MTH 443: Lab Problem Set 8

[1] Consider the dataset in "heart.csv". Rows of the data represent medical and lifestyle records of individuals on the following variables: (1) systolic BP, (2) tobacco intake level, (3) LDL cholesterol, (4) adiposity, (5) obesity, (6) alcohol intake level and (7) age. The classification variable is "coronary heart disease", which takes values 0 or 1; 0 for no coronary heart disease and 1 for presence of coronary heart disease in the medical history. Split the available data into 2 parts, keeping the last 10% records as out of sample test data and apply the following classifiers to build classification models:
(a) Fisher linear discriminant function, **(b)** Quadratic discriminant function.
Calculate the misclassification rates of the respective classifiers, separately for the learning data and the test set data.

[2] Consider the dataset in "currency_crisis.csv". The dataset gives monthly values of the following economic/financial leading indicators of currency crisis for India: (1) foreign exchange reserves, (2) foreign assets minus foreign liabilities, (3) imports, (4) exports, (5) sustainable foreign exchange reserves measured by the difference of monthly import and export as a percentage of previous year's average monthly foreign exchange reserves, (6) money multiplier of broad money, (7) ratio of broad money to gross international reserves, (8) real interest rate on deposits (deflated using consumer prices), (9) real interest rate differential (deflated using consumer prices), (10) ratio of lending to deposit interest rate, (11) stock of commercial bank aggregate deposits, (12) stock of commercial bank time deposits, (13) stock of commercial bank demand deposits, (14) index of industrial production, (15) deviation of REER from trend, (16) Foreign direct investments and (17) Crude oil price.
Except the variables on interest rates, the deviation of REER from trend and difference between foreign assets and foreign liabilities, for all the other variables, the value of the indicator on a given month is the percentage change in the level of the variable with respect to its value a year earlier.
The last column of the dataset is the classification variable, which is 0 or 1; the value is 1 if there is a currency crisis point in next 12 months and 0 otherwise.
Build classification models based on **(a)** Fisher linear discriminant function and **(b)** Quadratic discriminant function, under two different scenarios (i) last 10% data in test set; (ii) every 10^{th} observation in test set. Obtain the misclassification rates for training set and test set under both the scenarios and observe the difference in test set performance under (i) & (ii).

[3] Consider the wine data in "wine.csv". Each row of the data represents a sample of wine taken from one of two wineries (A and B) and gives values on the following variables: (1) Alcohol content, (2) Malic acid, (3) Ash, (4) Alkalinity of ash, (5) Magnesium, (6) Total phenols, (7) Flavanoids, (8) Nonflavanoid phenols, (9) Proanthocyanins, (10) Color intensity, (11) Hue, (12) OD280/OD315 of diluted wines and (13) Proline. The variable 'type' is the classfication variable and the remaining variables are the components of the feature vector. Split the available data into 2 parts, keeping the last 10% records as out of sample test data and apply the following classifiers to build classification models:

(a) Fisher Linear Discriminant Function, **(b)** Quadratic Discriminant function.

Calculate the misclassification rates of the respective classifiers, separately for the learning data and the test set data.