CS 6360: Advanced AI
Assignment 5

Due: 04/25/2017, at midnight

General Instructions:

If anything is ambiguous or unclear:

1. **Discuss possible interpretations with other students, your TA, and instructor**
2. **Make assumptions, state them explicitly, and then use them to work out the problem**
3. **Use Piazza for discussions among yourselves, and also for questions, also for questions and clarifications you need from the instructor and the TA. Piazza levels the playing field because the responses to questions asked are of interest to all, and shared by all. If you have very specific questions, you may email the TA**

Remember that after general discussions with others, you are required to work out the problems by yourself. All submitted work must be your own work. Please refer to the Honor code for clarifications.

Studying Classification Problems using Weka

Assignment (100 points)

For this assignment you will study classification algorithms using the Weka (Waikato Environment for Knowledge Analysis) package written in Java. The Weka package can be downloaded and installed from
http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html.

The following items from the Documentation page introduce the software and some of its capabilities:

- An introductory video by Brandon Weinberg (https://www.youtube.com/watch?v=m7kpIBGEdkI)
- A presentation demonstrating all graphical user interfaces (GUI) in Weka (uploaded to Blackboard as weka GUI.ppt).
- A presentation which explains how to use Weka for exploratory data mining (uploaded to Blackboard as Weka_a_tool_for_exploratory_data_mining.ppt).

Here are two data sets you will use for your assignment..

1. The first data set is Fisher's classic Iris study, in which 150 different instances of the Iris flower are classified into 3 distinct classes. Each flower is defined by four features: (1) *sepal length* in cm; (2) *sepal width* in cm; (3) *petal length* in cm; and (4) *petal width* in cm:

You may find a copy of this data set in CSV format and download it from http://archive.ics.uci.edu/ml/datasets/Iris. For this data set you will run and compare three classifier algorithms: (1) C4.5 decision tree analysis; (2) k-NN classifier, and (3) Naïve Bayes classifier.

2. The second data set is the Car Evaluation data set that can be downloaded from https://archive.ics.uci.edu/ml/datasets/Car+Evaluation. The data comes from four classes: *unacc, acc, good, vgood.* There are 1728 data object, each has 6 features: (1) **buying**: *vhigh, high, med, low*; (2**) maint**: *vhigh, high, med, low*; (3) **doors**: *2, 3, 4, 5 or more*; (4) **persons**: 2, 4, more; (5) **lug_boot***: small, med, big*; and (6) ***safety***: low, med, high. For this second, much larger data set you will compare the results of C4.5 decision tree analysis with the Naïve Bayes classifier.

For both data sets, you will find a number of references in the UCI repository and otherwise that discuss the results of applying different classifiers to this problem. Also, remember to run routines form the Weka library, you need to have java installed on your system.

You task is study the three algorithms, and by running cross-validation studies, demonstrate which one produces the most accurate results. You should run 10-fold cross validation to generate all of your performance parameters for each data set. Be sure to create a table that reports the performance of each of the classifiers. Be sure to: (1) rank the classifiers in terms of accuracy on the training set; and (2) rank them in terms of accuracy in testing through cross-validation. You should discuss in your paper why you think the three algorithms perform differently from one another, if at all. For example, you should discuss which classifier has the greatest discrepancy between training set and cross-validation test set accuracy? Why might that be?

Also, for k-NN, as a first experiment with the Iris data, try the classifier with 3-, 5-, and 7-NNs, note which one generates the best results, and discuss why? Use the NN classifier with the best results for the comparison table.

Report. The written report you submit to report and discuss your results should be brief, i.e., about 3 single-spaced pages at the most. Divide the report into sections: (1) Introduction: A brief description of the three algorithms; (2) Experiments and Results: Explain your experimental procedure, and the results you generate, in a way that others can replicate the experiments you conducted. In addition to the table, if appropriate, use graphs that help you make your points; and (3) Discussion and Conclusion: Summarize and analyze your findings and draw conclusions. It is important that your conclusions are clearly separated from your results: Results are facts, your interpretation of them will to some extent be your opinion.