COMP47490 Assignment 2

Deadline

Sunday, December 5. If submitted later, late submission penalties will apply. No submissions allowed two weeks after deadline.

Instructions

Answer both questions. Submit your assignment as a single Jupyter notebook (**not a PDF/DOC/DOCX/ODT/ZIP file**) via the module Brightspace page. If you have any images, make sure that they are properly embedded into the Jupyter notebook. You need to put answers to both the questions in the Jupyter notebook.

Please keep the whole code in a single Jupyter notebook. In your notebook, please split the code and explanations into many little cells so it is easy to see and read the results of each step of your solution. Please remember to name your variables and methods with self-explanatory names. Please remember to write comments and where needed, justifications, for the decisions you make and code you write.
Your code and analysis is like a story that awaits to be read, make it a nice story please. Always start with an introduction about the problem and your understanding of the problem domain and the data analytics solution. Then describe the steps you take and your findings from each step. Aim to keep the notebook clear and concise, with the key code and discussion.

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.

Download the file US_census_<student_number>.csv from Brightspace (My Learning -> Datasets -> Assignment 2). 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 -- Feel free to take the exploration and the discussion deeper than what is asked to get bonus points.

(a) Carefully clean and prepare the dataset for machine learning analysis. You can do basic feature engineering to make your techniques scalable, but there is no need to go overboard with the dataset cleaning. Carefully consider the evaluation measure(s) that you use for this exercise and justify why you selected the particular evaluation measure(s). [Important: As much as possible, use this evaluation measure for the subsequent parts] [10 marks]

(b) Evaluate the performance of three basic classifiers on your dataset: a decision tree with depth at most 3, a neural network with at most 10 hidden nodes and 1-NN. You can do basic parameter tuning, but there is no need to go overboard. The goal in this step is simply to create better than random classifiers. [5 marks]

(c) Apply ensembles with *bagging* using the three classifiers from Task (b). 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. [10 marks]

(d) Apply ensembles with *random subspacing* using the three classifiers from Task (b). 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. [10 marks]

(e) 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. Discuss if there is enough diversity in your ensemble and what else could you have done to improve the performance of your ensemble. [10 marks]

(f) Compare the ensemble classifiers with the tuned individual classifiers, e.g., decision trees with more depth, k-NN with a carefully chosen k and a neural network with a carefully chosen number of hidden layers and hidden nodes. [5 marks]

Question 2

Answers all parts below. Please provide answers **in your own words**.

(a) Emma took a rapid antigen test for SARS-CoV-2 (popularly called Covid-19) and her test came out positive. The particular brand of antigen test that she used claims that in their clinical study, the test showed a sensitivity of 0.825 and a specificity of 1.00. Assuming that their claimed numbers are true and given that she has been tested positive, what is the probability that she is actually positive for SARS-CoV-2? [10 marks]

(b) Comment on the interpretability of k-nearest neighbour, decision tree, SVM, random forest and a deep neural network. For each of these supervised learning techniques, (i) how easy or difficult it is to explain the reason behind predictions to a layman, (ii) can you easily find out which training examples need to be modified to change the prediction for a particular query and (iii) can you find out the weight of the different features in your model? [10 marks]

(c) What are the relative advantages and disadvantages of agglomerative and divisive hierarchical clustering algorithms with respect to each other? What are the advantages and disadvantages of the different cluster metrics used in the agglomerative approaches? [10 marks]

(d) Unlike the batch gradient descent, the stochastic gradient descent is not guaranteed to monotonically improve the cost function. And, yet, in applications involving large datasets, it is often preferred over the batch gradient descent. Why? [5 marks]

(e) Computing the exact solution for k-means clustering problem is NP-hard. How is it that k-means remains a highly popular clustering algorithm and is widely deployed in a range of applications and many different big data platforms? [5 marks]

(f) Consider a clustering task where you have three well-separated clusters, but the first cluster has ten times more items than the other two clusters combined. If you were using the random cluster centre initialisation in Lloyd's algorithm, where are the three initial cluster centres likely to be? What will be the impact of that on the final output of Lloyd's algorithm? How will the situation change if you were to use k-means++ for finding the initial cluster centres? [10 marks]

Grading

- Q1: 50 marks
- Q2: 50 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.

oOo