COMP47490 Assignment 1

Deadline

Submit before 5pm on Tuesday November 1st, 2017.

Instructions

Answer both questions. Submit your answers as a single PDF file (not a DOC/DOCX file) via the COMP47490 Moodle page. Include your full name and student ID number.

Question 1

The objective of this question is to use the feature selection functionality in Weka to identify useful features on a dataset which describes a set of bank customers, where the goal is to predict whether the creditworthiness of a customer is 'low' or 'high'.

You should download your personal dataset file from the URL:
http://mlg.ucd.ie/datasets/ml/bank/<STUDENT_NUMBER>.arff

For example, if your student number is 135023491, your dataset is at the URL:
http://mlg.ucd.ie/datasets/ml/bank/135023491.arff

When downloading your dataset, please ensure your student number is correct. Submissions using an incorrect dataset will receive a 0 grade.

Using your dataset, perform the tasks below. Each task carries equal marks.
(Total suggested page length for Q1 is 3-5 pages)

(a) Apply one filter and one wrapper feature selection technique from those available in Weka and report the feature subsets that they select. In the case of a filter, you should propose a way to choose a subset of the ranked features, rather than using the entire original set of features.

(b) Report and discuss the differences between the feature subsets produced by the filter and wrapper techniques from Task (a). Provide explanations for why the two techniques can potentially produce different results.

(c) Evaluate and discuss the performance of both of the above feature selection techniques, when each one is combined with two different classifiers of your choice available in Weka (i.e. there will be four experimental combinations). Which combination do you believe is most suitable for this dataset?

Question 2

Answers all parts below. Each parts carries equal marks.
(Total page length for Q2 should not exceed 2 pages)

(a) You are given the task of building a classification model. Explain in your own words how you would allocate data for training, testing, and validation.

(b) You realise that your data is actually a time-series. Do you expect that cross-validation will work well in this case? Explain your reasoning.

(c) Explain in your own words why the Naïve Bayes algorithm considered naïve?

(d) Having trained a new model which performs very well on your training data, you find it has surprisingly poor performance on new unseen data. Explain the reason for this problem in your own words, and suggest a few approaches you can take to address the problem.

Grading

- The assignment is marked out of 100:
  - Q1: 60 marks
  - Q2: 40 marks
- Assignments should be completed individually. Any evidence of plagiarism will result in a 0 grade.
- Penalties will apply for late submissions after 5pm on Tuesday November 1st:
  - 1-5 days late: 10% deduction from overall mark
  - 6-10 days late: 20% deduction from overall mark
  - Assignment will not be accepted after 10 days without extenuating circumstances form and/or medical certificate.