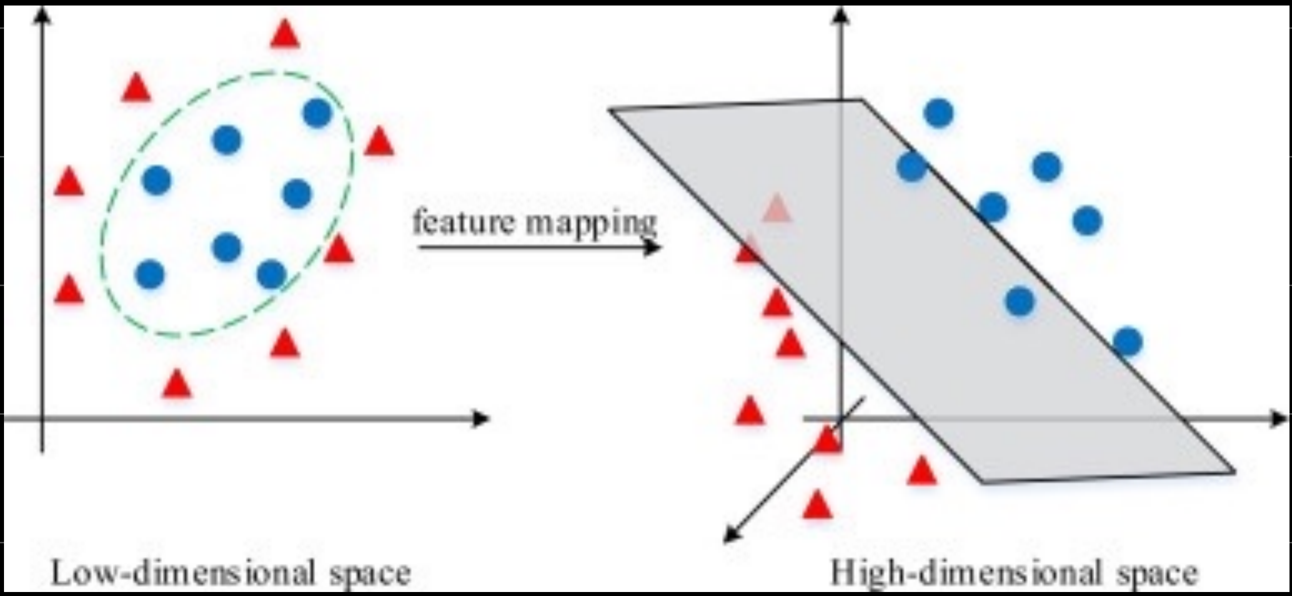
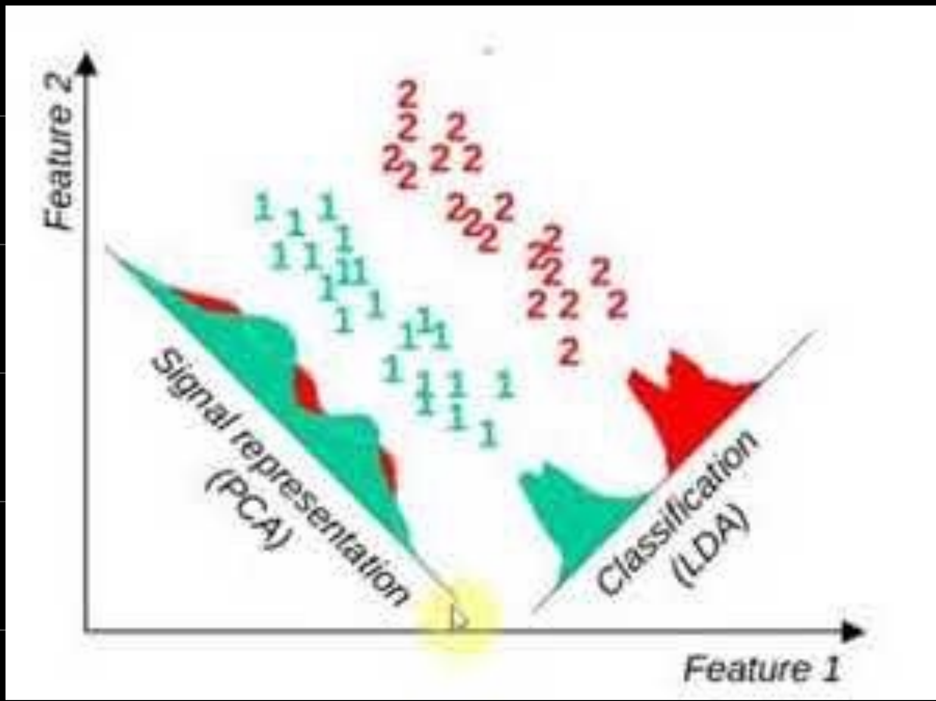
4.3 Linear Discriminant Analysis

4.3.1 Motivation.



Observation: high dimensions make data classification easier, but data representation harder. It would be ideal if we can classify data in high dimensions then represent it in low dimensions.



Observation: PCA may fail to preserve data separability.

Linear Discriminant Analysis (LDA): a method to reduce dimensions while preserving data separability.

Setting: sample pts with labels $(x^{(1)}, y_1) \cdots (x^{(n)}, y_n)$ with $y_i = \pm 1$.

Set $\mathcal{C}_1 \overset{\text{def}}{=} \{ x^{(i)} : y_i = 1 \}$

$$\mathscr{C}_2 \overset{\text{def}}{=} \{ x^{(i)} : y_i = -1 \}.$$

$n_i \overset{\text{def}}{=}$ # of sample pts in $\mathscr{C}_i$. $i = 1, 2$

Then the means of $\mathscr{C}_1$ and $\mathscr{C}_2$ are

$$\mu^{(1)} = \frac{1}{n_1} \sum_{x^{(i)} \in \mathscr{C}_1} x^{(i)}$$

$$\mu^{(2)} = \frac{1}{n_2} \sum_{x^{(i)} \in \mathscr{C}_2} x^{(i)}$$

The sample covariance matrices are

$$C^{(1)} = \frac{1}{n_1} \sum_{x^{(i)} \in \mathscr{C}_1} (x^{(i)} - \mu^{(1)})(x^{(i)} - \mu^{(1)})^T$$

$$C^{(2)} = \frac{1}{n_2} \sum_{x^{(i)} \in \mathscr{C}_2} (x^{(i)} - \mu^{(2)})(x^{(i)} - \mu^{(2)})^T$$

(If $\mathscr{C}_1$ is centered, then $\mu^{(1)} = 0$

and $C^{(1)} = \frac{1}{n_1} X^{(1)} X^{(1)T}$ where

$$X^{(1)} = \begin{pmatrix} 1 \\ x^{(i)} \\ 1 \end{pmatrix}_{x^{(i)} \in \mathcal{C}_1}$$ is the sample matrix)

Recall that $C^{(1)}$, $C^{(2)}$ are sym, pos semi-def.

First, we project $\mathcal{C}_1$, $\mathcal{C}_2$ to 1D.

Let u be a unit vector, then the projected sample means are

$$\tilde{\mu}^{(1)} \overset{\text{def}}{=} \mathrm{Proj}_{\mathrm{span}\{u\}} \mu^{(1)} = \left(\mu^{(1)T} u\right) u$$

$$\tilde{\mu}^{(2)} \overset{\text{def}}{=} \mathrm{Proj}_{\mathrm{span}\{u\}} \mu^{(2)} = \left(\mu^{(2)T} u\right) u.$$

The projected variances are

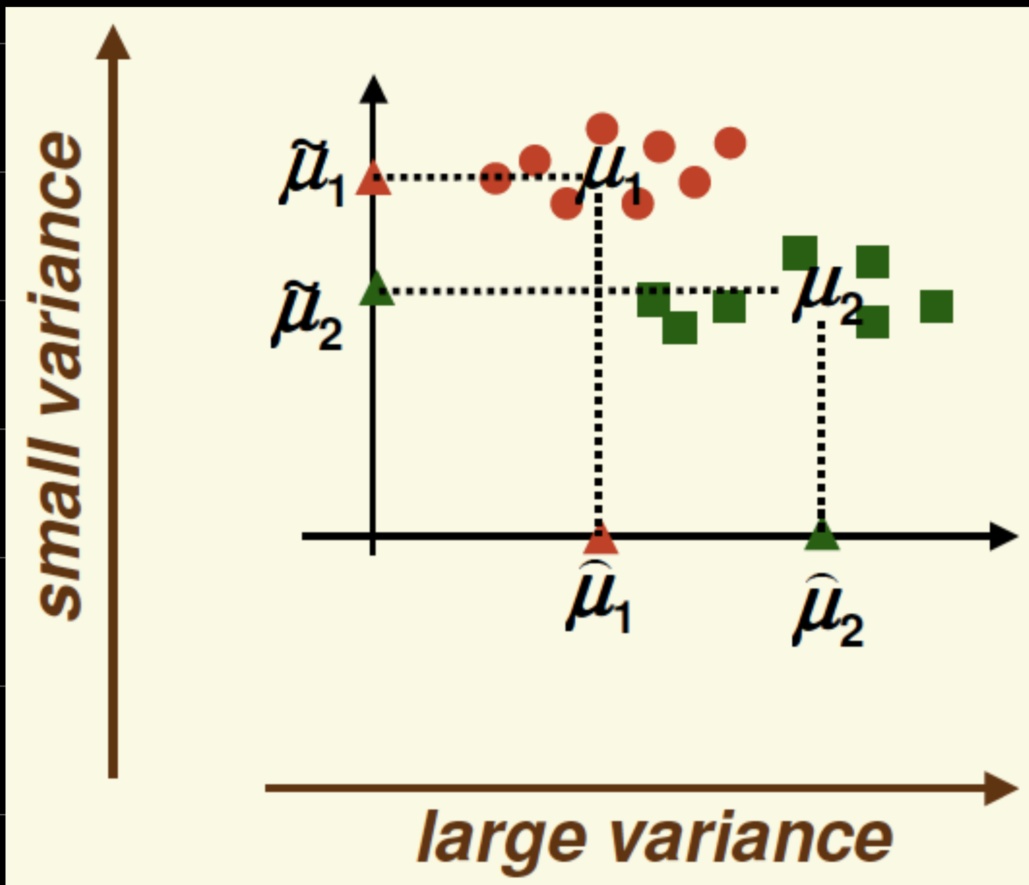
$$\tilde{C}^{(1)} = u^T C^{(1)} u$$

(see Lecture 12)

$$\tilde{C}^{(2)} = u^T C^{(2)} u$$

4.3.2 Theory.

Q: What is a good metric for data separability?



Observation: $\|\tilde{\mu}^{(1)} - \tilde{\mu}^{(2)}\|$ itself is not a good metric, as data in each class may be scattered.

Fisher's idea: find a unit vector u s.t.

$$\max_{\|u\|=1} \frac{\|\tilde{\mu}^{(1)} - \tilde{\mu}^{(2)}\|^2}{n_1 \widetilde{C}_1 + n_2 \widetilde{C}_2}$$

i.e. look for large $\|\tilde{\mu}^{(1)} - \tilde{\mu}^{(2)}\|$ with small scattering.

Notice: $$\frac{\|\tilde{\mu}^{(1)} - \tilde{\mu}^{(2)}\|^2}{n_1 \tilde{C}^{(1)} + n_2 \tilde{C}^{(2)}} \overset{\text{def of } \tilde{\mu}^{(i)} \text{ and } \tilde{C}^{(i)}}{=\!=}$$

$$\frac{\|(\mu^{(1)T} u - \mu^{(2)T} u) u\|^2}{n_1 u^T C^{(1)} u + n_2 u^T C^{(2)} u} \overset{\text{def of inner product}}{=}$$

$$\frac{u^T \overbrace{(\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})^T}^{S_B} u}{u^T \underbrace{(n_1 C_1 + n_2 C_2)}_{S_W} u} =$$

$$\frac{u^T S_B u}{u^T S_W u}$$ S_B, S_W are sym.