Lecture 9 Linear Discriminant Analysis I

STAT 441/505: Applied Statistical Methods in Data Mining

Linglong Kong

Department of Mathematical and Statistical Sciences
University of Alberta

Winter, 2016

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Outline

Introduction

Linear Discriminant Analysis for $p = 1$

Linear Discriminant Analysis for $p > 1$

Summary and Remark

Introduction

- Linear Discriminant Analysis (LDA) undertakes the same task as Logistic Regression. It classifies data based on categorical variables.
- Here the approach is to model the distribution of X in each of the classes separately, and then use Bayes theorem to flip things around and obtain $\Pr(Y|X)$.
- When we use normal (Gaussian) distributions for each class, this leads to linear or quadratic discriminant analysis.
- However, this approach is quite general, and other distributions can be used as well. We will focus on normal distributions.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Introduction

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- Linear discriminant analysis is popular when we have more than two response classes, because it also provides low-dimensional views of the data.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Bayes theorem

Thomas Bayes

Portrait used of Bayes in a 1936 book,[1] but it is doubtful whether the portrait is actually of him.[2] No earlier portrait or claimed portrait survives.

Born	c. 1701 London, England
Died	7 April 1761 (aged 59) Tunbridge Wells, Kent, England

Thomas Bayes was a famous mathematician whose name represents a big subfield of statistical and probabilistic modeling.

- Here we focus on a simple result, known as Bayes theorem:

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}.$$

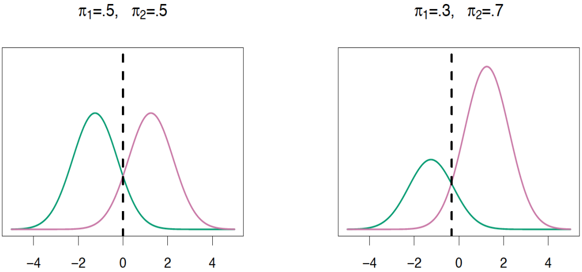
- In discriminant analysis, we have

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)},$$

- where $f_k(x) = \Pr(X = x|Y = k) \cdot \Pr(Y = k)$ is the density for X in class k — we use Gaussian in LDA.
- $\pi_k = \Pr(X = x)$ is the marginal or prior probability for class k.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Classification rule



- We classify a new point according to which density is highest.
- When the priors are different, we take them into account as well, and compare $\pi_k f_k(x)$.
- On the right, we favor the pink class — the decision boundary has shifted to the left.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Linear Discriminant Analysis for $p = 1$

- The Gaussian density has the form

$$f_k(x) = \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}},$$

  where μ_k is the mean and σ_k^2 is the variance in class k and we assume that $\sigma_k = \sigma$.
- Plugging this into Bayes formula, we get a rather complex expression for $p_k(x) = \Pr(Y = k|X = x)$

$$p_k(x) = \frac{\pi_k \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sigma_l\sqrt{2\pi}} e^{-\frac{(x-\mu_l)^2}{2\sigma_l^2}}}.$$

- This can be simplified to a linear equation of x.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Discriminant functions

- To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest.
- Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest discriminant score:

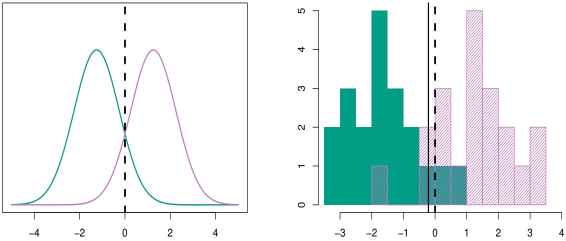
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k).$$

- Note that $\delta_k(x)$ is indeed a linear function of x.
- If there are $K = 2$ classes and $\pi_1 = \pi_2 = 0.5$, then one can see that the decision boundary is at

$$x = (\mu_1 + \mu_2)/2.$$

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Simulated Example



- 20 observations were drawn from each of the two classes with $\pi_1 = \pi_2 = 0.5$, $\mu_1 = -1.5$, $\mu_2 = 1.5$ and $\sigma = 1$.
- The dashed vertical line is the Bayes' decision boundary with error rate 10.6%
- The solid vertical line is the LDA decision boundary 11.1%

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Estimating the parameters

- Typically we don't know these parameters; we just have the training data. In that case we simply estimate the parameters and plug them into the rule.
- We use emprirical estimates,

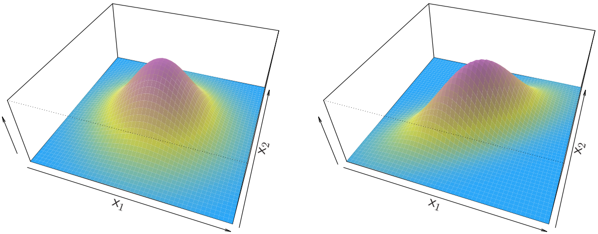
$$\pi_k = n_k/n, \ \mu_k = \sum_{i:y_i=k} x_i/n_k,$$

$$\hat{\sigma}^2 = \sum_{k=1}^{K} \sum_{i:y_i=k} \frac{(x_i - \mu_k)^2}{n - K} = \sum_{k=1}^{K} \frac{n_k - 1}{n - K} \cdot \hat{\sigma}_k^2,$$

- where $\hat{\sigma}_k^2 = \sum_{i:y_i=k} (x_i - \mu_k)^2 / (n_k - 1)$ is the usual formula for the estimated variance in the k-th class.

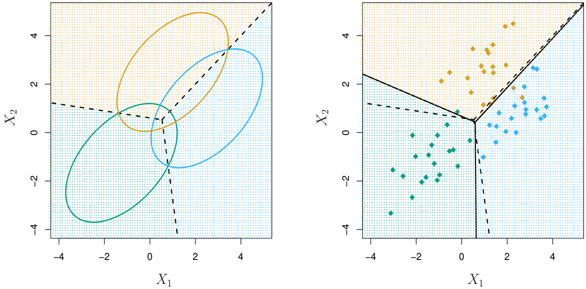
UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Linear Discriminant Analysis for $p > 1$



- Density function $f(x) = \frac{1}{(2\pi)^p|\Sigma|^{1/2}}e^{-1/2(s-\mu)^T\Sigma^{-1}(x-\mu)}$.
- Discriminant function:
  $\delta_k(x) = x^T\Sigma^{-1}\mu_k - 1/2\mu_k^T\Sigma^{-1}\mu_k + \log(\pi_k)$.
- Essentially, $\delta_k(x) = c_{k0} + c_{k1}x_1 + \cdots + c_{kp}x_p$ is a linear function.

Simulated example with $p = 2$ and $K = 3$



- 20 observations were generated from each class with $\pi_1 = \pi_2 = \pi_3 = 1/3$.
- The solid lines are LDA.
- The dashed lines are known as the Bayes decision boundaries.
- Were they known, they would yield the fewest misclassification errors, among all possible classifier.

Summary and Remark

- Linear Discriminant Analysis for $p = 1$
- Linear Discriminant Analysis for $p > 1$
- Read textbook Chapter 4 and `R` code
- Do `R` lab