

Homework 3

Due: Thursday, November 29

1. Bagging

- (a) Text 8.1
- (b) Perform a bagged estimation scheme using 100 bootstrap samples on the spam training data. Report the out-of-bag error for
 - i. Averaging probability estimates for each class
 - ii. Using a majority-vote classifier.

How do these compare to the test-error rates for your bagged classifier?

2. Boosting: Text Ex 10.1, 10.2

3. California Housing Data

The dataset `california.data` consists of aggregated data from 20,640 California census blocks (from the 1990 census). The goal is to predict the median house value in each neighborhood from the other attributes described in `california.info`.

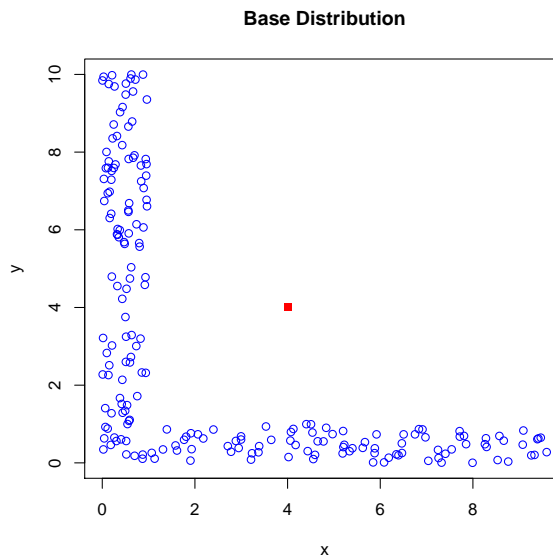
The `gbm` package in R implements Friedman's gradient boosting machine (MART in the text). Fit this model to the data and write a short report that should include *at least*

- (a) The prediction accuracy of `gbm` onto the dataset.
- (b) Identification of the most important variables.
- (c) comments on the dependence of the response on the most important variables (you may want to consider singleplots, pairplots etc).

4. Trees and Stability

- (a) Text, Ex 9.6
- (b) Obtain 5 versions of the California housing data data by sampling *with replacement*. Train a tree on each data set. Report on how stable the top three layers of the trees were.
- (c) One of the supposed advantages of trees is that they extrapolate as constants, so that prediction should not be too bad. To examine how reasonable this statement is, construct the following experiment
 - Generate 200 bivariate data points uniformly on arms of width 1 and length 10 along each axis. An example data set is plotted below.

- Generate a response for each data set given by $y_i = x_{1i} - x_{2i} + \epsilon_i$ with ϵ_i normal with unit variance.
- Use `rpart` to build a tree on these data. Repeat this 100 times and record the value of the prediction that your model gives at the point (4,4).
- Plot a histogram of the predictions. Does this appear to be "stable extrapolation"?



5. Partial Dependence Plots

- (a) Text 10.3
 - (b) Partial dependence plots integrate across the features that are not of interest. For the trees from Question 4c, plot the partial dependence on x_1 . Compare their stability and accuracy with the conditional dependence at $x_2 = 0$.
 - (c) Repeat Question 4c for `gbm`. Are the predictions more stable? Are the partial dependence plots?
6. Attempt your best fit to the Netflix data. Describe your approach and report your best estimate of prediction accuracy. How did you arrive at this? Submit predictions for the test data electronically.