STAT 480 Statistical Computing Applications

Unit 4. Classification

Lecture 2. Linear Discriminant Analysis

Department of Statistics, Iowa State University

Spring 2019

Motivation

- 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.

Using Bayes' Theorem for Classification

- Let π_k be the overall or prior probability that a randomly chosen observation comes from the kth class.
- Let $f_k(x) \equiv Pr(X = x|Y = k)$ denote the density function of X for an observation that comes from the kth class.
- Then Bayes' theorem states that

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

Using Bayes' Theorem for Classification (cont')

- Supposing $f_k(x)$ is normal

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

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

- Now assume that σ_k^2 for $k = 1, \ldots, K$.
- Then we have

$$p_k(x) = \frac{\pi_k \exp\left\{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right\}}{\sum_{l=1}^{K} \pi_l \exp\left\{-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right\}}. \tag{1}$$

Bayes Classifier

- The Bayes classifier involves assigning an observation $X = x$ to the class for which (1) is the largest.
- This is equivalent to check which

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

  is the largest.
- The decision boundary between class k and l is:

$$\{x : \delta_k(x) = \delta_l(x)\}.$$

  or equivalently the following holds

$$\log\frac{\pi_k}{\pi_l} - \frac{1}{2\sigma^2}(\mu_k^2 - \mu_l^2) + x\frac{1}{\sigma^2}(\mu_k - \mu_l) = 0.$$

Linear Discriminant Analysis

- The Linear Discriminant Analysis (LDA) approximates the Bayes classifier by plugging estimates for π_k, μ_k, σ into (2).

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

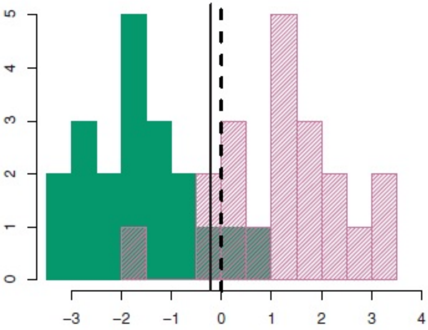
- The Linear Discriminant Analysis (LDA) classifier assigns an observation $X = x$ to the class for which

$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

  is the largest.

- The word linear in the classifier's name stems from the fact that the discriminant functions $\hat{\delta}_k(x)$ are linear functions of x.

Example



- 20 obs drawn from each of the two classes are shown as histograms.
- The Bayes decision boundary is the dashed vertical line. The solid vertical line is the LDA decision boundary estimated from the training data.

When $p > 1$: Bayes Classifier

- Bayes classifier assigns an obs $X = x$ to the class for which

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

  is the largest.

- The decision boundary between class k and l is:

$$\{x : \delta_k(x) = \delta_l(x)\}.$$

  or equivalently the following holds

$$\log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k - \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l) = 0.$$

When $p > 1$: LDA Classifier

- Let $\hat{\pi}_k = \frac{n_k}{n}$, and

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

- The LDA classifier assigns an obs $X = x$ to the class for which

$$\hat{\delta}_k(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log(\hat{\pi}_k)$$

  is the largest.

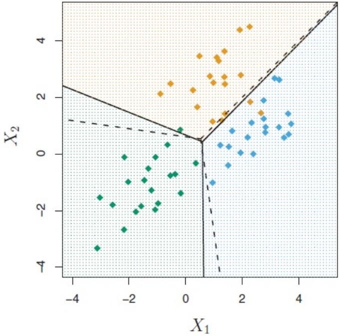
- The decision boundary between class k and l is:

$$\{x : \hat{\delta}_k(x) = \hat{\delta}_l(x)\}.$$

  or equivalently the following holds

$$\log \frac{\hat{\pi}_k}{\hat{\pi}_l} - \frac{1}{2}(\hat{\mu}_k - \hat{\mu}_l)^T \hat{\Sigma}^{-1}(\hat{\mu}_k - \hat{\mu}_l) + x^T \hat{\Sigma}^{-1}(\hat{\mu}_k - \hat{\mu}_l) = 0.$$

An Example with Three Classes



- Obs from each class are drawn from a multivariate normal with $p = 2$, with a class-specific mean vector and a common covariance matrix.
- The Bayes decision boundary is the dashed vertical line. The solid vertical line is the LDA decision boundary estimated from the training data.

Default Example

Confusion Matrix for the Default data set

		True default status		
		No	Yes	Total
Predicted	No	9644	252	9896
default status	Yes	23	81	104
	Total	9667	333	10000

- A confusion matrix compares the LDA predictions to the true `default` statuses for the 10,000 training observations.
- Diagonal elements: individuals whose `default` statuses were correctly predicted
- Off-diagonal elements: individuals that were misclassified.