ITU

Homework 3
Patricia Hoffman, PhD.

1) Once again check out wine quality data set described in the web page below:

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

Remember the Red Wine data set (winequality-red.csv) 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. This problem is to create an ordinary least squares linear model (use the lm function in R) for this data set using the first 1400 observations. Don’t forget to scale each column before you create the model. Next check the model’s performance on the last 199 observations. How well did the model predict the results for the last 199 observations? What measure did you use to evaluate how well the model did this prediction? Next use the model to predict the results for the whole data set and measure how well your model worked. (hint: use the r function lm and the regression example from class)

2) Perform a ridge regression on the wine quality data set from problem 1 using only the first 1400 observations. Compare the results of applying the ridge regression model to the last 199 observations with the results of applying the ordinary least square model to these observations. Compare the coefficients resulting from the ridge regression with the coefficients that were obtained in problem 1. What conclusions can you make from this comparison?

3) This problem uses the Iris Data Set. It only involves the Versicolor and Virginica species (rows 51 through 150). Use cross validated ridge regression to classify these two species. Create and plot a ROC curve for this classification method.

4) See if you can improve on regression-based classification of the iris data that we did in class. Classify the iris data set with second degree terms added using a ridge regression. (ie supplement the original 4 attributes x1, x2, x3, and x4 by including the 10 second degree terms (x1*x1, x1*x2, x1*x3, …) for a total of 14 attributes.) Use multiclass to classify the data and then compare the results with the results obtained in class.

It is fine to use brute force to add these attributes. For those who are adventurous, investigate the function mutate in the package plyr.

5) This is a multi-class problem. Consider the Glass Identification Data Set from the UC Irvine Data Repository. The Data is located at the web site:
http://archive.ics.uci.edu/ml/datasets/Glass%2BIdentification
This problem will only work with building and vehicle window glass (classes 1,2 and 3), so it only uses the first 163 rows of data. (Ignore rows 164 through 214) With this set up this is a three class problem. Use ridge regression to classify this data into the three classes: building windows float processed, building windows non float processed, and vehicle windows float processed.