COMP47460 Assignment 1

Deadline: Submit no later than **Sunday 15th Nov, 2020**.

Instructions

Answer both questions. Submit your assignment as one PDF file (not a DOC/DOCX/ODT/ZIP file) via the module Brightspace page. Include your full name and student ID number on the PDF.

This assignment should be completed individually. Any evidence of plagiarism will be reported to the CS plagiarism committee and it can result in a Fail grade.

Please stick to the suggested page lengths.

Question 1

You will analyse a dataset collected from shoppers at Tesco stores in the UK. The dataset has a number of features describing aggregated purchasing behaviour in different geographic locations. These features relate to the quantity of items purchased and their nutritional information, in addition to other features. Researchers have used this dataset to try to predict the prevalence of diabetes, and you might find the paper to be an interesting read. Some information on the features in the dataset can be found on Figshare. The last column in the dataset is a categorical feature describing the diabetes prevalence {low, mid, high}.

Your objective is use the methods you have learned in the module to make predictions about the diabetes prevalence based on the shopping behaviour.

Download the file tesco_<student_number>.arff or tesco_<student_number>.csv from Brightspace. So, if your student number is 12345678, then download tesco_12345678.arff. Please ensure that you are using your personalised dataset- **submissions using an incorrect dataset will receive a 0 grade.**

Using your dataset, perform the tasks below. In each task, summarise the differences in performance, and describe some factors which might explain the results. You should normalise and/or clean the dataset, as appropriate. Describe the cleaning steps you took in your submission to sufficient degree (e.g., Your description may look like "I did min-max normalization of feature X using the minimum on the feature values in training examples and maximum of feature values over all labelled examples. I manually removed feature Y because ...").

This is an open-ended assignment -- You do not have to restrict yourself to what is asked. You can take the exploration and the discussion deeper than what is asked.

Total suggested page length for Q1 is 4-5 pages.

(a) This dataset has a large number of features. Carefully identify the most discriminating features to predict the diabetes prevalence rate category in a ward using the filter and wrapper feature selection techniques. Report the feature subsets that these techniques select. In the case of a filter, you must propose a way to choose a subset of the ranked

features, rather than using the entire original set of features. You should justify your choice. In the case of wrapper techniques, carefully select features for at least one Decision Tree, one Naïve Bayes and one k-NN classifier. Report and discuss the differences between the feature subsets produced by the filter and wrapper techniques. [15 marks]

(b) Carefully consider the evaluation measure(s) that you use for this exercise and justify why you selected the particular evaluation measure(s). [5 marks]

(c) Evaluate the performance of various classifiers (including at least one Decision Tree, one Naïve Bayes and one k-NN classifier) on your dataset using the feature subset(s) identified in (a) and evaluation measure(s) identified in (b). Explore the effect of different parameter settings on these classifiers. Describe the evaluation procedure that you used in good detail. [25 marks]

(d) Carefully discuss the results obtained in part (c). To what extent are these results in line with or different from what you learnt about these classifiers in your lectures? For example, is your accuracy higher or lower on the dataset with reduced number of features as compared to the original dataset? Is the relative performance of different classifiers and configuration settings in line with your expectation? [15 marks]

(e) Plot the ROC curves for the "high" class and the different classification models? What do you learn from this ROC curve? Which classifier/configuration is best suited for this task? Are you satisfied with the performance? [10 marks]

Question 2

Answers all parts below. **Please provide answers in your own words.**

Total page length for Q2 should not exceed 2 pages

(a) Consider the nightmare situation in which you struggled hard to obtain a very high accuracy (>95%) on your training data for a binary classification task, but when your client ran it on their test data, the accuracy was very low (<50%). This is despite the fact that your dataset is reasonably balanced (majority class < 65%) and you are using a fairly complex learning algorithms with many parameters to fit your dataset. How do you explain this situation? What are the possible causes for this? How can you improve the testing accuracy in this situation? What precautions should you take in your evaluation procedure to avoid this situation? [15 marks]

(b) What is a ROC curve? What is the motivation behind using it? How do you interpret it? How do you use it to compare two different classification approaches? Why is the reference line consider to correspond to a random classifier? [15 marks]