University of Maryland Baltimore County
Department of Information Systems
Spring 2017

IS 698/800: Smart Home Health Analytics
Homework 3

(Handed Out: March 27, 2017 (Monday), Due: April 03, 2017 (Monday) Before Class)

General Instructions: Use printer paper for your answer sheets. Use blue or black ink. Number each page and write down the total number of pages on the upper right-hand corner of the first page. Thanks.

For this assignment you will compare the performance of the naive Bayes, nearest neighbor, and decision tree learners. You will also learn more about the decision tree algorithm.

1. Recall the loan.arff dataset that provides 18 training examples of whether or not a loan is approved for an applicant based on their income, debt and education. Using this data, compute the entropy for the entire dataset and the information gain for each of the three features (Income, Debt, Education) as the top-level feature in a decision tree. Also indicate which of the three features is the best choice for the top-level split feature. Show all your work.
2. Use WEKA to run the J48 decision-tree classifier on the loan.arff dataset. Use the default parameter settings for J48, and use the training set as the test option.
    a. Include in your report the printed results (tree and statistics) from WEKA.
    b. Draw graphically the decision tree classifier learned by J48.
    c. What is the percent accuracy of this tree on the training set?
3. WEKA's default parameter settings for J48 are -C 0.25 -M 2.
    a. Explain in your own words what these parameters mean.
    b. Find a setting for the -C and -M parameters so that the learned tree achieves 100% accuracy on the training set. Describe the difference between this tree and the one learned in problem 2.
4. Perform the same experiment as in Homework 2 Problem 4, except you should use the five classifiers: NaiveBayes, J48, and IBk (with k=1, k=3 and k=5). IBk is the nearest-neighbor classifier. Use default parameters for each (except the K parameter for IBk). As before, include in your report a table giving the percent correctly classified instances in the test split for the five classifiers on each dataset.
5. Compare the performance of the five classifiers based on the results from the previous problem. Specifically, which classifier performs better on which datasets and why. The "why" part should consider the characteristics of the data, the hypothesis space, and the learning algorithm.
6. Turn in your nicely-formatted report (PDF preferred) containing your responses to the above problems in class.