DS 303 Homework 11
Due: Nov. 27, 2023 on Canvas by 11:59 pm (CT)

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` output or code** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, a .R file, text file, Word file, or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: Survey

Please fill out the class survey on Canvas linked titled **HW 11 Class Survey**. It is part of this assignment and worth 10 points. You'll automatically receive full credit once taking the survey.

Problem 2: Boosting

We'll use the `Hitters` for this problem; it is part of the `ISLR2` library.

(a) Remove the observations for whom the salary information is unknown, and then log-transform the salaries.

(b) Create a training set consisting of the first 200 observations, and a test set consisting of the remaining observations.

(c) Perform a grid search on the training set using 10-fold cross-validation to decide the optimal number of trees, the optimal λ, and optimal depth. To keep things computationally simple, only considered 3 different values for each of the tuning parameters. You may decide what grid of values to use. Report your results for the optimal number of trees, the optimal λ, and optimal depth.

(d) Implement boosting on the training set with the tuning parameters you have selected from the grid search. Which variables appear to be the most important predictors in the boosted model?

(e) What is the test MSE of the boosted model from part (d)?

(f) Now apply random forest to the training set. How did you select your value for m? What is the test set MSE for this approach?

Problem 3: Bias of Trees

The `HW11script.R` simulates a training set from a population regression model.

(a) Run the code. Use this setup to simulate 1000 training sets. The true population regression line can be used to simulate $n = 100$ new Y values. There is no need to generate new X's. For each of these 1000 training sets, train a single decision tree (overfit, no pruning). Then store the predicted values $\hat{f}(X)$ when all X's equal 1 ($X_1 = 1, X_2 = 1, \ldots, X_{20} = 1$). Report the first predicted values for the first 5 iterations of your loop.

(b) Use your results from (a) to obtain the squared bias and variance of the decision trees when all predictor values equal 1 ($X_1 = 1, X_2 = 1, \ldots, X_{20} = 1$). Report both values here. Recall that for a fixed value x_0:

$$\text{Bias}(\hat{f}(x_0))^2 = [E(\hat{f}(x_0)) - f(x_0)]^2$$

$$\text{Va}r(\hat{f}(x_0)) = E[\hat{f}(x_0) - E(\hat{f}(x_0))]^2.$$

(c) Repeat (a) but now train a random forest for each training set. Set $m = 10$ and the number of trees to be 200. Again, store the predicted values $\hat{f}(X)$ when all X's equal 1. Report the first predicted values for the first 5 iterations of your loop. It may take a few minutes for all 1000 iterations to run.

(d) Use your results from (c) to obtain the squared bias and variance of the random forests when all predictor values equal 1. Report both values here.

(e) What is the change in the order of magnitude in (squared) bias between the two methods?

(f) What is the change in the order of magnitude in variance between the two methods?

(g) What can we conclude about how bias and variance behave for ensemble methods compared to a single decision tree?

Problem 4: Understanding K-Means

(a) Prove that

$$\frac{1}{|C_k|} \sum_{i,i^{\in} C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2,$$

where $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$.

(b) On the basis of this identity, argue that the K-means clustering algorithm decreases the total within-cluster variation (our objective function) at each iteration:

$$\min_{C_1,\ldots,C_K} \left(\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i^{\in} C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right).$$

(c) In this problem, you will perform K-means clustering 'manually', with $K = 2$ on a toy example with $n = 6$ observations and $p = 2$ features. The observations are as follow:

Obs	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

i. Plot the observations (using `R`) Show the plot here.

ii. Randomly assign a cluster label to each observation. You can use the `sample()` function in `R` to do this. Report the cluster labels for each observation.

iii. Compute the centroid for each cluster. Report those values here.

iv. Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

v. Repeat (iii) and (iv) until the answers obtained stop changing. Report the centroids and cluster labels for the first two iterations.

vi. In your plot from (i), color the observations according to the cluster labels obtained. Show that plot here.

Problem 5: Dendrogram

Suppose that we have four observations, for which we compute a dissimilarity matrix, given by

$$\begin{pmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4. & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{pmatrix}$$

For example, the dissimilarity between the 1st and 2nd observations is 0.3.

(a) Using the dissimilarity matrix, sketch the dendrogram that would result from carrying out hierarchical clustering on these four observations using complete linkage. Be sure to indicate on the dendrogram the height at which each fusion occurs.

(b) Repeat (a), this time using single linkage hierarchical clustering.

(c) Suppose that we cut the dendrogram obtained in (a) such that two clusters result. Which observations are in each cluster?

(d) Suppose that we cut the dendrogram obtained in (b) such that two clusters result. Which observations are in each cluster?

End of assignment