REPORT

수강과목	:	빅데이터통계분석
담당교수	:	선호근
학 과	:	통계학과
학 번	:	201611531
이 름	:	정호재
제출일자	:	2020.11.05.

Homework Assignment 03

The Due Date : By Thursday, November, 5th in class
Your solution should include R codes and the answer of each question.
You need to upload your R codes 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 `College` in the `R` package '`ISLR`'. The data information is available with `?College`. All of 17 variables except the response variable `Private` are used for predictors. For the training set, randomly generate 543 observations such that

```
> RNGkind(sample.kind = "Rounding")
> set.seed(1234)
> tran <- sample(dim(College)[1], floor(dim(College)[1]*0.7))
```

The other 234 observations are regarded as the test set.

```
> str(College)
'data.frame':    777 obs. of  18 variables:
 $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Apps       : num  1660 2186 1428 417 193 ...
 $ Accept     : num  1232 1924 1097 349 146 ...
 $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
 $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
 $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
 $ F.Undergrad: num  2885 2683 1036 510 249 ...
 $ P.Undergrad: num  537 1227 99 63 869 ...
 $ Outstate   : num  7440 12280 11250 12960 7560 ...
 $ Room.Board : num  3300 6450 3750 5450 4120 ...
 $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
 $ Personal   : num  2200 1500 1165 875 1500 ...
 $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
 $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
 $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
 $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
 $ Expend     : num  7041 10527 8735 19016 10922 ...
 $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

```
> dim(x)
[1] 777  17
```

1. Build 4 classifiers such as LR (Logistic Regression), LDA(Linear Discriminant Analysis), QDA(Quadratic Discriminant Analysis) and NB(Naive Bayes) based on the training set. Compute the misclassification rate of the test set. Use the threshold of 0.5 for LR.

```
# LR
> table(pred1, test)
      test
pred1  No Yes
  No   50   7
  Yes   3 174
> mean(pred1!=test)
[1] 0.04273504
```

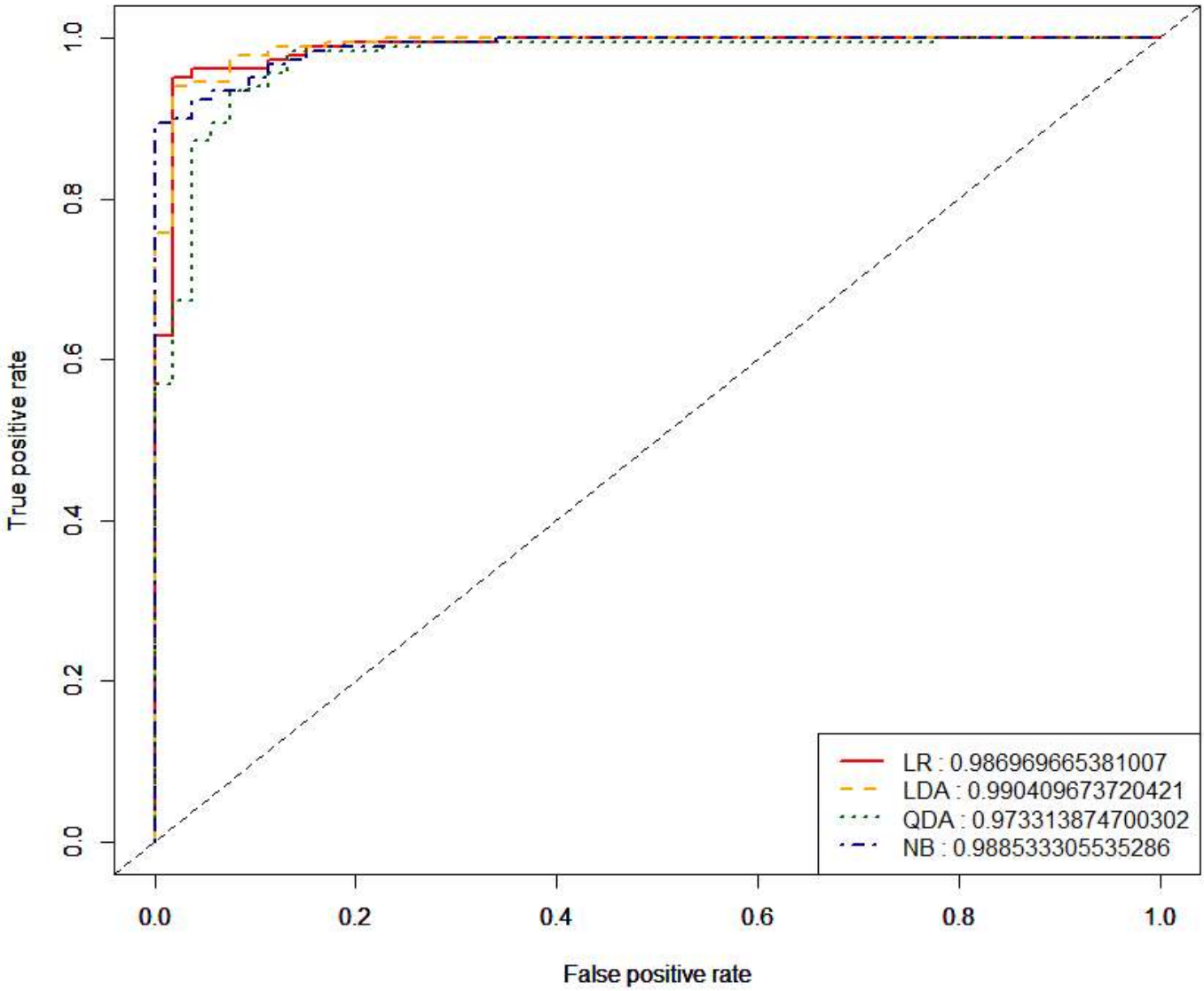
```
#LDA
> table(pred2, test)
      test
pred2  No Yes
  No   49   4
  Yes   4 177
> mean(pred2!=test)
[1] 0.03418803
```

```
#QDA
> table(pred3, test)
      test
pred3  No Yes
  No   43   3
  Yes  10 178
> mean(pred3!=test)
[1] 0.05555556
```

```
#NB
> table(pred4, test)
      test
pred4  No Yes
  No   44   3
  Yes   9 178
> mean(pred4!=test)
[1] 0.05128205
```

각각의 오분류율을 구하면 LR에서는 오분류율이 0.04273504, LDA는 0.03418803, QDA는 0.05555556, NB는 0.05128205으로 구해졌다.

2. Using 4 different classifiers built in Q1, draw ROC (Receiver Operating Characteristic) curves for the test set. You can use the package 'ROCR'. All of 4 curves should be displayed in the same plot so that we can directly compare their prediction performance for the test set. Also, compute AUC (Area Under the Curve) of 4 classifiers for the test set.




```
> cbind(AUC1,AUC2,AUC3,AUC4)
      AUC1      AUC2      AUC3      AUC4
[1,] 0.9869697 0.9904097 0.9733139 0.9885333
```

ROC 곡선을 비교해보았을 때 LDA일 때 가장 좌측상단에 가깝게 그려져 가장 좋은 성능을 보이는 것 같다. 또한 AUC값을 비교했을 때도 LDA에서 AUC값이 가장 높은 값을 가진다. 따라서 4가지 모델 중 LDA에서 성능이 가장 좋아보인다.

3. Separate 543 training observations into 10 folds such that

```
> RNGkind(sample.kind = "Rounding")
> set.seed(12345)
> N.lab <- sample(rep(seq(10), length=sum(College$Private[tran]=="No")))
> Y.lab <- sample(rep(seq(10), length=sum(College$Private[tran]=="Yes")))
> gr <- rep(0, length(tran))
> gr[College$Private[tran]=="No"] <- N.lab
> gr[College$Private[tran]=="Yes"] <- Y.lab
```

Perform 10-fold cross validation (CV) based on **gr** to find the optimal threshold of the prediction probability each classifier. For example, we can classify the jth observation as "Yes" if

$$P(y_j = \text{'Yes'}|x_j) > \alpha,$$

where $\alpha \in [0, 1]$ is a threshold. Let α starts from 0 to 1 increased by 0.01. The optimal threshold minimizes the misclassification rate of the 10-fold CV. If the minimum of the misclassification rate is achieved at multiple α values, the optimal α should be computed as the mean of the α values that minimize the misclassification rate. Find the optimal threshold for each classifier and then apply the optimal threshold to compute the misclassification rate of the test set. Provide the misclassification rate of the test set along with the optimal threshold for each classifier.

```
# LR
> min(misclass)
[1] 0.07758465
> which(misclass==min(misclass))
[1] 52
> thre[which(misclass==min(misclass))]
[1] 0.51
> table(pred1, test)
      test
pred1  No Yes
  No   50   7
  Yes   3 174
> mean(pred1!=test)
[1] 0.04273504
```

```
# LDA
> min(misclass)
[1] 0.07562925
> which(misclass==min(misclass))
[1] 63 67
> thre[which(misclass==min(misclass))]
[1] 0.62 0.66
> a <- mean(thre[which(misclass==min(misclass))])
> a
[1] 0.64
> table(pred2, test)
      test
pred2  No Yes
  No   49   8
  Yes   4 173
> mean(pred2!=test)
[1] 0.05128205
```

```
# QDA
> min(misclass)
[1] 0.2229109
> which(misclass==min(misclass))
[1] 88 92
> thre[which(misclass==min(misclass))]
[1] 0.87 0.91
> a <- mean(thre[which(misclass==min(misclass))])
> a
[1] 0.89
> table(pred3, test)
      test
pred3  No Yes
  No   47   8
  Yes   6 173
> mean(pred3!=test)
[1] 0.05982906
```

```
#NB
> min(misclass)
[1] 0.2083019
> which(misclass==min(misclass))
[1] 84 85
> thre[which(misclass==min(misclass))]
[1] 0.83 0.84
> a <- mean(thre[which(misclass==min(misclass))])
> a
[1] 0.835
> table(pred4, test)
      test
pred4  No Yes
  No   45   5
  Yes   8 176
> mean(pred4!=test)
[1] 0.05555556
```

10-fold CV에 의해서 LR에서는 임계값이 0.51, LDA는 0.62, 0.66, QDA는 0.87, 0.91, NB는 0.83, 0.84일 때 최소인 오분류율을 가졌다.
최적 α는 오분류율을 최소화하는 α 값의 평균으로 계산하여 LR에서는 임계값이 0.51, LDA는 0.64, QDA는 0.89, NB는 0.835일 때로 설정하였다. 그 후 train set으로 설정한 모델에 test set을 적용시켜 각각의 오분류율을 구하면 LR에서는 오분류율이 0.04273504, LDA는 0.05128205, QDA는 0.05982906, NB는 0.05555556으로 구해졌다.

4. Let us define Matthews correlation coefficient (MCC) by

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

where TP/TN means true positives/negatives and FP/FN means false positives/negatives. Note that MCC=0 if the denominator of MCC is 0. Repeat Q3 except that the optimal threshold maximizes MCC of the 10-fold CV. If the maximum of MCC is achieved at multiple α values, the optimal α should be computed as the mean of the α values that maximize MCC. Find the optimal threshold for each classifier and then apply the optimal threshold to compute MCC of the test set. Provide MCC of the test set along with the optimal threshold for each classifier.

```
#LR
> max(MCC_)
[1] 0.8355096
> which(MCC_==max(MCC_))
[1] 52
> thre[which(MCC_==max(MCC_))]
[1] 0.51
> ((RES[1]*RES[2])-(RES[3]*RES[4]))/sqrt(
+
(RES[1]+RES[3])*(RES[1]+RES[4])*(RES[2]+RES[3])*(RE
S[2]+RES[4]))
[1] 0.8822027
```

```
#LDA
> max(MCC_)
[1] 0.8448808
> which(MCC_==max(MCC_))
[1] 67
> thre[which(MCC_==max(MCC_))]
[1] 0.66
> ((RES[1]*RES[2])-(RES[3]*RES[4]))/sqrt(
+
(RES[1]+RES[3])*(RES[1]+RES[4])*(RES[2]+RES[3])*(RE
S[2]+RES[4]))
[1] 0.8584171
```

```
#QDA
> max(MCC_)
[1] 0.7547158
> which(MCC_==max(MCC_))
[1] 92
> thre[which(MCC_==max(MCC_))]
[1] 0.91
> ((RES[1]*RES[2])-(RES[3]*RES[4]))/sqrt(
+
(RES[1]+RES[3])*(RES[1]+RES[4])*(RES[2]+RES[3])*(RE
S[2]+RES[4]))
[1] 0.8317398
```

```
#NB
> max(MCC_)
[1] 0.7777142
> which(MCC_==max(MCC_))
[1] 84 85
> thre[which(MCC_==max(MCC_))]
[1] 0.83 0.84
> a <- mean(thre[which(MCC_==max(MCC_))])
> a
[1] 0.835
> ((RES[1]*RES[2])-(RES[3]*RES[4]))/sqrt(
+
(RES[1]+RES[3])*(RES[1]+RES[4])*(RES[2]+RES[3])*(RE
S[2]+RES[4]))
[1] 0.8387935
```

10-fold CV에 의해서 LR에서는 임계값이 0.51, LDA는 0.66, QDA는 0.91, NB는 0.83, 0.84일 때 최대인 MCC를 가졌다.
최적 α는 MCC를 최대화하는 α 값의 평균으로 계산하여 LR에서는 임계값이 0.51, LDA는 0.66, QDA는 0.91, NB는 0.835일 때로 설정하였다. 그 후 train set으로 설정한 모델에 test set을 적용시켜 각각의 MCC를 구하면 LR에서는 MCC가 0.8822027, LDA는 0.8584171, QDA는 0.8317398, NB는 0.8387935으로 구해졌다.

5. Let us define F1 score (F_1) by

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Repeat Q3 except that the optimal threshold maximizes F_1 of the 10-fold CV. If the maximum of F_1 is achieved at multiple α values, the optimal α should be computed as the mean of the α values that maximize F_1. Find the optimal threshold for each classifier and then apply the optimal threshold to compute F_1 of the test set. Provide F_1 of the test set along with the optimal threshold for each classifier.

```
# LR
> max(F1_)
[1] 0.9518359
> which(F1_==max(F1_))
[1] 52
> thre[which(F1_==max(F1_))]
[1] 0.51
> a <- mean(thre[which(F1_==max(F1_))])
> a
[1] 0.51
> (2*RES[1])/(2*RES[1]+RES[3]+RES[4])
[1] 0.972067
```

```
# LDA
> max(F1_)
[1] 0.952747
> which(F1_==max(F1_))
[1] 63
> thre[which(F1_==max(F1_))]
[1] 0.62
> a <- mean(thre[which(F1_==max(F1_))])
> a
[1] 0.62
> (2*RES[1])/(2*RES[1]+RES[3]+RES[4])
[1] 0.9693593
```

```
#QDA
> max(F1_)
[1] 0.9287476
> which(F1_==max(F1_))
[1] 75 76
> thre[which(F1_==max(F1_))]
[1] 0.74 0.75
> a <- mean(thre[which(F1_==max(F1_))])
> a
[1] 0.745
> (2*RES[1])/(2*RES[1]+RES[3]+RES[4])
[1] 0.9726776
```

```
#NB
> max(F1_)
[1] 0.935088
> which(F1_==max(F1_))
[1] 84 85
> thre[which(F1_==max(F1_))]
[1] 0.83 0.84
> a <- mean(thre[which(F1_==max(F1_))])
> a
[1] 0.835
> (2*RES[1])/(2*RES[1]+RES[3]+RES[4])
[1] 0.9643836
```

10-fold CV에 의해서 LR에서는 임계값이 0.51, LDA는 0.62, QDA는 0.74, 0.75, NB는 0.83, 0.84일 때 최대인 F1을 가졌다.
최적 α는 F1를 최대화하는 α 값의 평균으로 계산하여 LR에서는 임계값이 0.51, LDA는 0.62, QDA는 0.745, NB는 0.835일 때로 설정하였다. 그 후 train set으로 설정한 모델에 test set을 적용시켜 각각의 F1을 구하면 LR에서는 F1이 0.972067, LDA는 0.9693593, QDA는 0.9726776, NB는 0.9643836으로 구해졌다.

6. Summarize your result using the following tables

The optimal $\hat{\alpha}$

	LR	LDA	QDA	NB
Q3				
Q4				
Q5				

Predition Performance

	LR	LDA	QDA	NB	**Winner**
Q1 (Mis)					
Q2 (AUC)					
Q3 (Mis)					
Q4 (MCC)					
Q5 (F_1)					

Note that Mis stands for the misclassification rate. For each question, the winner of 4 classifiers should have the smallest misclassification rate, the largest AUC, the largest MCC, or the largest F_1.

1번부터 5번까지의 문제를 요약하면 다음과 같다.

the optimal alpha_hat

	LR	LDA	QDA	NB
Q3	0.51	0.64	0.89	0.835
Q4	0.51	0.66	0.91	0.835
Q5	0.51	0.62	0.745	0.835

Prediction Performance

	LR	LDA	QDA	NB	Winner
Q1(Mis)	0.04273504	0.03418803	0.05555556	0.05128205	LDA
Q2(AUC)	0.9869697	0.9904097	0.9733139	0.9885333	LDA
Q3(Mis)	0.04273504	0.05128205	0.05982906	0.05555556	LR
Q4(MCC)	0.8822027	0.8584171	0.8317398	0.8387935	LR
Q5(F1)	0.972067	0.9693593	0.9726776	0.9643836	QDA