# STAT 380 – kNN Regression Part 2 (Lecture 9)

<u>Selecting *k* for kNN</u>

IDEA: One of the important choices in kNN is selecting the number of nearest neighbors, *k*. There is not an agreed upon method or metric for making this choice. One way to make the decision is to use the holdout (validation) set approach for a variety of values of *k* (number of neighbors). Specifically, we will use calculate MSE (or RMSE) on the testing set for a variety of *k* values. We can then choose *k* based on testing set performance.

EXAMPLE 1: The file L08_bmd.csv which contains 169 records of bone densitometries (measurement of bone mineral density). The following variables are included:

| Variable | Meaning |
|---|---|
| id | Patient's identification number |
| age | Patient's age |
| fracture | A categorical variable indicating whether the patient has had a hip fracture |
| weight_kg | Patient's weight in kilograms |
| height_cm | Patient's height in centimeters |
| waiting_time | Time patient spent waiting for the densitometry in minutes |
| bmd | Patient's bone mineral density measurement in the hip |

a. After preparing the data (creating indicators, scaling X's, 85/15 training/testing split using seed of 123), use a for loop to run the kNN regression for predicting bmd using age, weight (in kg), and medication based on values of k from 1 to 75. Store the MSE and RMSE of the testing set for each value of k. NOTE: You should use the same Train and Test for all values of k.

Logic:

1. Prepare data (create indicators, scale data)

2. Train/Test split

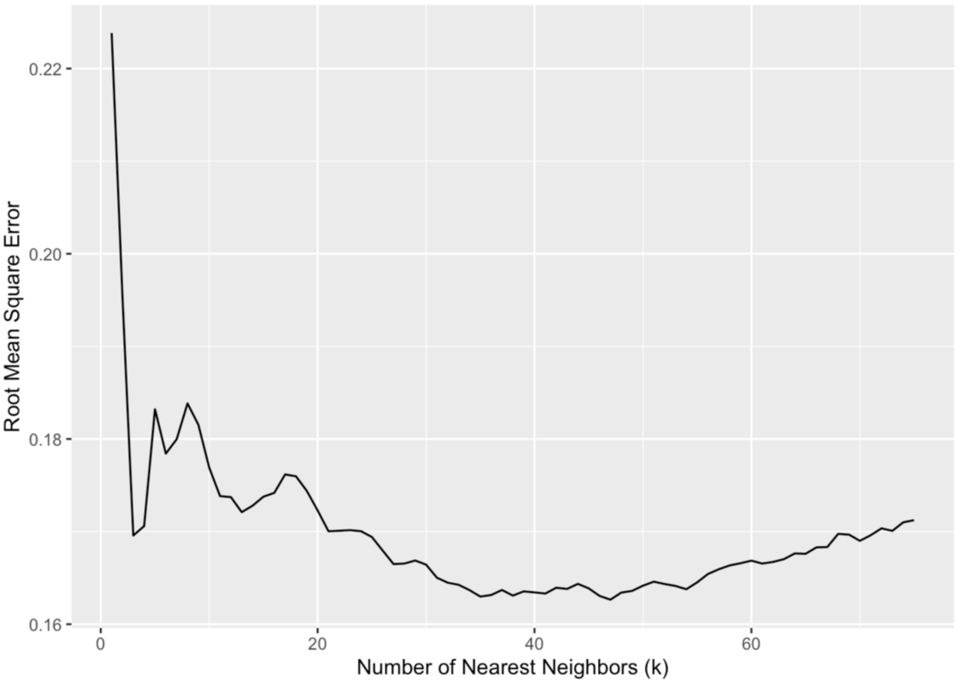
3. Loop

    a. Initialize objects for storing results

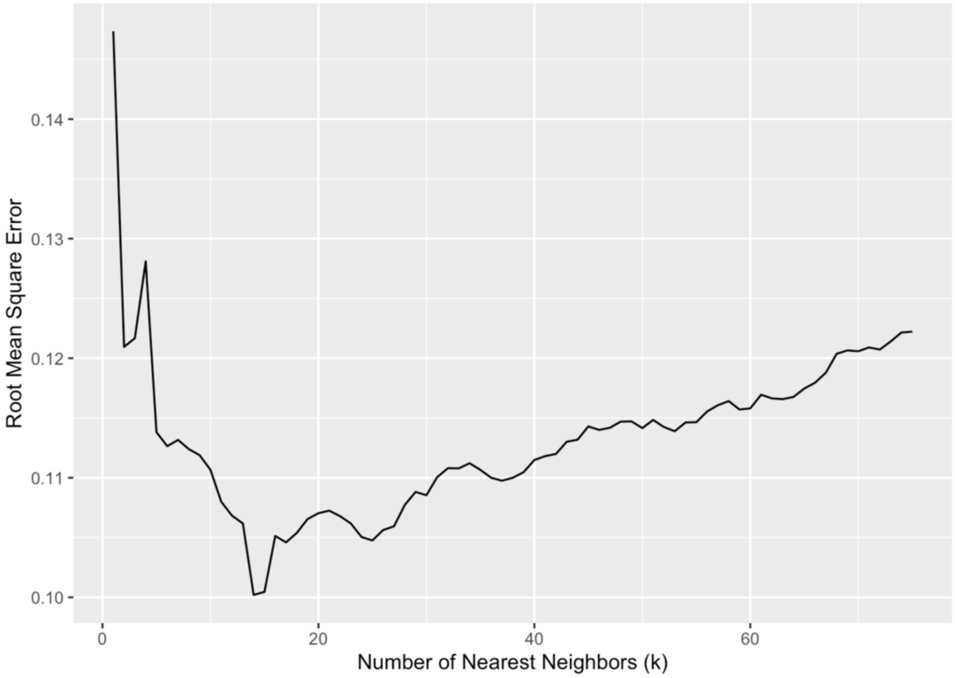
    b. Loop through values of k

NOTE: Loop structure in R

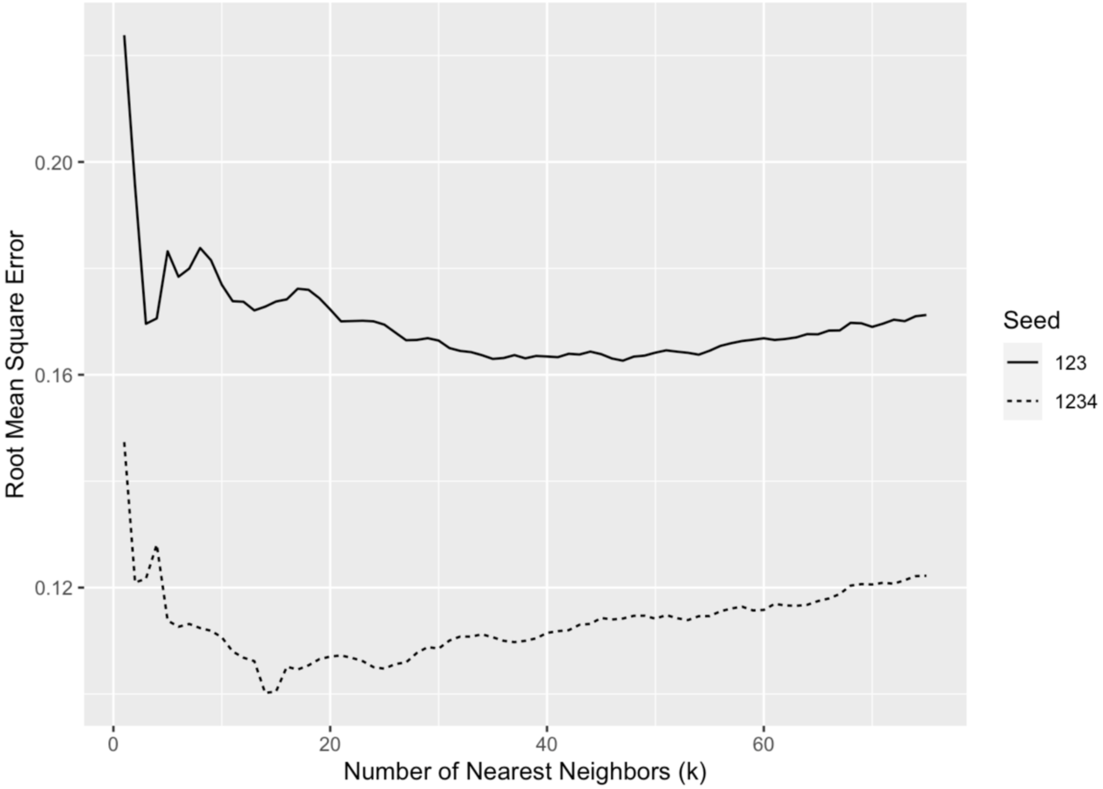
b. Create a plot showing the values of k against the RMSE values. How many nearest neighbors would you choose?



c. On small datasets, the holdout method set approach may be very sensitive to the random split. Repeat the process using a random seed of 1234. Which value of *k* would you choose?



d. Suppose you want to create a plot showing both line plot trajectories on the same plot, but use different line types based on the seed. Sketch the dataset required for this plot.


Root Mean Square Error
Number of Nearest Neighbors (k)
Seed