STAT 6340 (Statistical and Machine Learning, Fall 2024)

Mini Project 1

Instructions:

- Due date: Sep 4, 2024 (Use Submission Link on eLearning to upload by 11:59 pm)
- Total points = 20
- Submit a typed report.
- It is OK to consult with others regarding this project or use any publicly available resources, but you must write code and answers on your own. Violation of this policy will be considered academic dishonesty, and you will referred to appropriate university authorities. If you don't do the work by yourself, how will you learn?
- Do a good job.
- You must use the following template for your report:

  Mini Project #
  Name
  Section 1. Answers to the specific questions asked
  Section 2: Code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.
- Section 1 of the report must be limited to *two* pages. Also, only those software outputs should be provided in this section that are referred to in the report.
- You can code in `R` or `Python`.
- You are encouraged to use `R Markdown` or `Jupyter Notebook` to write your report.
- Bonus (4 points): Use *both* `R` and `Python` to code, compare the results, and provide an explanation if the results are not identical. This work should be described on a separate page titled Bonus Work with two sections — one to describe the comparison and the other for the code. The first section should be limited to one page.

---

1. (12 points) Consider the training and test data posted on eLearning in the files `1-tranining-data.csv` and `1-test-data.csv`, respectively, for a classification problem with two classes.

   (a) Fit KNN with $K = 1, 2, \ldots, 100$.

   (b) Plot training and test error rates against K. Explain what you observe. Is it consistent with what you expect from the class?

   (c) What is the optimal value of K? What are the training and test error rates associated with the optimal K?

   (d) Draw a plot of the training data that also shows the decision boundary for the optimal K. Comment on what you observe. Does the decision boundary seem sensible?

2. (8 points) Consider the following general model for the training data (Y_i, x_i), $i = 1, \ldots, n$ in a learning problem with quantitative response:

$$Y_i = f(x_i) + \epsilon_i,$$

where f is the true mean response function; and the random errors ϵ_i have mean zero, variance σ^2, and are mutually independent. We discussed this model in the class. Let $\hat{f}$ be the estimator of f obtained from the training data. Further, let (x_0, Y_0) be a test observation. In other words, x_0 is a future value of x at which we want to predict Y and Y_0 is the corresponding true value of Y. The test observation follows the same model as the training data, i.e.,

$$Y_0 = f(x_0) + \epsilon_0,$$

where ϵ_0 has the same distribution as the ϵ_i for the training data but ϵ_0 is independent of the ϵ_i. Let $\hat{Y}_0 = \hat{f}(x_0)$ be the predicted value of Y_0.

(a) Show that $\text{MSE}\{\hat{f}(x_0)\} = (\text{Bias}\{\hat{f}(x_0)\})^2 + \text{var}\{\hat{f}(x_0)\}$.

(b) Show that $E(\hat{Y}_0 - Y_0)^2 = (\text{Bias}\{\hat{f}(x_0)\})^2 + \text{var}\{\hat{f}(x_0)\} + \sigma^2$.