Advanced Computational Econometrics
Chapter 2: Linear classification

1 Classification of the `default_credit_card` data

1. Read the dataset `default_credit_card.csv`. Split the data into a training set of size 20,000 and a test set of size 10,000.
2. Build LDA, QDA, naive Bayes and logistic regression classifiers for these data. Print the confusion matrices and the test error rates.
3. Using function `roc` in package pROC, plot the ROC curve of the four classifiers built in the previous question.

2 Estimation of the Bayes error rate

We consider a classification problem with $K = 3$ classes and $p = 2$ input variables. The marginal distribution of Y is defined by the following prior probabilities :

$$\pi_1 = 0.3, \quad \pi_2 = 0.3, \quad \pi_3 = 0.4,$$

and the conditional densities of $\mathbf{X}$ given $Y = k$, $k = 1, 2, 3$ are multivariate normal distributions $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with

$$\boldsymbol{\mu}_1 = (0,0)^T, \quad \boldsymbol{\mu}_2 = (0,2)^T, \boldsymbol{\mu}_3 = (2,0)^T,$$

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$$

1. Estimate de Bayes error rate for this problem (use function `dmvnorm` of package `mvtnorm` to compute the density of the multivariate normal distribution.
2. Generate training datasets of different sizes, and compare the error probability of the LDA classifier trained with this data to the Bayes error rate.