**CS434b/654b : Pattern Recognition**
**Prof. Olga Veksler**

# Lecture 8

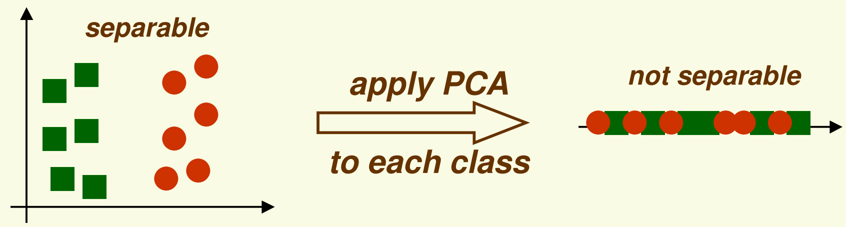
Fisher LDA and MDA

## Today

- Continue with Dimensionality Reduction
  - Last lecture: PCA
  - This lecture: Fisher Linear Discriminant

## Data Representation vs. Data Classification

- PCA finds the most accurate *data representation* in a lower dimensional space
  - Project data in the directions of maximum variance
- However the directions of maximum variance may be useless for classification

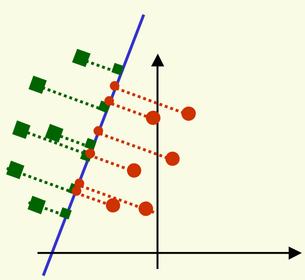



- Fisher Linear Discriminant project to a line which preserves direction useful for *data classification*

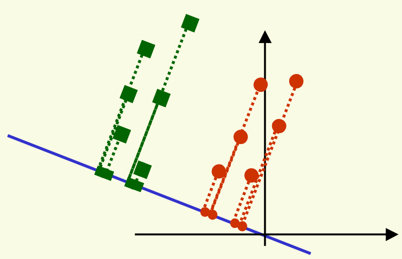
## Fisher Linear Discriminant

- Main idea: find projection to a line s.t. samples from different classes are well separated

**Example in 2D**



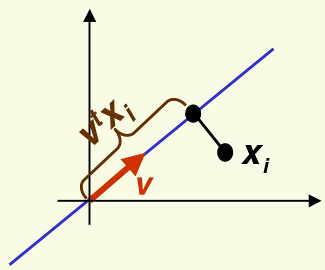
*bad line to project to, classes are mixed up*



*good line to project to, classes are well separated*

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

## Fisher Linear Discriminant

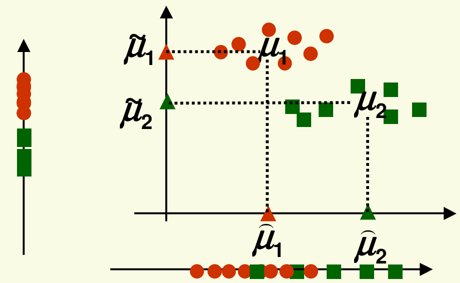
- Thus the projection of sample $\boldsymbol{x_i}$ onto a line in direction $\boldsymbol{v}$ is given by $\boldsymbol{v^t x_i}$
- How to measure separation between projections of different classes?
- Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ be the means of projections of classes 1 and 2
- Let $\mu_1$ and $\mu_2$ be the means of classes 1 and 2
- $|\tilde{\mu}_1 - \tilde{\mu}_2|$ seems like a good measure

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

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

## Fisher Linear Discriminant

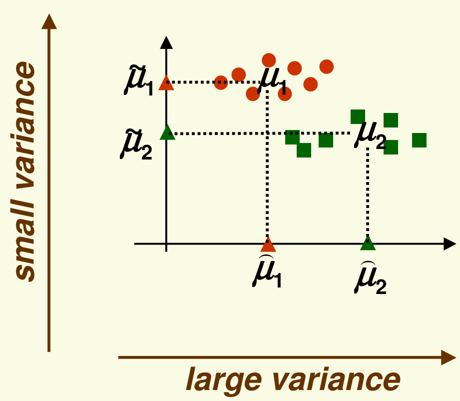
- How good is $|\tilde{\mu}_1 - \tilde{\mu}_2|$ as a measure of separation?
  - The larger $|\tilde{\mu}_1 - \tilde{\mu}_2|$, the better is the expected separation




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

## Fisher Linear Discriminant

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

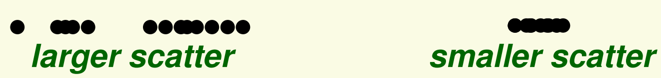


## Fisher Linear Discriminant

- We need to normalize $|\tilde{\mu}_1 - \tilde{\mu}_2|$ by a factor which is proportional to variance
- 1D samples $z_1, \dots, z_n$. 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




## Fisher Linear Discriminant

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

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

- Scatter for projected samples of class 2 is

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

## 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

***want projected means are far from each other***

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

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

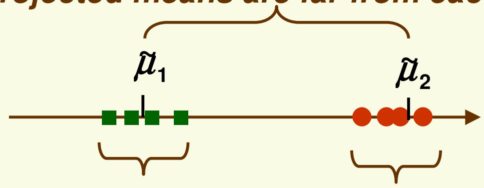
***want scatter in class 2 is as small as possible, i.e. samples of class 2 cluster around the projected mean $\tilde{\mu}_2$***

## 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***

## *Fisher Linear Discriminant Derivation*

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

- 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)

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

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

## *Fisher Linear Discriminant Derivation*

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

$$S_W = S_1 + S_2$$

- 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 $\tilde{\mu}_1 = v^t \mu_1$

$$\begin{aligned}
\tilde{s}_1^2 &= \sum_{y_i \in Class\ 1} \left(v^t x_i - v^t \mu_1\right)^2 \\
&= \sum_{y_i \in Class\ 1} \left(v^t (x_i - \mu_1)\right)^t \left(v^t (x_i - \mu_1)\right) \\
&= \sum_{y_i \in Class\ 1} \left((x_i - \mu_1)^t v\right)^t \left((x_i - \mu_1)^t v\right) \\
&= \sum_{y_i \in Class\ 1} v^t (x_i - \mu_1)(x_i - \mu_1)^t v = v^t S_1 v
\end{aligned}$$

## Fisher Linear Discriminant Derivation

- 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 Derivation

- Thus our objective function can be written:

$$J(\mathbf{v}) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{\mathbf{v}^t \mathbf{S}_B \mathbf{v}}{\mathbf{v}^t \mathbf{S}_W \mathbf{v}}$$

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

$$\frac{d}{d\mathbf{v}} J(\mathbf{v}) = \frac{\left(\frac{d}{d\mathbf{v}} \mathbf{v}^t \mathbf{S}_B \mathbf{v}\right) \mathbf{v}^t \mathbf{S}_W \mathbf{v} - \left(\frac{d}{d\mathbf{v}} \mathbf{v}^t \mathbf{S}_W \mathbf{v}\right) \mathbf{v}^t \mathbf{S}_B \mathbf{v}}{\left(\mathbf{v}^t \mathbf{S}_W \mathbf{v}\right)^2}$$

$$= \frac{(2\mathbf{S}_B \mathbf{v}) \mathbf{v}^t \mathbf{S}_W \mathbf{v} - (2\mathbf{S}_W \mathbf{v}) \mathbf{v}^t \mathbf{S}_B \mathbf{v}}{\left(\mathbf{v}^t \mathbf{S}_W \mathbf{v}\right)^2} = 0$$

## Fisher Linear Discriminant Derivation

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

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

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

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

## Fisher Linear Discriminant Derivation

$$S_B v = \lambda S_W v$$

- If $S_W$ has full rank (the inverse exists), can convert this to a standard eigenvalue problem

$$S_W^{-1} S_B v = \lambda v$$

- But $S_B x$ for any vector $x$, points in the same direction as $\mu_1 - \mu_2$

$$S_B x = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t x = (\mu_1 - \mu_2)\overbrace{\left((\mu_1 - \mu_2)^t x\right)}^{\alpha} = \alpha(\mu_1 - \mu_2)$$

- Thus can solve the eigenvalue problem immediately

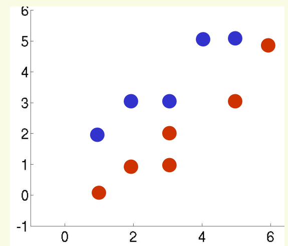
$$\boxed{v = S_W^{-1}(\mu_1 - \mu_2)}$$

$$S_W^{-1} S_B \underbrace{\left[S_W^{-1}(\mu_1 - \mu_2)\right]}_{v} = S_W^{-1}\left[\alpha(\mu_1 - \mu_2)\right] = \underbrace{\alpha}_{\lambda}\underbrace{\left[S_W^{-1}(\mu_1 - \mu_2)\right]}_{v}$$

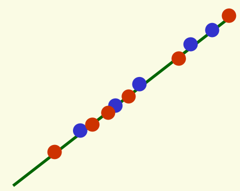
## Fisher Linear Discriminant Example

- 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



## Fisher Linear Discriminant Example

- First compute the mean for each class

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

- Compute scatter matrices $S_1$ and $S_2$ for each class

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

- Within the class scatter:

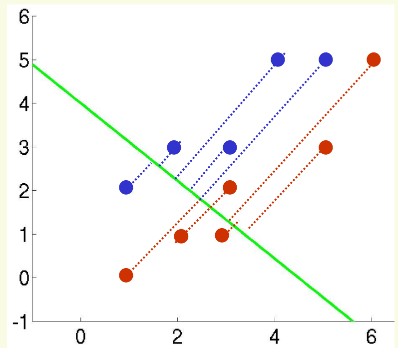
$$S_W = S_1 + S_2 = \begin{bmatrix} 27.3 & 24 \\ 24 & 23.2 \end{bmatrix}$$

  - it has full rank, don't have to solve for eigenvalues
- The inverse of $S_W$ is $S_W^{-1} = inv(S_W) = \begin{bmatrix} 0.39 & -0.41 \\ -0.41 & 0.47 \end{bmatrix}$
- Finally, the optimal line direction $v$

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

## Fisher Linear Discriminant Example

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



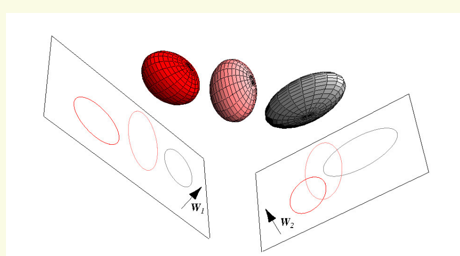

- Last step is to compute the actual ***1D*** vector ***y***. Let's do it separately for each class

$$Y_1 = v^t c_1^t = \begin{bmatrix} -0.79 & 0.89 \end{bmatrix} \begin{bmatrix} 1 & \cdots & 5 \\ 2 & \cdots & 5 \end{bmatrix} = \begin{bmatrix} 0.98 & \cdots & 0.48 \end{bmatrix}$$

$$Y_2 = v^t c_2^t = \begin{bmatrix} -0.79 & 0.89 \end{bmatrix} \begin{bmatrix} 1 & \cdots & 6 \\ 0 & \cdots & 5 \end{bmatrix} = \begin{bmatrix} -0.79 & \cdots & -0.31 \end{bmatrix}$$

## Multiple Discriminant Analysis (MDA)

- Can generalize FLD to multiple classes
- In case of ***c*** classes, can reduce dimensionality to 1, 2, 3,…, ***c***-1 dimensions
- Project sample $x_i$ to a linear subspace $y_i = V^t x_i$
  - ***V*** is called projection matrix



## Multiple Discriminant Analysis (MDA)

- Let
  - $n_i$ by the number of samples of class $i$
  - and $\mu_i$ be the sample mean of class $i$
  - $\mu$ be the total mean of all samples

$$\mu_i = \frac{1}{n_i} \sum_{x \in class\ i} x \qquad \mu = \frac{1}{n} \sum_{x_i} x_i$$

- Objective function: $J(V) = \frac{\det\left(V^t S_B V\right)}{\det\left(V^t S_W V\right)}$
- within the class scatter matrix $S_W$ is

$$S_W = \sum_{i=1}^{c} S_i = \sum_{i=1}^{c} \sum_{x_k \in class\ i} (x_k - \mu_i)(x_k - \mu_i)^t$$

- between the class scatter matrix $S_B$ is

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

maximum rank is c -1

## Multiple Discriminant Analysis (MDA)

- Objective function:

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

- It can be shown that “scatter” of the samples is directly proportional to the determinant of the scatter matrix
  - the larger ***det***(S), the more scattered samples are
  - ***det***(S) is the product of eigenvalues of S
- Thus we are seeking transformation ***V*** which maximizes the between class scatter and minimizes the within-class scatter

## *Multiple Discriminant Analysis (MDA)*

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

- First solve the ***generalized eigenvalue*** problem:

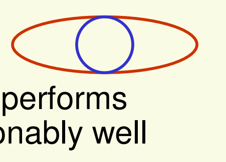
$$S_B v = \lambda S_W v$$

- At most $c$-$1$ distinct solution eigenvalues
- Let $v_1, v_2, \ldots, v_{c-1}$ be the corresponding eigenvectors
- The optimal projection matrix $V$ to a subspace of dimension $k$ is given by the eigenvectors corresponding to the largest $k$ eigenvalues
- Thus can project to a subspace of dimension at most $c$-$1$

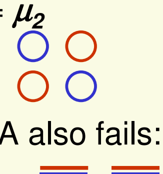
## *FDA and MDA Drawbacks*

- 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:

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