Homework Assignment 03

The Due Date : By 3:00pm, Monday, November 25^{th}.
Your solution should include R code and answer of each question.
You need to **upload** your homework on `http://plato.pusan.ac.kr`.

Open the data set `Auto` in the `R` package '`ISLR`'. The data information is available with `?Auto`. The data set includes 9 variables with $n = 392$ samples. The response is `mpg` while only 5 variables are used for predictors. Use the following `R` codes to generate the scaled predictor matrix `x` and the response `y`

```
> data(Auto)
> x <- scale(Auto[,3:7])
> y <- Auto$mpg
```

In order to answer each question, conduct the 5-fold cross validation (CV) based on the following `gr` variable.

```
> set.seed(13579)
> gr <- sample(rep(seq(5), length=length(y)))
```

For evaluating prediction performance, estimate the predicted value $\hat{y}_i$ of test sets of each fold and then compute a numerical value of R^2 (r-squared) of simple linear regression between y_i and $\hat{y}_i$ for $i = 1, \ldots, n$.

1. Apply both linear regression model (LM) and regression tree (RT) where all different combinations of predictors are considered. Since there are 5 predictors, the total number of different combinations of predictors is $2^5 - 1 = 31$, excluding an empty set. Find the largest R^2 among 31 models for each of LM and RT. Use an `R` function `tree(...)` to fit regression tree.

2. Apply a bagging (BG) method where all different combinations of predictors are considered like Q1. So, there is a total of 31 models based on 5 predictors. Use the following `R` codes to generate the bootstrap sequence for each fold.

   ```
   > set.seed(111)
   > boot <- vector(mode="list", length=max(gr))
   > for (k in 1:max(gr)) {
        mat <- matrix(1:sum(gr!=k), sum(gr!=k), 500)
        mat <- apply(mat, 2, function(t) sample(t, replace=TRUE))
        boot[[k]] <- mat
    }
   ```

   Each fold includes an index matrix where the number of rows is the same as the number of training set for the corresponding fold and the number of columns means the number of tress. Note that the bootstrap sequence does not represent the sample index but the index of the training set for each fold. Take an average of $\hat{y}_i$ over 500 trees and then find the largest R^2 among 31 models.

3. Apply local regression with the weight function

$$K_{i0}(\lambda) = D\left(\frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_i\|_2}{\lambda}\right),$$

where $\lambda > 0$ is a tuning parameter, $\boldsymbol{x}_0 = (x_{01}, \ldots, x_{05})^{\mathrm{T}}$ is a test sample, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{i5})^{\mathrm{T}}$ is a training sample and $\|\cdot\|_2$ is a l_2-norm. The weight function is defined as 4 different ways such as

$$D_1(t) = \frac{3}{4}(1-t^2)\mathbf{1}_{\{|t|<1\}}, \quad D_2(t) = \frac{71}{80}(1-t^3)^3\mathbf{1}_{\{|t|<1\}}, \quad D_3(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}, \quad D_4(t) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}t\right)\mathbf{1}_{\{|t|<1\}}.$$

Let us denote 4 local regression (LR) models with the different weight functions by LR_1, LR_2, LR_3, and LR_4, respectively. Use an `R` function `lm(..., weights=...)` to fit a weighted least squares regression. Set the tuning parameter λ to be equally spaced between $[\lambda_{\min}, \lambda_{\max}]$ with 30 different λ values, where $\lambda_{\min}$ is the 10th percentile of $\|\boldsymbol{x}_0 - \boldsymbol{x}_i\|_2$ for all training samples and $\lambda_{\max} = \max_i \|\boldsymbol{x}_0 - \boldsymbol{x}_i\|_2$. So, $[\lambda_{\min}, \lambda_{\max}]$ depends on a test sample $\boldsymbol{x}_0$. For each local regression model, find the largest R^2 among 30 different λ values.

4. Similar to the K-nearest neighbor (KNN) method, this time we assign the same weight for the γ nearest samples $\boldsymbol{x}_i$, so the weight function is

$$K_{i0}(\gamma) = \frac{1}{\gamma}\mathbf{1}_{\{i \in \mathcal{N}_\gamma(\boldsymbol{x}_0)\}},$$

where $\mathcal{N}_\gamma(\boldsymbol{x}_0)$ represents the γ training samples that are closest to a test sample $\boldsymbol{x}_0$. The distance to find nearest training samples is measured by the l_2-norm, i.e., Euclidean distance. Let us denote the local regression with the same weight by LR_5. Consider 40 different γ values such that $41 \le \gamma \le 80$. Find the largest R^2 among 40 different γ values.

5. Let us re-define the weight function as

$$K_{i0}(q, \gamma) = \begin{cases} D\left(\dfrac{\|\boldsymbol{x}_0 - \boldsymbol{x}_i\|_q}{\lambda(q,\gamma)}\right) & \text{if } i \in \mathcal{N}_\gamma(\boldsymbol{x}_0) \\ 0 & \text{if } i \notin \mathcal{N}_\gamma(\boldsymbol{x}_0), \end{cases}$$

where $\lambda(q, \gamma) = \max_{i \in \mathcal{N}_\gamma(\boldsymbol{x}_0)} \|\boldsymbol{x}_0 - \boldsymbol{x}_i\|_q$, $41 \le \gamma \le 80$, and $D(\cdot)$ is the same four functions used in Q3. Also, $\|\cdot\|_q$ is a l_q-norm, i.e., $\|z\|_q = \left(\sum_{i=1}^n |z_i|^q\right)^{1/q}$ for $q \in \{1, 2, 3, 4, 5\}$. Note that a weighted least squares regression for each $\boldsymbol{x}_0$ should be applied with only γ training samples since the weights are all equal to 0 for $i \notin \mathcal{N}_\gamma(\boldsymbol{x}_0)$. Let us denote 4 local regression models using the re-defined weight by LR_6, LR_7, LR_8, and LR_9, respectively. There are 5 different q values and 40 different γ values so a total of 200 different combinations of (q, γ). For each local regression model, find the largest R^2 among 200 different (q, γ) values.

6. We have considered a total of 12 different models including LM, RT, BG and 9 local regression models ($\mathrm{LR}_1 - \mathrm{LR}_9$). Finally, repeat from Ql to Q5 10 times based on a replication set `gr10` generated by

```
> set.seed(54321)
> gr10 <- matrix(rep(seq(5), length=length(y)), length(y), 10)
> gr10 <- apply(gr10, 2, sample)
```

Each column of `gr10` represents CV ID for the 5-fold CV. For each model, take an average of R^2 over 10 replications. Who is winner?