COMP47490 Assignment 2

Deadline

Friday, December 11 at 5pm. If submitted later, late submission penalties will apply. No submissions allowed two weeks after deadline.

Instructions

Answer both questions. Submit your assignment as one PDF file (not a DOC/DOCX/ ODT/ZIP file) via the module Brightspace page.

Question 1

The objective of this question is to use the ensemble learning functionality to identify the extent to which classification performance can be improved through the combination of multiple models. Experiments will be run on a dataset extracted from US Census data. The data contains 14 attributes including age, race, sex, marital status etc, and the goal is to predict whether the individual earns over $50k per year.

We have prepared a dataset for each student. Please download the dataset corresponding to your student id from Brightspace. 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 are free to normalise and/or clean the dataset, as appropriate. Describe the cleaning steps you took in your submission to sufficient degree. Also, note that this is a more realistic dataset -- There may be missing values and many other issues that you have to deal with.

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.

Hint: You can speed up the nearest neighbour classifier by setting its option for "nearestNeighbourSearchAlgorithm=KDTree".

Total suggested page length for Q1 is 5 pages.

(a) Evaluate the performance of three basic classifiers on your dataset:
Decision Tree, Neural Network and 1-NN. Carefully consider the evaluation measure(s) that you use for this exercise and justify why you selected the particular evaluation measure(s).

(b) Apply ensembles with *bagging* using the three classifiers from Task (a). Investigate the performance of these classifiers as the ensemble size increases (e.g., in steps of 2 from 2 to 20 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 *random subspacing* using the three classifiers from Task (a). Investigate the performance of these classifiers as the ensemble size increases (e.g., in steps of 2 from 2 to 20 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").

(d) Based on the lectures, which set of classifiers is expected to benefit from bagging techniques more and which set of classifiers is expected to benefit from random subspacing techniques more? For your dataset, determine the best ensemble strategy for each of these classifiers. Discuss if this is in line with what you expected.

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) Comment on the interpretability of different supervised learning techniques. How easy or difficult it is to explain the reason behind predictions to a layman? Can you easily find out which training examples need to be modified to change the prediction for a particular query? Can you easily find out the weight of the different features in your model?

(b) Explain the bias-variance tradeoff. Which classifiers generally suffer from high bias and which classifiers generally suffer from high variance? Which ensemble strategy helps you to deal with bias and which ensemble strategy helps you to deal with variance issues?

(c) What are the main limitations of k-means in general and Lloyd's algorithm in particular? Explain how the centroids are initialised in the k-means++ algorithm? What is the intuition for initialising the centroids in this way?

(d) What does R^2 measure represent in the case of linear regression?

(e) Explain the precision-recall tradeoff.

(f) Explain how the parameters are learnt in the training of neural networks.

Grading

- Q1: 40 marks
- Q2: 60 marks
- Assignments should be completed individually. Any evidence of plagiarism will be reported to the CS plagiarism committee and it can result in a 0 grade.