Statistical Data Mining I
Homework 4

Due: Monday December 2nd (11:59 pm)
45 points

1. (10 points) (Exercise 7.9) For the prostate data of Chapter 3, carry out a best-subset linear regression analysis, as in Table 3.3 (third column from the left). Compute the AIC, BIC, five- and tenfold cross-validation, and bootstrap .632 estimates of prediction error.

2) (10 points) A access the wine data from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/wine). These data are the results of a chemical analysis of 178 wines grown over the decade 1970-1979 in the same region of Italy, but derived from three different cultivars (Barolo, Grignolino, Barbera). The Babera wines were predominately from a period that was much later than that of the Barolo and Grignolino wines. The analysis determined the quantities MalicAcid, Ash, AlcAsh, Mg, Phenols, Proa, Color, Hue, OD, and Proline. There are 50 Barolo wines, 71 Grignolino wines, and 48 Barbera wines. Construct the appropriate-size classification tree for this dataset. How many testing samples fall into each node? Describe the resulting tree and your approach.

3) (10 points) Apply bagging, boosting, and random forests to a data set of your choice (not one used in the committee machines labs). Fit the models on a training set, and evaluate them on a test set. How accurate are these results compared to more simplistic (non-ensemble) methods (e.g., logistic regression, kNN, etc)? What are some advantages (and disadvantages) do committee machines have related to the data set that you selected?

4) (10 points ~ Exercise 15.6) Fit a series of random-forest classifiers to the SPAM data, to explore the sensitivity to m (the number of randomly selected inputs for each tree). Plot both the OOB error as well as the test error against a suitably chosen range of values for m.

5) (5 points) What is “random” about a random forest?