Training Multinomial Regression Model in R

Merve Topdemir

25 03 2021

This study aims to examine the relationship between drug choice of patients and characteristics such as age, sex and blood pressure levels using multinomial logistic regression. The "drug 200" data set were used in this study. The data set contains various information that effect the predictions like Age, Sex, BP, Cholesterol levels, Na to Potassium Ratio and finally the drug type.

Contents

Importing Data 1
Data Structure 2
Training a MLRM 3
Significant of The Features 4
Interpreting The Model Features 4
Predicted Probabilities of The Target Variable 5
Splitting Data 5
Train a MLRM on Train Set 6
Performance of The Model on Train And Test Set 7
Confusion Matrix 7

Importing Data

```
# Importing data into R

data <-read.csv("C:/Users/Merve/Desktop/drug200.csv",header =TRUE)
head(data,10)
```

```
##    Age  Sex    BP     Cholesterol  Na_to_K    Drug
## 1   23    F    HIGH        HIGH     25.355    DrugY
## 2   47    M    LOW         HIGH     13.093    drugC
## 3   47    M    LOW         HIGH     10.114    drugC
## 4   28    F    NORMAL      HIGH     7.798     drugX
## 5   61    F    LOW         HIGH     18.043    DrugY
## 6   22    F    NORMAL      HIGH     8.607     drugX
## 7   49    F    NORMAL      HIGH     16.275    DrugY
## 8   41    M    LOW         HIGH     11.037    drugC
## 9   60    M    NORMAL      HIGH     15.171    DrugY
## 10  43    M    LOW       NORMAL     19.368    DrugY
```

```
table(data$Drug) #To see the rate of drug variable
```

```
##
## drugA drugB drugC drugX DrugY
##    23    16    16    54    91
```

According to the observed values, there may be an imbalance problem.

Data Structure

The data set contains 200 observations and 6 variables.

Dependent (target) variable:

- Drug type

Independent (feature) variables:

- Age
- Sex
- Blood Pressure Levels (BP)
- Cholesterol Levels
- Na to Potassium Ration

```
str(data)
```

```
## 'data.frame':    200 obs. of  6 variables:
##  $ Age        : num  23 47 47 28 61 22 49 41 60 43 ...
##  $ Sex        : Factor w/ 2 levels "F","M": 1 2 2 1 1 1 1 2 2 2 ...
##  $ BP         : Factor w/ 3 levels "HIGH","LOW","NORMAL": 1 2 2 3 2 3 3 2 3 2 ...
##  $ Cholesterol: Factor w/ 2 levels "HIGH","NORMAL": 1 1 1 1 1 1 1 1 1 2 ...
##  $ Na_to_K    : num  25.4 13.1 10.1 7.8 18 ...
##  $ Drug       : Factor w/ 5 levels "drugA","drugB",..: 5 3 3 4 5 4 5 3 5 5 ...
```

```
summary(data)
```

```
##       Age          Sex       BP       Cholesterol      Na_to_K          Drug
##  Min.   :15.00   F: 96   HIGH  :77   HIGH  :103   Min.   : 6.269   drugA:23
##  1st Qu.:31.00   M:104   LOW   :64   NORMAL: 97   1st Qu.:10.445   drugB:16
##  Median :45.00           NORMAL:59                Median :13.937   drugC:16
##  Mean   :44.31                                    Mean   :16.084   drugX:54
##  3rd Qu.:58.00                                    3rd Qu.:19.380   DrugY:91
##  Max.   :74.00                                    Max.   :38.247
```

Factors are used to represent categorical data. Numerical variable is one that may take on any value within a finite or infinite interval.

- Sex is a factor variable, it has 2 levels, 96 females and 104 males.
- Blood Pressure is a factor variable, it has 3 levels, 77 high ,64 low and 59 normal.
- Cholesterol is a factor variable, it has 2 levels, 103 high and 97 normal.
- Drug is a factor variable, it has 5 levels, 23 drugA, 16 drugB, 16 drugC, 54 drugX and 91 DrugY.
- Age and Na_to_K are numerical variables.
- Age variable minimum value is 15, maximum value is 74, median is 45, mean is 44.31, first quartile is 31 and third quartile is 58. Na_to_K minimum value is 6.269, maximum value is 38.247, median is 13.937, mean is 16.084, first quartile is 10.445 and third quartile is 19.380.

Training a MLRM

Hypothesis

$H_0 = \beta_j = 0$ (All features is equal to zero)

$H_1 = \beta_j \neq 0$ (at least one of the features is different than zero)

```
#We will use drugX as a reference group.

data$Drug <- relevel(data$Drug, ref = "drugX")
#install.packages("nnet")

# Load the multinom package
library(nnet)
model <- multinom(Drug ~ ., data = data, trace = FALSE)
summary(model)

## Call:
## multinom(formula = Drug ~ ., data = data, trace = FALSE)
##
## Coefficients:
##       (Intercept)     Age       SexM     BPLOW    BPNORMAL    CholesterolNORMAL
## drugA   566.44559 -4.6634546  82.77775 -438.2499 -600.6267      -197.24762
## drugB   -69.86273  4.6141006  11.62788 -441.4286 -498.8625      -104.75904
## drugC   358.69295 -1.2509717  33.68212 -196.5789 -376.0936      -170.77067
## DrugY  -651.54070  0.2450448  61.98418 -251.6164 -342.7450       -72.00648
##         Na_to_K
## drugA -5.754367
## drugB  8.820351
## drugC -6.163332
## DrugY 63.466173
##
## Std. Errors:
##       (Intercept)    Age          SexM        BPLOW     BPNORMAL
## drugA   19.052119  739.1417 1.900795e+01 4.147925e-04 6.619479e-12
## drugB   19.338302  747.8732 7.875730e+00 1.811985e-50 1.305081e-54
## drugC    9.645832  541.3845 4.147925e-04 9.645832e+00 1.128745e-16
## DrugY  656.623434 1516.2045 6.455989e+02 3.017525e+03 2.393863e+03
```

```
##       CholesterolNORMAL   Na_to_K
## drugA      1.905159e+01  282.5283
## drugB      1.904925e+01  178.9801
## drugC      2.948798e-13  111.1163
## DrugY      7.672012e+02 4821.5404
## Residual Deviance: 8.430093e-05
## AIC: 56.00008
```

Significant of The Features

```
z <- summary(model)$coefficients/summary(model)$standard.errors
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
```

```
##       (Intercept)    Age          SexM      BPLOW    BPNORMAL  CholesterolNORMAL
## drugA 0.000000000 0.9949660 1.331263e-05  0.0000000  0.0000000   0.000000e+00
## drugB 0.000303071 0.9950774 1.398314e-01  0.0000000  0.0000000   3.811331e-08
## drugC 0.000000000 0.9981563 0.000000e+00  0.0000000  0.0000000   0.000000e+00
## DrugY 0.321071056 0.9998710 9.235123e-01  0.9335454  0.8861507   9.252235e-01
##         Na_to_K
## drugA 0.9837503
## drugB 0.9606952
## drugC 0.9557661
## DrugY 0.9894977
```

Interpreting The Model Features

```
exp(coef(model))
```

```
##         (Intercept)       Age           SexM           BPLOW        BPNORMAL
## drugA  1.009705e+246  9.433816e-03  8.910828e+35  4.682715e-191  1.416236e-261
## drugB  4.560401e-31   1.008970e+02  1.121822e+05  1.949903e-192  2.222236e-217
## drugC  6.002992e+155  2.862265e-01  4.245794e+14  4.235156e-86   4.619996e-164
## DrugY  1.095136e-283  1.277678e+00  8.305938e+26  5.301341e-110  1.405145e-149
##       CholesterolNORMAL      Na_to_K
## drugA  2.169946e-86        3.168912e-03
## drugB  3.189525e-46        6.770643e+03
## drugC  6.842912e-75        2.105226e-03
## DrugY  5.345452e-32        3.656021e+27
```

Predicted Probabilities of The Target Variable

```
predicted_probs <- predict(model, type = "probs")
head(predicted_probs,10)
```

```
##          drugX         drugA         drugB         drugC         DrugY
## 1   0.000000e+00  5.132545e-283 3.361849e-306  0.000000e+00  1.000000e+00
## 2   3.527665e-25  1.827354e-61  7.641011e-98   1.000000e+00  1.303218e-24
## 3   3.746230e-33  5.404039e-62  3.146383e-117  1.000000e+00  1.073795e-114
## 4   1.000000e+00  9.092325e-92  9.674785e-162  2.280188e-44  1.268631e-214
## 5   2.661117e-112 2.914811e-225 5.329907e-143  2.476768e-123 1.000000e+00
## 6   1.000000e+00  1.226837e-81  1.151729e-170  2.833088e-43  5.798021e-193
## 7   1.023733e-22  1.788599e-177 3.544254e-109  1.856392e-100 1.000000e+00
## 8   6.087641e-34  6.148776e-53  1.663669e-126  1.000000e+00  1.106378e-90
## 9   2.237291e-20  1.053230e-158 5.657592e-84   1.640782e-86  1.000000e+00
## 10  1.486938e-142 4.390751e-272 1.079869e-244  6.869213e-207 1.000000e+00
```

```
max<-colnames(predicted_probs)[apply(predicted_probs, 1,which.max)]
head(max,10)
```

```
##  [1] "DrugY" "drugC" "drugC" "drugX" "DrugY" "drugX" "DrugY" "drugC" "DrugY"
## [10] "DrugY"
```

Splitting Data

The following code splits 80% of the data selected randomly into training set and the remaining 20% sample into test set.

Training set is implemented to build up a model, while a test set is to validate the model built.

```
set.seed(123)
index <- sample(nrow(data), nrow(data) * 0.8)
train <- data[index,]
test <- data[-index,]
```

```
table(train$Drug)
```

```
##
## drugX drugA drugB drugC DrugY
##    35    16    14    14    81
```

```
table(test$Drug)
```

```
##
## drugX drugA drugB drugC DrugY
##    19     7     2     2    10
```

Train a MLRM on Train Set

```
model1 <- multinom(Drug ~ ., data = train, trace = FALSE)
summary(model1)
```

```
## Call:
## multinom(formula = Drug ~ ., data = train, trace = FALSE)
##
## Coefficients:
##       (Intercept)      Age      SexM      BPLOW      BPNORMAL CholesterolNORMAL
## drugA   280.1920  -2.3071426  43.325467  -257.08164  -322.4687    -92.79819
## drugB  -117.3046   3.3861257  1.267519   -278.72160  -309.5288    -74.92179
## drugC   157.5199  -0.9822237  25.897689  -36.51092   -213.9094    -131.11024
## DrugY  -422.4722   0.6572743  43.687092  -198.48918  -244.6364    -40.75083
##         Na_to_K
## drugA -1.316207
## drugB  11.538349
## drugC -4.829829
## DrugY  40.738771
##
## Std. Errors:
##       (Intercept)      Age         SexM         BPLOW       BPNORMAL
## drugA   6777.1729  2340.41159  2.706903e+03  4.669338e-08  2.268871e-07
## drugB   7446.2663  2488.87221  3.532718e+03  3.214258e-12  3.087462e-06
## drugC   17.1421    67.15912    1.383413e-04  1.714210e+01  6.192123e-30
## DrugY   1023.4282  477.08571   7.980024e+03  3.520961e+03  3.688230e+03
##       CholesterolNORMAL    Na_to_K
## drugA   6.777182e+03     9805.0306
## drugB   5.589089e+03     10228.9388
## drugC   1.429330e-21     327.2049
## DrugY   5.502868e+03     2441.9757
##
## Residual Deviance: 0.0001094693
## AIC: 56.00011
```

Performance of The Model on Train And Test Set

```
predicted_probs_train <- predict(model1, type = "probs")
predicted_class_train <- colnames(predicted_probs_train)[apply(predicted_pr
obs_train, 1,which.max)]
mean(predicted_class_train == train$Drug)
```

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

```
predicted_probs_test <- predict(model1, test, type = "probs")
predicted_class_test <- colnames(predicted_probs_test)[apply(
predicted_probs_test, 1, which.max)]
mean(predicted_class_test == test$Drug)
```

```
## [1] 0.925
```

Confusion Matrix

```
confmatr <- table(predicted = predicted_class_test, actual = test$Drug)
confmatr
```

```
##           actual
## predicted  drugX drugA drugB drugC DrugY
##     drugA     0     4     0     0     0
##     drugB     0     1     2     0     0
##     drugC     0     0     0     2     0
##     drugX    19     0     0     0     0
##     DrugY     0     2     0     0    10
```

```
accuracy <- sum(diag(confmatr)) / sum(confmatr)
accuracy
```

```
## [1] 0.275
```