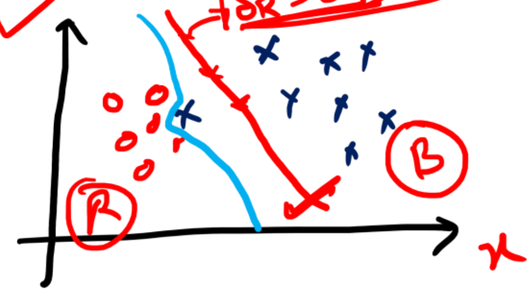
* Linear Discriminant Analysis

2D → 1D



① Discriminant function $\delta_R(x)$, $\delta_B(x)$

Red class ← $\delta_R(x_i) > \delta_B(x_i)$

$\delta_R(x_i) < \delta_B(x_i)$ → class B

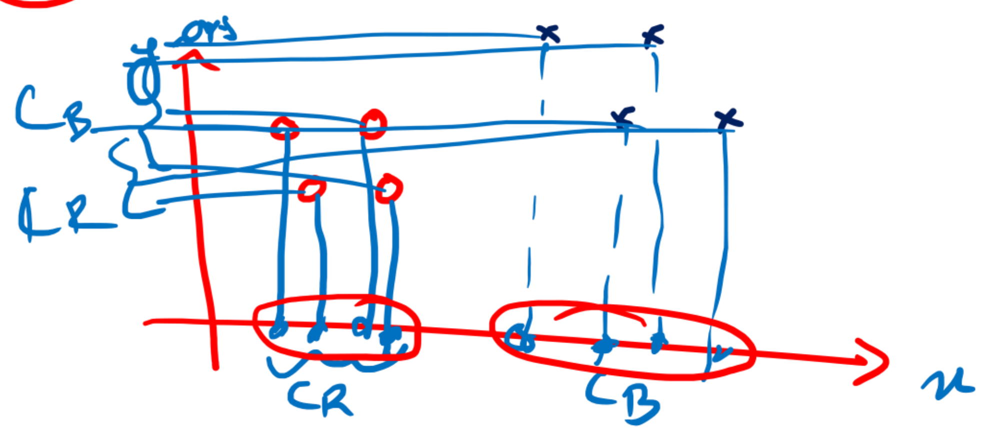
② Directly model hyperplane

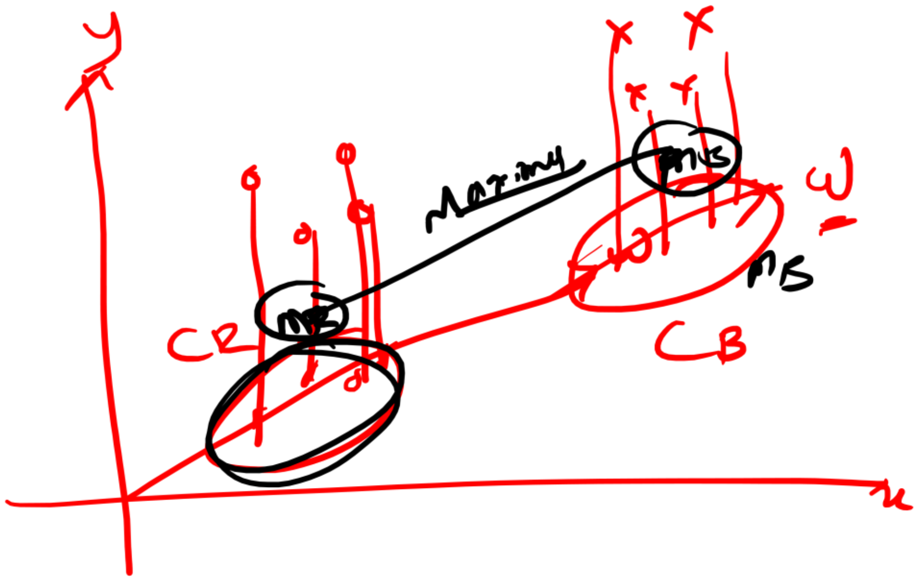
* PCA → unsupervised classifier

* LDA → supervised classifier

* LDA is used for Dimensionality Reduction

3D → 2D → 1D





(1) Maximize the distance betn means of two clusters eg CR, CB

(2) Minimize the variance within each class

Objectives

- Maximize Between-Class Variance: Ensures that different classes are as far apart as possible.
- Minimize Within-Class Variance: Ensures that data points within the same class are as close to each other as possible.

Assumptions

LDA operates under several key assumptions:

1. **Normal Distribution:** Features are assumed to follow a multivariate normal distribution within each class.
2. **Equal Covariance Matrices:** All classes share the same covariance matrix, implying that they have similar shapes in feature space.
3. **Linearly Separable Classes:** Classes can be separated by linear boundaries.

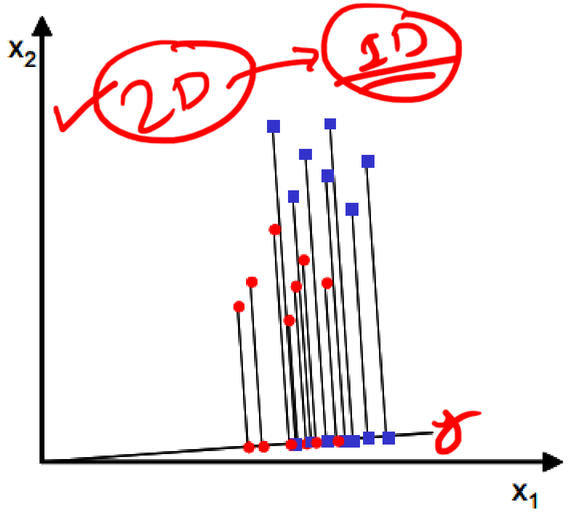
- **The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible**

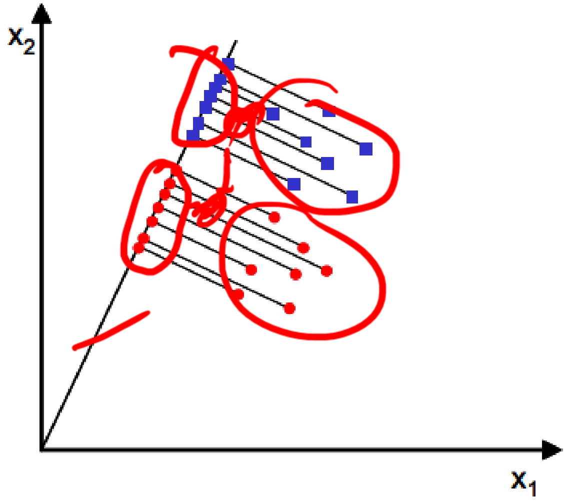
D=2

  - Assume we have a set of D-dimensional samples $\{x^{(1}, x^{(2}, \ldots, x^{(N}\}$, N_1 of which belong to class ω_1, and N_2 to class ω_2. We seek to obtain a scalar ***y*** by projecting the samples ***x*** onto a line

$$\boxed{y = w^T x}$$

  - 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





* Steps

1. Class Mean of dependent variable
2. Covariance matrix of class variable $\boxed{S_1 = \sum_{x \in \omega_1} (x-\mu_1)(x-\mu_1)^T}$
3. within class scatter matrix, $S_w = S_1 + S_2$
4. Between class scatter matrix, $\boxed{S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T}$
5. Compute eigen values & eigen vectors, S_w & S_B
6. obtain LDA by taking dot product of eigen vector and original data

$m_R \quad m_B$

$[\mu_R, \mu_B] = \frac{1}{N} \sum_{x \in Red} x$

(Binary class)

$3D \rightarrow 2D$

3

$\lambda_1 = 8$

$\lambda_2 = 3$

$\lambda_3 = 6.5$

After steps: k-class

$\Rightarrow$ sort the eigen values & select top 'k' values

- **Compute the Linear Discriminant projection for the following two-dimensional dataset**
  - $X1=(x_1,x_2)=\{(4,1),(2,4),(2,3),(3,6),(4,4)\}$
  - $X2=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$
- **SOLUTION (by hand)**
  - The class statistics are:

$$S_1=\begin{bmatrix}0.80 & -0.40\\ -0.40 & 2.60\end{bmatrix};\quad S_2=\begin{bmatrix}1.84 & -0.04\\ -0.04 & 2.64\end{bmatrix}$$

$$\mu_1=[3.00 \quad 3.60];\qquad \mu_2=[8.40 \quad 7.60]$$

  - The within- and between-class scatter are

$$S_B=\begin{bmatrix}29.16 & 21.60\\ 21.60 & 16.00\end{bmatrix};\quad S_W=\begin{bmatrix}2.64 & -0.44\\ -0.44 & 5.28\end{bmatrix}$$

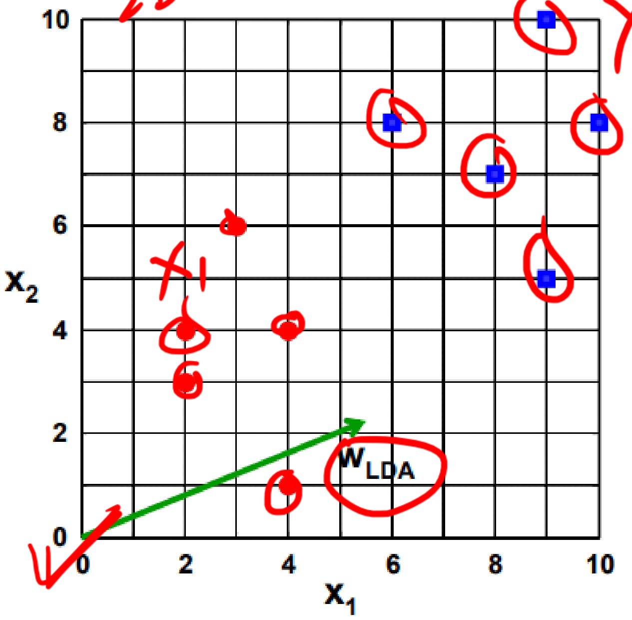
  - The LDA projection is then obtained as the solution of the generalized eigenvalue problem

$$S_W^{-1}S_B v=\lambda v \Rightarrow \left|S_W^{-1}S_B-\lambda I\right|=0 \Rightarrow \begin{vmatrix}11.89-\lambda & 8.81\\ 5.08 & 3.76-\lambda\end{vmatrix}=0 \Rightarrow \lambda=15.65$$

$$\begin{bmatrix}11.89 & 8.81\\ 5.08 & 3.76\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix}=15.65\begin{bmatrix}v_1\\ v_2\end{bmatrix}\Rightarrow \boxed{\begin{bmatrix}v_1\\ v_2\end{bmatrix}=\begin{bmatrix}0.91\\ 0.39\end{bmatrix}}$$

  - Or directly by

$$w^*=S_W^{-1}(\mu_1-\mu_2)=[-0.91 \quad -0.39]^T$$



$S_w = S_1 + S_2$

$(\mu_1-\mu_2)$

$\begin{bmatrix}3\\3.6\end{bmatrix}-\begin{bmatrix}8.4\\7.6\end{bmatrix}=\begin{bmatrix}-5.4\\-4.\end{bmatrix}$

$\begin{bmatrix}-5.4\\-4.0\end{bmatrix}[-5.4 \quad -4.0]$

$\begin{bmatrix}29.14\\21.6\end{bmatrix}$ 21.6 16

$\frac{15}{5}=3$ $\frac{42}{5}$

1m 2m

① μ ② S_1

$S_w = S_1 + S_2$

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

S_w S_B

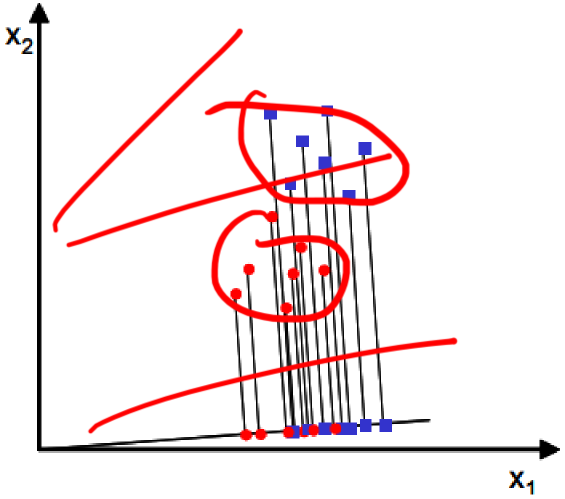
$w^* \Rightarrow$ optimal wt. vector

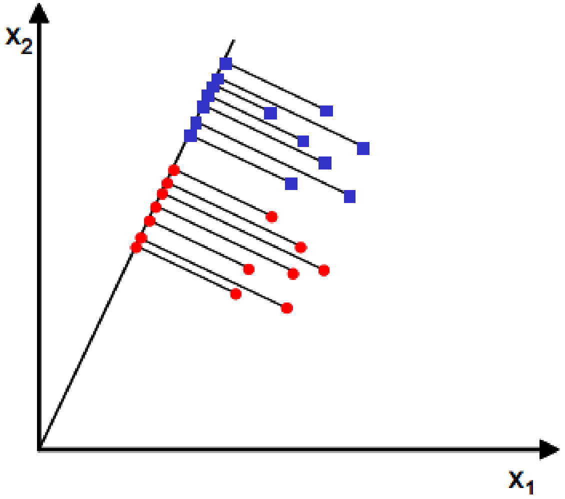
Fishers LDA

- **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 $\{x^{(1}, x^{(2}, \ldots, x^{(N}\}$, N_1 of which belong to class ω_1, and N_2 to class ω_2. 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 $\boldsymbol{x}$ and $\boldsymbol{y}$ feature space is

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

μ_i = mean vector

$\tilde{\mu}_i$ = projected mean vector

ω = Projection vector

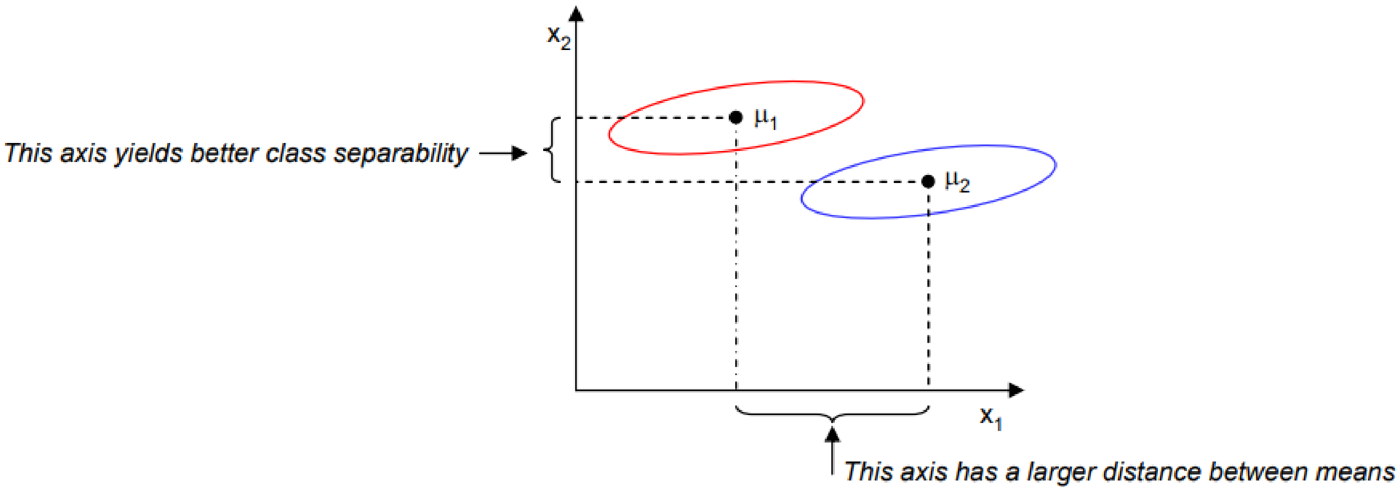
  - We could then choose the distance between the projected means as our objective function

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

→ measures the separation betn the means of two class after the projecting onto ω

  - 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

SB
SW



- 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 $\mathbf{w}^T\mathbf{x}$ that maximizes the criterion function

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

S1

S2

$S_w = S_1 + S_2$

$\tilde{S}_i$

diff betn projected means of two class.

total within class scatter

= betn class variance / within class variance

Scatter : measure the variance or spread of projected sample for class i

IMP

① SB is the outer product of two vectors, its rank is at most one (0, 1)

SB
Sw

- 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 S_W 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**

SB
Sw

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

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

① $= w^T S_1 w + w^T S_2 w =$

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

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

  - The matrix S_B is called the **between-class scatter**. Note that, since S_B is the outer product of two vectors, its rank is at most one
- We can finally express the Fisher criterion in terms of S_W and S_B as

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

2

- 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 $w^T S_W w$

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

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

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

~~Eigen values~~

$S_W^{-1} S_B w = J w$ → eigen vector of matrix $S_W^{-1} S_B$

↳ eigen value

- Solving the generalized eigenvalue problem ($S_W^{-1} S_B w = Jw$) yields

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

→ Imp

- This is know 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