ISEP
Machine Learning
September 16th 2024

Classification : Discriminant Analysis step by step

Patricia CONDE-CESPEDES

Objective

— Acquire the theoretical knowledge of Discriminant Analysis.

1 Exercises on Discriminant Analysis

1.1 Linear Discriminant Analysis with only one predictor

You have 20 observations of two variables X and Y. Y is a qualitative variable with categories *G,R* (*green* and *red*). X is a quantitative predictor. We assume X is a random variable that follows a normal law.

Given an observation x, we want to predict the class Y. In *Discriminant Analysis* we use the Bayes' theorem to estimate the probability of class k given an observation x :

$$P(Y = k|X = x) = p_k(X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{\kappa} \pi_\ell f_\ell(x)} \tag{1}$$

where :

— $f_k(x) = P(X = x|Y = k)$ is the density of X for an observation that comes from the kth class.

— $\pi_k = P(Y = k)$ represent the overall or prior probability that a randomly chosen observation comes from the kth class ;

The density of X in the class k is given by :

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2} \qquad k \in \{G, R\} \tag{2}$$

The parameters can be estimated with the following formula :

— $\hat{\pi}_k = \dfrac{n_k}{n}$, estimator of π_k.

— $\hat{\mu}_k = \dfrac{1}{n_k} \displaystyle\sum_{i:y_i=k} x_i$, estimator of μ_k.

— $\hat{\sigma}^2 = \dfrac{1}{n-\kappa} \displaystyle\sum_{k=1}^{\kappa} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$, estimator of $\sigma_k \forall k$.

Remind that LDA is based on the hypothesis that the variances are the same among all the classes, that is $\sigma_1 = \ldots \sigma_\kappa = \sigma$.

1. Given the following observations :
   $X = (-1.6, -0.8, -1.8, 0.6, -0.7, -1.8, -0.5, -0.3, -0.4, -1.3, 2.5, 1.4, 0.4, -1.2, 2.1, 1.0, 1.0, 1.9, 1.8, 1.6)$
   $Y = (G, G, G, G, G, G, G, G, G, G, R, R, R, R, R, R, R, R, R, R)$
   Estimate the paremeters of the densities of X in each class and the probabilities π_k of class k.

2. Plug in the estimated parameters into equation (1) and calculate the estimated probabilities $\hat{p}_k(x)$ for each value $x \in X$. You can start with only one observation x_i
3. Compare these probabilities to the values obtained using the function `lda()` in `R`. You can type `lda.pred$posterior[,1]` to obtain the posterior probabilities :

```
>lda.fit=lda(Y~X)
>lda.pred = predict(lda.fit)
>lda.pred$posterior[,1]
```

1.2 Measuring the accuracy with a *ROC* curve

In this section you are going to evaluate the accuracy of the classifier fitted in the previous section. Consider the posterior probabilities obtained in the previous exercise. For different values of threshold : from 0.1 to 1 by 0.1 follow the steps : *Hint : Use a for loop.*

— Calculate the predicted values of $\hat{y}$

— Calculate the confusion matrix between the estimated and predicted values. What is the value of the sensitivity, specificity ?

— Plot the obtained results in a *ROC* curve.

— Compare the results with those obtained with the function `roc()` of the package `pROC`

```
>install.packages("pROC")
>library(pROC)
>ROC.lda=roc(Y,lda.pred$posterior[,1],levels=c("R","G"))
>plot.roc(ROC.lda,print.auc =T,xlab="Specificity",col="red",axes=T)
```