Advanced Computational Econometrics
Chapter 1: K-nearest neighbor regression, bias-variance decomposition

Spring 2022

1 Linear and K-NN regression

The `MASS` library contains the `Boston` data set, which records `medv` (median house value) for 506 neighborhoods around Boston. We will seek to predict `medv` using 13 predictors such as `rm` (average number of rooms per house), `age` (average age of houses), and `lstat` (percent of households with low socioeconomic status).

1. Display the data using scatter plots (function `plot`) and boxplots (function `boxplot`). Which variables seem to explain the response variable `medv`?

2. Split the data into a learning set and a test set.

3. Predict `medv` for the test data from the 13 predictors using linear regression. Compute the MSE.

4. Predict `medv` for the test data from the 13 predictors using K-NN regression with different values of K. (Use function `knn.reg` in package `FNN` and normalize the input data using function `scale`). Compute the MSE.

5. Represent graphically the test mean-squared error as a function of K. Which value of K seems to be optimal?

2 Bias variance trade-off

We consider the following model:

$$Y = 1 + 5X^2 + \varepsilon \tag{1}$$

where $X \sim \mathcal{U}([0,1])$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.5$. Let $\widehat{f}_K(x)$ be the estimate of the regression function $f(x) = 1 + 5x^2$ obtained by computing the average of the y_i's for the K nearest neighbors of x. We recall the bias-variance decomposition formula:

$$\mathbb{E}_{\mathcal{L},Y}\left[(\widehat{f}_K(x) - Y)^2 \mid X = x\right] = \mathrm{Var}_{\mathcal{L}}\left(\widehat{f}_K(x)\right) + \left(\mathbb{E}_{\mathcal{L}}[\widehat{f}_K(x)] - f(x)\right)^2 + \mathrm{Var}_Y(\varepsilon \mid X = x). \quad (2)$$

1. Explain the different terms of this formula and prove it.
2. Check formula (2) by simulation, by randomly generating learning sets of size $n = 50$. For some value of x, plot the different terms of the formula as functions of K, for K ranging from 1 to 40. What do you observe? Repeat the experiment for different values of n and comment the results.