r-statistics.co (/) by Selva Prabhakaran

Search..

-

Tutorial

R Tutorial (R-Tutorial.html)

ggplot2

ggplot2 Short Tutorial (ggplot2-Tutorial-With-R.html)

ggplot2 Tutorial 1 - Intro (Complete-Ggplot2-Tutorial-Part1-With-R-Code.html)

ggplot2 Tutorial 2 - Theme (Complete-Ggplot2-Tutorial-Part2-Customizing-Theme-With-R-Code.html)

ggplot2 Tutorial 3 - Masterlist (Top50-Ggplot2-Visualizations-MasterList-R-Code.html)

ggplot2 Quickref (ggplot2-cheatsheet.html)

Foundations

Linear Regression (Linear-Regression.html)

Statistical Tests (Statistical-Tests-in-R.html)

Missing Value Treatment (Missing-Value-Treatment-With-R.html)

Outlier Analysis (Outlier-Treatment-With-R.html)

Feature Selection (Variable-Selection-and-Importance-With-R.html)

Model Selection (Model-Selection-in-R.html)

Logistic Regression (Logistic-Regression-With-R.html)

Advanced Linear Regression (Environments.html)

Advanced Regression Models

Advanced Regression Models (adv-regression-models.html)

Time Series

Time Series Analysis (Time-Series-Analysis-With-R.html)

Time Series Forecasting (Time-Series-Forecasting-With-R.html)

More Time Series Forecasting (Time-Series-Forecasting-With-R-part2.html)

High Performance Computing

Parallel computing (Parallel-Computing-With-R.html)

Strategies to Speedup R code (Strategies-To-Improve-And-Speedup-R-Code.html)

Useful Techniques

Association Mining (Association-Mining-With-R.html)

Multi Dimensional Scaling (Multi-Dimensional-Scaling-With-R.html)

Optimization (Profiling.html)

InformationValue package (Information-Value-With-R.html)

Stay up-to-date. Subscribe! (https://docs.google.com/forms/d/1xkMYkLNFU9U39Dd8S_2JC0p8B5t6_Yq6zUQjanQQJpY/viewform)

Chat! (https://docs.google.com/forms/d/13GrkCFcNa-TOIllQghsz2SIEbc-YqY9eJX02B19l5Ow/viewform)

Contents

Multinomial Regression

> Multinomial regression is much similar to logistic regression but is applicable when the response variable is a nominal categorical variable with more than 2 levels.

Introduction

Multinomial logistic regression can be implemented with `mlogit()` from mlogit package and `multinom()` from `nnet` package. We will use the latter for this example.

Example: Predict Choice of Contraceptive Method

In this example, we will try to predict the choice of contraceptive preferred by women *(1=No-use, 2=Long-term, 3=Short-term)*. We have the education, work, religion, number of children, media exposure and standard of living as variables available in the cmc data (http://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice). In this example, we will model the choice of contraceptive method `cmc` as a function of all these variables.

Import Data

```
cmcData <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-databases/cmc/cmc.da
ta", stringsAsFactors=FALSE, header=F)
colnames(cmcData) <- c("wife_age", "wife_edu", "hus_edu", "num_child", "wife_rel", "wife
_work", "hus_occu", "sil", "media_exp", "cmc")
head(cmcData)
#>   wife_age wife_edu hus_edu num_child wife_rel wife_work hus_occu sil media_exp cmc
#> 1       24        2       3         3        1         1        2   3         0   1
#> 2       45        1       3        10        1         1        3   4         0   1
#> 3       43        2       3         7        1         1        3   4         0   1
#> 4       42        3       2         9        1         1        3   3         0   1
#> 5       36        3       3         8        1         1        3   2         0   1
#> 6       19        4       4         0        1         1        3   3         0   1
```

Convert Numerics to Factors

```
cmcData$wife_edu <- factor(cmcData$wife_edu, levels=sort(unique(cmcData$wife_edu)))
cmcData$hus_edu <- factor(cmcData$hus_edu, levels=sort(unique(cmcData$hus_edu)))
cmcData$wife_rel <- factor(cmcData$wife_rel, levels=sort(unique(cmcData$wife_rel)))
cmcData$wife_work <- factor(cmcData$wife_work, levels=sort(unique(cmcData$wife_work)))
cmcData$hus_occu <- factor(cmcData$hus_occu, levels=sort(unique(cmcData$hus_occu)))
cmcData$sil <- factor(cmcData$sil, levels=sort(unique(cmcData$sil)))
cmcData$media_exp <- factor(cmcData$media_exp, levels=sort(unique(cmcData$media_exp)))
cmcData$cmc <- factor(cmcData$cmc, levels=sort(unique(cmcData$cmc)))
```

Create Training and Test Data

```r
# Prepare Training and Test Data
set.seed(100)
trainingRows <- sample(1:nrow(cmcData), 0.7*nrow(cmcData))
training <- cmcData[trainingRows, ]
test <- cmcData[-trainingRows, ]
```

Build Multinomial Model

```
library(nnet)
multinomModel <- multinom(cmc ~ ., data=training) # multinom Model
summary (multinomModel) # model summary

#> Call:
#> multinom(formula = cmc ~ ., data = training)
#>
#> Coefficients:
#>   (Intercept)    wife_age  wife_edu2 wife_edu3 wife_edu4  hus_edu2  hus_edu3
#> 2  -1.5937363 -0.04360644 1.07871567 2.0445226  2.835641 -1.407238 -1.268765
#> 3   0.4376064 -0.10923832 0.03095292 0.4308403  0.979347  1.073331  1.150374
#>     hus_edu4 num_child  wife_rel1 wife_work1  hus_occu2   hus_occu3 hus_occu4
#> 2 -1.3102661 0.3060657 -0.4455628  0.1165996 -0.4943500 -0.40723995 1.2664442
#> 3  0.8607095 0.3376620 -0.2072181  0.3427517 -0.1950799  0.04609764 0.5596847
#>         sil2      sil3      sil4 media_exp1
#> 2 0.81445361 1.2655842 1.3311827 -0.2440084
#> 3 0.03657688 0.3155116 0.5562075 -0.9285685

#> Std. Errors:
#>   (Intercept)   wife_age wife_edu2 wife_edu3 wife_edu4  hus_edu2  hus_edu3
#> 2   0.9964378 0.01485064 0.5520832 0.5649966 0.5834594 0.6270468 0.5823429
#> 3   0.9225193 0.01400097 0.3181759 0.3368472 0.3629088 0.6885676 0.6837955
#>    hus_edu4  num_child wife_rel1 wife_work1 hus_occu2 hus_occu3 hus_occu4
#> 2 0.5886178 0.05094430 0.2391401  0.2001434 0.2473945 0.2444405 0.6986301
#> 3 0.6915629 0.04595659 0.2373718  0.1814554 0.2302729 0.2226137 0.6189151
#>        sil2      sil3      sil4 media_exp1
#> 2 0.5462033 0.5229496 0.5268553  0.4951397
#> 3 0.3106383 0.2907037 0.2943716  0.3819526
#>
#> Residual Deviance: 1930.658
#> AIC: 2002.658
```

Predict on Test Data

```
predicted_scores <- predict (multinomModel, test, "probs") # predict on new data
#>              1          2          3
#> 6    0.2699230 0.18691129 0.54316572
#> 9    0.3626476 0.08523814 0.55211422
#> 10   0.7564912 0.19409005 0.04941879
#> 12   0.7680439 0.05851352 0.17344257
#> 14   0.8961808 0.04747638 0.05634281
#> 17   0.6677357 0.23683800 0.09542632
#> .
#> .
#> 1464 0.5523515 0.02851988 0.4191287
#> 1471 0.1816340 0.41055467 0.4078114
#> 1472 0.5369837 0.16864237 0.2943739

predicted_class <- predict (multinomModel, test)
#> [1] 3 3 1 1 1 1
#> Levels: 1 2 3
```

Confusion Matrix and Misclassification Error

```
table(predicted_class, test$cmc)
#> predicted_class   1   2   3
#>               1 112  26  58
#>               2  19  37  21
#>               3  55  39  75

mean(as.character(predicted_class) != as.character(test$cmc))
#=> 0.4932127
```

A misclassification error of 49.3% is probably too high. May be it can be improved by improving the model terms or may be the variables are not as good in explaining the contraceptive method used. Either ways, I would encourage the investigator to try other ML approaches as well for this problem.