Homework 04

Zach Stecher

Due: 11/1/16

Problem 1) Modify the code provided to generate 1000 data points and, using 10-fold CV, report the best values of k-neighbors that yield the best CV E_{out}. Report the best E_{out}.

The only modification I needed to make to the provided code was to change the value of genDataSet(100) to genDataSet(1000). I also removed the lines that plotted the data as they became unnecessary. I modified the k-neighbors code we worked on in class to report the score and the three best k-neighbors values and ran the program multiple times. On the final run, my E_{out} score ended up being a 0.137986913985, with k-neighbors values of 216, 208 and 214. These values fluctuated slightly with each run because of the random data generator used.

Problem 2) Modify the code to repeat the experiment 100 times, saving the best thee k-neighbors values in every single trial, and plot a histogram of all the values of k that were saved.

For this problem, I wrapped the code used in problem 1 inside a for loop to run 100 times, only generating the data once at the start. I also added in an array named bestklist to save each k-neighbors value that the regressor determined to be the best for each iteration. I tried this two ways, once with the data generator outside the for-loop so that it would use the same exact data every time, and once with the data generator inside the for-loop so that each of the 100 experiments would have it's own generated data. When I kept the generator outside the for-loop, it resulted in achieving the same

three k-neighbor values 100 times each. By moving the generator into the for-loop, I ended up with a slightly more even spread, however the majority of the results still hovered around the 250-350 range:

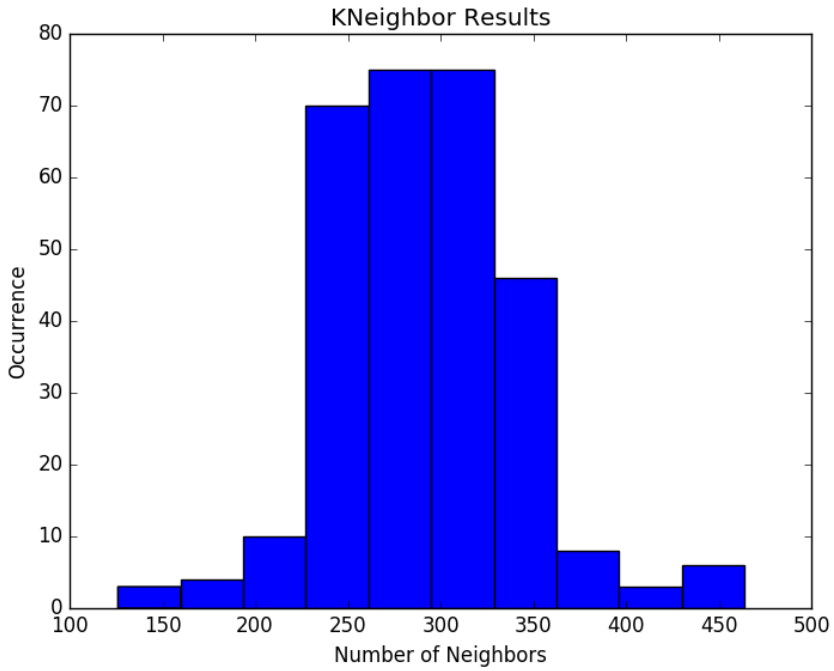

Figure 1: Histogram for k-neighbor results.