LINEAR DISCRIMINANT ANALYSIS

Generative Models for Classification

The Bayes classifier requires knowledge of the distribution on \( (X, Y) \). However, this distribution 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 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, plug the result into the formula for the Bayes classifier. For this reason, such methods are also called "plug-in" methods.
In LDA, we assume

\[ X \mid y = k \sim \mathcal{N}(\mu_k, \Sigma), \quad k = 1, \ldots, K \]

Here \( \mathcal{N}(\mu, \Sigma) \) is the multivariate Gaussian/normal distribution with parameters \( \mu, \Sigma \), and pdf

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

**Note**: each class has the same covariance matrix \( \Sigma \)

### Parameter Estimation

\[
\hat{\pi}_k = \frac{\left| \{ i : y_i = k \} \right|}{n}
\]

\[
\hat{\mu}_k = \frac{1}{\left| \{ i : y_i = k \} \right|} \sum_{i : y_i = k} x_i
\]

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

"pooled covariance estimate"
The LDA classifier is then

\[ f(x) = \arg \max_k \pi_k \cdot \phi(x; \hat{\mu}_k, \hat{\Sigma}) \]

\[ = \arg \max_k \log \pi_k + \log \phi(x; \hat{\mu}_k, \hat{\Sigma}) \]

\[ = \arg \min_k (x - \hat{\mu}_k)^T \hat{\Sigma}^{-1} (x - \hat{\mu}_k) - 2 \log \pi_k \]

Squared Mahalanobis distance between \( x \) and \( \hat{\mu}_k \).

Case \( K = 2 \)

\[ (x - \hat{\mu}_1)^T \hat{\Sigma}^{-1} (x - \hat{\mu}_1) - 2 \log \pi_1 \overset{!}{\leq} (x - \hat{\mu}_2)^T \hat{\Sigma}^{-1} (x - \hat{\mu}_2) - 2 \log \pi_2 \]

\[ \iff \quad a^T x + b \overset{!}{\geq} 0 \]

Linear classifier

where

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

\[ b = -\frac{1}{2} \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1 + \frac{1}{2} \hat{\mu}_2^T \hat{\Sigma}^{-1} \hat{\mu}_2 + \log \frac{\pi_1}{\pi_2} \]
Denote the Mahalanobis distance
\[ d_M(x; \mu, \Sigma) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \]

Recall that the contour
\[ \{ x : d_M(x; \mu, \Sigma) = c \} \]

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

  $(\text{if } n \text{ is too small, we could})$

- Generative model is not valid
Example (of a structured covariance matrix)

\[ \Sigma = \sigma^2 I \Rightarrow \hat{\Sigma}_2 = \]

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

\[ \frac{1}{\sigma^2} \| x - \hat{\mu}_1 \|^2 \geq \frac{1}{\sigma^2} \| x - \hat{\mu}_2 \|^2 \]

\[ \Leftrightarrow \| x - \hat{\mu}_1 \|^2 \geq \frac{1}{1 - \hat{\pi}_1} \| x - \hat{\mu}_2 \|^2 \]

which is called the \underline{classifier}. 

![Graph showing LDA decision boundary with two classes and means \( \mu_1 \) and \( \mu_2 \).]
QDA

Generative model with

\[ X | y = k \sim N(\mu_k, \Sigma_k), \quad k = 1, \ldots, K \]

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

Decision boundaries are now \textit{quadratic}

\[ \boxed{\text{Key}} \]

A. ellipse

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

\[ \text{\textcopyright}{\text{covariance}} \quad \text{\textcopyright}{\text{class prior probs.}} \]

\[ \text{could} \]

- first apply PCA
- assume a structured covariance matrix

C. \[ \hat{\Sigma} = \frac{1}{d} \text{tr}\{\hat{\Sigma}\} \]

nearest centroid classifier