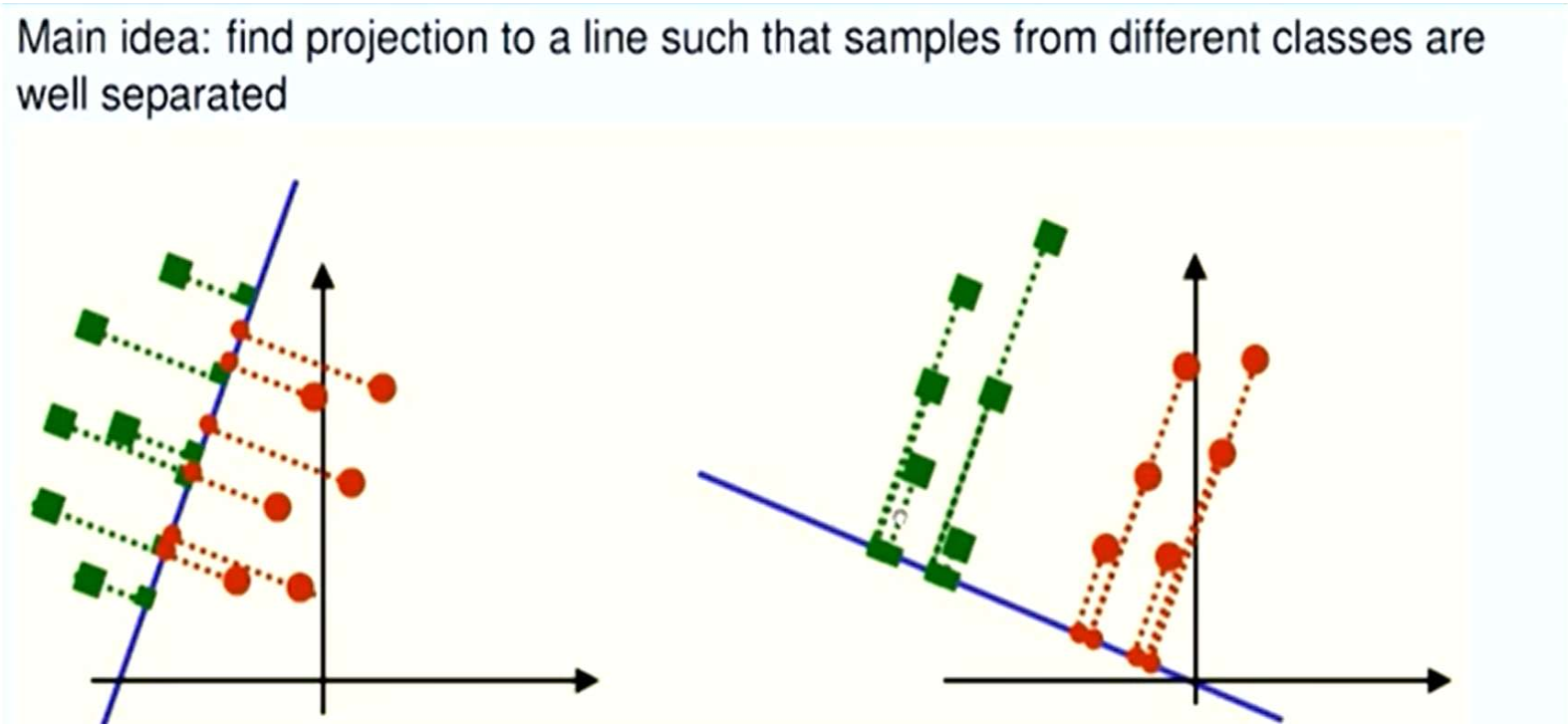


LDA:
maximizing the component axes for class-separation



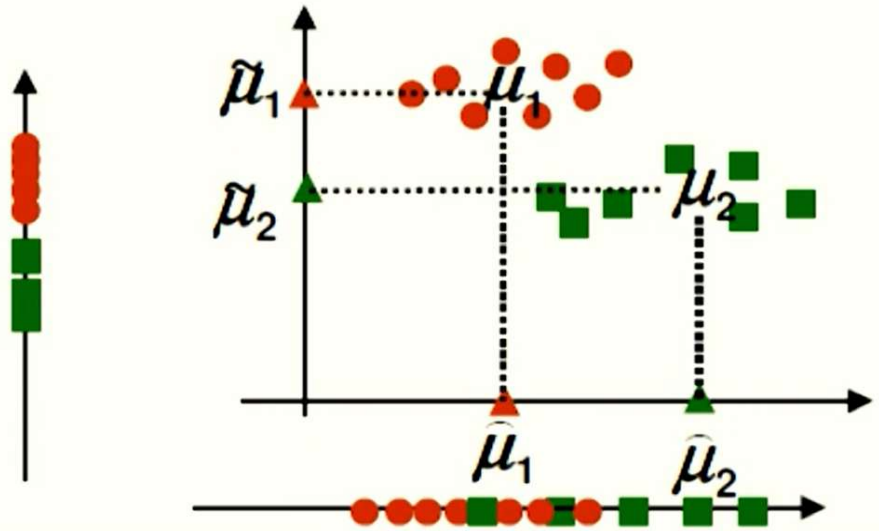
Linear Discriminant Analysis (LDA)

Main idea: find projection to a line such that samples from different classes are well separated

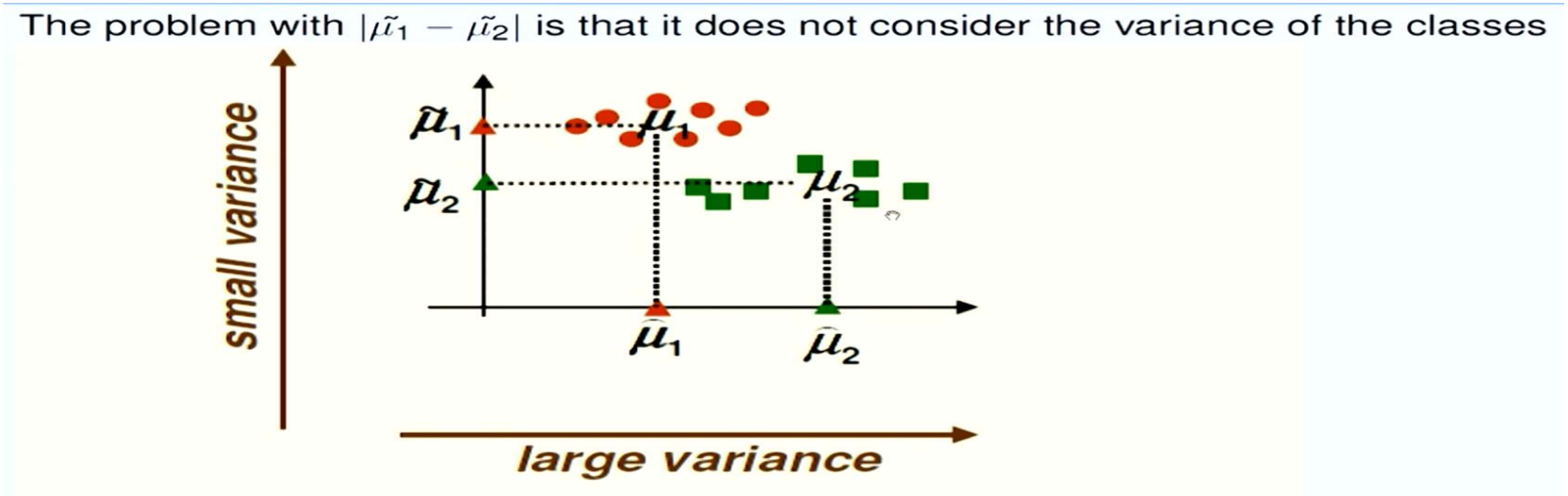


Mathematics behind LDA

- How good is $|\tilde{\mu}_1 - \tilde{\mu}_2|$ as a measure of separatic
  - The larger $|\tilde{\mu}_1 - \tilde{\mu}_2|$, the better is the expected separa



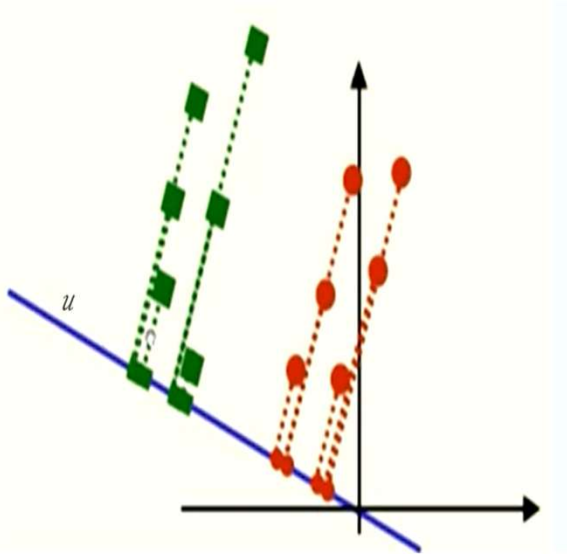
Mathematics behind LDA (Contd.....)



Therefore, separation between the projected means and the variance of the data points within classes on the projected axis both are important factors for finding the new axis.

Mathematics behind LDA – Two Classes Problem

Consider a data set X (of shape n×d; n be total number of data points and d be number of features) where each data points belong to one of the two classes C_1 and C_2. Let n_1 and n_2 be number of data points belonging to class C_1 and C_2 respective. Let μ_1 and μ_2 (of shape 1×d) be the mean of data points belonging to class C_1 and C_2 respectively (before projection). Let s_1^2 and s_2^2 be the variance of data points belonging to class C1 and C2 respectively (before projection).



Let $\vec{u}$ be a unit vector (of shape d ×1) and corresponds to a direction along which there is maximum separability of classes (as shown in Fig). Let xi be any data point of shape 1×d.

Projection of x_i on $\vec{u} = |\overrightarrow{x_i}|\cos\theta = \frac{\overrightarrow{x_i}.\vec{u}}{|\vec{u}|} = \overrightarrow{x_i}.\vec{u} = x_i u$

Mathematics behind LDA – Two Classes Problem (Contd…)

Let $\widetilde{\mu_1}$ and $\widetilde{\mu_2}$ be the mean of data points belonging to class C_1 and C_2 after projecting on $\vec{u}$. Let $\widetilde{s_1^2}$ and $\widetilde{s_2^2}$ be the variance of data points belonging to class C_1 and C_2 after projecting on $\vec{u}$.

Therefore, objective function of LDA (L) = Axis that maximizes class separability = Axis along which there is maximum distance between projected means and minimum variance with in projected data points.

Therefore, $\boldsymbol{L(u) = max_u \frac{|\widetilde{\mu_1} - \widetilde{\mu_2}|}{\widetilde{s_1^2} + \widetilde{s_2^2}} = max_u \frac{(\widetilde{\mu_1} - \widetilde{\mu_2})^2}{\widetilde{s_1^2} + \widetilde{s_2^2}}}$

Mathematics behind LDA – Two Classes Problem (Contd…)

$\widetilde{\mu_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i u = u\frac{1}{n_1}\sum_{i=1}^{n_1} x_i = \mu_1 u$

Similarly, $\widetilde{\mu_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i u = u\frac{1}{n_2}\sum_{i=1}^{n_2} x_i = \mu_2 u$

$(\widetilde{\mu_1} - \widetilde{\mu_2})^2 = (\mu_1 u - \mu_2 u)^2 = [(\mu_1 - \mu_2)u]^2$

$= u^T(\mu_1 - \mu_2)^T(\mu_1 - \mu_2)u$

$(because (\mu_1 - \mu_2)u \text{ is a scalar and square can be taken by multiplying one with transpose of other})$

$= u^T S_B u$

Therefore, $(\widetilde{\mu_1} - \widetilde{\mu_2})^2 = \boldsymbol{u^T S_B u}$

where $\boldsymbol{S_B} = (\mu_1 - \boldsymbol{\mu_2})^T(\mu_1 - \boldsymbol{\mu_2})$ and is called **between – class scatter matrix.**

Mathematics behind LDA – Two Classes Problem (Contd…)

$$\widetilde{s_1^2} = \sum_{i=1}^{n_1}(x_i u - \widetilde{\mu_1})^2 = \sum_{i=1}^{n_1}(x_i u - \mu_1 u)^2 = \sum_{i=1}^{n_1}[(x_i - \mu_1)u]^2$$

$$= \sum_{i=1}^{n_1} u^T (x_i - \mu_1)^T (x_i - \mu_1)\, u = u^T \left[\sum_{i=1}^{n_1}(x_i - \mu_1)^T (x_i - \mu_1)\right] u = u^T S_{w_1} u$$

Similarly, $\widetilde{s_2^2} = \sum_{i=1}^{n_2}(x_i u - \widetilde{\mu_2})^2 = \sum_{i=1}^{n_2}(x_i u - \mu_2 u)^2 = \sum_{i=1}^{n_2}[(x_i - \mu_2)u]^2$

$$= \sum_{i=1}^{n_2} u^T (x_i - \mu_2)^T (x_i - \mu_2)\, u = u^T \left[\sum_{i=1}^{n_2}(x_i - \mu_2)^T (x_i - \mu_2)\right] u = u^T S_{w_2} u$$

Total variance within class $= \widetilde{s_1^2} + \widetilde{s_2^2} = \boldsymbol{u^T S_{w_1} u + u^T S_{w_2} u = u^T (S_{w_1} + S_{w_2}) u = u^T S_w u}$

Where $\boldsymbol{S_w = \sum_{j=1}^{|C|} S_{w_j}}$ ***is called within − class scatter matrix***

and $\boldsymbol{S_{w_j} = \sum_{i=1}^{n_j}(x_i - \mu_2)^T (x_i - \mu_j)}$

Mathematics behind LDA – Two Classes Problem (Contd…)

$$L(u) = max_u \frac{(\widetilde{\mu_1} - \widetilde{\mu_2})^2}{\widetilde{s_1^2} + \widetilde{s_2^2}} = \frac{u^T S_B u}{u^T S_w u} = max_u \, u^T S_B u \;\; subject\ to\ u^T S_w u = 1$$

$$L(\lambda, u) = = max_u [u^T S_B u - \lambda(\, u^T S_w u - 1)]$$

$$\frac{\partial L(\lambda,u)}{\eth u} = 2S_B u - \lambda \times 2S_w u$$

For L to be maximum or minimum, $\frac{\partial L(u)}{\eth u} = 0$

$$2S_B u - \lambda \times 2S_w u = 0$$

$$S_B u = \lambda S_w \, u$$

$$(S_w)^{-1} S_B u = \lambda u$$

Mathematics behind LDA – Two Classes Problem (Contd…)

Therefore λ is the eigen values corresponding to $(S_w)^{-1}S_B$.

For k features, $(S_w)^{-1}S_B$ is of order kXk matrix and hence there are k eigen values

i.e., $\lambda_1 > \lambda_2 > \cdots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots > \lambda_k$

$$\frac{\partial^2 L(u)}{\partial u^2} = (S_w)^{-1}S_B - \lambda I$$

$$\frac{\partial^2 L(u)}{\partial u^2}_{\lambda_1} = (S_w)^{-1}S_B - \lambda_1 I$$

Therefore eigen values of $(S_w)^{-1}S_B - \lambda_1 I$ is $\lambda_i - \lambda_1$ which are all negative and zero.

Thus, the axis that captures maximum separability of input data points are the eigen vectors of the $(S_w)^{-1}S_B$ of the matrix corresponding to largest eigen value.

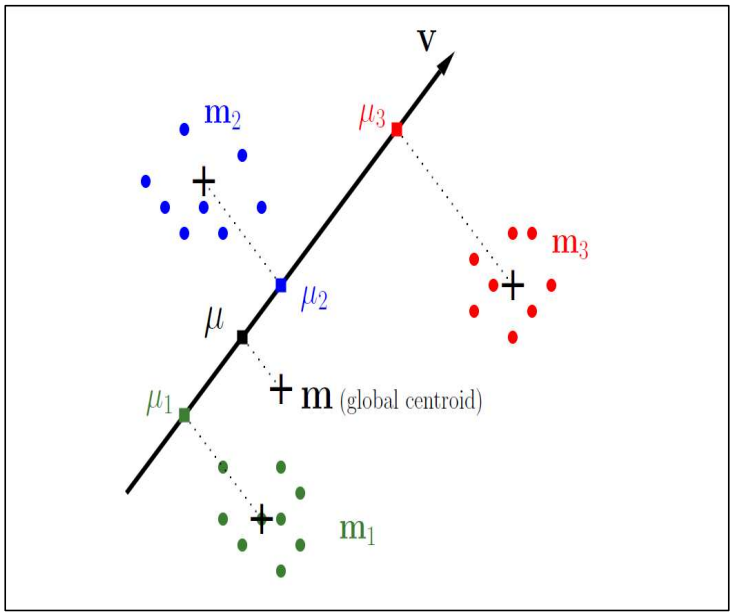
Mathematics behind LDA –Multi-Class Problem (Contd…)

In case of multi class problem,

Between class scatter-ness is computed as a weighted mean of the projected means.

$$S_B = \sum_{i=1}^{p} S_{Bi}$$

where $S_{Bi} = n_i(\mu_i - \mu)^T(\mu_i - \mu)$



Step-by-Step Working of LDA

- Consider a labelled dataset with n datapoints that is classified into p classes. Let each class contain n_1, n_2 , n_3,……n_p datapoints.

1. Compute the **mean vector** of each class (μ_1, μ_2, μ_3……. μ_p) and the global mean vector (μ).
2. Compute **the between-class scatter matrix** S_B given by:

$$S_B = \sum_{i=1}^{p} S_{Bi}$$

where $S_{Bi} = n_i(\mu_i - \mu)^T(\mu_i - \mu)$

Step-by-Step Working of LDA

3. Compute Within-class scatter matrix Sw given by:

$$S_w = \sum_{j=1}^{p} S_{wj}$$

Where $S_{wj} = \sum_{i=1}^{n_j} (x_{ij} - \mu_j)^T (x_{ij} - \mu_j)$

x_{ij} represents the i^{th} sample in the j^{th} class

4. Compute the scatter matrix from between-class and within-class variance given by $S_W^{-1} S_B$

Step-by-Step Working of LDA

5. Compute the eigen values of the scatter matrix ($S_W^{-1}S_B$) i.e. find λ's such that det($S_W^{-1}S_B$ - λI)=0.
6. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a p×k dimensional matrix W (where every column represents an eigenvector) .
7. Use this p X k eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: **Y=X×W** (where X is a n×p matrix representing the n samples and p classes, and Y are the transformed n×k-dimensional samples in the new subspace)

LDA Example

Given two different classes, $\omega_1(5\times2)$ and $\omega_2(6\times2)$ have $(n_1 = 5)$ and $(n_2 = 6)$ samples, respectively. Each sample in both classes is represented by two features as follows:

$$w_1 = \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 3 \\ 4 & 5 \\ 5 & 5 \end{pmatrix} \text{ and } w_2 = \begin{pmatrix} 4 & 2 \\ 5 & 0 \\ 5 & 2 \\ 3 & 2 \\ 5 & 3 \\ 6 & 3 \end{pmatrix}$$

Calculate the lower dimension space using LDA.

LDA Example

Step 1: Compute the mean of each class (μ_1 and μ_2) and the global mean vector (μ).

The values of the mean of each class and the total mean are shown below:

$$\mu_1 = (3 \quad 3.6)$$

$$\mu_2 = (4.67 \quad 2)$$

$$\mu = (3.91 \quad 2.72)$$

LDA Example

Step 2: Compute the between-class scatter matrix:

The values of the between-class variance of the first class (SB_1) is equal to,

$$S_{B1} = n_1(\mu_1 - \mu)^T(\mu_1 - \mu) =$$
$$= 5(-0.91 \quad 0.87)^T(-0.91 \quad 0.87) = \begin{pmatrix} 4.13 & -3.97 \\ -3.97 & 3.81 \end{pmatrix}$$

The values of the between-class variance of the second class (SB_2) is

$$S_{B2} = n_2(\mu_2 - \mu)^T(\mu_2 - \mu) =$$
$$= 6(0.76 \quad -0.72)^T(0.76 \quad -0.72) = \begin{pmatrix} 3.44 & -3.31 \\ -3.31 & 3.17 \end{pmatrix}$$

$$S_B = S_{B1} + S_{B2} = \begin{pmatrix} 7.58 & -7.27 \\ -7.27 & 6.98 \end{pmatrix}$$

LDA Example

Step 3: Compute the within-class scatter matrix:

To calculate the within-class matrix, first subtract the mean of each class from each sample in that class, and it is calculated as follows, $d_i = \omega_i - \mu_i$. The values of d_1 and d_2 are as follows:

$$d_1 = w_1 - \mu_1 = \begin{pmatrix} -2 & -1.6 \\ -1 & -0.6 \\ 0 & -0.6 \\ 1 & 1.4 \\ 2 & 1.4 \end{pmatrix} \text{ and } d_2 = w_2 - \mu_2 = \begin{pmatrix} -0.67 & 0 \\ 0.33 & -2 \\ 0.33 & 0 \\ -1.67 & 0 \\ 0.33 & 1 \\ 1.33 & 1 \end{pmatrix}$$

After centering the data, the within-class variance for each class (SW_j) is calculated as follows, $SW_j = d^T{}_j * d_j$,

LDA Example

The values of within- class matrix for each class and the total within-class matrix are as follows:

$$S_{w1} = \begin{pmatrix} -2 & -1.6 \\ -1 & -0.6 \\ 0 & -0.6 \\ 1 & 1.4 \\ 2 & 1.4 \end{pmatrix}^T \begin{pmatrix} -2 & -1.6 \\ -1 & -0.6 \\ 0 & -0.6 \\ 1 & 1.4 \\ 2 & 1.4 \end{pmatrix} = \begin{pmatrix} 10 & 8 \\ 8 & 7.2 \end{pmatrix}$$

$$S_{w2} = \begin{pmatrix} -0.67 & 0 \\ 0.33 & -2 \\ 0.33 & 0 \\ -1.67 & 0 \\ 0.33 & 1 \\ 1.33 & 1 \end{pmatrix}^T \begin{pmatrix} -0.67 & 0 \\ 0.33 & -2 \\ 0.33 & 0 \\ -1.67 & 0 \\ 0.33 & 1 \\ 1.33 & 1 \end{pmatrix} = \begin{pmatrix} 5.33 & 1 \\ 1 & 6 \end{pmatrix}$$

$$S_W = S_{W1} + S_{W2} = \begin{pmatrix} 15.33 & 9 \\ 9 & 13.2 \end{pmatrix}$$

LDA Example

Step 4: Compute the final scatter matrix given by $(S_W^{-1}S_B)$

$$S_w^{-1} = \frac{1}{121.356}\begin{pmatrix} 13.2 & -9 \\ -9 & 15.33 \end{pmatrix} = \begin{pmatrix} 0.11 & -0.07 \\ -0.07 & 0.13 \end{pmatrix}$$

$$S_w^{-1}S_B = \begin{pmatrix} 0.11 & -0.07 \\ -0.07 & 0.13 \end{pmatrix}\begin{pmatrix} 7.58 & -7.27 \\ -7.27 & 6.98 \end{pmatrix} = \begin{pmatrix} 1.36 & -1.31 \\ -1.48 & 1.42 \end{pmatrix}$$

Step 5: Compute the eigen values for the final scatter matrix.

i.e. $|S_w^{-1}S_B - \lambda I| = 0$ i.e. $\begin{vmatrix} 1.36-\lambda & -1.31 \\ -1.48 & 1.42-\lambda \end{vmatrix} = 0$

Thus the eigen values for the above characteristic equations are: 2.78, 0.0027

LDA Example

Step 6: Ignore the second eigen value as it is nearer to zero. So the only selected eigen value is 2.78

So for λ=2.78; $[S_W^{-1}S_B\text{-}2.78\text{I}]V_1=0$

$$\begin{pmatrix} -1.42 & -1.31 \\ -1.48 & -1.36 \end{pmatrix} V_1=0$$

Therefore eigen vector is $V_1 = \begin{pmatrix} 0.68 \\ -0.74 \end{pmatrix}$

LDA Example

Step 7: The original data is projected on the lower dimensional space, as follows, $y_i = \omega_i V_1$,

$$y_1 = w_1 V_1 = \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 3 \\ 4 & 5 \\ 5 & 5 \end{pmatrix} \begin{pmatrix} 0.68 \\ -0.74 \end{pmatrix} = \begin{pmatrix} -0.79 \\ -0.85 \\ -0.18 \\ -0.97 \\ -0.29 \end{pmatrix}$$

$$y_2 = w_2 V_1 = \begin{pmatrix} 4 & 2 \\ 5 & 0 \\ 5 & 2 \\ 3 & 2 \\ 5 & 3 \\ 6 & 3 \end{pmatrix} \begin{pmatrix} 0.68 \\ -0.74 \end{pmatrix} = \begin{pmatrix} 1.24 \\ 3.39 \\ 1.92 \\ 0.56 \\ 1.18 \\ 1.86 \end{pmatrix}$$