Machine Learning

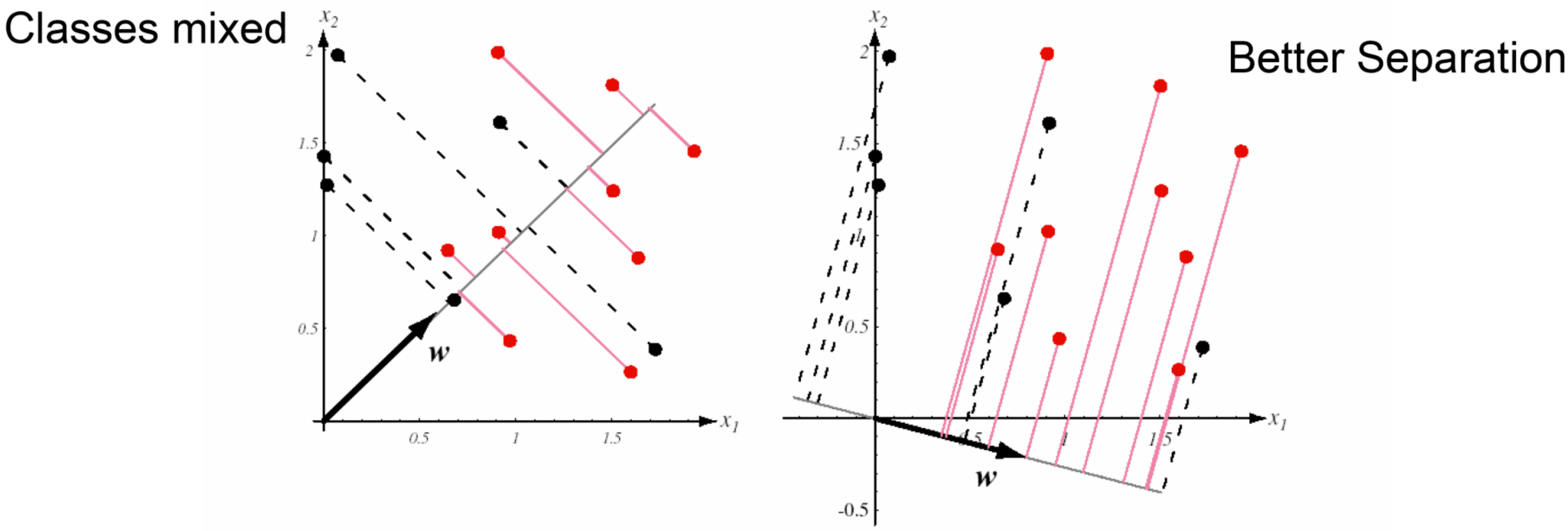
CSE 6363 (Fall 2016)

Lecture 7 Fisher Linear Discriminant Analysis

Heng Huang, Ph.D.

Department of Computer Science and Engineering

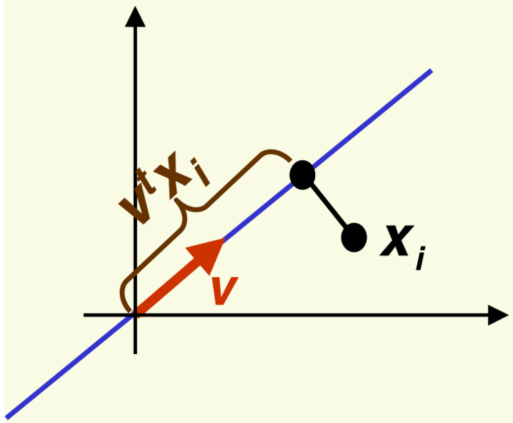
Fisher Linear Discriminant



The figure on the right shows greater separation between subsets, one set of the points with dashed line, another with solid line.

Fisher Linear Discriminant

- Suppose we have 2 classes and $\boldsymbol{d}$-dimensional samples $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ where
  - $\boldsymbol{n}_1$ samples come from the first class
  - $\boldsymbol{n}_2$ samples come from the second class
- consider projection on a line
- Let the line direction be given by unit vector $\boldsymbol{v}$



- Scalar $\boldsymbol{v}^t\boldsymbol{x}_i$ is the distance of projection of $\boldsymbol{x}_i$ from the origin
- Thus it $\boldsymbol{v}^t\boldsymbol{x}_i$ is the projection of $\boldsymbol{x}_i$ into a one dimensional subspace

Ref: Olga Veksler

Fisher Linear Discriminant

- Thus the projection of sample $\boldsymbol{x_i}$ onto a line in direction $\boldsymbol{v}$ is given by $\boldsymbol{v^t x_i}$
- How to measure separation between projections of different classes?
- Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ be the means of projections of classes 1 and 2
- Let μ_1 and μ_2 be the means of classes 1 and 2
- $|\tilde{\mu}_1 - \tilde{\mu}_2|$ seems like a good measure

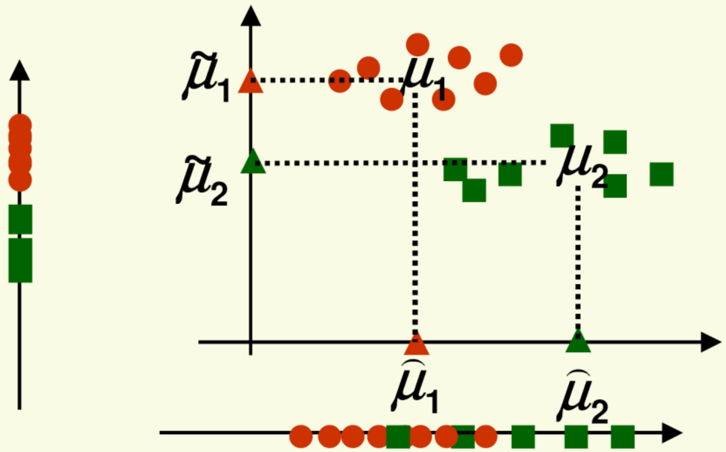
$$\tilde{\mu}_1 = \frac{1}{n_1}\sum_{x_i \in C1}^{n_1} v^t x_i = v^t\left(\frac{1}{n_1}\sum_{x_i \in C1}^{n_1} x_i\right) = v^t \mu_1$$

similarly, $\tilde{\mu}_2 = v^t \mu_2$

Ref: Olga Veksler

Fisher Linear Discriminant

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

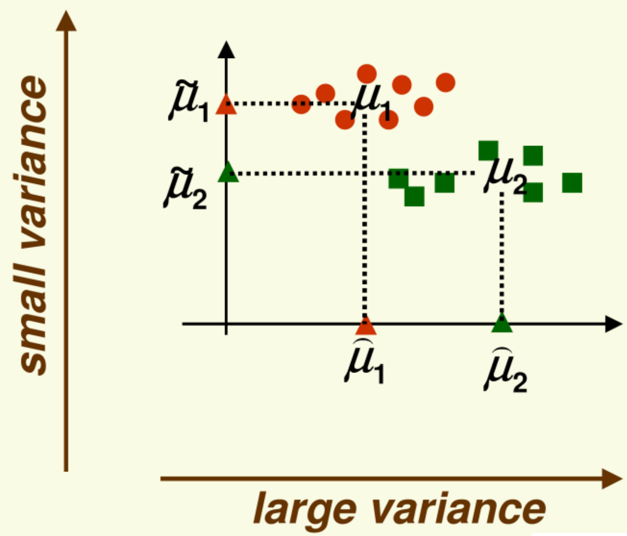


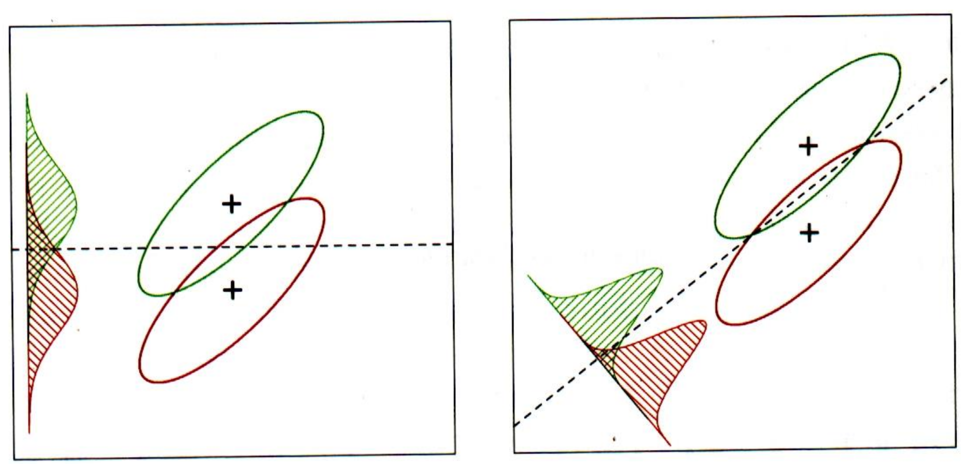
- the vertical axes is a better line than the horizontal axes to project to for class separability
- however $|\hat{\mu}_1 - \hat{\mu}_2| > |\tilde{\mu}_1 - \tilde{\mu}_2|$

Ref: Olga Veksler

Fisher Linear Discriminant

- The problem with $\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|$ is that it does not consider the variance of the classes





Ref: Olga Veksler

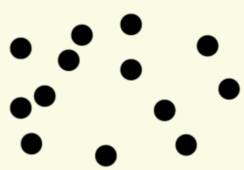
Fisher Linear Discriminant

- We need to normalize $|\tilde{\mu}_1 - \tilde{\mu}_2|$ by a factor which is proportional to variance
- Have samples $z_1, \ldots, z_n$. Sample mean is $\mu_z = \frac{1}{n}\sum_{i=1}^{n} z_i$
- Define their ***scatter*** as

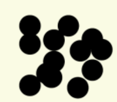
$$s = \sum_{i=1}^{n} (z_i - \mu_z)^2$$

- Thus scatter is just sample variance multiplied by ***n***
  - scatter measures the same thing as variance, the spread of data around the mean
  - scatter is just on different scale than variance

larger scatter:



smaller scatter:



Ref: Olga Veksler

Fisher Linear Discriminant

- Fisher Solution: normalize $|\tilde{\mu}_1 - \tilde{\mu}_2|$ by scatter
- Let $\boldsymbol{y}_i = \boldsymbol{v}^t\boldsymbol{x}_i$, i.e. $\boldsymbol{y}_i$'s are the projected samples
- Scatter for projected samples of class 1 is

$$\tilde{s}_1^2 = \sum_{y_i \in Class\ 1} (\boldsymbol{y}_i - \tilde{\mu}_1)^2$$

- Scatter for projected samples of class 2 is

$$\tilde{s}_2^2 = \sum_{y_i \in Class\ 2} (\boldsymbol{y}_i - \tilde{\mu}_2)^2$$

Ref: Olga Veksler

Fisher Linear Discriminant

- We need to normalize by both scatter of class 1 and scatter of class 2
- Thus Fisher linear discriminant is to project on line in the direction ***v*** which maximizes

$$J(v) = \frac{\overbrace{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}^{\text{want projected means are far from each other}}}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

want scatter in class 1 is as small as possible, i.e. samples of class 1 cluster around the projected mean $\tilde{\mu}_1$

want scatter in class 2 is as small as possible, i.e. samples of class 2 cluster around the projected mean $\tilde{\mu}_2$

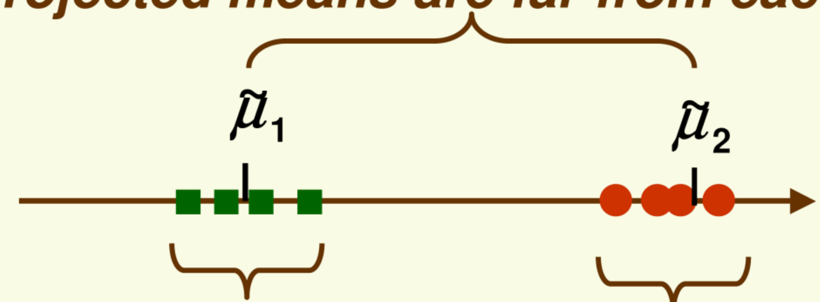
Ref: Olga Veksler

Fisher Linear Discriminant

$$J(v) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

- If we find v which makes $J(v)$ large, we are guaranteed that the classes are well separated

projected means are far from each other



small $\tilde{s}_1$ implies that projected samples of class 1 are clustered around projected mean

small $\tilde{s}_2$ implies that projected samples of class 2 are clustered around projected mean

Ref: Olga Veksler

Fisher Linear Discriminant

$$J(v) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

- All we need to do now is to express ***J*** explicitly as a function of ***v*** and maximize it
  - straightforward but need linear algebra and Calculus
- Define the separate class scatter matrices S_1 and S_2 for classes 1 and 2. These measure the scatter of original samples x_i (before projection)

$$S_1 = \sum_{x_i \in Class\ 1} (x_i - \mu_1)(x_i - \mu_1)^t$$

$$S_2 = \sum_{x_i \in Class\ 2} (x_i - \mu_2)(x_i - \mu_2)^t$$

Ref: Olga Veksler

Fisher Linear Discriminant

- Now define the ***within*** the class scatter matrix

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

- Recall that $\tilde{s}_1^2 = \sum_{y_i \in Class\ 1} (y_i - \tilde{\mu}_1)^2$

- Using $y_i = v^t x_i$ and $\tilde{\mu}_1 = v^t \mu_1$

$$\begin{aligned}
\tilde{s}_1^2 &= \sum_{y_i \in Class\ 1} \left(v^t x_i - v^t \mu_1\right)^2 \\
&= \sum_{y_i \in Class\ 1} \left(v^t (x_i - \mu_1)\right)^t \left(v^t (x_i - \mu_1)\right) \\
&= \sum_{y_i \in Class\ 1} \left((x_i - \mu_1)^t v\right)^t \left((x_i - \mu_1)^t v\right) \\
&= \sum_{y_i \in Class\ 1} v^t (x_i - \mu_1)(x_i - \mu_1)^t v = v^t \mathbf{S}_1 v
\end{aligned}$$

Ref: Olga Veksler

Fisher Linear Discriminant

- Similarly $\tilde{s}_2^2 = \mathbf{v}^t \mathbf{S}_2 \mathbf{v}$
- Therefore $\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{v}^t \mathbf{S}_1 \mathbf{v} + \mathbf{v}^t \mathbf{S}_2 \mathbf{v} = \mathbf{v}^t \mathbf{S}_W \mathbf{v}$
- Define between the class scatter matrix

$$\mathbf{S}_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t$$

- $\mathbf{S}_B$ measures separation between the means of two classes (before projection)
- Let's rewrite the separations of the projected means

$$\begin{aligned}(\tilde{\mu}_1 - \tilde{\mu}_2)^2 &= \left(\mathbf{v}^t \mu_1 - \mathbf{v}^t \mu_2\right)^2 \\ &= \mathbf{v}^t (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t \mathbf{v} \\ &= \mathbf{v}^t \mathbf{S}_B \mathbf{v}\end{aligned}$$

Fisher Linear Discriminant

- Thus our objective function can be written:

$$J(v) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{v^t S_B v}{v^t S_W v}$$

- **Maximize** $J(v)$ by taking the derivative w.r.t. v and setting it to 0

$$\frac{d}{dv} J(v) = \frac{\left(\frac{d}{dv} v^t S_B v\right) v^t S_W v - \left(\frac{d}{dv} v^t S_W v\right) v^t S_B v}{\left(v^t S_W v\right)^2}$$

$$= \frac{(2 S_B v) v^t S_W v - (2 S_W v) v^t S_B v}{\left(v^t S_W v\right)^2} = 0$$

Ref: Olga Veksler

Fisher Linear Discriminant

- Need to solve $v^t S_W v(S_B v) - v^t S_B v(S_W v) = 0$

$$\Rightarrow \frac{v^t S_W v(S_B v)}{v^t S_W v} - \frac{v^t S_B v(S_W v)}{v^t S_W v} = 0$$

$$\Rightarrow S_B v - \underbrace{\frac{v^t S_B v}{v^t S_W v}}_{=\lambda}(S_W v) = 0$$

$$\Rightarrow \underbrace{S_B v = \lambda S_W v}_{\textit{generalized eigenvalue problem}}$$

Ref: Olga Veksler

Fisher Linear Discriminant

$$\mathbf{S}_B \mathbf{v} = \lambda \mathbf{S}_W \mathbf{v}$$

- If $\mathbf{S}_W$ has full rank (the inverse exists), can convert this to a standard eigenvalue problem

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{v} = \lambda \mathbf{v}$$

- But $\mathbf{S}_B \mathbf{x}$ for any vector $\mathbf{x}$, points in the same direction as $\mu_1 - \mu_2$

α

$$\mathbf{S}_B \mathbf{x} = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t \mathbf{x} = (\mu_1 - \mu_2)\left((\mu_1 - \mu_2)^t \mathbf{x}\right) = \alpha(\mu_1 - \mu_2)$$

- Thus can solve the eigenvalue problem immediately

$$\boxed{\mathbf{v} = \mathbf{S}_W^{-1}(\mu_1 - \mu_2)}$$

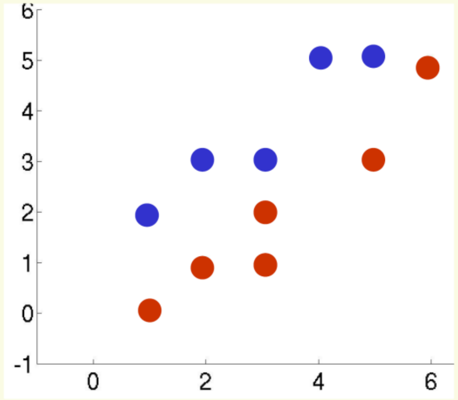
$$\mathbf{S}_W^{-1} \mathbf{S}_B \underbrace{\left[\mathbf{S}_W^{-1}(\mu_1 - \mu_2)\right]}_{\mathbf{v}} = \mathbf{S}_W^{-1}\left[\alpha(\mu_1 - \mu_2)\right] = \underbrace{\alpha}_{\lambda} \underbrace{\left[\mathbf{S}_W^{-1}(\mu_1 - \mu_2)\right]}_{\mathbf{v}}$$

Ref: Olga Veksler

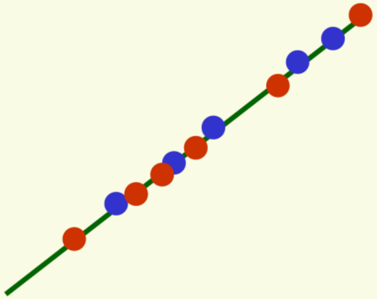
Fisher Linear Discriminant Example

- Data
  - Class 1 has 5 samples c_1=[(1,2),(2,3),(3,3),(4,5),(5,5)]
  - Class 2 has 6 samples c_2=[(1,0),(2,1),(3,1),(3,2),(5,3),(6,5)]
- Arrange data in 2 separate matrices

$$c_1 = \begin{bmatrix} 1 & 2 \\ \vdots & \vdots \\ 5 & 5 \end{bmatrix} \qquad c_2 = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 6 & 5 \end{bmatrix}$$



- Notice that PCA performs very poorly on this data because the direction of largest variance is not helpful for classification



Ref: Olga Veksler

Fisher Linear Discriminant Example

- First compute the mean for each class

$$\mu_1 = \textit{mean}(c_1) = [3 \quad 3.6] \qquad \mu_2 = \textit{mean}(c_2) = [3.3 \quad 2]$$

- Compute scatter matrices S_1 and S_2 for each class

$$S_1 = 4 * \textit{cov}(c_1) = \begin{bmatrix} 10 & 8.0 \\ 8.0 & 7.2 \end{bmatrix} \qquad S_2 = 5 * \textit{cov}(c_2) = \begin{bmatrix} 17.3 & 16 \\ 16 & 16 \end{bmatrix}$$

- Within the class scatter:

$$S_W = S_1 + S_2 = \begin{bmatrix} 27.3 & 24 \\ 24 & 23.2 \end{bmatrix}$$

  - it has full rank, don't have to solve for eigenvalues

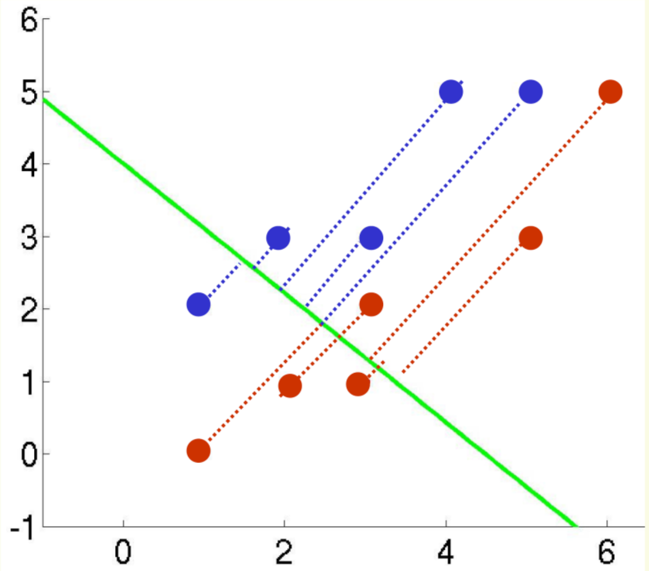
- The inverse of S_W is $S_W^{-1} = \textit{inv}(S_W) = \begin{bmatrix} 0.39 & -0.41 \\ -0.41 & 0.47 \end{bmatrix}$
- Finally, the optimal line direction v

$$v = S_W^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} -0.79 \\ 0.89 \end{bmatrix}$$

Ref: Olga Veksler

Fisher Linear Discriminant Example

- Notice, as long as the line has the right direction, its exact position does not matter



- Last step is to compute the actual ***1D*** vector ***y***. Let's do it separately for each class

$$Y_1 = v^t c_1^t = \begin{bmatrix} -0.65 & 0.73 \end{bmatrix} \begin{bmatrix} 1 & \cdots & 5 \\ 2 & \cdots & 5 \end{bmatrix} = \begin{bmatrix} 0.81 & \cdots & 0.4 \end{bmatrix}$$

$$Y_2 = v^t c_2^t = \begin{bmatrix} -0.65 & 0.73 \end{bmatrix} \begin{bmatrix} 1 & \cdots & 6 \\ 0 & \cdots & 5 \end{bmatrix} = \begin{bmatrix} -0.65 & \cdots & -0.25 \end{bmatrix}$$

Ref: Olga Veksler

Within-class Covariance Matrix

- The projection is from a d-dimensional space to a (c-1) dimensional space.
- The within-class scatter is

$$S_w = \sum_{i=1}^{c} S_i$$

$$S_i = \sum_{X \in D_i} (X - m_i)(X - m_i)^t$$

$$m_i = \frac{1}{n_i} \sum_{X \in D_i} X$$

Total Covariance Matrix

- Total mean

$$m = \frac{1}{n}\sum_X X$$

- Total scatter matrix

$$S_T = \sum_X (X-m)(X-m)^t$$

$$S_T = \sum_{i=1}^{c}\sum_{X\in D_i}(X-m_i+m_i-m)$$
$$(X-m_i+m_i-m)^t$$

$$S_T = \sum_{i=1}^{c}\sum_{X\in D_i}(X-m_i)(X-m_i)^t +$$
$$\sum_{i=1}^{c}\sum_{X\in D_i}(m_i-m)(m_i-m)^t$$

Total Covariance Matrix

- Note

$$\sum_{i=1}^{c}\sum_{X\in D_i}(X-m_i)(m_i-m)^t =$$

$$\sum_{i=1}^{c}[\sum_{X\in D_i}(X-m_i)](m_i-m)^t = [0]$$

$$0-matrix$$

$$S_T = S_W + \sum_{i=1}^{c} n_i(m_i-m)(m_i-m)^t$$

$$Define$$

$$S_B = \sum_{i=1}^{c} n_i(m_i-m)(m_i-m)^t$$

$$S_T = S_W + S_B$$

Multiple Discriminant Analysis

- The projection from a d-dimensional space to a (c-1)-dimensional space is done by c-1 discriminant functions

$$y_i = w_i^t X \quad i = 1,\ldots,c$$

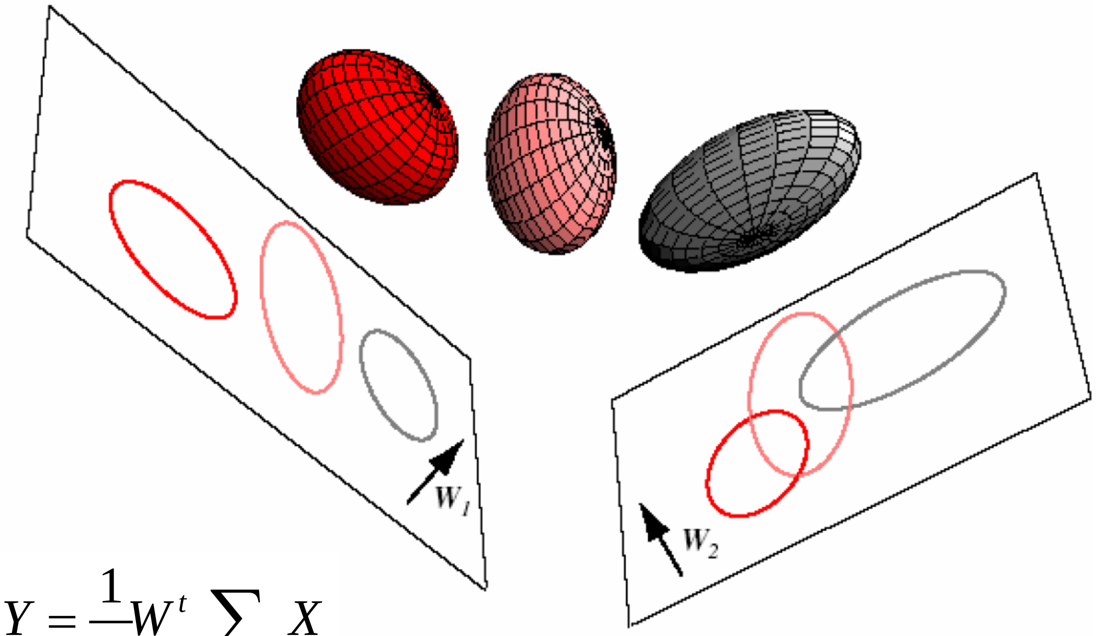
y_i can be viewed as a component of a vector Y

- w_i are viewed as columns of a d x (c-1) matrix $Y = W^t X$

- The samples $X_1,\ldots,X_n$ (d-dimensional) are mapped to a set of $y_1,\ldots,y_n$ (c-1)-dimensional which can be described by their own mean vectors and scatter matrices

Multiple Discriminant Analysis

Mapping from d-dimensional space to c-dimensional space d=3, c=3



$$Y = W^t X \qquad \text{and} \qquad \tilde{m}_i = \frac{1}{n_i}\sum_{Y \in y_i} Y = \frac{1}{n_i} W^t \sum_{X \to y_i} X$$

It can be shown

$$\tilde{S}_W = W^t S_W W$$

$$\tilde{S}_B = W^t S_B W$$

Multiple Discriminant Analysis

- We want to find a transform matrix W that maximizes the ratio of the determinants of the between-class scatter to the within-class scatter:

$$J(W) = \frac{|\tilde{S}_B|}{|\tilde{S}_W|} = \frac{|W^t S_B W|}{|W^t S_W W|}$$

- The columns of an optimal W are the generalized eigenvectors corresponding to the largest eigenvalues

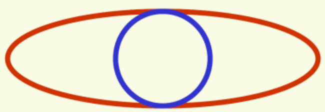
$$S_B w_i = \lambda_i S_W w_i$$

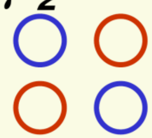
- If S_W is nonsingular, $S_W^{-1} S_B w_i = \lambda_i w_i$

FDA and MDA Drawbacks

- Reduces dimension only to $k = c-1$ (unlike PCA)
  - For complex data, projection to even the best line may result in unseparable projected samples
- Will fail:

1. $J(v)$ is always 0: happens if $\mu_1 = \mu_2$



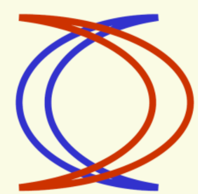


PCA performs reasonably well here:

PCA also fails:

$$J(V) = \frac{\det\left(V^t S_B V\right)}{\det\left(V^t S_W V\right)}$$

2. If $J(v)$ is always large: classes have large overlap when projected to any line (PCA will also fail)



Ref: Olga Veksler