University of Bristol
EFIMM0095 (Data Science for Economics)
Professor Vincent Han

Coursework 2

Due: 18 March at 11:00am

- This coursework is a team project. The output will be evaluated as a team, that is, all team members will receive the same score. Therefore, cooperation among team members is important. If there is a free-rider problem, try to figure out how to resolve it. Please show work ethic to your team members.
- All the questions should be answered using Python. Report all the codes used in answering the questions.
- Please submit one pdf file per group with the names of all the members. Please indicate the group number in the file name.
- All questions are equally weighted. This assignment is one of three assignments that consist of 50% of your grade.

1. Consider the data set "Heart.csv," which consists of a sample of patients. It contains information about whether he or she has heart disease (a variable called AHD). It also contains information about the possible predictors of heart disease, such as age, sex, cholesterol level, and other heart and lung function measurements. We want to explore this sample and conduct a classification analysis using a logistic regression.

   (a) Summarize the data set. If there are any missing observations, remove them. For classification (and also many other machine learning methods), it is helpful to first normalize the explanatory variables. Except for the categorical variables, standardize all the other variables (i.e., $RestBP, Chol, MaxHR$) so that their means are zero and standard deviations are one. For all the questions in 1 and 2 below, we will use this preprocessed data set.

      * You want to quantify all the categorical variables before proceeding. Think about the most appropriate way of doing it. For Sex, the dataset doesn't contain information about whether 1 is female or male. For this missing information, assume that a male is more likely to have heart diseases than a female (which is a well-known fact), and conclude how to interpret the value of Sex.

   (b) Run a logistic regression. Report the coefficient estimates and p-values and interpret the results (without necessarily knowing the meaning of each predictor).

   (c) Set up a classification rule based on (b) and justify why you use this rule.

(d) Using the classifier in (c), classify whether the following patient has a heart disease or not: This patient is a 55-year-old woman who has a typical *ChestPain, Thal* is normal, and $RestBP = 130, Chol = 246, Fbs = 0, RestECG = 2, MaxHR = 150, ExAng = 1, Oldpeak = 1, Slope = 2, Ca = 0$.

(e) Based purely on (b)–(d), can you assess the accuracy of the classification in (d)? If yes, assess the result. If no, discuss why you can't.

(f) Randomly split the sample (by **randomly** selecting rows of the data frame) into a training set (of size 207) and a test set (with remaining observations).

(g) Given (f), rerun the logistic regression using the training set and calculate the classification error rate using the test set.

(h) Consider the following random classification: With probability 1/2, randomly classify a patient into having a heart disease. Using the test set obtained in (f), derive the classification error rate of this classifier.

(i) Compare the error rates in (g) and (h) and discuss.

2. For the same data set as in 1, we want to use the KNN, LDA and QDA as alternative classification methods.

(a) Use the training and test sets in 1-(f). Calculate the test error rates for the KNN classification with $K = 5$ vs. $K = 10$.

(b) Using the training and test sets in 1-(f). Calculate the test error rates for the LDA and QDA.

(c) Based on your previous answers, how do you compare the logistic regression, KNN-5, KNN-10, LDA and QDA? Discuss.

(d) This time, use the 10-fold cross validation to calculate the test error rates for those methods. Compare the error rates across the methods, and discuss how the comparison is different from (c).

(e) Classify the patient in 1-(d) using the method that has the smallest error rate in (d).

(f) For this question and all below, consider the classification method chosen in (e). Construct a confusion matrix (see Table 4.4 in ISLR) and interpret the results.

(g) Calculate the true positive rate and false positive rate for each of the following choice of a threshold level in the classifier: $\{0, 0.25, 0.5, 0.75, 1\}$. Discuss the results.

(h) Confirm your answers in (g) by drawing the ROC curve (using a built-in option).