Advanced Regression

Rahul Roy and Sancharee Basak

July-December 2024

1 Problem Set 3 : KNN Smoother

1. Write the following lines of code in R.

```
set.seed(1234)
X<-runif(100,5,15)
Y<-5*sin(X)+23*cos(X)^2+rnorm(100,0,5)
Sim.1 <- data.frame(X=X,Y=Y)
reg <- function(x){
  5*sin(x)+23*cos(x)^2
}
```

   The function "reg" represents the truth (i.e., the regression function) of the simulated dataset.

2. Divide the X values into parts. Call the first 80 values the training test. Denote them by X.train. The rest of the 20 values constitute the test set. Call them X.test.

3. Divide the Y values in training set (Y.train) and test set (Y.test) in the similar way.

4. Plot the data such that the points corresponding to the training set and the test set are represented with different symbols.

5. Plot the regression function in the same graph.

6. Apply the KNN smoother algorithm on the training dataset with the following numbers of nearest neighbours (k)

   (a) k=1

   (b) k=2.

   (c) k=5.

   (d) k=10.

   (e) k=20.

   (f) k=40.

(g) k=80.

For each case, plot the estimates. Comment on your findings (i.e., how do the estimates change with changing k?)

7. For each case, compute the training and test errors.

8. Use the *knnreg* function in the *caret* library in R to compare the results with your code.

9. Repeat the procedure 50 times. Each time, the training set and the test set should be chosen randomly from the given sample. Find the average of the training errors and the testing errors thus obtained for each case and comment.

10. Consider the "Boston" dataset in the "MASS" library in R and let the variable "medv" be the response (Y) and the variable "lstat" be the predictor (X). Use 22-fold cross-validation method to find the best choice of k for applying the KNN smoother.