STATS 415 Homework 9

Due Thursday March 29, 2018

Please include your name, uniqname, and lab section (number or time or GSI). A point will be taken off homework without the section info. Turn in a printout of your homework in the lecture or in your GSI's mailbox across room 305A West Hall, no later than 5pm on the due date.

1. Draw an example partition of a two-dimensional feature space, with each feature defined on the interval (0,1), that could result from a tree with binary splits. Include at least 5 regions. Draw a decision tree corresponding to the partition you drew, labeling every tree branch. Show terminal nodes as rectangles and interior nodes as circles. Do not include predictions at the terminal nodes, just the branch labels (e.g. $X_1 < 0.5$).

2. This question uses the `crabs` data, available through the R package `MASS`. The data contain five size-related measurements on two different species of crabs, blue and orange, with 50 male and 50 female crabs of each species measured.

   (a) Set the random seed to 45678 and randomly select 80% of the data as your training data. Make sure you select the same number from each species/sex combination. Set the remaining 20% aside to use as test data.

   (b) Train a classification tree to predict Species from the five numerical measurements and sex, selecting the optimal size by cross-validation but using no more than 8 splits. Plot the tree. Comment on which variables are used by the tree. Compute training and test errors.

   (c) Now train random forests on the data, using three randomly selected predictors at each split, and 1000 trees total. Make a variable importance plot and compare with your results for a single tree. Compute training and test errors.

(d) Finally, train AdaBoost on the data. Plot the training and test errors as a function of the number of trees M constructed by boosting, for a range of values of M up to 1000. Report training and test errors for a value of M of your choice.

(e) Comment on which method appears to perform best for this dataset, and how consistent the results are across methods.

Please limit your solution to Problem 2 to at most 6 pages.