INFO411/911: Data Mining and Knowledge Discovery
Assignment 2 (15%)

Autumn 2021
Due 11:55 pm, Friday, 28 May 2021, via Moodle

- Submit a single PDF document which contains your answers to the questions. All questions are to be answered.
- The PDF must contain typed text of your answer (do not submit a scan of a handwritten document, any handwritten document will be ignored). The document can include computer generated graphic (hand drawn graphic and illustrations will be ignored).
- Make sure you do the questions in order and use the following format: R-code. R-output, and interpretations/calculations.
- The PDF document of your answers should be no more than 10 pages including all graphs. If it is over 10 pages, only the first 10 pages will be marked. The size limit for this PDF document is 20MB.
- Late submission will not be accepted without academic consideration being granted.

Questions

1. (3 marks) In this assignment we make use of the data `creditworthiness.csv` which was used in Task 2 of Assignment 1. As before, we wish to predict the credit rating that would be assigned to each individual. Recall that data on 2500 customers have been collected, and credit rating for 1962 of them has been assessed as either A, B, or C, coded as 1, 2, or 3, respectively, with the remaining 538 needing to be classified. Write the code to split the dataset into 50% training set and 50% test set and only include the data with known ratings.

2. Using default settings, fit a decision tree to the training set predict the credit ratings of customers using all of the other variables in the dataset.

   (a) *(2 marks)* Report the resulting tree.

   (b) *(2 marks)* Based on this output, predict the credit rating of a hypothetical "median" customer, i.e., one with the attributes listed in Table 1, showing the steps involved.

   (c) *(2 marks)* Produce the confusion matrix for predicting the credit rating from this tree on the test set, and also report the overall accuracy rate.

   (d) *(5 marks)* What is the numerical value of the gain in entropy corresponding to the first split at the top of the tree? (Use logarithms to base 2, and show the details of the calculation rather than just providing a final answer.)

   (e) *(2 marks)* Fit a random forest model to the training set to try to improve prediction. Report the R output.

   (f) *(2 marks)* Produce the confusion matrix for predicting the credit rating from this forest on the test set, and also report the overall accuracy rate.

3. Using default settings for `svm()` from the `e1071` package, fit a support vector machine to predict the credit ratings of customers using all of the other variables in the dataset.

   (a) *(2 marks)* Predict the credit rating of a hypothetical "median" customer, i.e., one with the attributes listed in Table 1. Report decision values as well.

   (b) *(2 marks)* Produce the confusion matrix for predicting the credit rating from this SVM on the test set, and also report the overall accuracy rate.

   (c) *(2 marks)* Automatically or manually tune the SVM to improve prediction over that found in 3b. Report the resulting SVM settings and the resulting confusion matrix for predicting the test set. (Any amount of improvement is acceptable.)

4. Fit the Naive Bayes model to predict the credit ratings of customers using all of the other variables in the dataset.

   (a) *(2 marks)* Predict the credit rating of a hypothetical "median" customer, i.e., one with the attributes listed in Table 1. Report predicted probabilities as well.

(b) *(2 marks)* Reproduce the first 20 or so lines of the R output for the Naive Bayes fit, and use them to explain the steps involved in making this prediction.

(c) *(2 marks)* Produce the confusion matrix for predicting the credit rating using Naive Bayes on the test set, and also report the overall accuracy rate.

5. Based on the confusion matrices reported in the preceding parts,

(a) *(2 marks)* Which of the classifiers look to be the best? (Be specific, and specify the figures you used to answer this question.)

(b) *(2 marks)* Which look to be the worst? (Be specific, and specify the figures you used to answer this question.)

(c) *(2 marks)* Are there any categories that all classifiers seem to have trouble with?

6. Consider a simpler problem of predicting whether a customer gets a credit rating of A or not.

(a) *(2 marks)* Fit a logistic regression model to predict whether a customer gets a credit rating of A using all of the other variables in the dataset, with no interactions.

(b) *(2 marks)* Report the summary table of the logistic regression model fit.

(c) *(2 marks)* Which predictors of credit rating appear to be significant at 5% significance level?

(d) *(2 marks)* Fit an SVM model of your choice to the training set.

(e) *(3 marks)* Produce an ROC chart comparing the logistic regression and the SVM results of predicting the test set. Comment on any differences in their performance.

Table 1: Attributes of the median person in the credit-worthiness dataset.

functionary	0
re-balanced (paid back) a recently overdrawn current acount	1
FI3O credit score	1
gender	0
0. accounts at other banks	3
credit refused in past?	0
years employed	3
savings on other accounts	3
self employed?	0
max. account balance 12 months ago	3
min. account balance 12 months ago	3
avrg. account balance 12 months ago	3
max. account balance 11 months ago	3
min. account balance 11 months ago	3
avrg. account balance 11 months ago	3
max. account balance 10 months ago	3
min. account balance 10 months ago	3
avrg. account balance 10 months ago	3
max. account balance 9 months ago	3
min. account balance 9 months ago	3
avrg. account balance 9 months ago	3
max. account balance 8 months ago	3
min. account balance 8 months ago	3
avrg. account balance 8 months ago	3
max. account balance 7 months ago	3
min. account balance 7 months ago	3
avrg. account balance 7 months ago	3
max. account balance 6 months ago	3
min. account balance 6 months ago	3
avrg. account balance 6 months ago	3
max. account balance 5 months ago	3
min. account balance 5 months ago	3
avrg. account balance 5 months ago	3
max. account balance 4 months ago	3
min. account balance 4 months ago	3
avrg. account balance 4 months ago	3
max. account balance 3 months ago	3
min. account balance 3 months ago	3
avrg. account balance 3 months ago	3
max. account balance 2 months ago	3
min. account balance 2 months ago	3
avrg. account balance 2 months ago	3
max. account balance 1 months ago	3
min. account balance 1 months ago	3
avrg. account balance 1 months ago	3