Machine Learning

Worth: 5% of total grade [5 marks]
Due Date: Friday March 27th 2020, 23:59

How do Ensembles Work? (5 marks):

You will have to run the Experimenter in Weka. You will need to run Randomforest, AdaBoostM1, and Bagging. You will have to run all three algorithms on 3 datasets: Iris, Car, and Balance-scale. Make sure you run 10-fold cross validation so you can tell which algorithm is better on which dataset. Run the default settings on the 3 algorithms. It is important that you input the datasets in this order:

1. iris.arff
2. car.arff
3. balance-scale.arff

It is important you enter the algorithms in this order:

1. RandomForest
2. AdaBoostM1
3. Bagging

This is to make the marking of your assignments more automated. If you do them in another order for this task you will not get full marks. In addition, it is important to use the default settings for all the algorithms, please do not change them.

In the second part of the assignment you will need to run the algorithms in two more orders so that each ensemble algorithm is the first algorithm once. But leave the datasets in the same order as above. This means you will run the experiment 3 times in total (once for the first task, and twice more for this task).

What you need to turn in:
You need to turn in a .pdf file of 2 page MAXIMUM which should include:

1) The 3 runs of the two-tailed t-test (corrected) output from Weka "Analyse" in the Experimenter. Include a screen shot of this in your .pdf
2) A paragraph about each of the 3 datasets saying why you think a particular algorithm did the best on each dataset. What is it about this particular dataset that made a particular algorithm do better or worse? Use your knowledge of the algorithms, to hypothesis what must be true about the datasets.
3) A paragraph explaining why you get different results, about which algorithms are statistically significantly different from each other, depending on which algorithm you run first.
4) A final paragraph saying which algorithm you think is more reliable and why.

Marking is based on the following material:
½ mark for having the correct Weka output
1.5 marks (0.5 for each paragraph about each dataset saying "why" you think a particular algorithm did the best on each dataset)

1.5 marks for explaining why you get different results when you run the algorithms in a different order
1 mark for the final paragraph saying which algorithm is more reliable
½ mark for good presentation and English

The datasets can be found at
https://canvas.auckland.ac.nz/courses/45892/files/folder/Pat/Assign2

The assignment must be submitted to Canvas. It will be run through Turnitin.