Classification

Shiyou Yan

2023-02-16

Logistic regression is used to handle the classification work. In logistic regression, it estimates the probability of an event occurring and provides discreet output. Then it will fit the line values to sigmoid curve. It is supervised machine learning as linear regression. The strengths of logistic regression are efficient, easy to classify, and less over-fitting. The weakness of logistic regression are only used for predict discrete functions and the number of observations cannot lesser than the number of features.

#Read data set

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00640/Occupancy_Estimation.csv")
```

1.Divide into 80/20 train/test

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00640/Occupancy_Estimation.csv")
set.seed(1234)

sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.8,0.2))
train  <- data[sample, ]
test   <- data[!sample, ]
```

2.Use at least 5 R functions for data exploration, using the training data

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00640/Occupancy_Estimation.csv")
set.seed(1234)
sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.8,0.2))
train  <- data[sample, ]
test   <- data[!sample, ]
dim(train)
```

```
## [1] 8120   19
```

```
str(train)
```

```
## 'data.frame':    8120 obs. of  19 variables:
##  $ Date                : chr  "2017/12/22" "2017/12/22" "2017/12/22" "2017/12/22" ...
##  $ Time                : chr  "10:49:41" "10:50:12" "10:50:42" "10:51:13" ...
##  $ S1_Temp             : num  24.9 24.9 25 25 25 ...
##  $ S2_Temp             : num  24.8 24.8 24.8 24.8 24.8 ...
##  $ S3_Temp             : num  24.6 24.6 24.5 24.6 24.6 ...
##  $ S4_Temp             : num  25.4 25.4 25.4 25.4 25.4 ...
##  $ S1_Light            : int  121 121 121 121 121 120 121 122 101 122 ...
##  $ S2_Light            : int  34 33 34 34 34 34 34 35 34 35 ...
##  $ S3_Light            : int  53 53 53 53 54 54 54 56 57 57 ...
##  $ S4_Light            : int  40 40 40 40 40 40 41 43 43 43 ...
##  $ S1_Sound            : num  0.08 0.93 0.43 0.41 0.13 1.39 0.09 0.09 3.84 2.2 ...
##  $ S2_Sound            : num  0.19 0.05 0.11 0.1 0.06 0.32 0.06 0.05 0.64 0.31 ...
##  $ S3_Sound            : num  0.06 0.06 0.08 0.1 0.06 0.43 0.09 0.06 0.48 0.33 ...
##  $ S4_Sound            : num  0.06 0.06 0.06 0.09 0.07 0.06 0.05 0.13 0.39 0.21 ...
##  $ S5_CO2              : int  390 390 390 390 390 390 390 390 390 390 ...
##  $ S5_CO2_Slope        : num  0.769 0.646 0.519 0.388 0.165 ...
##  $ S6_PIR              : int  0 0 0 0 0 1 0 0 1 1 ...
##  $ S7_PIR              : int  0 0 0 0 0 0 0 0 1 1 ...
##  $ Room_Occupancy_Count: int  1 1 1 1 1 1 1 1 1 1 ...
```

```
head(train,5)
```

	Date <chr>	Time <chr>	S1_Temp <dbl>	S2_Temp <dbl>	S3_Temp <dbl>	S4_Temp <dbl>	S1_Light <int>	S2_Light <int>	S3_Light <int>
1	2017/12/22	10:49:41	24.94	24.75	24.56	25.38	121	34	53
2	2017/12/22	10:50:12	24.94	24.75	24.56	25.44	121	33	53
3	2017/12/22	10:50:42	25.00	24.75	24.50	25.44	121	34	53
4	2017/12/22	10:51:13	25.00	24.75	24.56	25.44	121	34	53
6	2017/12/22	10:52:14	25.00	24.81	24.56	25.44	121	34	54

5 rows | 1-10 of 20 columns

```
tail(train,5)
```

	Date <chr>	Time <chr>	S1_Te... <dbl>	S2_Te... <dbl>	S3_Te... <dbl>	S4_Te... <dbl>	S1_Light <int>	S2_Light <int>	S3_Light <int>
10124	2018/01/11	08:57:36	25.06	25.13	24.69	25.25	6	7	33
10125	2018/01/11	08:58:07	25.06	25.13	24.69	25.31	6	7	33
10126	2018/01/11	08:58:37	25.06	25.06	24.69	25.25	6	7	34
10127	2018/01/11	08:59:08	25.13	25.06	24.69	25.25	6	7	34
10129	2018/01/11	09:00:09	25.13	25.06	24.69	25.25	6	7	34

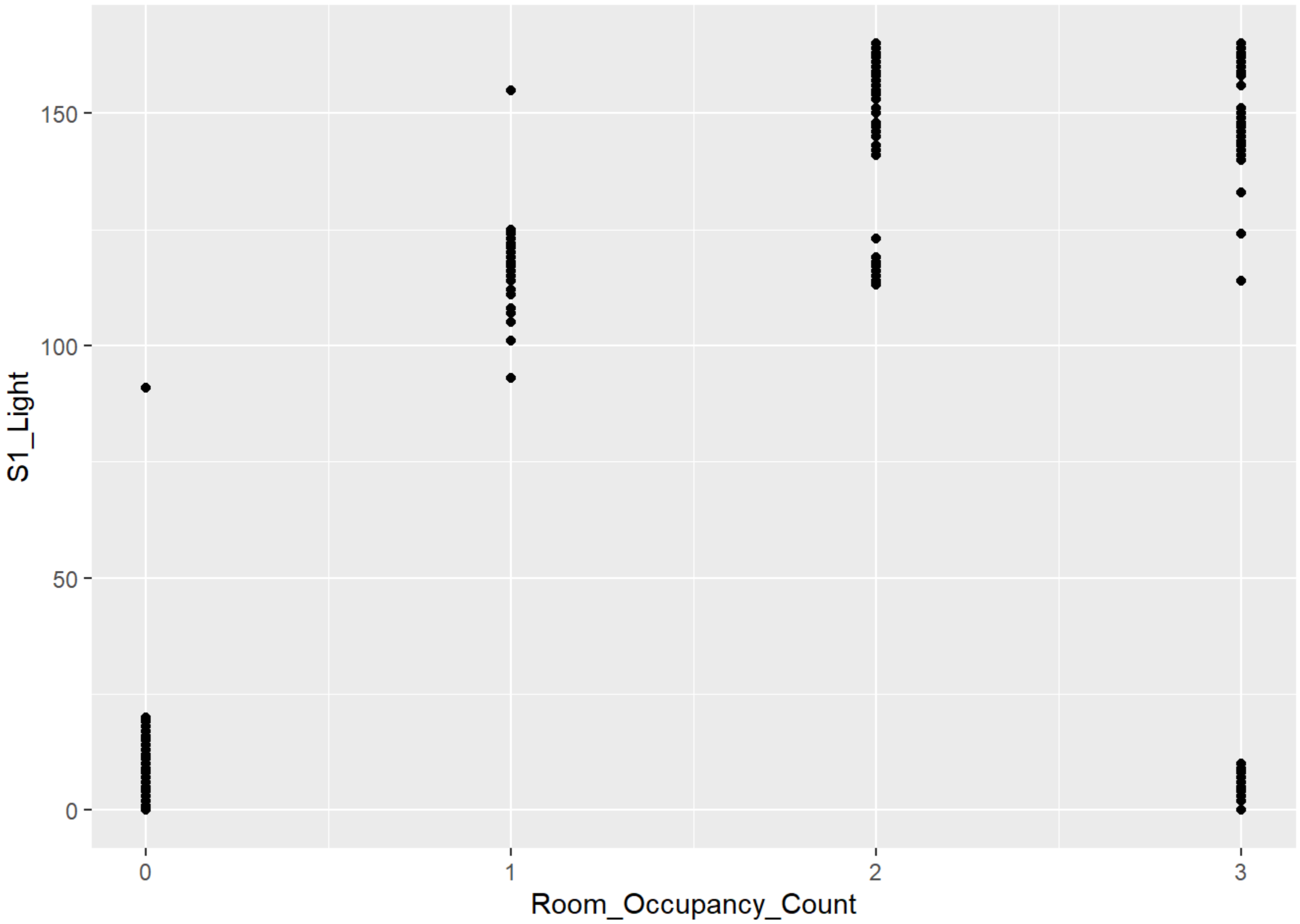
5 rows | 1-10 of 20 columns

```
summary(train[3])
```

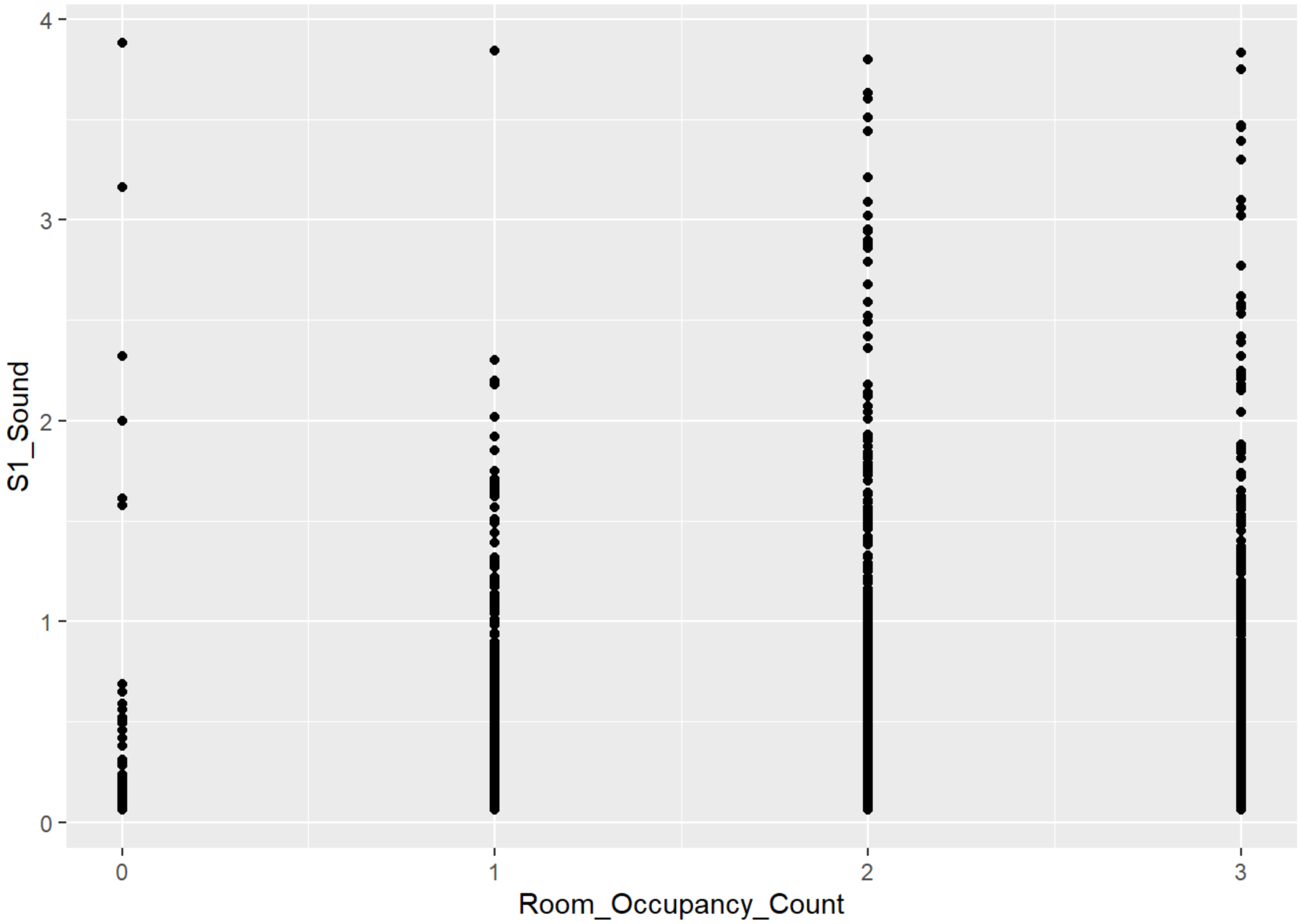
```
##     S1_Temp
##  Min.   :24.94
##  1st Qu.:25.19
##  Median :25.38
##  Mean   :25.45
##  3rd Qu.:25.63
##  Max.   :26.38
```

3.Create at least 2 informative graphs, using the training data

```
library(ggplot2)
ggplot(data = data,mapping = aes(x=Room_Occupancy_Count,y=S1_Light)) + geom_point()
```



```
library(ggplot2)
ggplot(data = data,mapping = aes(x=Room_Occupancy_Count,y=S1_Sound)) + geom_point()
```



4.Build a logistic regression model and output the summary. Write a thorough explanation of the information in the model summary

```
data <- read.csv ("https://archive.ics.uci.edu/ml/machine-learning-databases/00640/Occupancy_Estimation.csv")
sum(is.na(data))
```

```
## [1] 0
```

```
result <- glm(Room_Occupancy_Count ~ S7_PIR + S6_PIR + S5_CO2_Slope + S5_CO2 + S4_Sound + S3_Sound + S2_Sound + S1_Sound + S4_Light + S3_Light + S2_Light + S1_Light + S4_Temp + S3_Temp + S2_Temp + S1_Temp, data = data, x = TRUE, y = TRUE)
summary(result)
```

```
##
## Call:
## glm(formula = Room_Occupancy_Count ~ S7_PIR + S6_PIR + S5_CO2_Slope +
##     S5_CO2 + S4_Sound + S3_Sound + S2_Sound + S1_Sound + S4_Light +
##     S3_Light + S2_Light + S1_Light + S4_Temp + S3_Temp + S2_Temp +
##     S1_Temp, data = data, x = TRUE, y = TRUE)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.03322  -0.12133   0.00224   0.07669   2.43127
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.005e+00  5.231e-01 -15.305  < 2e-16 ***
## S7_PIR        4.139e-01  1.593e-02  25.990  < 2e-16 ***
## S6_PIR        1.800e-01  1.424e-02  12.640  < 2e-16 ***
## S5_CO2_Slope  1.896e-01  3.460e-03  54.789  < 2e-16 ***
## S5_CO2        3.420e-05  3.768e-05   0.908  0.36404
## S4_Sound     -3.485e-01  3.670e-02  -9.495  < 2e-16 ***
## S3_Sound     -7.635e-02  1.140e-02  -6.696 2.25e-11 ***
## S2_Sound      1.957e-01  1.617e-02  12.104  < 2e-16 ***
## S1_Sound      1.009e-01  1.321e-02   7.642 2.34e-14 ***
## S4_Light     -4.042e-03  2.618e-04 -15.440  < 2e-16 ***
## S3_Light      2.606e-03  1.247e-04  20.900  < 2e-16 ***
## S2_Light      8.937e-04  9.764e-05   9.153  < 2e-16 ***
## S1_Light      5.368e-03  1.787e-04  30.039  < 2e-16 ***
## S4_Temp      -6.959e-01  2.056e-02 -33.853  < 2e-16 ***
## S3_Temp       7.861e-01  2.911e-02  27.003  < 2e-16 ***
## S2_Temp       1.166e-01  9.973e-03  11.690  < 2e-16 ***
## S1_Temp       1.335e-01  4.479e-02   2.981  0.00288 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.08495938)
##
##     Null deviance: 8088.02  on 10128  degrees of freedom
## Residual deviance:  859.11  on 10112  degrees of freedom
## AIC: 3790
##
## Number of Fisher Scoring iterations: 2
```

5. Build a naive Bayes model and output what the model learned. Write a thorough explanation of the data.

The code has some problems library(e1071) data <- read.csv (“https://archive.ics.uci.edu/ml/machine-learning-databases/00640/Occupancy_Estimation.csv (https://archive.ics.uci.edu/ml/machine-learning-databases/00640/Occupancy_Estimation.csv)”) result <- naiveBayes(Room_Occupancy_Count, data = data, x = TRUE, y = TRUE) plot(result)

6.Write a paragraph listing the strengths and weaknesses of Naive Bayes and Logistic Regression The strengths of Naive Bayes are no need lots of training data. It is easy to complete and faster work compare to other models. In addition, it solves both continuous data, discrete data, and multiple class prediction. The weaknesses of Naive Bayes is hard to implement in the real life because those cases are rarely happening. The strengths of logistic regression are efficient and classify fast when solving unknown data. It also avoid over fit as a linear model. The weakness of logistic regression are unable to predict a continuous outcome, the data size cannot small, and over depends on independent variables