Assignment 2

Due date: July 29, 2024

* Create a report discussing all of the parts below in pdf format. (including your codes)

1 Cross Validation

1. What are the advantages and disadvantages of k-fold cross-validation relative to:

   (1) The validation set approach?

   (2) LOOCV?

2. Use the simulation dataset ($n = 500$) posted on the blackboard. This dataset is created by
$$y_i = f(x) + \epsilon_i, \quad -1 < x < 4.$$
where $f(x)$ is unknown and $\epsilon_i \sim N(0, 2)$.

3. Fit polynomial regression models from degree 1 to 9.

4. Create R or Python codes for LOOCV and k-fold cross-validation to estimated your fitted regression models. Do not use the pre-installed functions in R and Python packages for LOOCV and k-fold cross-validation. Creat your own functions. You can choose the k value as you want.

5. Display your cross-validation results numerically and graphically.

6. What degree of the polynomial regression model will you choose to fit the data? Why?

2 Bagging and Random Forest models

1. Discuss the disadvantage of the traditional Tree model and describe how the Bagging model can improve this problem.

2. Discuss the disadvantage of the Bagging model and describe how the Random Forest model can improve this problem.

3. In Bagging and Random Forest, how to to estimate the test error without the need to perform cross-validation? Explain in terms of out-of-bag (OOB) observations.