Statistical Data Mining I
Homework 5

1) Construct an appropriate-size classification tree for the vehicle data (ublearns).

2) For the prostate data of Chapter 3 (ElemStatLearn, carry out a best subset 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.

3) 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. Apply an ensemble technique (e.g., random forests or boosting). Compare the performance.

   b) Construct an LDA model, and compare your performance to part (A).

Extra Credit:

4) The "covertype" data (ub learns) was obtained from the US Forest Service and are concerned with seven different types of forest cover. The data can be found on UB learns. There are 581,012 observations (each a 30x30 meter cell) on 54 input variables (10 quantitative variables, 4 binary wilderness areas, and 40 binary soil type variables). Divide these data randomly into a training set and a test set. Use any of the methods we discussed to develop a model that predicts the forest cover type. Report the training and test error rates.