Statistics 745

Assignment 7

1. **Individual Assignment:** The AIRQUALITY data consists of daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973 with the following variables:

   (a) Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island

   (b) Solar.R: Solar radiation in Langleys in the frequency band 4000 to 7700 Angstroms from 0800 to 1200 hours at Central Park

   (c) Wind: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport

   (d) Temp: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

   Program a regression tree with $M = 3$ for the air quality data with $Ozone^{1/3}$ as the response.

2. **Individual Assignment**

   (a) This problem concerns the pima data set in the faraway package. Use test as the response. Break the data with the first 50% as training and the remaining as testing. Transform the response so that it is either -1 or 1 as opposed to 0 or 1, i.e. $\text{sign}(f)$ is the classification rule. Write a bagging routine in **R** to fit trees with $M = 10$ and $ntree = 100$. Plot iteration against testing/training error.

   (b) Implement least squares boosting and apply it to the data. Choose $ntree = 500$.

   (c) Define $\nu \in (0, 1]$ as the learning rate and modify the update step as $f(x) = f(x) + \nu\beta h_j(x)$. Note $\nu = 1$ is the boosting algorithm. Implement this with $\nu = 0.05, 0.1, 0.5, 0.75$. What values of ν perform well? Also is there a relationship to M?