LINEAR DISCRIMINANT ANALYSIS

Generative Models for Classification

The Bayes classifier requires knowledge of the distribution on (X, Y). However, this distrib. is often unknown, and we only have training data $(x_1, y_1), \ldots, (x_n, y_n)$, $x_i \in \mathbb{R}^d$, $y_i \in \{1, \ldots, K\}$.

A <u>generative model</u> is an assumption about the true form of this unknown distribution. Generative models are typically <u>parametric</u>. To build a classifier, we may estimate the model parameters using the training data, a plug the result into the formula for the Bayes classifier. For this reason, such methods are also called "plug-in" methods.

LDA In LDA, we assume

$$X \mid Y = k \sim N(\mu_k, \Sigma), \quad k = 1, \ldots, K$$

Here $N(\mu, \Sigma)$ is the multivariate Gaussian / normal distribution with parameters μ, Σ, and pdf

$$\phi(x; \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

Note: each class has the same covariance matrix Σ

Parameter Estimation

$$\hat{\pi}_k = \frac{|\{i : y_i = k\}|}{n}$$

$$\hat{\mu}_k = \frac{1}{|\{i : y_i = k\}|} \sum_{i : y_i = k} x_i$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i : y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

"pooled covariance estimate"

Interpretations The LDA classifier is then

$$f(x) = \arg\max_k \ \hat{\pi}_k \cdot \phi(x; \hat{\mu}_k, \hat{\Sigma})$$

$$= \arg\max_k \ \log \hat{\pi}_k + \log \phi(x; \hat{\mu}_k, \hat{\Sigma})$$

$$= \arg\min_k \ \underbrace{(x-\hat{\mu}_k)^T \hat{\Sigma}^{-1} (x-\hat{\mu}_k)}_{\text{squared Mahalanobis distance between } x \text{ and } \hat{\mu}_k} - 2\log \hat{\pi}_k$$

squared Mahalanobis distance between x and $\hat{\mu}_k$

Case $K=2$

$$(x-\hat{\mu}_1)^T \hat{\Sigma}^{-1} (x-\hat{\mu}_1) - 2\log \hat{\pi}_1 \underset{2}{\overset{1}{\lessgtr}} (x-\hat{\mu}_2)^T \hat{\Sigma}^{-1} (x-\hat{\mu}_2) - 2\log \hat{\pi}_2$$

$$\iff a^T x + b \underset{2}{\overset{1}{\gtrless}} 0$$

linear classifier

where

$$a = \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2)$$

$$b = -\frac{1}{2} \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1 + \frac{1}{2} \hat{\mu}_2 \hat{\Sigma}^{-1} \hat{\mu}_2 + \log \frac{\hat{\pi}_1}{\hat{\pi}_2}$$

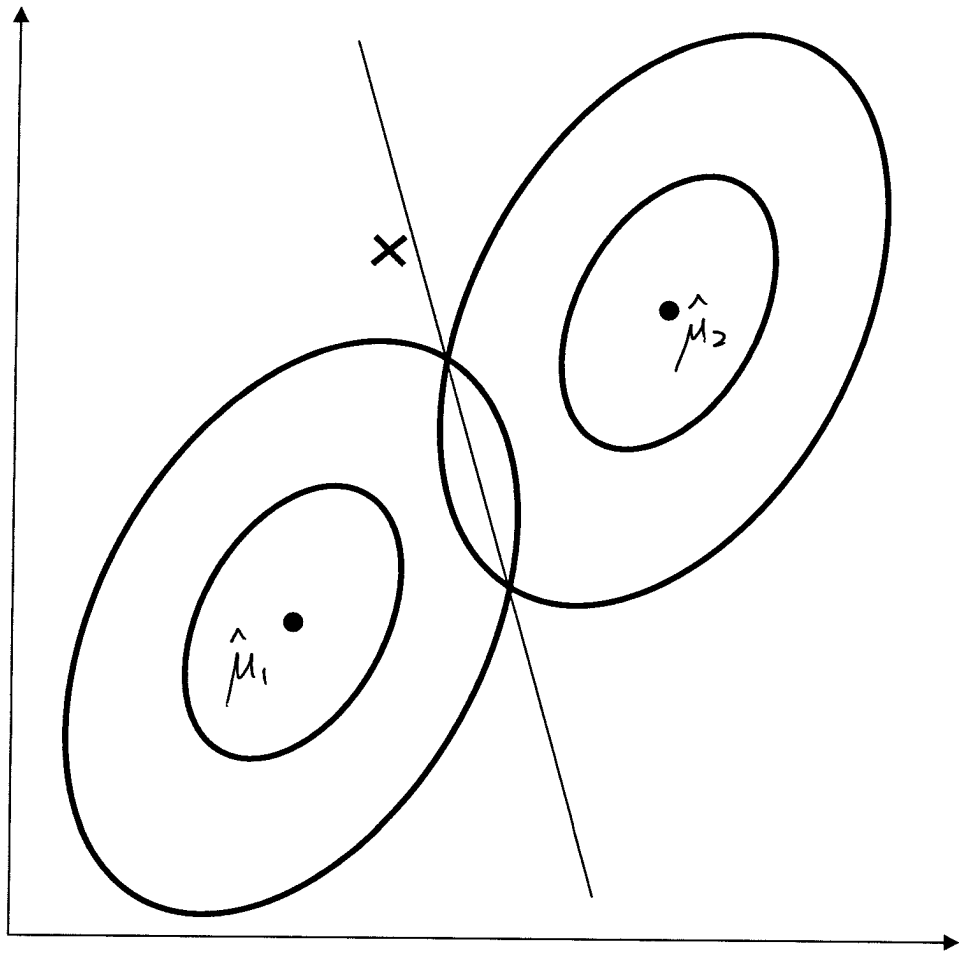
Denote the Mahalanobis distance

$$d_M(x; \mu, \Sigma) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Recall that the contour

$$\{x : d_M(x; \mu, \Sigma) = c\}$$

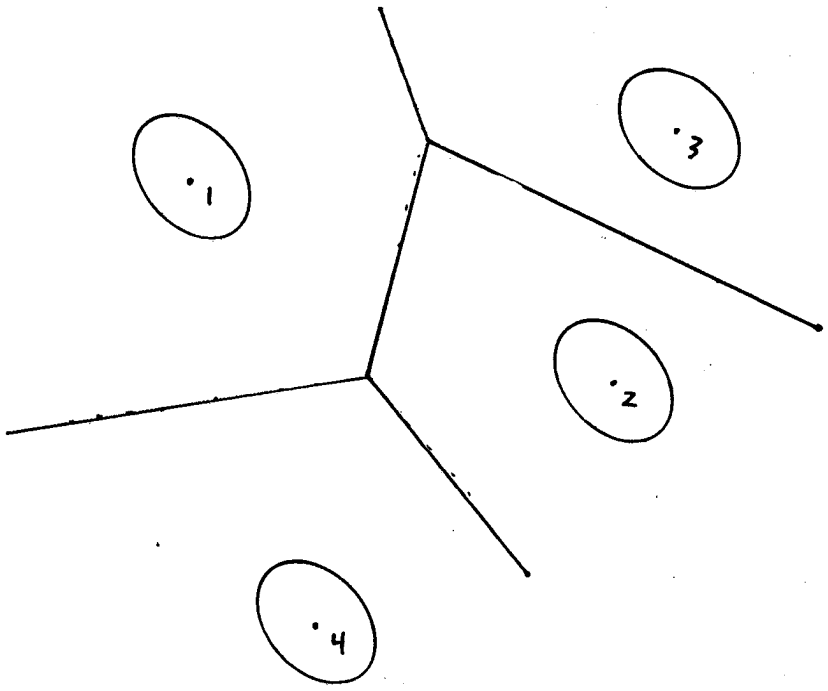
(A) is an



This picture assumes $\hat{\pi}_1 = \hat{\pi}_2$

Case $K > 2$

The decision regions are convex polytopes (intersections of linear half-spaces)



Issues w/ LDA

The number of parameters to be estimated is

(B)

if n is too small, we could

Example | (of a structured covariance matrix)

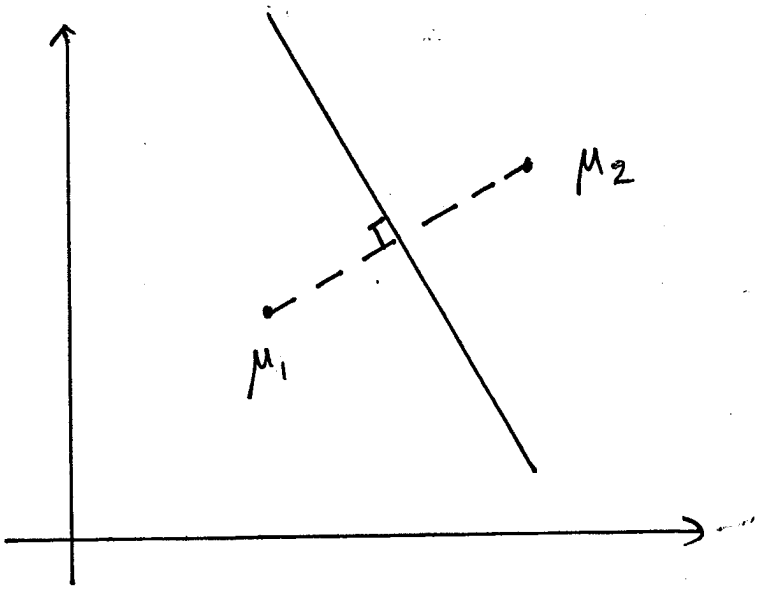
(c) $\Sigma = \sigma^2 I \Rightarrow \hat{\sigma}^2 =$

Then LDA becomes (assuming $K=2$, $\hat{\pi}_1 = \hat{\pi}_2$)

$$\frac{1}{\hat{\sigma}^2}\|x-\hat{\mu}_1\|^2 \underset{1}{\overset{2}{\gtrless}} \frac{1}{\hat{\sigma}^2}\|x-\hat{\mu}_2\|^2$$

$$\Longleftrightarrow \quad \|x-\hat{\mu}_1\|^2 \underset{1}{\overset{2}{\gtrless}} \|x-\hat{\mu}_2\|^2$$

which is called the _____ _____ classifier.



QDA Generative model with

$$X \mid Y = k \sim N(\mu_k, \Sigma_k), \quad k = 1, \ldots, K$$

$$\hat{\Sigma}_k = \frac{1}{|\{i : y_i = k\}|} \sum_{i : y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

Decision boundaries are now <u>quadratic</u>

Key A. ellipse

B. $Kd + \frac{d(d+1)}{2} + K - 1$

↑ means ↑ covariance ↑ class prior probs.

could

- first apply PCA
- assume a structured covariance matrix

C. $\hat{\sigma}^2 = \frac{1}{d} tr\{\hat{\Sigma}\}$

nearest centroid classifier