ASSIGNMENT 10.3

Siddhartha Bhaumik

2022-05-21

Add Citations

- R for Everyone
- Discovering Statistics Using R

```
## Load the `data/binary-classifier-data` to
class_df <- read.csv("/Users/siddharthabhaumik/Documents/GitHub/dsc520/data/binary-classifier-data.csv")

## Viewing Sample data
head(class_df)
```

```
##   label        x        y
## 1     0 70.88469 83.17702
## 2     0 74.97176 87.92922
## 3     0 73.78333 92.20325
## 4     0 66.40747 81.10617
## 5     0 69.07399 84.53739
## 6     0 72.23616 86.38403
```

```
# Fit a linear model
class.glm <- glm(label ~ x + y, data = class_df, family = "binomial")

# Print the model
class.glm
```

```
##
## Call:  glm(formula = label ~ x + y, family = "binomial", data = class_df)
##
## Coefficients:
## (Intercept)            x            y
##    0.424809    -0.002571    -0.007956
##
## Degrees of Freedom: 1497 Total (i.e. Null);  1495 Residual
## Null Deviance:       2076
## Residual Deviance: 2052  AIC: 2058
```

```
# View the summary of your model
summary(class.glm)
```

```
##
## Call:
## glm(formula = label ~ x + y, family = "binomial", data = class_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.3728  -1.1697  -0.9575   1.1646   1.3989
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.424809   0.117224   3.624  0.00029 ***
## x           -0.002571   0.001823  -1.411  0.15836
## y           -0.007956   0.001869  -4.257 2.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2075.8  on 1497  degrees of freedom
## Residual deviance: 2052.1  on 1495  degrees of freedom
## AIC: 2058.1
##
## Number of Fisher Scoring iterations: 4
```

```
## What is the accuracy of your model?

set.seed(1234)
#load necessary packages
library(caTools)
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(InformationValue)
```

```
##
## Attaching package: 'InformationValue'
```

```
## The following objects are masked from 'package:caret':
##
##     confusionMatrix, precision, sensitivity, specificity
```

```
library(ISLR)

# 70:30 data split into training and validation
data_split = sample.split(class_df, SplitRatio = 0.7)

training_data = subset(class_df, data_split==TRUE)
test_data = subset(class_df, data_split==FALSE)

# Print data frames
print(dim(training_data))
```

```
## [1] 999   3
```

```
print(dim(test_data))
```

```
## [1] 499   3
```

```
# Fit a linear model
new_class_glm = glm( label ~ . , family="binomial", data = training_data)
```

```r
#Summary
summary(new_class_glm)
```

```
##
## Call:
## glm(formula = label ~ ., family = "binomial", data = training_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3733  -1.1714  -0.9584   1.1637   1.3962
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.426096   0.143309   2.973 0.002946 **
## x           -0.002613   0.002229  -1.172 0.241028
## y           -0.007897   0.002279  -3.465 0.000531 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1384.4  on 998  degrees of freedom
## Residual deviance: 1368.6  on 996  degrees of freedom
## AIC: 1374.6
##
## Number of Fisher Scoring iterations: 4
```

```r
#Prediction based on test data
predict_data = predict(new_class_glm, newdata = test_data, type = "response")

# Confusion matrix on test set
confusionMatrix(test_data$label,predict_data)
```

```
##     0   1
## 0 143  96
## 1 113 147
```

```r
#calculate sensitivity: The "true positive rate"
sensitivity(test_data$label,predict_data)
```

```
## [1] 0.6049383
```

```r
#calculate specificity: The "true negative rate"
specificity(test_data$label,predict_data)
```

```
## [1] 0.5585938
```

```r
#calculate total mis-classification error rate
misClassError(test_data$label,predict_data)
```

```
## [1] 0.4188
```

```r
## The total misclassification error rate is 41.88% for this model which indicates its not a accurate m
```