ACTSC 632 – Project for Module 4 – due on July 10

This project will consider the German Credit data set available in the `CASdatasets` package in which it is named `credit`. As a team of data analysts for a credit card company, you have been asked to compare different classification methods and provide a recommendation to your manager for which one is the best to identify the "good" (i.e., creditworthy, no non-payments) vs the "bad" (not credit-worthy, having existing non-payments) credit files.

This project is to be complete in teams. Information about teams' membership will be emailed on LEARN.

1. Determine the number of predictors in this data set, and whether they are quantitative of qualitative. How many observations are there? How many observations are classified as a "bad credit"? "good credit"?

2. To explore the data further, produce the following plots: (i) for each of the predictors age and duration, make a box plot showing the distribution of the observations, separately for the "good" and "bad" observations.; (ii) for the pairs duration & savings and duration & credit history, plot the observations (using duration on the x axis) and use different symbols for the "good" and "bad" observations. Comments on the plots you obtained.

3. Initially, you should work with the following predictors: age, duration, purpose, credit_history, and savings. For each of (i) logistic regression (ii) linear discriminant analysis (iii) quadratic discriminant analysis, do the following:

   (a) Determine the confusion matrix, the overall error rate, Type-I error, Type-II error.

   (b) Plot the ROC curve for the three methods considered and determine the AUC (area under the curve).

4. The manager who asked you to recommend a classifier for this problem says you'd be allowed to use up to two more predictors to build your classifier. Choose one or two additional predictors and determine how their inclusion affects the performance of the three classifiers (by repeating steps (a) and (b) in the previous question). Explain your choice of predictors.

5. Suppose the credit card issuer assesses that it looses 4 times more money for each "bad credit" that is incorrectly identified (false negative) as it does for a "good credit" that is incorrectly identified (false positive). Based on this, and using your results based on logistic regression with your choice of predictors from the previous question, propose a decision threshold to identify a "bad credit" (via the estimated probability $\hat{P}(Y = \text{bad}|X)$) that minimizes the financial loss due to wrong classification. Then for all three methods, compute the overall (training) error rate, the type-I error and the type-II error.

6. As an alternative to the previous question, suppose now that you wish to take a more conservative view point that does not rely on the above cost function, and instead what to choose the threshold that maximize Youden's J statistic (or equivalently, that maximizes the sum of the specificity and sensitivity). As in the previous question, compute the overall (training) error rate, the type-I error and the type-II error for all three methods using this new threshold. Comment on the results obtained using the three different thresholds used so far.

7. Using again what you believe is the best set of predictors from part 4, use 10-fold cross-validation to estimate the overall error rate and type-II error. Is it consistent with the training error rate? Make sure to make your comparisons using the same threshold.

8. Next you wish to use the K nearest neighbours (KNN) as a classifier for this problem, using the three predictors age, duration and credit_amount. Use a random sample of 750 observations as your training set and for each of $K = 1, 3, 5$, apply the KNN approach. Produce the confusion table, determine the overall error rate, Type-I error and Type-II error. Which choice of K seems the best?

9. Now use 10-fold cross validation to estimate the overall error rate and Type-II error for each of $K = 1, 3, 5$. Do you reach the same conclusion as in the previous question about the choice of K?

10. What would be your final recommendation for the best classifier to use for this problem?

Your report should be professionally organized, with a brief introduction.

You should not include R code or output in the main part of the report. However you should upload to the LEARN dropbox all code used.