Advanced Regression

Rahul Roy and Sancharee Basak

July-December 2024

1 Problem Set 2 : Bin 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 bin smoother algorithm on the simulated dataset with the following number of bins (k)

   (a) k=2.

   (b) k=5.

   (c) k=10.

   (d) k=20.

   For each case, plot the bins and the estimates for each bin. Comment on your findings.

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

8. 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.

9. 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). Divide the observations into 23 strata, each of equal size. At the i-th stage, consider the i-th stratum to be the test set, and all the other 22 strata consists of the training set. Fit Bin-Smoothers of different sizes to the training set and compute the validation error and based on this find out the best model for the given dataset.