Data Mining HW 2

202055364 황 성 윤

Exercises for Classification

1. 주식에 대한 배당금을 지급할 지 여부('Yes', 'No')를 예측하기 위해, 작년 한 해 동안의 수익률을 설명변수 X로 사용하고자 한다. 많은 수의 회사를 조사해보니 배당금을 지급한 회사의 주식 수익률의 평균은 $\overline{X}=10$이고, 지급하지 않은 회사의 주식 수익률의 평균은 $\overline{X}=0$이다. 두 그룹에서 수익률의 분산은 두 그룹 모두 $\sigma^2=36$이다. X가 정규분포를 따른다고 가정하자. 어떤 회사의 작년 수익률 $X=4$였다면, 이 회사가 배당금을 지급할 확률은 얼마인가?

solve) 작년 한 해 동안의 수익률과 주식에 대한 배당금을 지급할 지의 여부에 대한 확률변수를 각각 X와 Y로 설정하자. 이 때 문제에서 주어진 조건들은 다음과 같다.

$$P(Y='yes')=P(Y='no')=0.5$$

$$P(X=x|Y='yes')=\frac{1}{\sqrt{2\pi\times 36}}exp\left[-\frac{(x-10)^2}{2\times 36}\right]$$

$$P(X=x|Y='no')=\frac{1}{\sqrt{2\pi\times 36}}exp\left[-\frac{(x-0)^2}{2\times 36}\right]$$

따라서 Bayes theorem에 의하여 문제에서 요구하는 확률을 다음과 같이 계산할 수 있다.

$$P(Y='yes'|X=4)=\frac{P(Y='yes'\ X=4)}{P(X=4)}$$

$$=\frac{P(Y='yes')P(X=4|Y='yes')}{P(Y='yes')P(X=4|Y='yes')+P(Y='no')P(X=4|Y='no')}$$

$$=\frac{0.5\times\frac{1}{\sqrt{2\pi\times 36}}exp\left[-\frac{(4-10)^2}{2\times 36}\right]}{0.5\times\frac{1}{\sqrt{2\pi\times 36}}exp\left[-\frac{(4-10)^2}{2\times 36}\right]+0.5\times\frac{1}{\sqrt{2\pi\times 36}}exp\left[-\frac{(4-0)^2}{2\times 36}\right]}$$

$$\approx 0.4310$$

2. ‘Auto.csv’ 데이터를 이용하여 자동차의 연비가 높을지 낮을지에 대해 예측을 하고자 한다. 다음 물음에 답하여라. (NA 처리 후 분석)

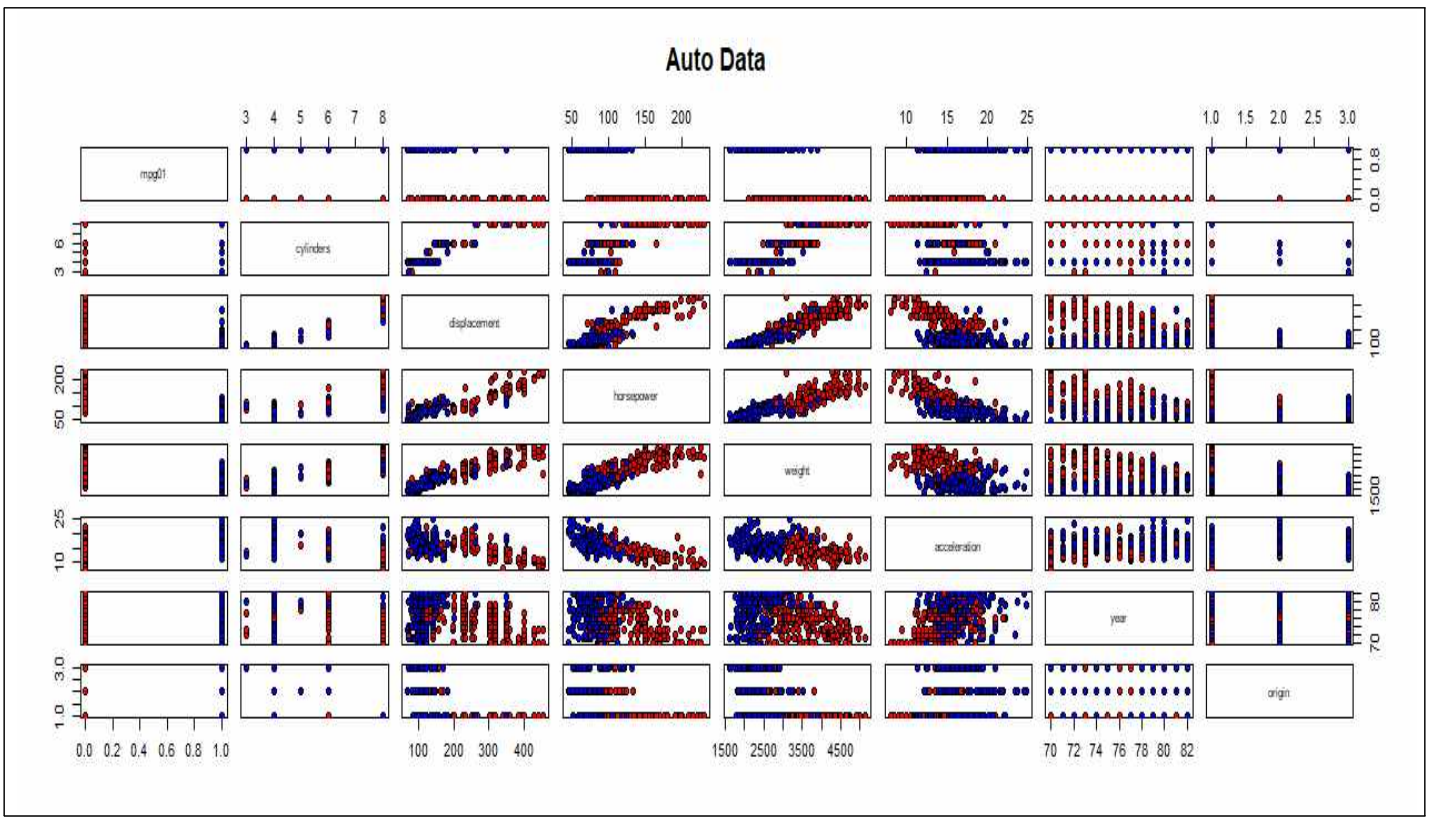
(a) mpg01 변수를 생성하여라. 이 변수는 mpg가 mpg의 중앙값보다 크면 1의 값을 갖고 아니면 0의 값을 갖는 변수이다. (중앙값 : median())
solve) 다음과 같은 R code를 통하여 Auto 데이터를 불러들이고 새로운 변수 mpg01을 생성하였다.

```
auto <- read.csv("C:/Users/HSY/Desktop/Auto.csv",sep=",",header=T,na.strings="?",stringsAsFactors=T)
auto <- na.omit(auto)
auto$mpg01 <- ifelse(auto$mpg > median(auto$mpg), 1, 0)
```

Auto 데이터에 포함된 변수는 다음과 같다.

```
mpg : miles per gallon
cylinders : Number of cylinders between 4 and 8
displacement : Engine displacement (cu. inches)
horsepower : Engine horsepower
weight : Vehicle weight (lbs.)
acceleration : Time to accelerate from 0 to 60 mph (sec.)
year : Model year (modulo 100)
origin : Origin of car (1. American, 2. European, 3. Japanese)
name : Vehicle name
```

(b) 시각화를 통하여 mpg01과 다른 변수들 사이의 관계를 확인하고 설명하여라. 어떤 변수가 mpg01을 예측하는 데 가장 유용할 것으로 생각되는가?
solve) 우선 데이터에 있는 모든 변수를 연속형이라고 가정하고 분석하도록 한다. mpg01과 다른 변수들 사이의 관계를 확인해보기 위해 산점도를 그려보면 다음과 같다.



그리고 상관계수를 계산해보면 다음과 같다.

	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg01	-0.7592	-0.7535	-0.6671	-0.7578	0.3468	0.4299	0.5137

위의 산점도에서 점의 색깔은 변수 mpg01의 값이 1인지 아니면 0인지를 나타내는 것이다. 전체적으로 상관계수까지 참고해서 봤을 때 변수 displacement, horsepower, weight 이 정도가 변수 mpg01에 대하여 어느정도 유의한 영향력을 미치는 것으로 판단되며, 이 3가지 변수의 값이 증가할수록 mpg의 값은 감소하는 경향을 보이고 있다. 그러므로 이 3가지 설명변수를 데이터 분석 시 눈여겨보면 좋을 것 같다.

(c) 데이터를 training data (60%)와 testing data(40%)로 나누어라.
solve) 우선 변수 mpg01과 origin은 연속형이 아닌 이산형이기 때문에 이 2가지 변수에 대해서는 함수 as.factor()를 통해 이산형으로 처리하였으며, 다음과 같은 R code를 통해 데이터를 임의로 train:test=6:4 의 비율로 나누었다.

```
set.seed(55364)
id_train <- sample(x=1:nrow(auto),size=round(0.6*nrow(auto),0),replace=F)
auto_train <- auto[id_train,]
auto_test <- auto[-id_train,]
```

(d) (b)에서 연관이 있다고 생각되는 변수들을 이용하여, mpg01을 예측하기 위한 LDA를 수행하여라. test 오분류율은 얼마인가?
solve) (b)에서 선택한 설명변수들을 이용하여 train data에 대해 LDA를 수행하면 다음과 같은 결과를 얻게 된다.

```
Call:
lda(mpg01 ~ displacement + horsepower + weight, data = auto_train)

Prior probabilities of groups:
        0         1
0.4808511 0.5191489

Group means:
  displacement horsepower    weight
0     274.5752   133.01770 3635.646
1     113.8361    78.81148 2323.172

Coefficients of linear discriminants:
                    LD1
displacement -0.009987186
horsepower    0.010432709
weight       -0.001175822
```

위의 결과를 통해 각 group에 대한 사전확률, mpg01의 값에 따라 3가지 설명변수들의 평균, 그리고 판별함수에 대한 계수는 어떻게 추정되었는지 확인할 수 있다. 추정된 판별함수는 다음과 같이 적을 수 있으며, 이 함수의 값이 0보다 작으면 mpg01의 예측값을 0으로, 그렇지 않으면 1로 예측하게 된다.

$$\widehat{lda} = -0.00999 \times displacement + 0.01043 \times horsepower - 0.00118 \times weight$$

그리고 이 LDA 모형을 통하여 다음과 같은 confusion matrix를 얻을 수 있다.

		Predicted class (mpg01)		Sum Total
		0	1	
True class (mpg01)	0	67	16	83
	1	6	68	74
Sum		73	84	157

그리고 이를 통해서 계산한 test 오분류율의 값은 약 0.1401이다.

(e) (b)에서 연관이 있다고 생각되는 변수들을 이용하여, mpg01을 예측하기 위한 QDA를 수행하여라. test 오분류율은 얼마인가?

solve) (b)에서 선택한 설명변수들을 이용하여 train data에 대해 QDA를 수행하면 다음과 같은 결과를 얻게 된다.

```
Call:
qda(mpg01 ~ displacement + horsepower + weight, data = auto_train)

Prior probabilities of groups:
        0         1
0.4808511 0.5191489

Group means:
  displacement horsepower    weight
0     274.5752   133.01770 3635.646
1     113.8361    78.81148 2323.172
```

위의 결과를 통해 각 group에 대한 사전확률, mpg01의 값에 따라 3가지 설명변수들의 평균을 확인할 수 있다. 이 QDA 모형을 통하여 다음과 같은 confusion matrix를 얻을 수 있다.

		Predicted class (mpg01)		Sum Total
		0	1	
True class (mpg01)	0	72	11	83
	1	7	67	74
Sum		79	78	157

그리고 이를 통해서 계산한 test 오분류율의 값은 약 0.1146이다.

(f) (b)에서 연관이 있다고 생각되는 변수들을 이용하여, mpg01을 예측하기 위한 logistic regression을 수행하여라. test 오분류율은 얼마인가?

solve) (b)에서 선택한 설명변수들을 이용하여 train data에 대해 logistic regression을 수행하면 다음과 같은 결과를 얻게 된다.

```
Call:
glm(formula = mpg01 ~ displacement + horsepower + weight, family = binomial,
    data = auto_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5932  -0.1599   0.1227   0.3329   3.4521

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  11.8663332  2.1892375   5.420 5.95e-08 ***
displacement -0.0153949  0.0080790  -1.906   0.0567 .
horsepower   -0.0314625  0.0171342  -1.836   0.0663 .
weight       -0.0021798  0.0009746  -2.237   0.0253 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 325.43  on 234  degrees of freedom
Residual deviance: 115.81  on 231  degrees of freedom
AIC: 123.81

Number of Fisher Scoring iterations: 7
```

위의 결과를 통하여 다음과 같은 logistic regression model을 적을 수 있다.

$$\log\left(\frac{\hat{p}(mpg01=1)}{1-\hat{p}(mpg01=1)}\right) = 11.8663 - 0.0154 \times displacement - 0.0315 \times horsepower - 0.0022 \times weight$$

이 logistic regression model을 통하여 다음과 같은 confusion matrix를 얻을 수 있다. 단, mpg01을 예측하기 위한 추정확률의 cut-off value는 0.5로 설정하였다.

		Predicted class (mpg01)		Sum Total
		0	1	
True class (mpg01)	0	69	14	83
	1	5	69	74
Sum		74	83	157

그리고 이를 통해서 계산한 test 오분류율의 값은 약 0.1210이다.

(g) (b)에서 연관이 있다고 생각되는 변수들을 이용하여, mpg01을 예측하기 위한 KNN을 수행하여라. KNN을 수행할 때 몇 개의 k 값을 선택하여 분석하여라. test 오분류율은 얼마인가? 어떤 k를 선택했을 때 결과가 가장 좋았는가?
solve) (b)에서 선택한 설명변수들을 이용하여 train data에 대해 $k=1,2,3$에 대하여 KNN을 수행하면 다음과 같은 결과를 얻을 수 있다.

$k=1$

confusion matrix

		Predicted class (mpg01)		Sum Total
		0	1	
True class (mpg01)	0	70	13	83
	1	13	61	74
Sum		83	74	157

misclassification rate : 0.1656

$k=2$

confusion matrix

		Predicted class (mpg01)		Sum Total
		0	1	
True class (mpg01)	0	70	13	83
	1	8	66	74
Sum		78	79	157

misclassification rate : 0.1338

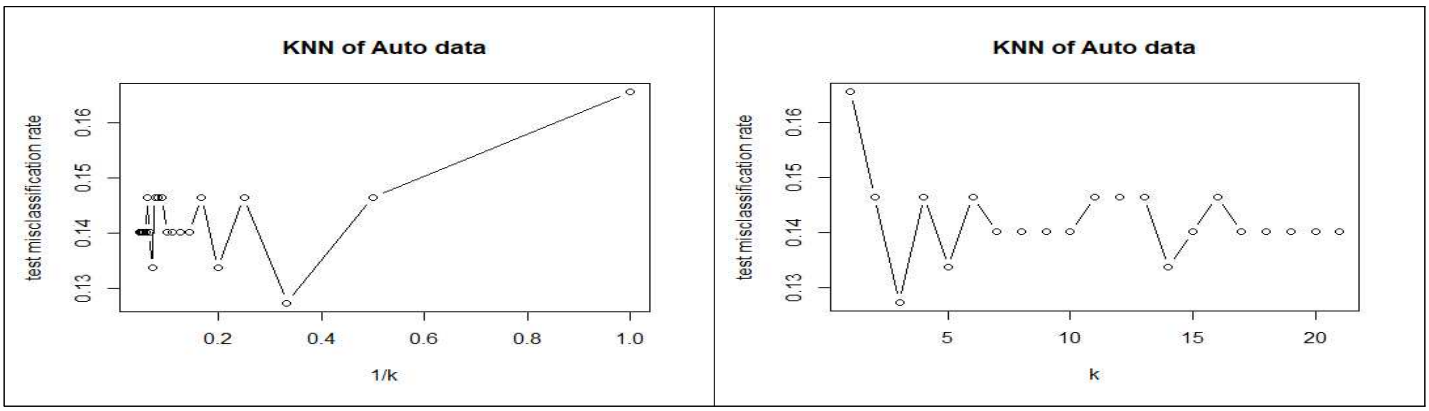
$k=3$

confusion matrix

		Predicted class (mpg01)		Sum Total
		0	1	
True class (mpg01)	0	70	13	83
	1	7	67	74
Sum		77	80	157

misclassification rate : 0.1274

위의 3가지 결과만 놓고 본다면 test 오분류율이 가장 작은 $k=3$이 가장 바람직하다고 판단할 수 있다. 하지만 이는 한정적인 결과이므로 $k=1,\cdots,21$까지 범위를 확장해서 KNN을 수행한 뒤 각각에 대하여 test 오분류율을 계산한 뒤 그래프로 표현해보면 다음과 같다.



결과적으로 이 경우에 대해서는 $k=3$일 때 KNN이 가장 좋은 성능을 보여준다는 사실을 알 수 있다.

3. ‘Boston.csv’ 데이터를 이용하여, 어떤 지역의 범죄율이 중앙값 이상인지 아닌지를 예측하기 위한 분류 모형 (logistic regression, LDA, KNN)을 적합하여라.

- 변수설명
crim : per capita crime rate by town.
zn : proportion of residential land zoned for lots over 25,000 sq.ft.
indus : proportion of non-retail business acres per town.
chas : Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
nox : nitrogen oxides concentration (parts per 10 million).
rm : average number of rooms per dwelling.
age : proportion of owner-occupied units built prior to 1940.
dis : weighted mean of distances to five Boston employment centres.
rad : index of accessibility to radial highways.
tax : full-value property-tax rate per $10,000.
ptratio : pupil-teacher ratio by town.
lstat : lower status of the population (percent).
medv : median value of owner-occupied homes in $1000s.

solve) 우선 Boston 데이터를 불러들이고 범죄율이 중앙값 이상이면 1, 그렇지 않으면 0인 변수 crim01을 추가한 뒤 임의로 training data (60%)와 testing data(40%)로 나눈다. 이를 바탕으로 각각의 분류 모형을 다음과 같이 적합할 수 있다. 설명변수는 변수 crim과 crim01을 제외한 모든 변수로 설정하였다.

1) logistic regression

모형 적합 결과

```
Call:
glm(formula = crim01 ~ ., family = binomial, data = boston_train)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.91177  -0.12678  -0.00038    0.00190    2.50735

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -37.909344   8.735688  -4.340 1.43e-05 ***
zn           -0.107159   0.059190  -1.810  0.07023 .
indus        -0.061166   0.062076  -0.985  0.32446
chas          0.633695   1.195763   0.530  0.59615
nox          48.388493  10.561233   4.582 4.61e-06 ***
rm            0.388487   1.039347   0.374  0.70857
age           0.036974   0.017594   2.101  0.03560 *
dis           0.417976   0.283781   1.473  0.14078
rad           0.735444   0.224652   3.274  0.00106 **
tax          -0.010213   0.004407  -2.318  0.02047 *
ptratio       0.205827   0.182762   1.126  0.26008
lstat         0.009348   0.058768   0.159  0.87362
medv          0.060085   0.097716   0.615  0.53862
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 420.59  on 303  degrees of freedom
Residual deviance: 121.71  on 291  degrees of freedom
AIC: 147.71

Number of Fisher Scoring iterations: 9
```

이를 바탕으로 하여 다음과 같은 결과를 얻을 수 있다.

confusion matrix

		Predicted class (crim01)		Sum Total
		0	1	
True class (crim01)	0	89	4	93
	1	14	95	109
Sum		103	99	202

test misclassification rate : 0.0891

2) LDA

모형 적합 결과

```
Call:
lda(crim01 ~ ., data = boston_train)

Prior probabilities of groups:
        0         1
0.5263158 0.4736842

Group means:
            zn     indus       chas       nox       rm      age      dis      rad
0 21.3062500   7.09225 0.04375000 0.4728044 6.389875 50.52375 5.064829  4.10625
1  0.9722222 15.54347 0.09027778 0.6438889 6.158972 87.03681 2.431185 15.52778
        tax   ptratio     lstat      medv
0 305.6313  17.97312   9.21775  24.91250
1 521.6597  19.15139  15.86049  19.32986

Coefficients of linear discriminants:
                  LD1
zn      -0.0010188564
indus    0.0160859284
chas     0.0504097117
nox      7.0707746045
rm       0.0829011402
age      0.0138146989
dis     -0.0233015796
rad      0.0794203051
tax     -0.0007562095
ptratio  0.0155866141
lstat    0.0112281648
medv     0.0241193260
```

이를 바탕으로 하여 다음과 같은 결과를 얻을 수 있다.

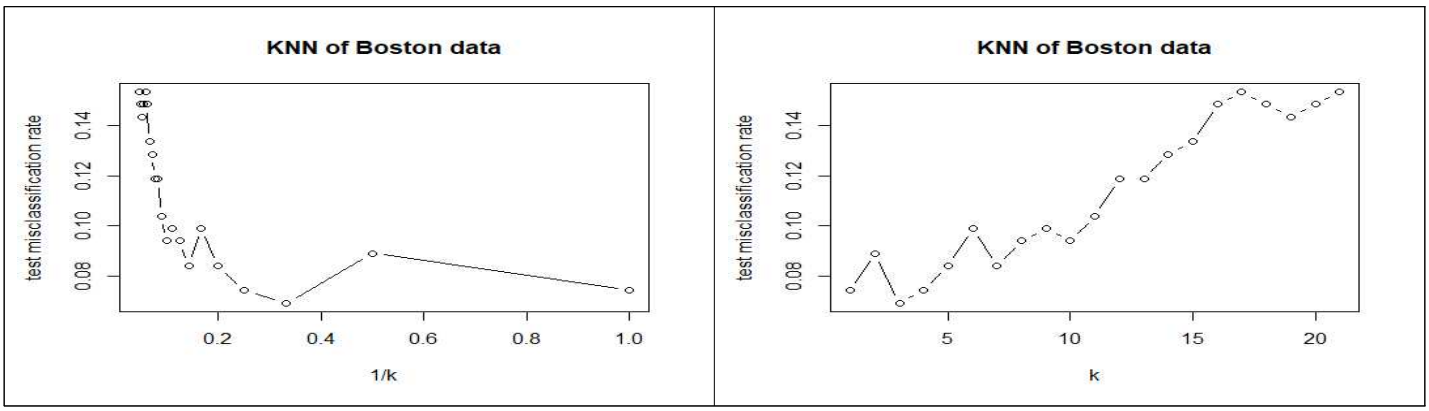
confusion matrix

		Predicted class (crim01)		Sum Total
		0	1	
True class (crim01)	0	85	8	93
	1	27	82	109
Sum		112	90	202

test misclassification rate : 0.1733

3) KNN

최적의 모수 k 설정



최적의 모수 : $k = 3$

이를 바탕으로 하여 다음과 같은 결과를 얻을 수 있다.

confusion matrix

		Predicted class (crim01)		Sum Total
		0	1	
True class (crim01)	0	90	3	93
	1	11	98	109
Sum		101	101	202

test misclassification rate : 0.0693

* 결과적으로 본 상황에 대해서는 KNN의 경우가 test 오분류율이 가장 작게 산출되었다.