

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 $(x_1, y_1), \dots, (x_n, y_n)$, $x_i \in \mathbb{R}^d$, $y_i \in \{1, \dots, K\}$.

A generative model is an assumption about the true form of this unknown distribution.

Generative models are typically parametric.

To build a classifier, we may estimate the model parameters using the training data, and 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

$$X | Y = k \sim \mathcal{N}(\mu_k, \Sigma), \quad k = 1, \dots, K$$

Here $\mathcal{N}(\mu, \Sigma)$ 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

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

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

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

Case K=2

$$(x - \hat{\mu}_1)^T \hat{\Sigma}^{-1} (x - \hat{\mu}_1) - 2 \log \hat{\pi}_1 \stackrel{!}{\leq} (x - \hat{\mu}_2)^T \hat{\Sigma}^{-1} (x - \hat{\mu}_2) - 2 \log \hat{\pi}_2$$

$$\Leftrightarrow a^T x + b \stackrel{!}{\leq} 0$$

linear
classifier

where

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

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

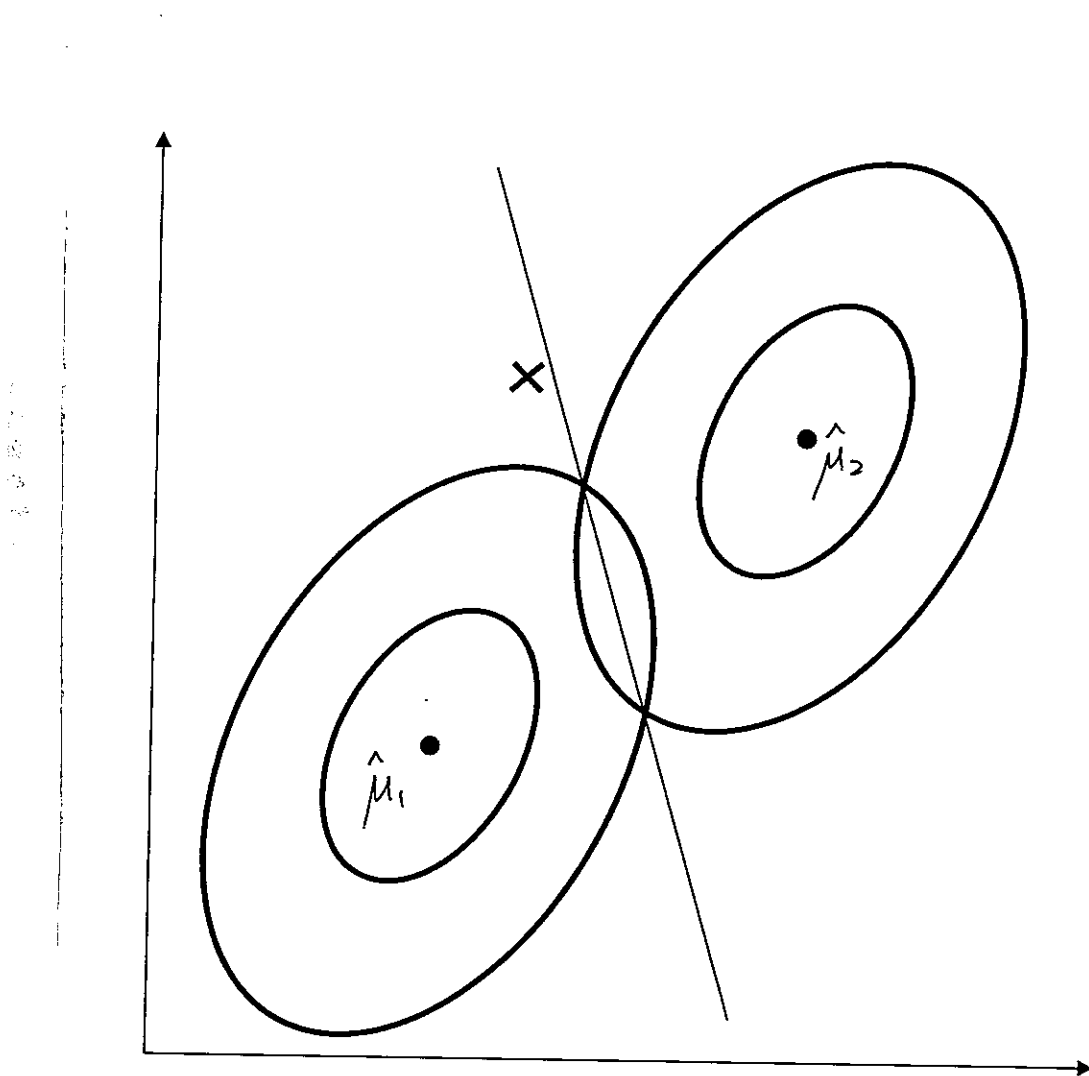
Denote the Mahalanobis distance

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

Recall that the contour

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

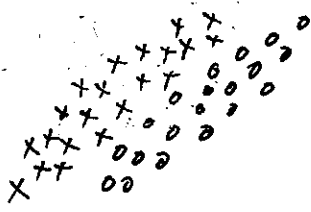
(A) is an



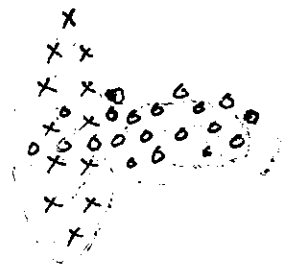
This picture assumes $\hat{\pi}_1 = \hat{\pi}_2$

Would LDA be appropriate:

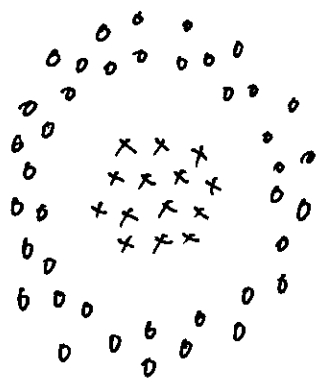
(i)



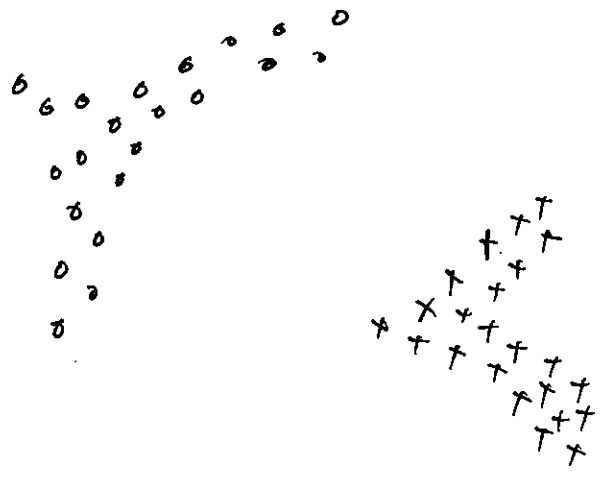
(ii)



(iii)

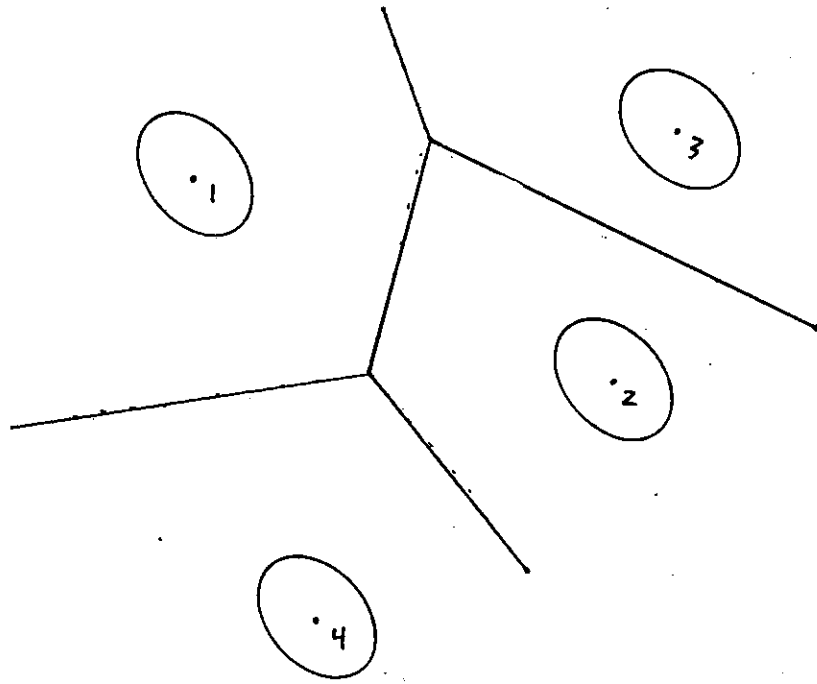


(iv)



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

- Generative model is rarely valid

Example | (of a structured covariance matrix)

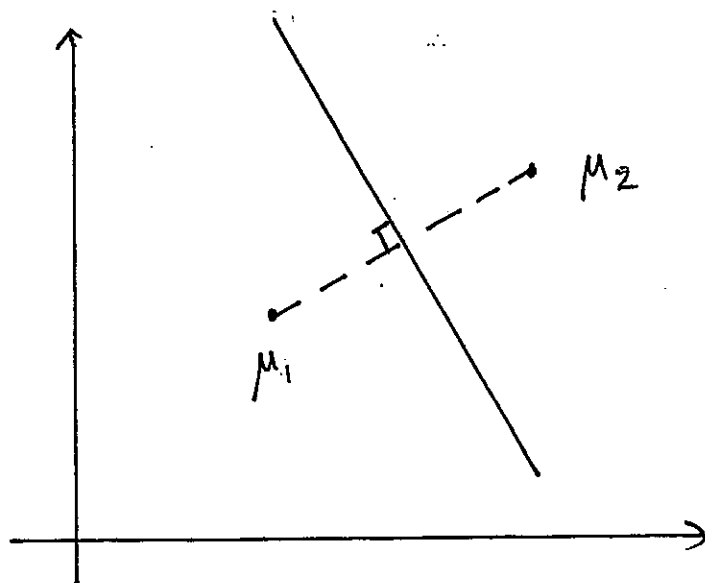
① $\Sigma = \sigma^2 I \Rightarrow \hat{\sigma}_1^2 =$

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

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

$$\Leftrightarrow \|x - \hat{\mu}_1\|^2 \stackrel{2}{\underset{1}{>}} \|x - \hat{\mu}_2\|^2$$

which is called the _____ classifier.



QDA

Generative model with

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

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

Decision boundaries are now quadratic

Key

A. ellipse

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

↑ means

↑ covariance

↑ class prior probs.

could

- first apply PCA
- assume a structured covariance matrix

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

nearest centroid classifier