Data Mining Assignment 3, [10 Marks]

Your name goes here

March 23, 2016

I declare that all material in this assessment task is my own work except where there is clear acknowledgement or reference to the work of others and I have complied and agreed to the UIS Academic Integrity Policy at the University website URL: http://www.uis.edu/academicintegrity

Student Name: Your name goes here UID: Your UID Date: March 23, 2016

Tree classifiers

- 1. Test the ID3 classifier on the weather data (in weather.nominal.arff.)
  - Report (250 words) your observations and the methodology of your test.
- 2. Invent a heuristic to deal with missing values in breast-cancer.arff and implement it, you may need to export the arff file data to a spreadsheet implement your heuristic as a formula, then re-export the data to make a new arff file named "UID- breast-cancer.arff". Test it with ID3 and compare the results to those ID3 produces when classifying the data in breast-cancer.arff.
  - Report the heuristic algorithm and your test methodology (250 words max) and your observations on classifying the new file with ID3.
- 3. Run a series of tests on the soybean data (in soybean.arff) to determine ID3's ability to generalize. Divide the data into training and a test set. Run experiments with different numbers of training instances, in each case using several random splits to be sure that the results are not fortuitous.
  - Report (250 words) your observations and the methodology of your test.

- 4. Make a copy of the soybean.arff file. In a separate experimental process use the following technique. Split the original training data into a small training set UID-soybean-train.arff and a large test set UID-soybean-test.arff. If the ID3 tree classifier trained from the training set classifies test objects incorrectly, add some of the incorrectly classified objects to the training set and re-retrain the classifier. Consider the following claim: "Accurate decision trees (here tree classifier ID3) are usually trained more quickly by this iterative method than by training directly from the entire training data set", do your observations support this claim?
    - Report (250 words) your observations and the methodology of your test.
- 5. Present your results so that they are easy to understand (may be use graphical presentation) and explain them carefully.
    - Use only relevant data from WEKA (including unnecessary data here may adversely affect your grade, I am assessing your classifier training experimental work not your weka outputs)
- 6. Attempt to reproduce your best result from 3 with UserClassifier tree classifier?
    - Report on your attempt (250 words) and include a copy of your tree (use visualize tree)

Pleasel submit in 2 different drop boxes as follows: Your report shall be submitted via the turnitin link only, reports submitted elsewhere will not be graded (Max similarity 15%).

Your data files namely "UID- breast-cancer.arff, UID-soybean-test.arff, UID-soybean-train.arff", where UID is your own UID in an *uncompressed*.TAR archive, in the drop-box marked data.
