Logistic Regression

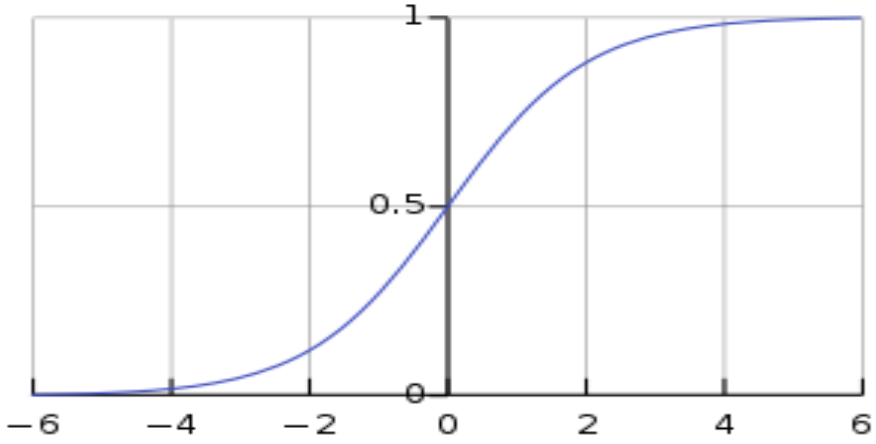
Logistic Regression

- This algorithm is used for classification type problems
- Types of Logistic Regression:
  - Binary
  - Multinomial
  - Ordinal
- We are going to cover Binary Logistic Regression

Logistic Response Function

- Standard logistic function on 2-dimensional plane is given by the following expression given on the right.
- From the graph, it is evident that the value of the f(x) ranges between 0 and 1.
- This function is also called $\mathbf{sigmoid}$ function and has a wide usage in various other algorithms such as neural network.

$$y = f(x) = \frac{1}{1 + e^{-x}}$$



Sane's STATS Academy of Statistics

Logistic Response Function

- The same function in the m-dimensional space can be written in the following way:

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}}$$

Where

$\beta_0, \beta_1, \beta_2, \ldots \beta_m$: Coefficients of the variables in m-dimensional space

- For any values of $\beta_0, \beta_1, \beta_2, \ldots \beta_m$ and $x_1, x_2, \ldots x_m$, the value of y always between 0 and 1.
- We can denote y by probability p.

Odds

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m)}}$$

$$1-p = \frac{e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m)}}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m)}}$$

$$\frac{p}{1-p} = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m)}$$

- The ratio $\frac{p}{1-p}$ is called odds. For any values of $\beta_0, \beta_1, \beta_2, \ldots \beta_m$ and $x_1, x_2, \ldots x_m$, odds always ranges from 0 to ∞.

Sane's STATS Academy of Statistics

Interpreting Logistic Function

- In our binary classification, let us consider 0 and 1 as two possible outcomes, with 0 as non-occurrence of a particular event and 1 as occurrence of the particular event.
- p in our expression, is considered as probability of occurrence of the event and $1 - p$ as non-occurrence of the event
- Hence, the ratio $\frac{p}{1-p}$ is ratio of probability of occurrence to the probability of non-occurrence of the event.

Logit Function

$$\frac{p}{1-p} = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_m x_m)}$$

$$\log(odds) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

- The ratio $\log\left(\frac{p}{1-p}\right)$ is called logit function. For any values of $\beta_0, \beta_1, \beta_2, \ldots \beta_m$ and $x_1, x_2, \ldots x_m$, log(odds) always range from -∞ to ∞.

Sane's STATS Academy of Statistics

Parameter Calculation

- Parameters $\beta_0, \beta_1, \beta_2, \ldots \beta_m$ are calculated with the help of maximum likelihood method.
- In R, we make use of the function glm() with the following options:

Syntax :

glm(formula, data , family = binomial(link="logit", …)

Assumptions

- Logistic regression does **not** make many of the key assumptions of linear regression and general linear models - particularly regarding
  - Linearity
  - Normality
  - Homoscedasticity
- So we can apply logistic regression to any data for which we have categorical response and mixture of categorical and numerical predictors

Example

- We consider here a dataset given at Kaggle (https://www.kaggle.com/ludobenistant/hr-analytics) with the following variables:
  - satisfaction_level : Employee satisfaction level
  - last_evaluation : Last evaluation
  - number_project : Number of projects
  - average_montly_hours : Average monthly hours
  - time_spend_company : Time spent at the company
  - work_accident : Whether they have had a work accident (0=No, 1=Accident)
  - promotion_last_5years : Whether they have had a promotion in the last 5 years
  - sales : Department (Categorical)
  - salary : Relative Level of Salary (Categorical)
  - left : Whether the employee has left (Response Variable)
- Here we want to build a model for the response variable left.

Sane's STATS Academy of Statistics

R Program & Output – With only numeric variable

```
fit.lg <- glm(left ~ satisfaction_level , data = hr , family = binomial())
summary(fit.lg)
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.97388    0.04935   19.73   <2e-16 ***
satisfaction_level -3.83248    0.08720  -43.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The equation in this case, the following will be the equation,

$$P(Left = \text{"Left"}) = \frac{1}{1+e^{0.97388-3.832448*satisfaction\_level}}$$

Sane's STATS Academy of Statistics

R Program & Output – With only numeric variable

```
testdf <- data.frame(satisfaction_level = c(0,0.25,0.5,0.75,1))
pred.lg <- predict.glm(fit.lg , newdata = testdf, type = "response")
testdf <- data.frame(testdf, pred.lg)
```

- Now, we plug in some values 0,0.25,0.5, 0.75 and 1 to the satisfaction_level component in the equation and following are the values obtained

satisfaction_level	pred.lg
0.00	0.72589294
0.25	0.50394089
0.50	0.28042469
0.75	0.13005462
1.00	0.05423869

- Here, we see that as satisfaction_level increases from 0 to 1, the probability of employee leaving the company decreases.

R Program & Output – With only categorical variable

```
fit.lg <- glm(left ~ Work_accident , data = hr , family = binomial())
summary(fit.lg)
```

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.01932    0.02000  -50.97   <2e-16 ***
Work_accidentHappened -1.45168    0.08257  -17.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The equation in this case, the following will be the equation,

$$P(Left = \text{"Left"}) = \frac{1}{1+e^{-1.01932-1.45168*Work\_accidentHappened}}$$

Sane's STATS Academy of Statistics

R Program & Output – With only numeric variable

```
testdf <- data.frame(Work_accident = factor(c(0,1), levels = c(0,1),
                                            labels = c("Not Happened","Happened")))
pred.lg <- predict.glm(fit.lg , newdata = testdf, type = "response")
testdf <- data.frame(testdf, pred.lg)
```

- Now, we plug in some values 0 and 1 to the Work_accident component in the equation and following are the values obtained

Work_accident	pred.lg
Not Happened	0.26515978
Happened	0.07791609

- Here, we see that as the employee with whom the work accident not happened, the probability of employee leaving the company is more than that of if it happenned.

Sane's STATS Academy of Statistics

R Program & Output – With multiple variables

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.4762862  0.1938373  -7.616 2.61e-14 ***
satisfaction_level       -4.1356889  0.0980538 -42.178  < 2e-16 ***
last_evaluation           0.7309032  0.1491787   4.900 9.61e-07 ***
number_project           -0.3150787  0.0213248 -14.775  < 2e-16 ***
average_montly_hours      0.0044603  0.0005161   8.643  < 2e-16 ***
time_spend_company        0.2677537  0.0155736  17.193  < 2e-16 ***
Work_accidentHappened    -1.5298283  0.0895473 -17.084  < 2e-16 ***
promotion_last_5years    -1.4301364  0.2574958  -5.554 2.79e-08 ***
saleshr                   0.2323779  0.1313084   1.770  0.07678 .
salesIT                  -0.1807179  0.1221276  -1.480  0.13894
salesmanagement          -0.4484236  0.1598254  -2.806  0.00502 **
salesmarketing           -0.0120882  0.1319304  -0.092  0.92700
salesproduct_mng         -0.1532529  0.1301538  -1.177  0.23901
salesRandD               -0.5823659  0.1448848  -4.020 5.83e-05 ***
salessales               -0.0387859  0.1024006  -0.379  0.70486
salessupport              0.0500251  0.1092834   0.458  0.64713
salestechnical            0.0701464  0.1065379   0.658  0.51027
salarylow                 1.9440627  0.1286272  15.114  < 2e-16 ***
salarymedium              1.4132244  0.1293534  10.925  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sane's STATS Academy of Statistics

Calculating Odds Ratios

```
> exp(coef(fit.lg))
          (Intercept)    satisfaction_level       last_evaluation        number_project
           0.22848466            0.01599164            2.07695560            0.72973146
 average_montly_hours    time_spend_company Work_accidentHappened promotion_last_5years
           1.00447026            1.30702513            0.21657284            0.23927628
              saleshr               salesIT       salesmanagement         salesmarketing
           1.26159637            0.83467078            0.63863409            0.98798460
     salesproduct_mng              salesRandD              salessales          salessupport
           0.85791269            0.55857528            0.96195663            1.05129748
       salestechnical               salarylow            salarymedium
           1.07266519            6.98708012            4.10918362
```

- The function coef() extracts the coefficients and exp() function calculates the odds of success(1) class over the failure(0) class.

Sane's STATS Academy of Statistics

Interpreting Odds

```
> exp(coef(fit.lg))
          (Intercept)     satisfaction_level        last_evaluation         number_project
            0.22848466             0.01599164             2.07695560             0.72973146
 average_montly_hours    time_spend_company Work_accidentHappened promotion_last_5years
            1.00447026             1.30702513             0.21657284             0.23927628
               saleshr                salesIT       salesmanagement        salesmarketing
            1.26159637             0.83467078             0.63863409             0.98798460
     salesproduct_mng            salesRandD             salessales          salessupport
            0.85791269             0.55857528             0.96195663             1.05129748
       salestechnical              salarylow          salarymedium
            1.07266519             6.98708012             4.10918362
```

- Let us interpret odds for some of the variables
  - average_montly_hours: 1.00447026 indicates that probability of employee leaving the company is 1.00447026 times more than probability of employee not leaving the company with the average_montly_hours variable increased by 1, keeping all other values constant.
  - Work_accidentHappened : 0.21657284 indicates that probability of employee with work accident taken place leaving the company is 1/0.21657284 = 4.617384 times less than probability of employee not leaving the company with the Work accident not taken place. That means, that with work accident taking place, employees are not willing to leave.

Predicting on Logistic Regression

```
library(caret)
set.seed(1992)
intrain<-createDataPartition(y=hr$left , p=0.7,list=FALSE)
training   <- hr[ intrain , ]
validation <- hr[-intrain , ]
fit.lg <- glm(left ~ . , data = training , family = binomial())
pred.lg <- predict(fit.lg, newdata = validation , type = "response")
pred.lg.cat <- factor(ifelse(pred.lg < 0.5 , "Stayed" , "Left"),
                      levels = c("Stayed","Left"))

confusionMatrix(pred.lg.cat , validation$left)
```

```
Confusion Matrix and Statistics

          Reference
Prediction Stayed Left
    Stayed   3180  718
    Left      248  353

               Accuracy : 0.7853
                 95% CI : (0.773, 0.7972)
    No Information Rate : 0.7619
    P-Value [Acc > NIR] : 0.0001086
```

Sane's STATS Academy of Statistics