ST563 - Homework 1

Wenbin Lu

1 Instructions

Please follow the instructions below when you prepare and submit your assignment.

- Include a cover page with your homework. It should contain
  1. Full name
  2. Course #
  3. HW #
  4. Submission date
- Assignments should be submitted using **Gradescope**.
- Neatly typed work should be submitted. All R code/output should be well commented on, with relevant outputs highlighted and discussed.
- When you solve a particular problem, do not only give the final answer. Instead, show your work and the steps you used (with proper explanation) to arrive at your answer to get full credit.
- Submission in the PDF format is preferred; Please convert other formats such as doc or docx into PDF.

Problem 1 (30 points)

Consider the model: for $i = 1, \dots, n$,

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

where $\epsilon_i \sim N(0, \sigma^2)$ independent over i. Assume that $x_1, \dots, x_n$ are non-random. Suppose we apply KNN regression method, with a pre-specified value K, to estimate $f(\cdot)$ as a fixed point x_0.

(a) Compute the variance of $\widehat{f}(x_0)$. [Hint: Recall, $\widehat{f}(x_0)$ is just an average of Ys.]

(b) Using your answer in (a), can you verify that more flexible procedures tend to have more variance?

(c) Suppose we run two KNN regressions, one with $K = 10$ and the other with $K = 30$. For each model, we compute training MSE, and test MSE using an independent test sample. Which regression will have lower training MSE? Which one will have lower test MSE? Explain you answers.

Problem 2 (50 points)

Consider the `Boston` data discussed in lectures. Now we want to build a prediction model for `medv` based on the remaining numeric variables in the dataset, excluding `chas`.

(a) Use the KNN regression method to build a predictive model where the hyperparameter is tuned using 5-fold cross-validation?

(b) Estimate the test error of your model using holdout method.

(c) Predict `medv` for the following new datapoint:

```
## # A tibble: 1 x 12
##    crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio lstat
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl>
## 1 0.257     0  9.69     0 0.538  6.21  77.5  3.21     5   330    19.0  11.4
```

(d) How much does `medv` change if `lstat` changes from 5 to 10, while keeping the other variables fixed at their median value?

(e) Do you expect to see the same amount of change in `medv` for each increase of 5 units in `lstat`? Explain.

Problem 3 (20 points)

Do Problem 2 in Chapter 2.4 in “An Introduction to Statistical Learning”, second edition".