Homework Assignment 02

The Due Date : By 3:00pm, Monday, November 4^{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 `College` in the `R` package '`ISLR`'. The data information is available with `?College`. The response is `Private` while the other 17 variables are predictors. Use the following `R` codes to generate the scaled predictor matrix `x` and the response `y`

```
> data(College)
> x <- scale(College[,-1])
> y <- College$Private
```

Next, randomly generate a training set (`tran`), a validation set (`vald`) and a test set (`test`) such that

```
> set.seed(12345)
> set <- sample(rep(c("tran", "vald", "test"), length=nrow(x)))
```

Note that the training set is for building a classification model, the validation set is for finding the best model, and the test set is for evaluating classification performance. Let us define two numerical measurements to evaluate classification perpermance such as

$$D_1 = -\frac{1}{2m}\sum_{i=1}^{m}\left[I(y_i = \text{“Yes”})\log(\hat{p}_i) + I(y_i = \text{“No”})\log(1-\hat{p}_i)\right].$$

and

$$D_2 = \frac{1}{m_0(m-m_0)}\sum_{i=1}^{m_0}\sum_{j=m_0+1}^{m}\left[I(\hat{p}_i > \hat{p}_j) + \frac{1}{2}I(\hat{p}_i = \hat{p}_j)\right],$$

where $\hat{p}_i$ is the i-th prediction probability to estimate $P(y_i = \text{“Yes”}|x_i)$ and m is the total number of the test (validation) set observations. Also, assume that the first m_0 observations of the test (validation) sets have a label “No”, and the other $m - m_0$ observations have a label “Yes”. The smaller D_1 and D_2, the better classification performance. For valid computation, limt the prediction probability to have the lower and upper bounds such that

$$10^{-10} \le \hat{p}_i \le 1 - 10^{-10} \quad \text{for} \quad i = 1, \ldots, m.$$

1. Build 4 classifiers including logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and naive Bayes (NB) from the training set. For each method, compute both D_1 and D_2 of the test set. Note that the validation set is not necessary to answer this question.

2. Estimate the ridge (RG) regression coefficients of the logistic model from the training set, where the following 100 different λ values should be used

   ```
   > lam1 <- 10^seq(-3, -1.5, length.out=100)
   ```

   Next, compute D_1 and D_2 of the validation set for each λ value and then find the optimal $\hat{\lambda}_{D_1}$ and $\hat{\lambda}_{D_2}$ that minimize D_1 and D_2, respectively. Let us denote two best models by RG_1 with $\hat{\lambda}_{D_1}$ and RG_2 with $\hat{\lambda}_{D_2}$. Finally, compute D_1 and D_2 of the test set using each of two best models (RG_1 and RG_2). When you compute the prediction probability of the ridge regression, use `predict(..., type="response")`.

3. Replace the ridge regression by lasso (LS) and repeat Q2, using the following 100 different λ values

```
> lam2 <- 10^seq(-0.5, -3.5, length.out=100)
```

Let us denote the best models by LS_1 and LS_2 instead of RG_1 and RG_2, respectively. Finally, compute D_1 and D_2 of the test set using each of two best models (LS_1 and LS_2).

4. Find the nonzero coefficients of two best models (LS_1 and LS_2) in Q3. Next, select the corresponding predictors with the nonzero coefficients. Then, build 4 classifiers used in Q1 with only selected predictors. Let us denote 4 models based on LS_1 by LR_1, LDA_1, QDA_1 and NB_1, respectively, Also, denote another 4 models based on LS_2 by LR_2, LDA_2, QDA_2 and NB_2, respectively, Finally, compute D_1 and D_2 of the test set using each of these 8 models.

5. Build a KNN classifier from the training set with $k = 1, 2, \ldots, 100$. Next, use the validation set to find two optimal values of $\hat{k}_1$ and $\hat{k}_2$ that minimize D_1 and D_2, respectively. If more than one k value has the minimum value, take the smallest k value as the optimal value. Let us denote two best models by KNN_1 with $\hat{k}_1$ and KNN_2 with $\hat{k}_2$, Finally, compute D_1 and D_2 of the test set using each of two best models (KNN_1 and KNN_2). Use an R function `knn3Train(...)` in the package '`caret`' to compute the prediction probability.

6. In Q1 – Q5, both D_1 and D_2 of the single test set have been computed. In this question, repeat Q1 – Q5 for 100 different test sets. First, randomly generate 100 different training, validation and test sets such that

```
> set.seed(111222)
> set100 <- matrix(rep(c("tran", "vald", "test"), length=nrow(x)), nrow(x), 100)
> set100 <- apply(set100, 2, sample)
```

Each column of the matrix `set100` represents a replication set. For each replication set, compute D_1 and D_2 of the test set just like Q1 – Q5. There are a total of 18 classification models including 4 models (LR, LDA, QDA and NB) in Q1, 2 models (RG_1 and RG_2) in Q2, 2 models (LS_1 and LS_2) in Q3, 8 models (LR_1, LDA_1, QDA_1, NB_1, LR_2, LDA_2, QDA_2 and NB_2) in Q4, and 2 models (KNN_1 and KNN_2) in Q5. For each model, compute the average values of D_1 and D_2 over 100 replication sets. Finally, summarize your result using the following table

	Model	D_1	D_2
1	LR		
2	LDA		
3	QDA		
⋮	⋮		
16	NB_2		
17	KNN_1		
18	KNN_2		

Who is winner in terms of D_1 and D_2, respectively?