

STAT 4360 (Introduction to Statistical Learning, Spring 2023)

Mini Project 1

Instructions:

- Due date: Feb 2, 2023.
 - Total points = 30
 - Submit a typed report.
 - It is OK to discuss the project with other students in the class, but each student must write their own code and answers. If your submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will be referred to appropriate university authorities.
 - 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: R code (or Python 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 output should be provided in this section that are referred to in the report.
-

1. (20 points) Consider the training and test data included in the zip file (`1-training-data.csv` and `1-test-data.csv`), respectively, for a classification problem with two classes.
 - (a) Fit KNN with $K = 1, 6, 11, \dots, 200$.
 - (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) Make a plot of the training data that also shows the decision boundary for the optimal K .
 - (e) Based on the decision boundary you obtained in (d), what phenomenon do you observe for points near the decision boundary? what phenomenon do you observe for points far from the decision boundary? (Are the points from one side of the decision boundary fall into the same class?) Do you think this decision boundary is sensible?
2. (10 points) Consider the following general model for the training data (Y_i, x_i) , $i = 1, \dots, n$ in a learning problem:

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

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 $E(\hat{Y}_0 - Y_0)^2 = (\text{Bias}\{\hat{f}(x_0)\})^2 + \text{var}\{\hat{f}(x_0)\} + \sigma^2$.
- (c) Based on the phenomenon of “bias and variance trade off”, explain why there is usually a “U” shape for the test MSE when the model flexibility increases.