University of Maryland Baltimore County
Department of Information Systems
Spring 2017

IS 698/800: Smart Home Health Analytics
Homework 4

(Handed Out: April 3, 2017 (Monday), Due: April 17, 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 learn how to perform a statistical comparison of learning algorithms both by hand and using WEKA.

1. Consider the following error rates made by the hypotheses learned by two different learning algorithms L1 and L2 using a 10-fold cross-validated paired t-test.

Trial	L1	L2
1	0.36	0.25
2	0.36	0.24
3	0.26	0.20
4	0.23	0.20
5	0.25	0.21
6	0.28	0.22
7	0.33	0.25
8	0.33	0.24
9	0.28	0.21
10	0.30	0.20

   a. What is the 95% confidence interval around the true error for learner L1? Show all your work.
   b. What is the 95% confidence interval around the true error for learner L2? Show all your work.
   c. Can we conclude with 95% confidence that learner L2 is better than learner L1 on this domain? Show all work used to justify your answer.

2. For this problem, we will use WEKA to generate ROC curves for J48 and NaiveBayes on the labor dataset. First, we need to generate and save ROC curve data.

a. Using the WEKA Explorer open the labor dataset under the Preprocess tab.
b. Under the Classify tab, choose the J48 classifier with default settings and click Start to perform the default 10-fold cross-validation test.
c. In the Result list window, right-click on the J48 entry and choose Visualize Threshold Curve and class "good". The visualization window will appear.
d. Verify the X axis to be False Positive Rate, and the Y axis to be True Positive Rate. You should now see the ROC curve.
e. Click Save and store the results to a file in ARFF format.
f. Exit the visualization window and repeat the above for the NaiveBayes classifier with default settings.

Now, we need to load the data into Excel (or some other charting software) to visualize the ROC curves for both classifiers at once. Here's an outline of the process for Excel 2007.

g. Edit the two ARFF files containing the threshold curve results saved above and remove everything above and including the "@data" line. Note that the False Positive Rate and True Positive Rate values are the sixth and seventh entries, respectively, in each line.
h. Open Excel and choose Data -> Get External Data -> From Text. Browse to the first ARFF file and load it as a Delimited file using comma as the delimiter. Do the same for the second ARFF file.
i. Insert a chart of type Scatter with Straight Lines and put two lines on the plot: one is TP vs. FP for J48, and one is TP vs. FP for NaiveBayes.
j. This chart will now show the two ROC curves for J48 and NaiveBayes on the labor dataset.
k. Nicely format your chart with a title, correct axis titles, correct legend titles, and proper ranges on X and Y axes.

3. Discuss your conclusions about the performance of J48 vs. NaiveBayes on the labor dataset based on the appearance of the ROC curves.

4. Print a file containing the following and submit in the class.
    a. Raw threshold curve data for J48 and NaiveBayes on the labor dataset (the two files you saved in step 2e above).
    b. Nicely-formatted report (MSWord or PDF) containing:
        - Answers from problem 1.
        - Nicely-formatted plot of the two ROC curves (question 2).
        - Discussion of performance comparison based on the ROC curves (question 3).