Homework Assignment 02

The Due Date : By 1:30pm, November, 8^{th} (Tuesday)
Your solution should include R codes and the answer of each question.
You need to upload your homework on `http://plato.pusan.ac.kr` for full credits.
You may collaborate on this problem but you must write up your own solution.

Open the data set `OJ` in the `R` package `ISLR`. The data information is available with `?OJ`. Let us begin with the following `R` commands

```
> data(OJ)
> y <- OJ[, 1]
> x <- scale(OJ[, -c(1,11:14,17)])
```

A matrix `x` consists of $n = 1,070$ and $p = 12$, and a binary response `y` has either `CH` or `MM`. Let us regard `CH` as "negative" and `MM` as "positive." Next, randomly generate training, validation and tests samples, using the following `R` commands.

```
> set.seed(1111)
> M <- sample(rep(c(-1, 0, 1), c(600, 370, 100)))
```

The vector `M` consists of 600 training samples (`-1`), 370 validation samples (`0`) and 100 test samples (`1`). In order to assess classification performance, consider 3 different scores which are accuracy (ACC), F_1 score and Matthews correlation coefficient (MCC). They are

$$\text{ACC} = \frac{TP + TN}{TP + FP + TN + FN}, \qquad F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

and

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}},$$

respectively. Note that MCC = 0 if the denominator is equal to 0.

1. Apply a logistic regression (LR) for the training samples and then predict the class labels of the validation samples, where the prediction probability of '$y =$ `MM`'

   $$P(y = \texttt{MM}|x) > c$$

   indicates $\hat{y} =$ `MM`; otherwise $\hat{y} =$ `CH`. The threshold c starts from 0 to 1 increased by 0.001. Based on the validation samples, find 3 optimal thresholds $\hat{c}_1$, $\hat{c}_2$ and $\hat{c}_3$ that maximize ACC, F_1 and MCC, respectively. If multiple thresholds have the same largest score, the optimal threshold should be the average of the multiple thresholds. Provide a single plot with 3 lines representing ACC, F_1 and MCC, respectively. In the plot, the thresholds are on the x-axis and the scores are on the y-axis. Also, include the numerical values of $\hat{c}_1$, $\hat{c}_2$ and $\hat{c}_3$.

2. With $\hat{c}_1$, $\hat{c}_2$ and $\hat{c}_3$ obtained by Q1, find ACC, F_1 and MCC of the test samples. LR should be applied to compute ACC of the test samples with $\hat{c}_1$, F_1 score of the test samples with $\hat{c}_2$, and MCC of the test samples with $\hat{c}_3$.

3. Repeat Q1 and Q2 with linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and naive Bayes (NB) classification methods. Note that the prediction probability is equivalent to the posterior probability of 3 methods. You don't need to provide a line plot and the optimal thresholds here. For each classification method, just find the ACC, F_1 score and MCC of the test samples.

4. Repeat Q1 and Q2 with a K-nearest neighbor (KNN) classification method, where $K = 1, 3, 5, \ldots, 197, 199$. First, find the optimal K values that maximize ACC, F_1 score and MCC of the validation samples, respectively. If multiple K values have the same largest score, the optimal K should be the smallest one among them. Provide a single plot with 3 lines representing ACC, F_1 and MCC, respectively. In the plot, the values of K are on the x-axis and the scores are on the y-axis. Finally, find ACC, F_1 and MCC of the test samples, using the corresponding optimal thresholds.

5. Next, randomly generate training, validation and test samples 100 times, using the following `R` commands.

```
> set.seed(1234)
> M <- rep(c(-1, 0, 1), c(600, 370, 100))
> M <- apply(matrix(M, length(M), 100), 2, sample)
```

For each column of the matrix `M`, $1,070$ samples consist of 600 training samples (`-1`), 370 validation samples (`0`) and 100 test samples (`1`). Since we have 100 different training, validation and test samples, you need to compute ACC, F_1 score and MCC of test sets 100 times. That is to say, you have to repeat Q1-Q4 for each set, where 5 classification methods such as LR, LDA, QDA, NB and KNN should be applied. Note that the optimal threshold or K can be determined using the validation set. For each method, find the average ACC, average F_1 score and average MCC of the test samples over 100 different sets. Summarize your answer using the following table

(test)	LR	LDA	QDA	NB	KNN
ACC					
F_1					
MCC					

Which method is a winner?