Data Mining I

Homework 6

Select 2 problems – a third may be done for extra credit

1) (a) 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.

   b) How accurate are these results compared to more simplistic (non-ensemble) methods (e.g., logistic regression, kNN, etc)? Use the same test/training as in part A.

   c) What are some advantages (and disadvantages) do committee machines have related to the data set that you selected?

2) Consider the pima data. Use boosting, random forests and a single tree (CART model). Comment on your performance. Explore the partial dependence plots for those variables that are have high ranking "variable importance".

3) (ESL 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.