CMSE 381

Fall Semester 2021

For the coding project, you can find the data on http://www.statlearning.com/resources-first-edition. Please write your codes using Google style.

1. Exercise 2.4.7

2. Exercise 2.4.8

3. Load dat1.csv file from D2L, and write a Nearest neighbor prediction function in R to predict the values of Y for $X = -1.9, -1.8, -1.7, \ldots, 1.8, 1.9$ with radius $r = 0.1$ or 0.3 (namely, we define the neighborhood of a point x as the interval $[x - r, x + r]$). Make a scatter plot of original data and the predicted results. Explain the difference between the two sets of result. Which one is more flexible?

4. Exercise 2.4.10. you can find the data on http://www.statlearning.com/resources-first-edition. Please write your codes using Google style.

5. What are the advantages and disadvantages of very flexible (vs less flexible) approach for regression or classification? When would be a more flexible approach preferable? What about a less-flexible approach?

6. (Challenging problem, not required) Given nine data point in descending order $\{x_1, x_2, \ldots, x_9\}$, prove that the best point, a, to represent these nine data with the smallest absolute error is the median of these nine data points, which is x_5 . Namely, a minimizes

$$\sum_{i=1}^{9} |x_i - a|$$

7. For simple linear regression, we assume that $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ and X is fixed (not random). We collect n i.i.d. training sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Prove that the $(\hat{\beta}_0, \hat{\beta}_1)$ estimated through minimizing RSS equals to the one through maximizing likelihood.

8. Using equation (3.4) in textbook, prove that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

9. Exercise 3.7.8

10. Download the Regression.csv file from D2L, and write a modified KNearest Neighbor prediction function to predict the values of Y for $X = -2.30, -2.29, -2.28, \ldots, 2.28, 2.29, 2.30$ with $K = 1, 5, 10,$ and 25 (namely, we define the neighborhood of a point x as the K points in the training set with smallest Euclidean distance to x, and then calculate the median of the observed y at these points).

a Make a scatter plot of original training data and the predicted results (you can use 'lines' in R).

b Calculate the MSE for the training data and the testing data (the third column in the downloaded data) for different Ks and plot it. (similar to Figure 2.10 in the textbook). Describe the pattern for the two MSE. What is the optimal 'K' you will choose?

11. Download wine.csv and 'wine.R' file from D2L. Run the scripts in wine.R to generate the training and testing sets. Write a K-Nearest Neighbor classifier function.

a For $k = 1, 2, \dots, 10, 20, 30$, calculate the training error rate and the testing error rate for different Ks and plot it. (similar to Figure 2.17 in the textbook). Describe the pattern for the two error rates. What is the optimal 'K' you will choose?

b For $K = 10$, we fix $x_2 = 4$ and vary the x_1 from 10 to 30 to estimate the $P(Y = 1|X_1 = x_1, X_2 = x_2)$. Approximately, find out the value of x_1 for the decision boundary with $x_2 = 4$.