Statistical Data Mining I
Homework 4

Due: Wednesday December 6th (11:59 pm)
50 points

Directions: Select only FIVE exercises. Submit all source codes with write up.

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 training and 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) (10 points; Exercise 11.7) Fit a neural network to the spam data of Section 9.1.2. The data is available through the package "ElemStatLearn". Use cross-validation or the hold out method to determine the number of neurons to use in the layer. Compare your results to those for the additive model given in the chapter. When making the comparison, consider both the classification performance and interpretability of the final model.

(6) (10 points) Take any classification data set and divide it up into a learning set and an independent test set. Change the value of one observation on one input variable in the learning set so that the value is now a univariate outlier. Fit

separate single-hidden-layer neural networks to the original learning-set data and to the learning-set data with the outlier. Use cross-validation or the hold out method to determine the number of neurons to use in the layer. Comment on the effect of the outlier on the fit and on its effect on classifying the test set. Shrink the value of that outlier toward its original value and evaluate when the effect of the outlier on the fit vanishes. How far away must the outlier move from its original value that significant changes to the network coefficient estimates occur?

(7) (10 points; ISLR modified Ch9ex8) This problem involves the OJ data set in the ISLR package. We are interested in the prediction of "Purchase". Divide the data into test and training.

(A) Fit a support vector classifier with varying cost parameters over the range [0.01, 10]. Plot the training and test error across this spectrum of cost parameters, and determine the optimal cost.

(B) Repeat the exercise in (A) for a support vector machine with a radial kernel. (Use the default parameter for gamma). Repeat the exercise again for a support vector machine with a polynomial kernel of degree=2. Reflect on the performance of the SVM with different kernels, and the support vector classifier, i.e., SVM with a linear kernel.