STA 141C: Homework 4

Due 5/23/2025, 8PM

This homework assignment has to be submitted electronically on Gradescope by the due date. Please submit one PDF file containing all your answers and the code used for your data analysis. It is strongly recommended to type your answers rather than submitting handwritten work. If handwritten, please ensure that it is legible and neat.

Problem 1. (2/25 points)
Boosting using depth-one trees (or stumps) can aid in interpretability, since it leads to an additive model: that is, a model of the form

$$f(X) = \sum_{j=1}^{p} f_j\left(X_j\right),$$

where $f_j(X_j)$ is a function of the jth predictor variable. In other words, the predicted value for a vector X is the sum of functions of its arguments. Explain why this is the case.

Hint: The boosted model is $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$ (Chapter 8, slide 42).

Problem 2. (5/25 points)
In the lab, a classification tree was applied to the `Carseats` data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable. In R, the data are in the ISLR2 package. In Python, you can download the data (csv format) from Canvas.

(a) Split the data set into a training set and a test set.

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

(d) (Python version) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the `feature_importance_` values to determine which variables are most important.
(R version) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the `importance() function` to determine which variables are most important.

(e) (Python version) Use random forests to analyze this data. What test MSE do you obtain? Use the `feature_importance_` values to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.
(R version) Use random forests to analyze this data. What test MSE do you obtain? Use the `importance() function` to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

Problem 3. (5/25 points)
We now use boosting to predict Salary in the `Hitters` data set. In R, the data are in the ISLR2 package. In Python, you can download the data (csv format) from Canvas.

(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 boosting on the training set with 1,000 trees for a range of values of the shrinkage parameter λ. Produce a plot with different shrinkage values on the x-axis and the corresponding training set MSE on the y-axis.

(d) Produce a plot with different shrinkage values on the x-axis and the corresponding test set MSE on the y-axis.

(e) Compare the test MSE of boosting to the test MSE that results from applying two of the regression approaches seen in Chapters 3 and 6.

(f) Which variables appear to be the most important predictors in the boosted model?

(g) Now apply bagging to the training set. What is the test set MSE for this approach?

Problem 4. (5/25 points)
Suppose that we have four observations, for which we compute a dissimilarity matrix, given by

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

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

(a) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

(b) Repeat (a), this time using single linkage 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?

(e) It is mentioned that at each fusion in the dendrogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in (a), for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.

Problem 5. (4/25 points)
This problem involves the K-means clustering algorithm.

(a) Prove equation (12.18) in the textbook.

(b) On the basis of this identity, argue that the K-means clustering algorithm (Algorithm 12.2 in the book) decreases the objective (12.17) at each iteration.

Problem 6. (4/25 points) Write a R/Python function to perform matrix completion as in Algorithm 12.1 in the book, and as outlined in Section 12.5.2. In each iteration, the function should keep track of the relative error, as well as the iteration count. Iterations should continue until the relative error is small enough or until some maximum number of iterations is reached (set a default value for this maximum number). Furthermore, there should be an option to print out the progress in each iteration.

Test your function on the `Boston` data. First, standardize the features to have mean zero and standard deviation one using the `scale()/StandardScaler()` function. Run an experiment where you randomly leave out an increasing (and nested) number of observations from 5% to 30%, in steps of 5%. Apply Algorithm 12.1 with $M = 1, 2, ..., 8$. Display the approximation error as a function of the fraction of observations that are missing, and the value of M, averaged over 10 repetitions of the experiment.

Hint: The last 7 slides for Chapter 12 also provide a good reference.