COMP47460 Assignment 1

Deadline

Submit no later than Friday November 2nd, 2018.

Instructions

Answer both questions. Submit your assignment as one PDF file (not a DOC/DOCX/1 ODT/ZIP file) via the COMP47460 Moodle page. Include your full name and student ID number on the PDF.

Grading

- Both questions carry equal marks:
  - Q1: 50 marks
  - Q2: 50 marks
- Assignments should be completed individually. Any evidence of plagiarism will result in a 0 grade.
- Penalties will apply for any late submissions after Friday November 2nd:
  - 1-5 days late: 10% deduction from overall mark
  - 6-10 days late: 20% deduction from overall mark
  - Assignments later than 10 days will not be accepted without proof of extenuating circumstances (i.e. a medical certificate).

Question 1

Use the feature selection functionality in Weka to identify informative features on a dataset related to a bank marketing campaign. Given a set of demographic details for a group of bank customers, the objective is to predict whether or not these customers will subscribe to a new service being offered by the bank ("yes" or "no"). See the Appendix of this document for a description of the features in the data.

http://claritytrec.ucd.ie/~alawlor/comp47460/datasets/ml/marketing/<StudentID>.arff

For example, if your student number is 145023491, your dataset is at the URL:

*http://claritytrec.ucd.ie/~alawlor/comp47460/datasets/ml/marketing/*145023491*.arff*

When downloading your dataset, please ensure that 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 3-4 pages)

(a) Apply <u>one filter</u> and <u>one wrapper</u> feature selection strategy from those available in Weka and report the feature subsets that they 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.

(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 <u>two different classifiers</u> 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. Please provide answers in your own words.

Each parts carries equal marks. Total page length for Q2 should not exceed 2 pages

(a) Explain what is meant by *overfitting* in the context of classification. Why is overfitting considered a problem? Briefly explain some the techniques you might use to avoid or mitigate overfitting.

(b) Explain the difference between a *test set* and a *validation set*?

(c) Describe the F-measure used in the context of evaluating classifier performance.

(d) Explain the difference between *feature selection* and *feature transformation* approaches for dimension reduction. Give one example of each.

(e) Explain the use of *entropy* and *information gain* in the decision tree model.

(f) Explain how you choose the best value of *k* when building a kNN classifier?

Appendix

Details of features present in the dataset related to the bank marketing campaign, for assignment Q1:

#	Name	Feature Description	Type
1	age	Customer’s age	numeric
2	job	Customer's job type	categorical
3	status	Customer’s marital status	categorical
4	education	Education level attained by customer	categorical
5	has_defaulted	Does the customer have credit in default?	binary
6	balance	Average yearly balance, in euros	numeric
7	housing_loan	Does the customer have a housing loan?	binary
8	personal_loan	Does the customer have a personal loan?	binary
9	contact	How was the customer contacted by the bank?	categorical
10	contact_day	Contact day of the month	numeric
11	contact_month	Contact month of the year	categorical
12	contact_duration	Contact duration, in seconds	numeric
13	num_contacts	Number of contacts during this campaign for this customer	numeric
14	num_days	Days since the customer was last contacted for the last campaign	numeric
15	num_last	Number of contacts in the last marketing campaign for this customer	numeric
16	last_outcome	outcome of the last marketing campaign	categorical