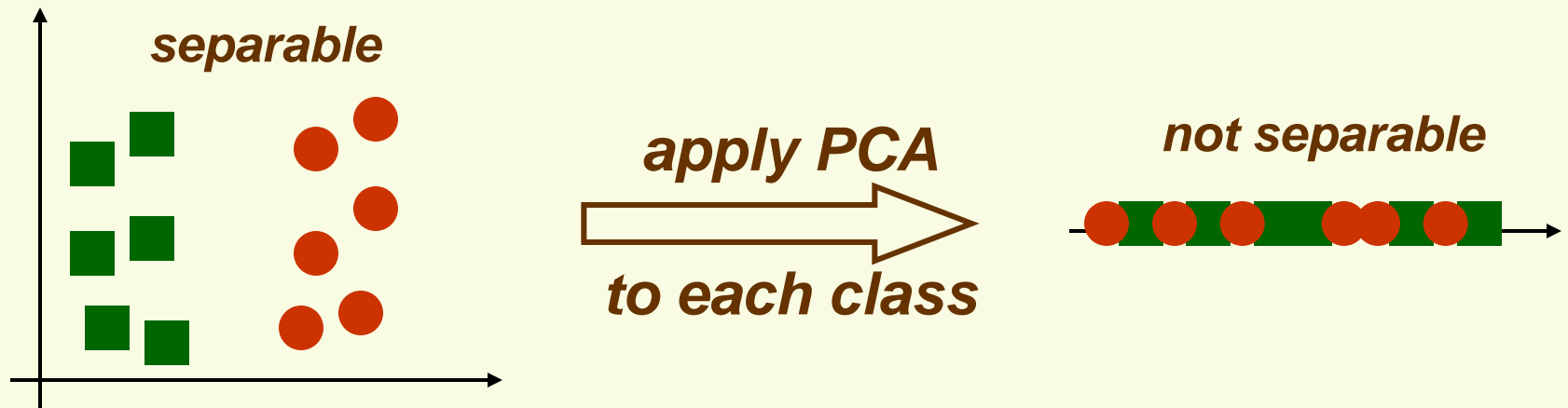


Fisher LDA
MDA

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

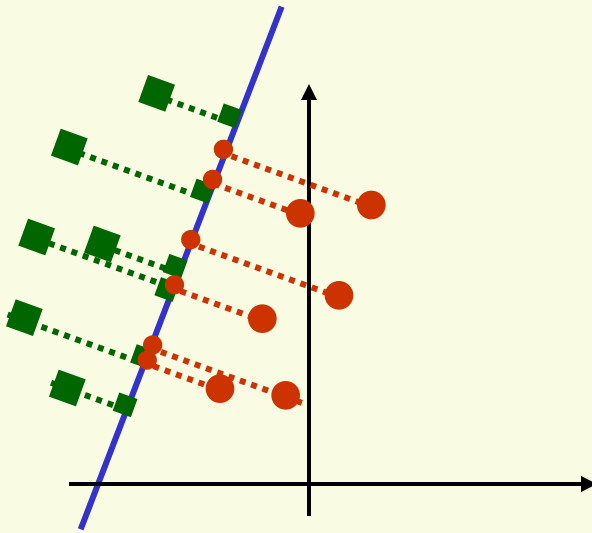


- Fisher Linear Discriminant projects to a line which preserves direction useful for *data classification*

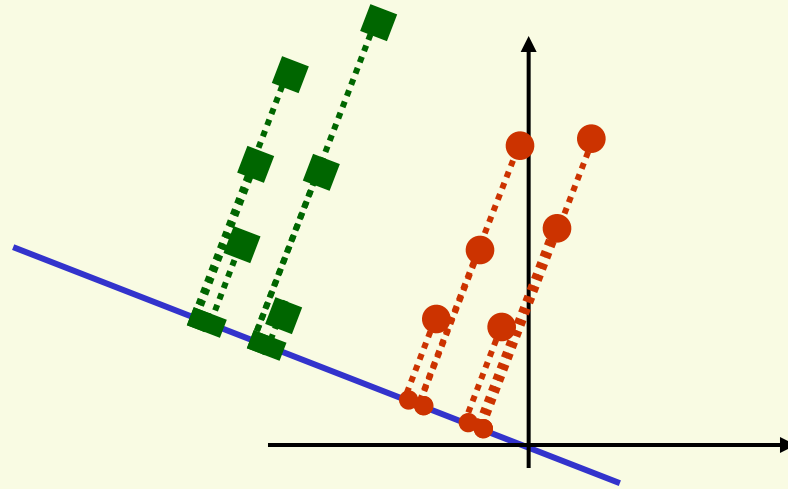
Fisher Linear Discriminant

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

Example in 2D



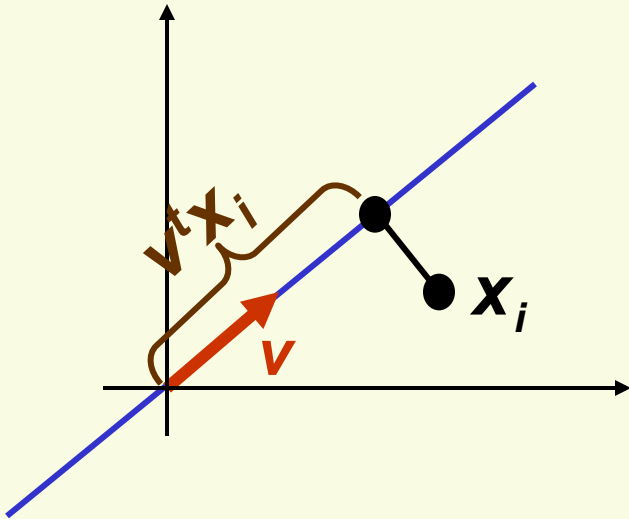
*bad line to project to,
classes are mixed up*



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

Fisher Linear Discriminant

- Suppose we have 2 classes and d -dimensional samples $\mathbf{x}_1, \dots, \mathbf{x}_n$ where
 - n_1 samples come from the first class
 - n_2 samples come from the second class
- consider projection on a line
- Let the line direction be given by unit vector \mathbf{v}



- Thus the projection of sample \mathbf{x}_i onto a line in direction \mathbf{v} is given by $\mathbf{v}^t \mathbf{x}_i$

Fisher Linear Discriminant

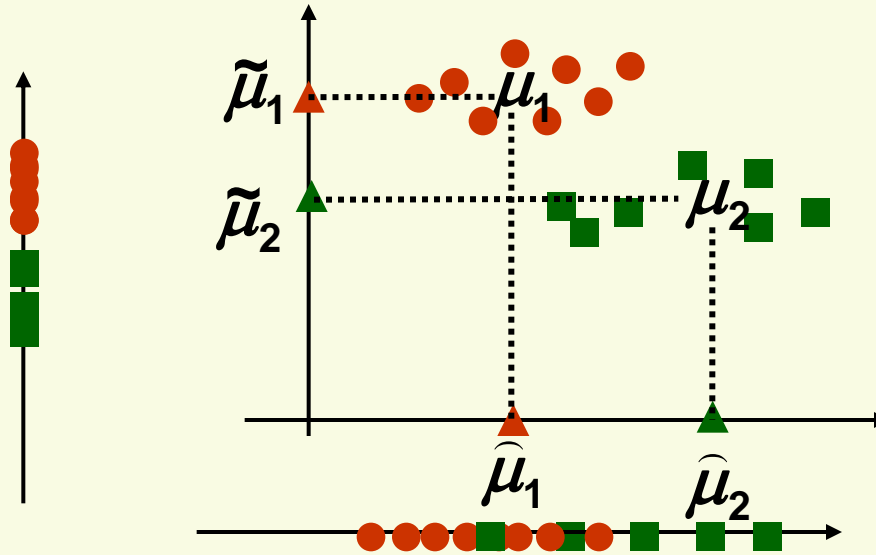
- How to measure separation between projections of different classes?
- Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ be the means of projections of classes 1 and 2
- Let μ_1 and μ_2 be the means of classes 1 and 2
- $|\tilde{\mu}_1 - \tilde{\mu}_2|$ seems like a good measure

$$\tilde{\mu}_1 = \frac{1}{n_1} \sum_{x_i \in C_1} v^t x_i = v^t \left(\frac{1}{n_1} \sum_{x_i \in C_1} x_i \right) = v^t \mu_1$$

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

Fisher Linear Discriminant

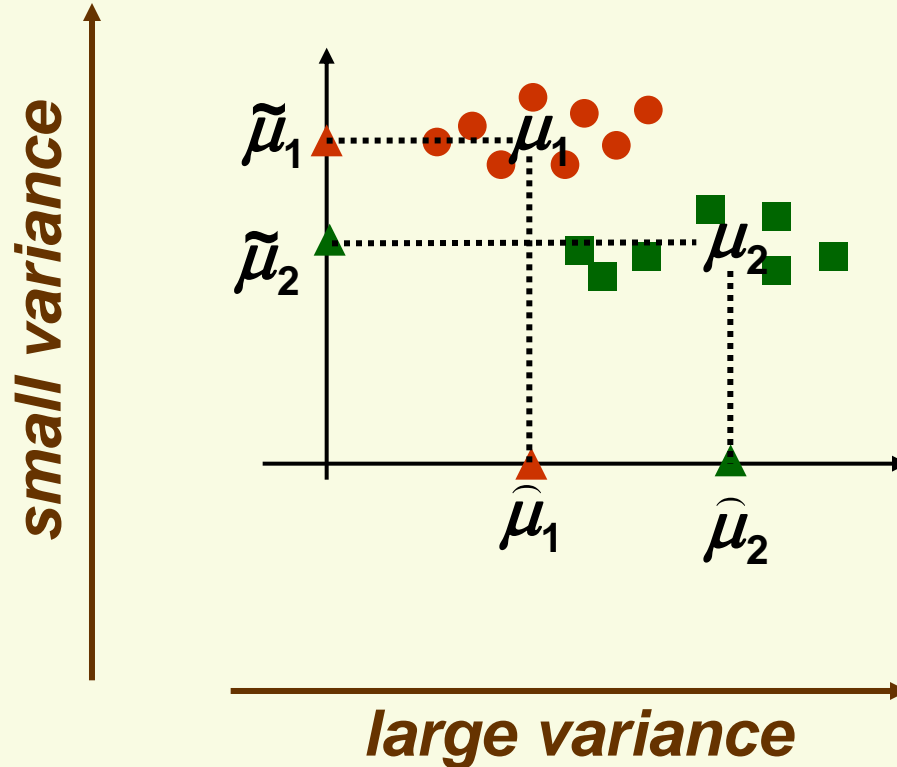
- How good is $|\tilde{\mu}_1 - \tilde{\mu}_2|$ as a measure of separation?
 - The larger $|\tilde{\mu}_1 - \tilde{\mu}_2|$, the better is the expected separation



- 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 $\mathbf{z}_1, \dots, \mathbf{z}_n$. 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

Fisher Linear Discriminant

- **Fisher Solution**: normalize $|\tilde{\mu}_1 - \tilde{\mu}_2|$ by scatter
- Let $\mathbf{y}_i = \mathbf{v}^t \mathbf{x}_i$, i.e. \mathbf{y}_i 's are the projected samples

- Scatter for projected samples of class 1 is

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

- Scatter for projected samples of class 2 is

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

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 \mathbf{v} which maximizes

want projected means are far from each other

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

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

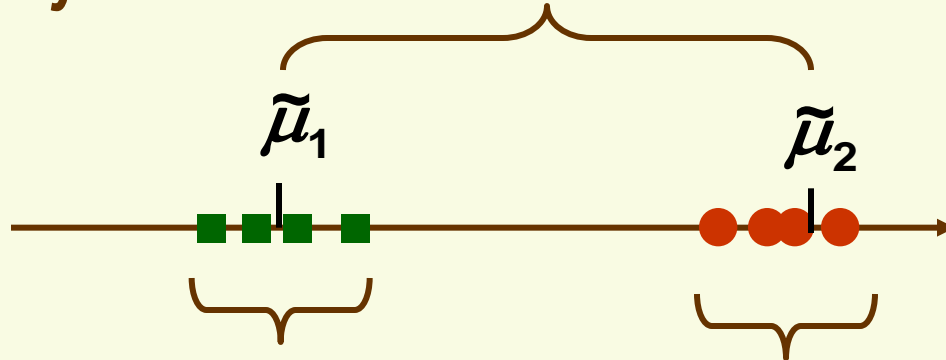
want scatter in class 2 to be as small as possible, i.e. samples of class 2 cluster around the projected mean $\tilde{\mu}_2$

Fisher Linear Discriminant

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

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

projected means are far from each other



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

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

Fisher Linear Discriminant Derivation

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

- All we need to do now is to express J explicitly as a function of \mathbf{v} and maximize it
 - straightforward but need linear algebra and Calculus (the derivation is shown in the next few slides.)
 - The solution is found by **generalized eigenvalue problem** $\Rightarrow \mathbf{S}_B \mathbf{v} = \lambda \mathbf{S}_W \mathbf{v}$

between class scatter matrix $\mathbf{S}_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^t$

within the class scatter matrix $\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$

$$\mathbf{S}_1 = \sum_{x_i \in \text{Class1}} (\mathbf{x}_i - \mu_1)(\mathbf{x}_i - \mu_1)^t$$

$$\mathbf{S}_2 = \sum_{x_i \in \text{Class2}} (\mathbf{x}_i - \mu_2)(\mathbf{x}_i - \mu_2)^t$$

Fisher Linear Discriminant Derivation

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

- Define the separate class scatter matrices \mathbf{S}_1 and \mathbf{S}_2 for classes 1 and 2. These measure the scatter of original samples \mathbf{x}_i (before projection)

$$\mathbf{S}_1 = \sum_{\mathbf{x}_i \in \text{Class1}} (\mathbf{x}_i - \mu_1)(\mathbf{x}_i - \mu_1)^t$$

$$\mathbf{S}_2 = \sum_{\mathbf{x}_i \in \text{Class2}} (\mathbf{x}_i - \mu_2)(\mathbf{x}_i - \mu_2)^t$$

Fisher Linear Discriminant Derivation

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

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

- Recall that $\tilde{\mathbf{S}}_1^2 = \sum_{y_i \in \text{Class 1}} (\mathbf{y}_i - \tilde{\boldsymbol{\mu}}_1)^2$

- Using $\mathbf{y}_i = \mathbf{v}^t \mathbf{x}_i$ and $\tilde{\boldsymbol{\mu}}_1 = \mathbf{v}^t \boldsymbol{\mu}_1$

$$\begin{aligned}\tilde{\mathbf{S}}_1^2 &= \sum_{y_i \in \text{Class 1}} (\mathbf{v}^t \mathbf{x}_i - \mathbf{v}^t \boldsymbol{\mu}_1)^2 \\ &= \sum_{y_i \in \text{Class 1}} (\mathbf{v}^t (\mathbf{x}_i - \boldsymbol{\mu}_1))^t (\mathbf{v}^t (\mathbf{x}_i - \boldsymbol{\mu}_1)) \\ &= \sum_{y_i \in \text{Class 1}} ((\mathbf{x}_i - \boldsymbol{\mu}_1)^t \mathbf{v})^t ((\mathbf{x}_i - \boldsymbol{\mu}_1)^t \mathbf{v}) \\ &= \sum_{y_i \in \text{Class 1}} \mathbf{v}^t (\mathbf{x}_i - \boldsymbol{\mu}_1) (\mathbf{x}_i - \boldsymbol{\mu}_1)^t \mathbf{v} = \mathbf{v}^t \mathbf{S}_1 \mathbf{v}\end{aligned}$$

Fisher Linear Discriminant Derivation

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

$$\mathbf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t$$

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

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

Fisher Linear Discriminant Derivation

- Thus our objective function can be written:

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

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

$$\begin{aligned} \frac{d}{d\mathbf{v}} J(\mathbf{v}) &= \frac{\left(\frac{d}{d\mathbf{v}} \mathbf{v}^t \mathbf{S}_B \mathbf{v} \right) \mathbf{v}^t \mathbf{S}_W \mathbf{v} - \left(\frac{d}{d\mathbf{v}} \mathbf{v}^t \mathbf{S}_W \mathbf{v} \right) \mathbf{v}^t \mathbf{S}_B \mathbf{v}}{\left(\mathbf{v}^t \mathbf{S}_W \mathbf{v} \right)^2} \\ &= \frac{(2\mathbf{S}_B \mathbf{v}) \mathbf{v}^t \mathbf{S}_W \mathbf{v} - (2\mathbf{S}_W \mathbf{v}) \mathbf{v}^t \mathbf{S}_B \mathbf{v}}{\left(\mathbf{v}^t \mathbf{S}_W \mathbf{v} \right)^2} = 0 \end{aligned}$$

Fisher Linear Discriminant Derivation

- Need to solve $\mathbf{v}^t \mathbf{S}_W \mathbf{v} (\mathbf{S}_B \mathbf{v}) - \mathbf{v}^t \mathbf{S}_B \mathbf{v} (\mathbf{S}_W \mathbf{v}) = 0$

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

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

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

generalized eigenvalue problem

Fisher Linear Discriminant Derivation

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

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

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

- Turn's out that we don't have to solve for eigenvalues; the solution is:

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

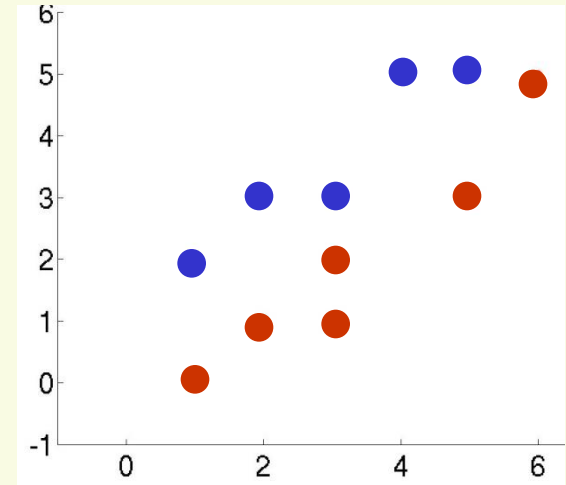
Fisher Linear Discriminant Example

- Data

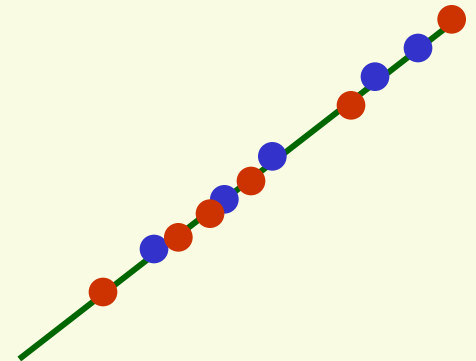
- Class 1 has 5 samples $\mathbf{c}_1 = [(1,2), (2,3), (3,3), (4,5), (5,5)]$
- Class 2 has 6 samples $\mathbf{c}_2 = [(1,0), (2,1), (3,1), (3,2), (5,3), (6,5)]$

- Arrange data in 2 separate matrices

$$\mathbf{c}_1 = \begin{bmatrix} 1 & 2 \\ \vdots & \vdots \\ 5 & 5 \end{bmatrix} \quad \mathbf{c}_2 = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 6 & 5 \end{bmatrix}$$



- 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

$$\mu_1 = \text{mean}(c_1) = [3 \quad 3.6]$$

$$\mu_2 = \text{mean}(c_2) = [3.3 \quad 2]$$

- Compute scatter matrices \mathbf{S}_1 and \mathbf{S}_2 for each class

$$\mathbf{S}_1 = 4 * \text{cov}(c_1) = \begin{bmatrix} 10 & 8.0 \\ 8.0 & 7.2 \end{bmatrix}$$

$$\mathbf{S}_2 = 5 * \text{cov}(c_2) = \begin{bmatrix} 17.3 & 16 \\ 16 & 16 \end{bmatrix}$$

- Within the class scatter:

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2 = \begin{bmatrix} 27.3 & 24 \\ 24 & 23.2 \end{bmatrix}$$

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

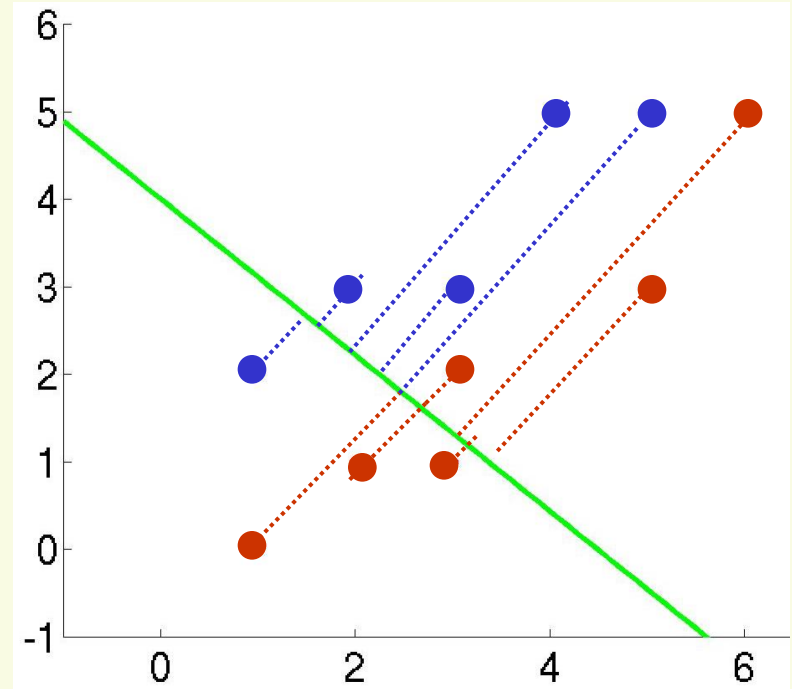
- The inverse of \mathbf{S}_W is $\mathbf{S}_W^{-1} = \text{inv}(\mathbf{S}_W) = \begin{bmatrix} 0.39 & -0.41 \\ -0.41 & 0.47 \end{bmatrix}$

- Finally, the optimal line direction \mathbf{v}

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

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 \mathbf{y} . Let's do it separately for each class

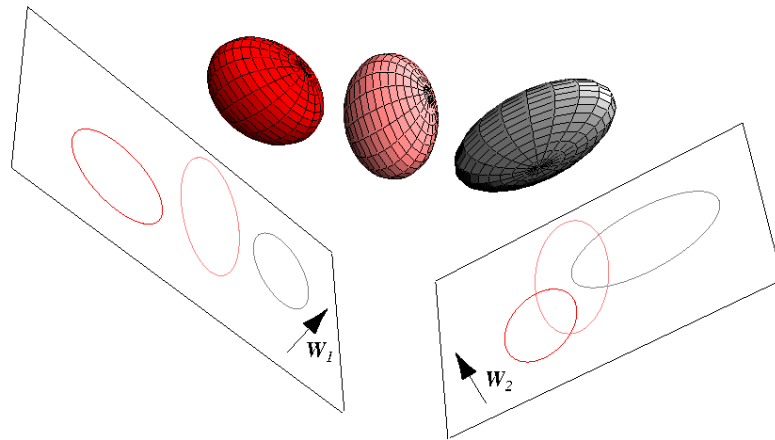


$$Y_1 = \mathbf{v}^t \mathbf{c}_1^t = [-0.79 \quad 0.89] \begin{bmatrix} 1 & \cdots & 5 \\ 2 & \cdots & 5 \end{bmatrix} = [0.98 \quad \cdots \quad 0.48]$$

$$Y_2 = \mathbf{v}^t \mathbf{c}_2^t = [-0.79 \quad 0.89] \begin{bmatrix} 1 & \cdots & 6 \\ 0 & \cdots & 5 \end{bmatrix} = [-0.79 \quad \cdots \quad -0.31]$$

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 \mathbf{x}_i to a linear subspace $\mathbf{y}_i = \mathbf{V}^t \mathbf{x}_i$
 - \mathbf{V} is called projection matrix



Multiple Discriminant Analysis (MDA)

- Let
 - n_i by the number of samples of class i
 - and μ_i be the sample mean of class i
 - μ be the total mean of all samples

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

- Objective function: $J(V) = \frac{\det(V^t S_B V)}{\det(V^t S_W V)}$

- within the class scatter matrix S_W is

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

- 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:

$$J(\mathbf{V}) = \frac{\det(\mathbf{V}^t \mathbf{S}_B \mathbf{V})}{\det(\mathbf{V}^t \mathbf{S}_W \mathbf{V})}$$

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

Multiple Discriminant Analysis (MDA)

$$J(\mathbf{V}) = \frac{\det(\mathbf{V}^t \mathbf{S}_B \mathbf{V})}{\det(\mathbf{V}^t \mathbf{S}_W \mathbf{V})}$$

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

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

- At most **$c-1$** eigenvalues are nonzero.
- Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{c-1}$ be the corresponding eigenvectors
- The optimal projection matrix \mathbf{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 **$c-1$**