Machine Learning 201

Homework 1

Mike Bowles, PhD & Patricia Hoffman, PhD

Check out the web page which describes a wine quality data set:

http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality.names

Use the Red Wine data set: winequality-red.csv This data set contains 1599 observations of 11 attributes. The median score of the wine tasters is given in the last column. Note also that the delimiter used in this file is a semi colon and not a comma.

1) Use forward stepwise regression on the Red Wine data set with 10 fold cross validation. What is the least squared error? Which variables are included in the optimum solution?

2) Next use backward stepwise regression on the Red Wine data set 10 fold cross validation. What is the least squared error? Which variables are included in the optimum solution?

3) Now use all subset selection on the Red Wine data set 10 fold cross validation. What is the least squared error? Which variables are included in the optimum solution? Was there an improvement?

4) Finally install the lars package and use both lar and lasso on the Red Wine data set with 10 fold cross validation.

5) Create a chart with each of the techniques from the previous problems given in the first column. The first row should have the headings: method, least square error, and coefficients in play. Fill in the chart with the values obtained from doing problems 1 - 4 (this can be done by hand). Plot (using r) the solutions for problems 1, 2, & 3 all on the same graph. This plot should graph squared error vs. subset size.