CMSE 381

Spring Semester 2021

1. Assume $Y = f(X) + \epsilon$. From training data, we obtain the estimate $\hat{f}(x)$ of $f(x)$. Now assume both $\hat{f}$ and $X = x$ are fixed. prove that

$$E((Y - \hat{f}(X))^2 | X = x) = [f(x) - \hat{f}(x)]^2 + \text{Var}(\epsilon).$$

2. Assuming the same setting as in Question 1 and suppose we have fit a model $\hat{f}(x)$ from some training data Tr. Let (x_0, y_0) be a test observation drawn from the same distribution of training data. Prove that

$$E\left[(y_0 - \hat{f}(x_0))^2\right] = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon).$$

3. (Challenging problem, not required) For a classification problem with $Y \in$ {Apple, Orange, Coconut} and $X \in \mathbb{R}$, we want to find a classifier $C(X) : \mathbb{R} \to$ {Apple, Orange, Coconut} minimizing the following Loss

$$E\left[I(Y \neq C(X)) | X = x\right].$$

   Prove that this ideal classifier is the Bayes Classifier.

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

5. 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.

6. (Challenging problem, not required) It is claimed in the book that in the case of simple linear regression of Y onto X (one predictor), the R^2 statistic (formula 3.17 in the book) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

7. Exercise 3.7.8

8. 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 (dat3.csv) 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?

9. 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 $k = 1, 2, \ldots, 10, 20, 30$, 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 $x_2 = 4$.