Example2_LogisticRegression

Shaurya Jauhari (Email: shauryajauhari@gzhmu.edu.cn)

2019-05-31

```
#Import
# *file.choose()* function can be used to manually browse for desired files.
data <- read.csv(file = "/Users/mei/Desktop/Machine-Learning_Lab_Workshops/Machine_Learning_Logistic_Re
labels <- read.csv(file = "/Users/mei/Desktop/Machine-Learning_Lab_Workshops/Machine_Learning_Logistic_

# Shorten dimension of the data frame
data <- data[,2:11]
labels <- labels[,-1]

# Removing columns with majority null entries.
data <- data[, -c(1,6,9,10)]

# Assign new column for class labels
data$Class <- labels

# Consider only data with two labels for logistic regression
# Extracting just two labels BRCA and PRAD out of 5.
data <- data[data$Class == c("BRCA", "PRAD"),]
```

```
## Warning in `==.default`(data$Class, c("BRCA", "PRAD")): longer object
## length is not a multiple of shorter object length
```

```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

```
data$Class <- factor(data$Class, levels = c("BRCA", "PRAD"))

#stats::glm

model1 <- glm(formula = as.factor(data$Class) ~ .,
              data = data,
              family = "binomial")
summary(model1)
```

```
##
## Call:
## glm(formula = as.factor(data$Class) ~ ., family = "binomial",
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4157  -0.6716  -0.3446   0.7368   2.5088
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.43130    5.37810   0.266   0.7901
## gene_1       0.02139    0.20136   0.106   0.9154
## gene_2       1.64729    0.29363   5.610 2.02e-08 ***
```

```
## gene_3       0.10983    0.36117   0.304   0.7611
## gene_4      -0.91280    0.43345  -2.106   0.0352 *
## gene_6       0.10668    0.20865   0.511   0.6091
## gene_7      -0.95858    0.40971  -2.340   0.0193 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 281.32  on 225  degrees of freedom
## Residual deviance: 194.90  on 219  degrees of freedom
## AIC: 208.9
##
## Number of Fisher Scoring iterations: 5
```

```
# Multicollinearity check
VIF(model1)
```

```
##   gene_1   gene_2   gene_3   gene_4   gene_6   gene_7
## 1.442830 1.347412 1.072369 1.178884 1.198789 1.273080
```

```
#Prediction
y_estimate <- predict(model1,
                      data,
                      type = "response")

# Converting probabilities to labels
prediction_probabilities <- ifelse(y_estimate>0.5, "BRCA", "PRAD")

# Confusion matrix
confusion_matrix<- table(Predicted = prediction_probabilities, Actual = data$Class)
print(confusion_matrix)
```

```
##           Actual
## Predicted BRCA PRAD
##      BRCA   17   44
##      PRAD  138   27
```

```
#Misclassification error
misclassfication_error <- 1- sum(diag(confusion_matrix))/sum(confusion_matrix)
cat("The misclassification error in test data is",
    (round(misclassfication_error*100)), "percent")
```

```
## The misclassification error in test data is 81 percent
```