Jiayuan Guo -- HW4

1. **Chapter 8, Problem 2**

   1) Set $\hat{f}(x) = 0$ and $r_i = y_i$

   2)$\hat{f}^1(x) = c_1 I(x_1 < t_1) + c_1' = \frac{1}{\lambda} f_1(x_1)$

   3) Set $\hat{f}(x) = \lambda \hat{f}^1(x)$ and $r_i = y_i - \lambda \hat{f}^1(x_i)$

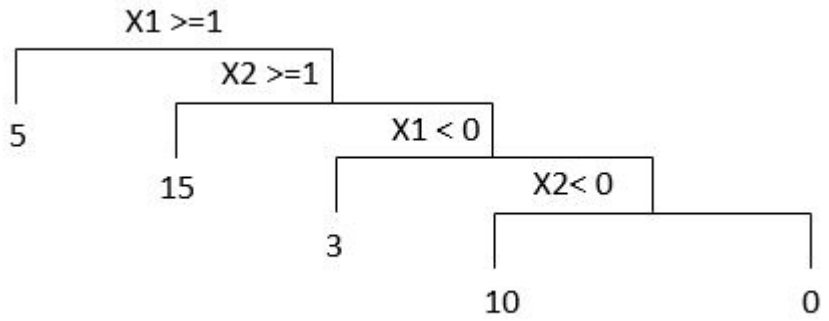
   4) $\hat{f}^2(x) = c_2 I(x_2 < t_2) + c_2' = \frac{1}{\lambda} f_2(x_2)$

   5) Maximize the fit to the residuals, another distinct stump must be fit. $\hat{f}(x) = \lambda \hat{f}^1(x) + \lambda \hat{f}^2(x)$ and $r_i = y_i - \lambda \hat{f}^1(x_i) - \lambda \hat{f}^2(x_i)$

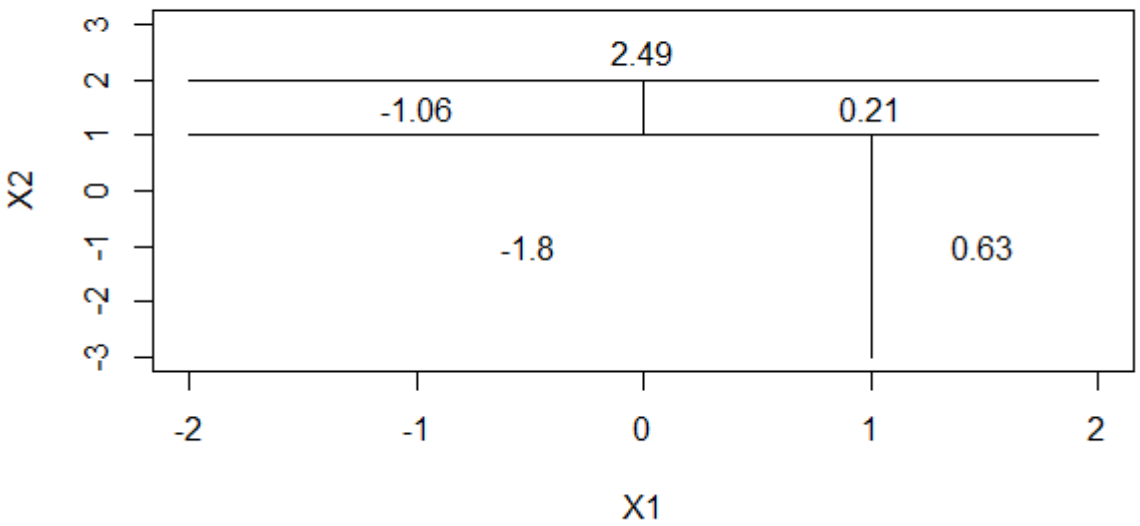
   6)Iterate and finally get: $\hat{f}(x) = \sum_{j=1}^{p} f_j(x_j)$

2. **Chapter 8, Problem 4**

(a)



(b)



3. **Chapter 8, Problem 5**

1)For the majority vote approach: 6 estimates for red and 4 estimates for green. The number of red predictions is greater than the number of green predictions, so we classify X as red.

2)For the average probability approach: the mean of the 10 estimates is 0.45, so we classify X as green.

4. **This problem uses the diabetes data in lars package. Specifically, use the x2 design matrix, and the y outcome. First split the data into training and test sets of size 300 and 142 each, using 1234 as the random seed. Make sure to fit the models on a training set and to evaluate their performance on a test set.**
   (a) Apply boosting, bagging, and random forests to predict the outcome. Compare the performance of these methods to linear regression with and without penalties.
   (b) Repeat the previous task 10 times, using random seeds 1 to 10. Summarize the ranking of the methods in the 10 different runs.

   Will finish it sooner