Homework Assignment 01

The Due Date : By 3:00pm, Monday, October 7^{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 `Boston` in the `R` package `MASS`. The data information is available with `?Boston`. It has a total of 506 observations with 14 variables, where the variable `crim` is considered as a response and only 11 variables are considered as predictors. We exclude two integer variables. Also, we scale all predictors such that each predictor has a mean of 0 and a standard deviation of 1. So, you can make the predictor `x` and the response `y` using the following `R` codes

```
> data(Boston)
> y <- Boston[, 1]
> x <- Boston[, -c(1, 4, 9)]
> x <- as.matrix(scale(x))
```

1. We first define 3 different distance functions. For the p-dimensional predictor, the distance between the i-th training observation $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ and the test observation $\boldsymbol{x}_0 = (x_{01}, \ldots, x_{0p})^{\mathrm{T}}$ can be measured by three different vector distances such that

$$d_1(\boldsymbol{x}_i, \boldsymbol{x}_0) = \sum_{j=1}^{p} |x_{ij} - x_{0j}|, \quad d_2(\boldsymbol{x}_i, \boldsymbol{x}_0) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{0j})^2}, \quad \text{and} \quad d_3(\boldsymbol{x}_i, \boldsymbol{x}_0) = \sum_{j=1}^{p} \frac{|x_{ij} - x_{0j}|}{|x_{ij}| + |x_{0j}|}$$

The predicted value of the test observation ($\boldsymbol{x}_0$) is computed by

$$\hat{f}_{l,K}(\boldsymbol{x}_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_K(\boldsymbol{x}_0; d_l)} y_i,$$

where $\mathcal{N}_K(\boldsymbol{x}_0; d_l)$ represents the K training observations that are closest to a test observation $\boldsymbol{x}_0$ in terms of the distance measured by d_l for $l = 1, 2$ and 3. Suppose that the first observation is a test sample ($\boldsymbol{x}_0$) and the other 505 observations are a training set. Compute $\hat{f}_{1,K}(\boldsymbol{x}_0)$, $\hat{f}_{2,K}(\boldsymbol{x}_0)$ and $\hat{f}_{3,K}(\boldsymbol{x}_0)$ when $K = 10$.

2. The prediction error (PE) of the m test observations $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)$ is defined as

$$\mathrm{PE}_{l,K} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(y_i - \hat{f}_{l,K}(\boldsymbol{x}_i) \right)^2},$$

for $l = 1, 2, 3$ and $K = 1, 2, \ldots, 100$. Use the following `R` codes to randomly select a 400 training set.

```
> set.seed(12345)
> tran <- sample(nrow(x), 400)
```

The other 106 observations are considered as a test set. For each l, find the optimal value of K that minimizes $\mathrm{PE}_{l,K}$ among $K = 1, 2, \ldots, 100$. Also, include the minimum value of the prediction error for each l. Finally, provide a plot including 3 lines, where the x-axis is K, the y-axis is PE, and each line represents a distance function.

3. Perform 10-fold cross-validation (CV) to compute the smallest test error among $K = 1, \ldots, K$. Use the following R codes to generate the fold ID.

```
> set.seed(1234)
> foldID <- sample(rep(1:10, length=nrow(x)))
```

Let us denote PE for the k-th fold by $\text{PE}_{l,K}(C_k)$. Then, CV error (CVE) based on 10-fold CV is computed by

$$\text{CVE}_{l,K} = \sqrt{\frac{1}{506}\sum_{k=1}^{10} m_k \left(\text{PE}_{l,K}(C_k)\right)^2},$$

where m_k is the number of test observations in the k-th fold. For each l, find the smallest value of CVE and the corresponding optimal value of K. Provide a CVE plot over K, including 3 lines just like Q2.

4. The predicted value of the test observation ($\boldsymbol{x}_0$) is re-defined by

$$\hat{g}_{l,K}(\boldsymbol{x}_0) = \text{median}\,(y_i) \;\text{ for } i \in \mathcal{N}_K(\boldsymbol{x}_0; d_l).$$

In R, use the `median(...)` function to compute the median value. Repeat Q3 based on the predict function $\hat{g}_{l,K}$. Find the smallest value of CVE and the corresponding optimal value of K. Provide a CVE plot over K, including 3 lines just like Q3.

5. The predicted value of the test observation ($\boldsymbol{x}_0$) is re-defined by

$$\hat{h}_{l,K}(\boldsymbol{x}_0) = \frac{1}{D_{l,K}} \sum_{i \in \mathcal{N}_K(\boldsymbol{x}_0; d_l)} \delta_{i,l} y_i,$$

where

$$D_{l,K} = \sum_{i \in \mathcal{N}_K(\boldsymbol{x}_0; d_l)} \delta_{i,l} \quad \text{and} \quad \delta_{i,l} = \exp\left(-\left(d_l(\boldsymbol{x}_i, \boldsymbol{x}_0) - \min_{i \in \mathcal{N}_K(\boldsymbol{x}_0; d_l)} d_l(\boldsymbol{x}_i, \boldsymbol{x}_0)\right)^2\right).$$

Repeat Q3 based on the predict function $\hat{h}_{l,K}$. Find the smallest value of CVE and the corresponding optimal value of K. Provide a CVE plot over K, including 3 lines just like Q3.

6. In order to predict the value of `crim`, we have considered 9 statistical models which consist of 3 different predict functions ($\hat{f}$, $\hat{g}$ and $\hat{h}$) and 3 different distance functions (d_1, d_2 and d_3). Based on Q3, Q4 and Q5, summarize your results in the following table;

	$\hat{f}$		$\hat{g}$		$\hat{h}$	
	K	CVE	K	CVE	K	CVE
d_1						
d_2						
d_3						

Who is winner?