Advanced Computational Econometrics
Chapter 2: Linear and quadratic classification

1 Classification of the `default_credit_card` data

The file `default_credit_card.csv` contains data about customers' default payments in Taiwan.

Attribute Information :

— Y : default payment (Yes = 1, No = 0).

— X1 : Amount of the given credit (NT dollar) : it includes both the individual consumer credit and his/her family (supplementary) credit.

— X2 : Gender (1 = male ; 2 = female).

— X3 : Education (1 = graduate school ; 2 = university ; 3 = high school ; 4 = others).

— X4 : Marital status (1 = married ; 2 = single ; 3 = others).

— X5 : Age (year).

— X6 - X11 : History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows : X6 = the repayment status in September, 2005 ; X7 = the repayment status in August, 2005 ; . . . ;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is : -1 = pay duly ; 1 = payment delay for one month ; 2 = payment delay for two months ; . . . ; 8 = payment delay for eight months ; 9 = payment delay for nine months and above.

— X12-X17 : Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005 ; X13 = amount of bill statement in August, 2005 ; . . . ; X17 = amount of bill statement in April, 2005.

— X18-X23 : Amount of previous payment (NT dollar). X18 = amount paid in September, 2005 ; X19 = amount paid in August, 2005 ; . . . ;X23 = amount paid in April, 2005.

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. Test the statistical significance of the differences in error rates using McNemar tests.

4. Using function `roc` in package pROC, plot the ROC curves of the four classifiers built in the previous question.

2 Estimation of the Bayes error rate

We consider a classification problem with $c = 3$ classes and $p = 2$ predictors. 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, and function `rmvnorm` for randomly drawing vectors from the multivariate normal distribution).
2. Generate training datasets of different sizes, and compare the error probabilities of the LDA and QDA classifiers trained with this data to the Bayes error rate.

3 Ordinal regression

We consider again the `Boston` data set from the `MASS` package.

1. Define an ordinal variable `medv_ord` with three levels : `low`, `medium` and `high` by discretizing variable `medv` (median house value) using the 33% and 66% quantiles as cut-points.
2. Partition the `Boston` data set into a training set and a test set. Fit ordered logit and probit regression models on these data, with `medv_ord` as the response. Compute the test error rates and compare the results obtained by the two models. (Use a McNemar test).
3. Fit a multinomial logistic regression model of these data and compare the results with those obtained in the previous question.