“인공지능 및 머신러닝 입문“ 리포트

제출일: 2019.06.24.
작성자: 신재익 (의학과, 박사과정, 학번 2018313168)

본 리프트에서는 과제로 제시한 데이터셋 중 하나인 cesarian 데이터셋을 분석하였음.

1. 데이터 설명
이 데이터셋은 cesarian section(이하 cesarian)을 시행한 환자에 대한 데이터로 feature(이하 피쳐)는 age, delivery_no, dielivery_time, bp, heart 가 존재함.
이 데이터셋의 피쳐들을 이용하여 예측하고 싶은것은 cesarian section 시행(가능)의 유무임.
본 데이터셋에는 각 환자에 대한 cesarian section 시행 유무가 라벨링되어 있음.

2. 본 연구의 목적
본 보고서는 cesarian에 대한 예측모델을 만들기 위하여 R을 이용한 분류기모델을 개발 및 테스트한 내용임.

3. 연구방법
3-1. 데이터 파일 읽기
(1) 주어진 피쳐의 값은 양의 정수이며 각 피쳐별 값의 범위는 다음과 같음.
Age(17~), DN(DeliveryNumber=1~4), DT(DeliveryTime=0~2), BP(0,1), HP(0,1)
(2) 주어진 데이터셋은 초기(홈페이지에 공개된 시점)에 arff 포맷으로 작성되었으며 R의 DataFrame으로 변환하기위해 foreign패키지의 read.arff 함수를 사용하엿음.

```
library("foreign")
rawdata=read.arff("./caesarian.csv.arff")
colnames(rawdata) = c("Age","DN","DT","BP","HP","CS")
str(rawdata)
```

```
'data.frame':   80 obs. of  6 variables:
 $ Age: Factor w/ 22 levels "17","18","19",..: 6 10 10 12 6 10 11 16 12 11 ...
 $ DN : Factor w/ 4 levels "1","2","3","4": 1 2 2 1 2 1 2 3 2 1 ...
 $ DT : Factor w/ 3 levels "0","1","2": 1 1 2 1 1 2 1 1 1 2 ...
 $ BP : Factor w/ 3 levels "0","1","2": 3 2 2 3 2 1 2 2 2 2 ...
 $ HP : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ CS : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 1 2 1 2 ...
```

(3) 그 결과 RDataFrame의 모든 피쳐(column, 행)가 factor타입으로 저장되었음.

3-2. 피쳐간의 correlation 확인

(1) correlation plot을 그리기 위하여 주어진 데이터의 값들을 모두 numeric 타입으로 변환하였음.

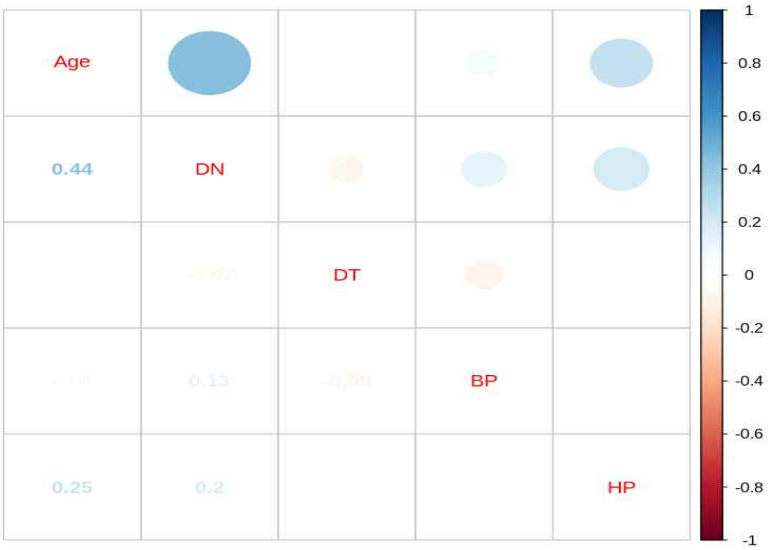
(2) corrplot을 이용하여 correlation plot을 그림. 값을 확인했을 때 강한 correlation을 확인하지못함

```
## check collinearity btw features
library(corrplot)
str(rawdata)
rawdata.num=rawdata
for ( i in 1:5)  { rawdata.num[,i] = as.numeric(rawdata[,i]) }
str(rawdata.num)

rawdata.cor = cor(rawdata.num[,1:5])
corrplot.mixed(rawdata.cor)
```

```
corrplot 0.84 loaded
```

```
'data.frame':   80 obs. of  6 variables:
 $ Age: Factor w/ 22 levels "17","18","19",..: 6 10 10 12 6 10 11 16 12 11 ...
 $ DN : Factor w/ 4 levels "1","2","3","4": 1 2 2 1 2 1 2 3 2 1 ...
 $ DT : Factor w/ 3 levels "0","1","2": 1 1 2 1 1 2 1 1 1 2 ...
 $ BP : Factor w/ 3 levels "0","1","2": 3 2 2 3 2 1 2 2 2 2 ...
 $ HP : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ CS : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 1 2 1 2 ...
'data.frame':   80 obs. of  6 variables:
 $ Age: num  6 10 10 12 6 10 11 16 12 11 ...
 $ DN : num  1 2 2 1 2 1 2 3 2 1 ...
 $ DT : num  1 1 2 1 1 2 1 1 1 2 ...
 $ BP : num  3 2 2 3 2 1 2 2 2 2 ...
 $ HP : num  1 1 1 1 1 1 1 1 1 1 ...
 $ CS : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 1 2 1 2 ...
```



3-2. 다변수 선형 regression

(1) 주어진 데이터를 7:3 비율로 트레이닝용:테스트용의 데이터로 나눔 (random seed=123)

In [13]:
```
## log reg from numeric data

traindata.num = rawdata.num[ind==1,]
str(traindata.num) ## 80datas * 70% = 56datas
testdata.num = rawdata.num[ind==2,]
str(testdata.num)
```

```
'data.frame':   56 obs. of  6 variables:
 $ Age: num  6 10 10 11 12 11 17 7 4 13 ...
 $ DN : num  1 2 1 2 2 1 1 1 1 1 ...
 $ DT : num  1 2 2 1 1 2 2 2 1 3 ...
 $ BP : num  3 2 1 2 2 2 1 2 2 1 ...
 $ HP : num  1 1 1 1 1 1 1 1 2 2 ...
 $ CS : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 1 1 2 ...
'data.frame':   24 obs. of  6 variables:
 $ Age: num  10 12 6 16 19 9 8 10 11 2 ...
 $ DN : num  2 1 2 3 1 1 1 1 1 1 ...
 $ DT : num  1 1 1 1 1 3 3 2 1 1 ...
 $ BP : num  2 3 2 2 2 1 1 2 1 2 ...
 $ HP : num  1 1 1 1 1 1 2 1 2 1 ...
 $ CS : Factor w/ 2 levels "0","1": 2 1 2 2 1 1 2 1 2 1 ...
```

(2) 모든 피처를 사용하여 multi variable linear regression을 수행함.

```
## multi-var linear reg.
test.lr = lm(CS ~., data=numdata)
summary(test.lr)
```

```
Call:
lm(formula = CS ~ ., data = numdata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9176 -0.4219  0.1192  0.4173  0.7420

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.31671    0.26246   5.017  3.5e-06 ***
Age         -0.00589    0.01244  -0.473  0.63741
DelN         0.06096    0.07511   0.812  0.41965
DelT        -0.10064    0.06529  -1.541  0.12750
BP          -0.05033    0.07526  -0.669  0.50576
HP           0.35679    0.11282   3.162  0.00227 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4702 on 74 degrees of freedom
Multiple R-squared:  0.1632,    Adjusted R-squared:  0.1067
F-statistic: 2.886 on 5 and 74 DF,  p-value: 0.01954
```

3-3. logistic regression

(1) 트레이닝용 데이터로 logistic regression 수행함.

(2) 모든 피쳐 사용함

(3) 팩터타입은 자동으로 dummy column생성됨

(4) DN=4 와 DT=1, BP=1, HP=1 이 큰 영향을 줌.

<logistic regression 결과>

all:

glm(formula = CS ~ ., family = binomial, data = traindata.num_age)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7656	-0.8930	0.2696	0.7518	1.9929

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.95446	1.44035	0.663	0.5075	
Age	0.03012	0.08426	0.357	0.7208	
DN2	0.04250	0.80266	0.053	0.9578	
DN3	-0.39535	1.33656	-0.296	0.7674	
DN4	16.89548	2150.24360	0.008	0.9937	
DT1	-1.12722	0.91685	-1.229	0.2189	
DT2	-0.96034	0.93098	-1.032	0.3023	
BP1	-1.99671	1.02044	-1.957	0.0504	.
BP2	0.18724	1.04530	0.179	0.8578	
HP1	1.73123	0.83863	2.064	0.0390	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 75.837 on 55 degrees of freedom

Residual deviance: 54.993 on 46 degrees of freedom

AIC: 74.993

Number of Fisher Scoring iterations: 16

(5) log. reg.로 얻은 모델로 트레이닝용 데이터에 predict 함. -> confusion matrix 얻음 -> accuracy 75%

onfusion Matrix and Statistics

	Reference	
Prediction	0	1
0	15	6
1	8	27

```
               Accuracy : 0.75
                 95% CI : (0.6163, 0.8561)
    No Information Rate : 0.5893
    P-Value [Acc > NIR] : 0.009055

                  Kappa : 0.4766
 Mcnemar's Test P-Value : 0.789268

            Sensitivity : 0.8182
            Specificity : 0.6522
         Pos Pred Value : 0.7714
         Neg Pred Value : 0.7143
             Prevalence : 0.5893
         Detection Rate : 0.4821
   Detection Prevalence : 0.6250
      Balanced Accuracy : 0.7352
```

(7) log reg로 얻은 모델로 테스트용 데이터에 predict 함. -> confusion matrix 얻음 -> accuracy 58%

```
Confusion Matrix and Statistics
          Reference
Prediction 0 1
         0 5 4
         1 6 9

               Accuracy : 0.5833
                 95% CI : (0.3664, 0.7789)
    No Information Rate : 0.5417
    P-Value [Acc > NIR] : 0.4213

                  Kappa : 0.1489
 Mcnemar's Test P-Value : 0.7518

            Sensitivity : 0.6923
            Specificity : 0.4545
         Pos Pred Value : 0.6000
         Neg Pred Value : 0.5556
             Prevalence : 0.5417
         Detection Rate : 0.3750
```

```
   Detection Prevalence : 0.6250
      Balanced Accuracy : 0.5734

       'Positive' Class : 1
```

3-4. best subset regression 수행
(1) best subset reg를 위해 모든 데이터를 numeric 타입으로 전환함.
-> numeric데이터로 log reg 시도
-> HP 가 큰영향을줌

```
Call:
glm(formula = CS ~ ., family = binomial, data = traindata.num)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9359  -1.0293   0.5149   0.8963   1.5201
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.78921    1.76248  -1.583  0.11353
Age          0.05993    0.07465   0.803  0.42208
DN          -0.13158    0.44714  -0.294  0.76856
DT          -0.14475    0.38001  -0.381  0.70327
BP           0.24278    0.45556   0.533  0.59409
HP           1.79571    0.67783   2.649  0.00807 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 75.837  on 55  degrees of freedom
Residual deviance: 65.160  on 50  degrees of freedom
AIC: 77.16
Number of Fisher Scoring iterations: 4
```

(2) 트레이닝용 데이터에 numeric log reg 모델 적용시 accuracy 67%

```
Confusion Matrix and Statistics
          Reference
Prediction  0  1
         0 17 12
         1  6 21

               Accuracy : 0.6786
                 95% CI : (0.5404, 0.7971)
    No Information Rate : 0.5893
```

P-Value [Acc > NIR] : 0.1098

Kappa : 0.3612
Mcnemar's Test P-Value : 0.2386

Sensitivity : 0.6364
Specificity : 0.7391
Pos Pred Value : 0.7778
Neg Pred Value : 0.5862
Prevalence : 0.5893
Detection Rate : 0.3750
Detection Prevalence : 0.4821
Balanced Accuracy : 0.6877

'Positive' Class : 1

(3) 테스트용 데이터에 numeric logreg 모델 적용시 accuracy 50%
Confusion Matrix and Statistics
Reference
Prediction 0 1
0 6 7
1 5 6

Accuracy : 0.5
95% CI : (0.2912, 0.7088)
No Information Rate : 0.5417
P-Value [Acc > NIR] : 0.7313

Kappa : 0.0069
Mcnemar's Test P-Value : 0.7728

Sensitivity : 0.4615
Specificity : 0.5455
Pos Pred Value : 0.5455
Neg Pred Value : 0.4615
Prevalence : 0.5417
Detection Rate : 0.2500
Detection Prevalence : 0.4583
Balanced Accuracy : 0.5035

'Positive' Class : 1

(4) best subset reg 을 트레이닝용 데이터에 적용 -> CV는 에러, BIC에서는 피쳐 HP 만 사용한 모델을 얻음

BIC
BICq equivalent for q in (0.0581156714516717, 0.852591422895177)
Best Model:
```
             Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -2.168908  0.8937965 -2.426624 0.015240046
HP           1.863526  0.6532164  2.852847 0.004332954
```

(5) 피쳐 HP만 이용하여 log reg 함

```
Call:
glm(formula = CS ~ HP, family = binomial, data = traindata.num)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8704  -1.0508   0.6181   0.7910   1.3095
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.1689     0.8938  -2.427  0.01524 *
HP            1.8635     0.6532   2.853  0.00433 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 75.837  on 55  degrees of freedom
Residual deviance: 66.241  on 54  degrees of freedom
AIC: 70.241
Number of Fisher Scoring iterations: 4
```

(6) 트레이닝 데이터에서 accruacy 67%

```
Confusion Matrix and Statistics
          Reference
Prediction  0  1
         0 19 14
         1  4 19

               Accuracy : 0.6786
                 95% CI : (0.5404, 0.7971)
    No Information Rate : 0.5893
```

P-Value [Acc > NIR] : 0.10983

Kappa : 0.377
Mcnemar's Test P-Value : 0.03389

Sensitivity : 0.5758
Specificity : 0.8261
Pos Pred Value : 0.8261
Neg Pred Value : 0.5758
Prevalence : 0.5893
Detection Rate : 0.3393
Detection Prevalence : 0.4107
Balanced Accuracy : 0.7009

'Positive' Class : 1

(7) 테스트용 데이터에서 accruacy 58%

Confusion Matrix and Statistics

Prediction	Reference 0	Reference 1
0	9	8
1	2	5

Accuracy : 0.5833
95% CI : (0.3664, 0.7789)
No Information Rate : 0.5417
P-Value [Acc > NIR] : 0.4213

Kappa : 0.1946
Mcnemar's Test P-Value : 0.1138

Sensitivity : 0.3846
Specificity : 0.8182
Pos Pred Value : 0.7143
Neg Pred Value : 0.5294
Prevalence : 0.5417
Detection Rate : 0.2083
Detection Prevalence : 0.2917
Balanced Accuracy : 0.6014

'Positive' Class : 1

3-5. 원본데이터의 특성을 반영하여 데이터를 변화시킴 (이진화)
(1) 이진화: 앞의 결과를 봤을때 피쳐 DN를 <4(0) 와 ==4(1) 두개로 나누고. 또 DT도 ==1(0) 과 >1(1) 로 나누고, BP도 <2(0)과 ==2(1) 로 나눔

(2) 이진화된 트레이닝용 numeric 데이터로 log. reg. 함

```
Call:
glm(formula = y ~ ., family = binomial, data = traindata.bin)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7500  -0.8868   0.3002   0.7790   1.9638
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.43120    1.39253  -0.310   0.7568
Age           0.01776    0.07178   0.247   0.8046
DN1          16.97808 2147.45903   0.008   0.9937
DT1          -1.03625    0.77515  -1.337   0.1813
BP1          -2.09340    0.78030  -2.683   0.0073 **
HP            1.61217    0.73013   2.208   0.0272 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 75.837  on 55  degrees of freedom
Residual deviance: 55.147  on 50  degrees of freedom
AIC: 67.147
Number of Fisher Scoring iterations: 16
```

(3) numeric 데이터에 대한 log. reg. 모델은 트레이닝용 데이터에서 정확도 75% 달성함.

```
Confusion Matrix and Statistics
          Reference
Prediction  0  1
         0 15  6
         1  8 27

               Accuracy : 0.75
                 95% CI : (0.6163, 0.8561)
    No Information Rate : 0.5893
    P-Value [Acc > NIR] : 0.009055
```

```
                    Kappa : 0.4766
 Mcnemar's Test P-Value : 0.789268

             Sensitivity : 0.8182
             Specificity : 0.6522
          Pos Pred Value : 0.7714
          Neg Pred Value : 0.7143
              Prevalence : 0.5893
          Detection Rate : 0.4821
    Detection Prevalence : 0.6250
       Balanced Accuracy : 0.7352
```

(4) log reg 모델의 테스트용 데이터에서 정확도 58% 였음.

```
Confusion Matrix and Statistics
          Reference
Prediction 0 1
         0 5 4
         1 6 9

               Accuracy : 0.5833
                 95% CI : (0.3664, 0.7789)
    No Information Rate : 0.5417
    P-Value [Acc > NIR] : 0.4213

                  Kappa : 0.1489
 Mcnemar's Test P-Value : 0.7518

            Sensitivity : 0.6923
            Specificity : 0.4545
         Pos Pred Value : 0.6000
         Neg Pred Value : 0.5556
             Prevalence : 0.5417
         Detection Rate : 0.3750
   Detection Prevalence : 0.6250
      Balanced Accuracy : 0.5734
```

4. 결과

(1) 본 연구에서 사용된 데이터는 양이 매우 한정적이라서 logistic regression 만으로 수행하였음.
(2) 데이터의 80%를 트레이닝용으로 사용하고 20%를 테스트용으로 사용하였음.
(3) best subset을 통하여 도출한 최적의 feature조합과 모든 피쳐를 사용한 기존모델의 성능비교를 수행하였음.
(4) 원본 데이터의 특성을 무시하고 numeric으로 처리할때와 원본데이터를 고려하여 numeric이지만 이진화를 시행하여 factor에 가까운 데이터로 regression을 시도하였음.
(5) 최종적으로 트레이닝 데이터에서 accuracy 75%, sensitivity 81%, 테스트데이터에서 accuracy 58%, sensitivity 69%가 최대인 모델을 만들었음.

표. 각 모델별 결과

		train			test		
	data type	accuracy	sensitivity (precision)	recall	accuracy	sensitivity (precision)	recall
Full log reg	numeric	67	63	77	50	46	54
only hp log reg	numeric	67	57	82	58	38	71
full log reg	numeric 이진화	75	81	77	58	69	60
hp+BP log reg	numeric 이진화	69	78	72	58	69	60