MONASH University

ETC3250

Business Analytics

Week 4.
Linear discriminant analysis

17 August 2017

Outline

Week	Topic	Chapter	Lecturer
1	Introduction to business analytics & R	1	Souhaib
2	Statistical learning	2	Souhaib
3	Regression for prediction	3,7	Tas & David
4	Classification	4	Souhaib
	Logistic regression		Souhaib
	Linear discriminant analysis		Souhaib
5	Classification	4, 9	Souhaib
6	Resampling methods	5	Souhaib
7	Dimension reduction	6,10	Souhaib
8	Advanced regression	6	Souhaib
9	Advanced learning methods	8	Souhaib
	Semester break		
10	Clustering	10	Souhaib
11	Visualization		Souhaib
12	Data wrangling		Souhaib

Linear discriminant analysis

- ***Logistic regression*** involves directly modeling $\Pr(Y = k|X = x)$ using the logistic function. ***Linear discriminant analysis*** is a less direct approach: we first model the distribution of the predictors X separately in each of the response classes, and then use *Bayes' theorem* to compute $\Pr(Y = k|X = x)$.
- When the classes are well-separated, the parameter estimates for the ***logistic regression model*** are unstable. ***Linear discriminant analysis*** does not suffer from this problem.
- If the number of observation is small and the distribution of the predictors X is approximately normal in each of the classes, the ***linear discriminant analysis*** is again more stable than the ***logistic regression***.

Linear discriminant analysis (LDA)

- Let $\pi_k = \Pr(Y = k)$ represent the overall or prior probability that a randomly chosen observation comes from the kth class;
- Let $f_k(x) = f(x|Y = k)$ denote the density function of X for an observation that comes from the kth class
- Using Bayes' therorem:

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

$\rightarrow$ Instead of directly computing $p_k(x)$, we can *plug in* estimation of π_k and $f_k(X)$ into the expression above.

Estimation

- We can estimate π_k by computing the *fraction of the training observations* that belong to the kth class
- It is more challenging to estimate $f_k(X)$, unless we assume some *simple forms* for these densities.
- We refer to $p_k(x)$ as the *posterior probability* that an observation $X = x$ belongs to the kth class
- Recall that the Bayes classifier will classify an observation to the class for which $p_k(X)$ is *largest*. The Bayes classifier has the lowest possible error rate out of all classifiers.

LDA: univariate case

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

- We would like to obtain an estimate for $f_k(x)$ that we can plug into the expression above in order to estimate $p_k(x)$
- We can assume $f_k(x)$ is *Normal* or *Gaussian*:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

where μ_k and σ_k^2 are the mean and variance parameters for the kth class.

LDA: univariate case

For now, let us further assume that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$; then the conditional probabilities are

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} \exp\ \left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma}} \exp\ \left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

$$\begin{aligned}
&\text{maximize}_k\ p_k(x) \\
&\equiv \text{maximize}_k\ log(p_k(x)) \\
&\equiv \text{maximize}_k\ log\left(\pi_k \frac{1}{\sqrt{2\pi\sigma}} \exp\ \left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)\right) \\
&\equiv \text{maximize}_k\ \delta_k(x) := x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)
\end{aligned}$$

The discriminant functions $\delta_k(x)$ are linear functions of x.

LDA: univariate case

If $K = 2$ and $\pi_1 = \pi_2$, then we solve:

$$\text{maximize}_{k=1,2}\ \delta_k(x) := x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$

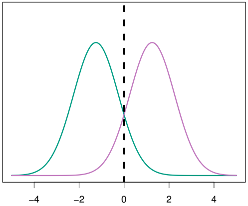
The Bayes classifier assigns the observation x to class 1 if

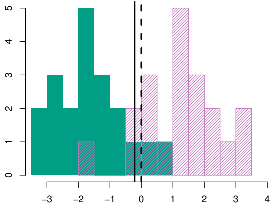
$$x \cdot \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} > x \cdot \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} \tag{1}$$

$$\implies 2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2 \tag{2}$$

and the class 2 otherwise. The Bayes decision boundary corresponds to the point where $x = \frac{\mu_1+\mu_2}{2}$.

LDA: univariate case

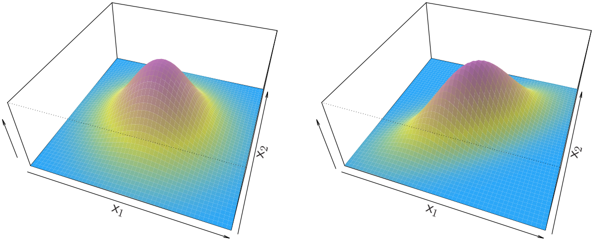




- $\mu_1 = -1.25$, $\mu_2 = 1.25$ and $\sigma_1^2 = \sigma_2^2 = 1$
- 20 observations were drawn from each of the two classes, and are shown as histograms
- The Bayes decision boundary (ashed vertical line). The LDA decision boundary (solid vertical line) estimated from data ($\pi_1 = \pi_2 = 0.5$). Class 1 if $x < 1$ and class 2 otherwise.
- In practice, we need $\hat{\mu}_k$, $\hat{\sigma}^2$, and $\hat{\pi}_k$

LDA: multivariate case

- We assume that $X = (X_1, \ldots, X_p)$ is drawn from a multivariate Gaussian (or multivariate normal) distribution, with a class-specific mean vector and a common covariance matrix.
- The multivariate Gaussian distribution assumes that each individual predictor follows a one-dimensional normal distribution with some correlation between each pair of predictors.



LDA: multivariate case

- To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution with $E[X] = \mu$ and $\text{Cov}(X) = \Sigma$, we write $X \sim N(\mu, \Sigma)$
- The multivariate Gaussian density is defined as

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)\right)$$

- In the multivariate case, the LDA classifier assumes that the observations in the kth class are drawn from a multivariate Gaussian distribution $N(\mu_k, \Sigma)$ where μ_k is a class-specific mean vector, and Σ is a covariance matrix that is common to all K classes

LDA: multivariate case

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

- Maximizing the conditional probabilities under the multivariate Gaussian density gives the following discriminant functions:

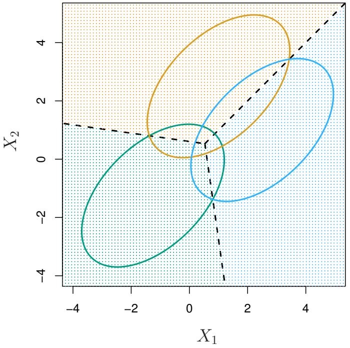
$$\delta_k(x) = x^T \boldsymbol{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \boldsymbol{\Sigma}^{-1} \mu_k + \log \pi_k$$

- The Bayes decision boundaries represent the set of values x for which $\delta_k(x) = \delta_l(x)$; i.e.

$$x^T \boldsymbol{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \boldsymbol{\Sigma}^{-1} \mu_k = x^T \boldsymbol{\Sigma}^{-1} \mu_l - \frac{1}{2} \mu_l^T \boldsymbol{\Sigma}^{-1} \mu_l$$

for $k \neq l$.

LDA: multivariate case



The dashed lines are the Bayes decision boundaries. Ellipses that contain 95% of the probability for each of the three classes are shown. 20 observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines.

Quadratic discriminant analysis

- Unlike LDA, quadratic discriminant analysis (QDA) assumes that each class has *its own covariance matrix*. That is, it assumes that an observation from the kth class is of the form $X \sim N(\mu_k, \Sigma_k)$ where Σ_k is a covariance matrix for the kth class

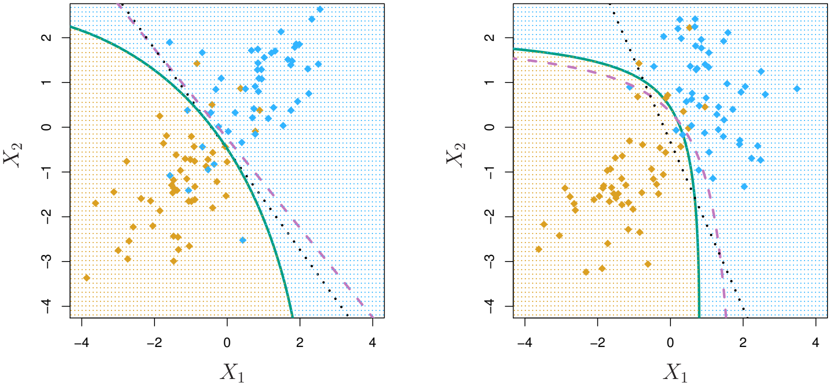
- Under this assumption, the Bayes classifier assigns an observation $X = x$ to the class for which

$$\begin{aligned} \delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T \boldsymbol{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log \pi_k \\ &= -\frac{1}{2}x^T \boldsymbol{\Sigma}_k^{-1} x + x^T \boldsymbol{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^T \boldsymbol{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log \pi_k \end{aligned}$$

  is largest.

$\rightarrow$ the quantity x appears as a **quadratic** function in $\delta_k(x)$.

QDA vs LDA



The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries.

Why would one prefer LDA to QDA, or vice-versa? The answer lies in the bias-variance trade-off.