Data Mining Assignment 3
Due date: April 9, 11:00PM

Your name goes here

March 31, 2017

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 31, 2017

Tree classifiers

- 1. Test the ID3, PRISM, JRIP classifiers on the Weather data "nominal". Repeat these experiments with the Vote dataset.
  - Report (250 words) your observations and the methodology of your tests. Be sure to discuss which classifiers do not work and why.
- 2. Weka has many filters to deal with missing data one or more will work well with the breast cancer dataset, investigate this and use the knowledge you gain to invent your own (heuristic) to deal with missing values in breast-cancer.arff and implement it, I leave the details of how you do this up to you just report what you did as the experimental methodology. For example, you may 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. Name your fixed up dataset "UID- breast-cancer.arff". Test it with ID3 and compare the results to those ID3 produces when classifying the data in breast-cancer.arff after applying the ReplaceMissingValues filter. Using a WEKA filter as your heuristic or copying a WEKA heuristic algorithm will forfeit the points available for this task.

- Report the heuristic algorithm, 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, we are exploring the concepts from week 9-10 theory. 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.

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)

Please 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", and report TeX file, where UID is your own UID in an *uncompressed*.TAR archive, in the drop-box marked data.

Please submit 1 report comprising 4 sections corresponding to the 4 tasks. The manuscript needs no abstract or introduction. Each section will contain a brief description of the methodology followed for the task, observations, and opinions.

Last updated: March 31, 2017

Typeset using TEXshop