School of Computing and Information Technology
University of Wollongong
INFO411 Data Mining and Knowledege Discovery

Assignment 1 (12 Marks in total)

Submission of the answers must be done online by using the submission link that is associated with this subject for assignment 1 on MOODLE . One PDF document is to be submitted. The PDF must contain typed text of your answer (do not submit a scan of a handwritten document). The document can include computer generated graphics and illustrations (hand drawn graphics and illustrations will be ignored). The size limit for this PDF document is 20MB. All questions are to be answered. A clear and complete explanation and analysis needs to be provided with each answer.

Submissions made after the due time will be assessed as late submissions and are counted in full day increments (i.e. 1 minute late counts as a 1 day late submission). There is a 25% penalty for each day after the due date. The submission site closes four days after the due date. No submission will be accepted after the submission site has closed.

This is an individual assignment. Plagiarism of any part of the assignment will result in having zero marks for all students involved.

You may need to do some research on background information for this assignment. For example, you may need to develop a deeper understanding of writing code in R.

What you need:
The R software package, the file **assignment1.zip** from the Moodle site.

Task

Preface: Banks are often posed with a problem to whether or not a client is credit worthy. Banks commonly employ data mining techniques to classify a customer into risk categories such as category A (highest rating) or category C (lowest rating).
A bank collects data from **past** credit assessments. The file **creditworthiness.csv** contains 2500 of such assessments. Each assessment lists 46 attributes of a customer. The last attribute (the 47-th attribute) is the result of the assessment. Open the file and study its contents. You will notice that the columns are coded by numeric values. The meaning of these values is defined in the file **definitions.txt**. For example, a value 3 in the 47-th column means that the customer credit worthiness is rated "C". Any value of attributes not listed in **definitions.txt** is "as is".
This poses a "prediction" problem. A machine is to learn from the outcomes of past assessments and, once the machine has been trained, to assess any customer who has not yet been assessed. For example, the value 0 in column 47 indicates that this customer has not yet been assessed.

Purpose of this task:
You are to start with an analysis of the general properties of this dataset by using suitable visualization and clustering techniques (i.e. Such as those introduced during the lectures), and you are to obtain an insight into the degree of difficulty of this prediction task. Then you are to design and deploy an appropriate supervised prediction model (i.e. MLP) to obtain a prediction of customer ratings.

Question 1: **(5 marks)**
Analyse the general properties of the dataset and obtain an insight into the difficulty of the prediction task. Create a statistical analysis of the attributes and their values, then list 5 of the most interesting (most valuable) attributes. Explain the reasons that make these attributes interesting.
Note: A set of R-script files are provided with this assignment (included in the **assignment1.zip** file). The scripts provided will allow you to produce some first results. However, virtually none of the parameters used

in these scripts are suitable for obtaining a good insight into the general properties of the given dataset. Hence your task is to modify the scripts such that informative results can be obtained from which conclusions about the learning problem can be made. Note that finding a good set of parameters is often very time consuming in data mining. An additional challange is to make a correct interpretation of the results.

This is what you need to do: Find a good set of parameters (i.e. Through a trial and error approach), obtain informative results then offer an interpretation of the results. Write down your approach to conducting the experiments, explain your results, and offer a comprehensive interpretation of the results. Do not forget that you are also to provide an insight into the degree of difficulty of this learning problem (i.e. From the results that you obtained, can it be expected that a prediction model will be able to obtain 100% prediction accuracy?). Always explain your answers.

Question 2: (7 marks)

Deploy a prediction model to predict the credit worthiness of customers which have not yet been assessed. The prediction capabilities of the MLP in the lab of “Classification” was very poor. Your task is to:

a.) Describe a valid strategy that maximises the accuracy of predicting the credit rating. Explain why your strategy can be expected to maximise the prediction capabilities.

b.) Use your strategy to train MLP(s) then report your results. Give an interpretation of your results. What is the best classification accuracy (expressed in % of correctly classified data) that you can obtain for data that were not used during training (i.e. The test set)?

What you need:

The R software package (Rstudio is optional) and the file **assignment1.zip**. Successful completion of the lab of “Classification”. You may use the R-script of the lab of “Classification” as a basis for attempting this question.

Note that in this assignment the term "prediction capabilities" refer to a model's ability to predict the credit rating of samples that were not used to train the model (i.e. samples in a test set).

The answers to this assignment should be provided with a single PDF document which is to be submitted. Submit one single PDF document that contains your answers to this assignment. Submit before the due date and follow the submission procedure as described in the header of this assignment.