COMP47490 Assignment 2

Deadline

Submit before 5pm on Friday December 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 ensemble learning functionality in Weka to identify the extent to which classification performance can be improved through the combination of multiple models. Experiments will be run on data where the objective is to classify vehicles into one of four classes, based on their visual descriptions.

You should download your personal dataset from the URL:
http://mlg.ucd.ie/datasets/ml/vehicles/<STUDENT_NUMBER>.arff
For example, if your student number is *135023491*, your dataset is at the URL:
http://mlg.ucd.ie/datasets/ml/vehicles/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 4-5 pages)

(a) Evaluate the performance of two basic classifiers on the data: Decision Trees (J48) and 3-NN.

(b) Apply ensembles with *bagging* using both classifiers from Task (a). Investigate the performance of both classifiers as the ensemble size increases, in steps of 20 from 20 to 100 members. Using the best performing ensemble size, investigate how changing the number of instances in the bootstrap samples affects classification performance (i.e. the "bag size").

(c) Apply ensembles with *boosting* using both classifiers from Task (a). Investigate the performance of both classifiers as the ensemble size increases, in steps of 20 from 20 to 100 members.

(d) Apply ensembles with *random subspacing* using both classifiers from Task (a). Investigate the performance of both classifiers as the ensemble size increases, in steps of 20 from 20 to 100 members. Using the best performing ensemble size, investigate how changing the number of features used when applying random subspacing affects classification performance (i.e. the "subspace size").

(e) Summarise the differences in the performance results from Tasks (a)-(d), and describe some factors which might explain these differences.

Question 2

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

(a) What is the curse of dimensionality? Why does it cause problems for certain machine learning algorithms? Describe some of the techniques we have seen to deal with the problem of the curse of dimensionality.

(b) In your own words, explain what you understand by the bias-variance tradeoff. How would you tell if your model had a) high bias? b) high variance? Explain how the bagging algorithm affects both bias and variance.

(c) Describe in your own words how we use the coefficients and the odds ratio to interpret a logistic regression model.

(d) In the Principal Components Analysis (PCA) algorithm we often centre the data to the mean. What might happen if we do not do this? With some algorithms we normalise the data beforehand. Explain why you would (or would not) normalise your data before applying PCA.

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 Friday December 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.