Assignment 4

Learning Objectives:

1. Gain insight into the concept of confidence intervals
2. Learn how to use Weka's Remove Folds filter and the Experimenter

Description:

Note that this assignment has 2 parts. For part one of this assignment you will work with the data you are working with for your term project. If for some reason you are not able to do this assignment with your own data, you need to set up an alternative arrangement with the instructor ahead of time.

In order to do this assignment, you will have to use the Remove Folds filter to create 5 train/test pairs (See slides from Thursday's lecture) from your project dataset. Use the stratified option so that you will have a roughly equivalent class value distribution across folds. If the problem you are doing is a numeric prediction problem, you should first transform it into a classification problem by selecting a threshold on the values in the class value distribution. The nominal version of your class value should then be "High" if the real value is above that threshold, and "Low" otherwise. For this assignment, remove the original column with the numeric version of the class value before doing the assignment. Do this transformation on the class value attribute before applying the Remove Folds filter.

(1) Using the train/test pairs you created, compute the success rate and 90% confidence interval using cross-validation over the whole set of data. To compute average performance, compute performance for each fold and then average across folds. The classifier you should use is Naïve Bayes. In preparation for problems (2) and (3) below, set up Weka to output the predictions before you do the cross validation. Save the result buffer on each fold.

Comment on what would happen to the confidence interval if the number of data points were doubled. (See Witten & Frank, chapter 5, section 2)

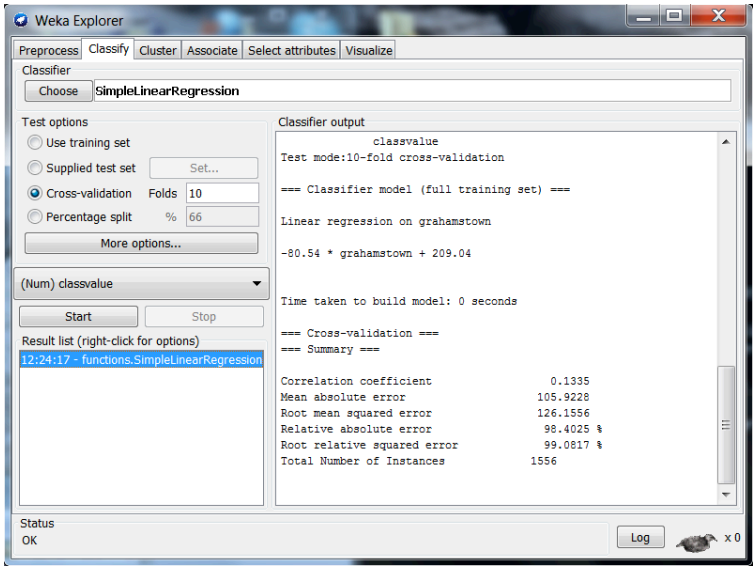
Success rate:
Number of trials:
Z = 1.65
Confidence interval: [_____ , _____]

(2) To set up for the error analysis in problem (3) below: Save each of your test files in csv format. Then using the text files you have saved the result buffers into, copy out the predictions, and copy them in to a column in the corresponding test csv file. Then combine the test csv files into one file so that every instance in your dataset is included, and you have one column for the actual class value and

a separate one for the predicted class value. Then open the file in Excel and sort the file by actual and predicted class values so that the instances that belong together in the same cell within the confusion matrix from your cross validation are in contiguous segments of the file.

(3) Now do an error analysis on the output file you constructed from problem (2). What are the systematic errors you see? i.e., Can you find any explanation for how subsets of instances ended up in error cells in the confusion matrix rather than being correctly classified? For example, if a set of examples that should have been classified as category A were misclassified as category B, what features make those examples look different from those that were correctly classified as A and similar to those that were correctly classified as B.

(4) For question 4, you will use the two data sets included in the Assignment 4 folder on blackboard. The two data sets are very similar. The only difference is that the less simple data set includes two features not included in the more complex dataset. If you run an experiment using the SimpleLinearRegression classifier on the simple version of the data set, you get the following result:



What you should be trying to do in your experimentation on this part of the assignment is to beat this score. You can use either dataset, and you can use any numeric prediction classifier **except for** the SimpleLinearRegression.

You will use both the Explorer and the Experimenter to prepare your solution to this question. First, in the Explorer, load the data set you chose from the two provided and evaluate its performance using cross validation and the numeric prediction classifier you have selected. The first thing you should insert into your solution document is a statement of which dataset you picked and which algorithm you are using. If you changed any parameter settings on the chosen algorithm, you should also mention that. After your description, insert a screen

shot with the performance of your experiment showing, just like you see with the simple data and the SimpleLinearRegression above. You need to answer the following questions: (1) Is the algorithm you picked significantly different from SimpleLinearRegression on the simple version of the dataset? (2) And for the algorithm you picked, does its performance differ significantly between the two versions of the dataset? You must use the Experimenter to answer these two questions, and you must include screen shots to document your answer (both the front panel where you set up the analysis and the analyze panel where you actually perform the t-test). Finally, explain how you know whether the performance differs significantly in both cases (See Witten & Frank, chapter 5, section 5).