Logistic Regression

Dian Ramadhani

08/01/2019

Logistic Regression

Regresi logistik adalah model prediksi seperti halnya regresi linear, namun dengan variabel terikat berskala dikotomi.

Install Packages

```
# Menginstall package(s)
install.packages("readr") # membaca file
install.packages("here") # menampilkan direktori
install.packages("tidyverse") # manipulasi data
```

Import Library

```
# Mengaktifkan package(s)
library(readr)
library(here)
library(tidyverse)
```

Menampilkan Direktori

```
# Mengetahui direktori proyek
here()
```

Import Data

```
# Mengimport data
df.insurance <- read_csv(here("data", "raw", "logreg_insurance.csv"))
```

```
## Parsed with column specification:
## cols(
##   Age = col_double(),
##   Sex = col_character(),
##   Bmi = col_double(),
##   Children = col_double(),
##   Smoker = col_character(),
##   Region = col_character(),
##   Charges = col_double(),
##   Claim = col_character()
## )
```

Data yang digunakan yaitu data asuransi. Data ini berisi tentang profil calon konsumen perusahaan asuransi dan keputusan pengambilan asuransi yang diambil masing masing orang.

Eksplorasi Data

Data yang telah diimpor selanjutnya dieksplorasi untuk mengetahui strukturnya.

```r
# Melihat attribute dan struktur data
names(df.insurance) # menampilkan nama kolom
```

```
## [1] "Age"      "Sex"      "Bmi"      "Children" "Smoker"   "Region"
## [7] "Charges"  "Claim"
```

```r
dim(df.insurance) # menampilkan dimensi tabel
```

```
## [1] 1338    8
```

```r
head(df.insurance) # menampilkan beberap data teratas
```

```
## # A tibble: 6 x 8
##     Age Sex      Bmi Children Smoker     Region    Charges Claim
##   <dbl> <chr>  <dbl>    <dbl> <chr>      <chr>       <dbl> <chr>
## 1    19 Female  27.9        0 Smoker     Southwest  16885. Yes
## 2    18 Male    33.8        1 Non Smoker Southeast   1726. Yes
## 3    28 Male    33          3 Non Smoker Southeast   4449. No
## 4    33 Male    22.7        0 Non Smoker Northwest  21984. No
## 5    32 Male    28.9        0 Non Smoker Northwest   3867. Yes
## 6    31 Female  25.7        0 Non Smoker Southeast   3757. No
```

```r
str(df.insurance) # menampilkan struktur data
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 1338 obs. of  8 variables:
##  $ Age     : num  19 18 28 33 32 31 46 37 37 60 ...
##  $ Sex     : chr  "Female" "Male" "Male" "Male" ...
##  $ Bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ Children: num  0 1 3 0 0 0 1 3 2 0 ...
##  $ Smoker  : chr  "Smoker" "Non Smoker" "Non Smoker" "Non Smoker" ...
##  $ Region  : chr  "Southwest" "Southeast" "Southeast" "Northwest" ...
##  $ Charges : num  16885 1726 4449 21984 3867 ...
##  $ Claim   : chr  "Yes" "Yes" "No" "No" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Age = col_double(),
##   ..   Sex = col_character(),
##   ..   Bmi = col_double(),
##   ..   Children = col_double(),
##   ..   Smoker = col_character(),
##   ..   Region = col_character(),
##   ..   Charges = col_double(),
##   ..   Claim = col_character()
##   .. )
```

```r
summary(df.insurance) # menampilkan rangkuman data
```

```
##       Age            Sex                 Bmi           Children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     Smoker             Region             Charges         Claim
##  Length:1338        Length:1338        Min.   : 1122   Length:1338
##  Class :character   Class :character   1st Qu.: 4740   Class :character
```

```
##  Mode  :character   Mode  :character   Median : 9382   Mode  :character
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

```r
# Mengetahui jumlah data kosong
sum(is.na(df.insurance))
```

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

Mengubah Data Char menjadi Angka

Kita akan menggunakan fungsi glm() dir r dimana hanya menerima input berupa numerik, sehingga target variable yang pada awalnya No dan Yes harus dirubah menjadi 0 dan 1.

```r
# Mengubah target Variable ke 0 dan 1  (No = 0, Yes = 1)
df.insurance2 <- df.insurance %>%
  mutate(Claim = ifelse(Claim == "No", 0, 1))
```

Membagi Data

Dalam regresi logistik, data dibagi menjadi dua yaitu data train untuk membuat model dan data test untuk menguji akurasi model. Biasanya data dibagi dengan proporsi 70% train dan 30% test.

```r
# Membangi data
split <- sample(1:nrow(df.insurance2), 0.7 * nrow(df.insurance2))

# Membuat tabel data train
df.train <- df.insurance2[split, ]

# Membut tabel data test
df.test <- df.insurance2[-split, ]
```

Model Building

```r
# Membuat model regresi logistik
reglog <- glm(Claim ~ ., data = df.train)

# Melihat hasil model regresi logistik
reglog
```

```
##
## Call:  glm(formula = Claim ~ ., data = df.train)
##
## Coefficients:
##     (Intercept)              Age          SexMale              Bmi
##      -5.062e-01        5.776e-03       -2.262e-02        3.539e-02
##        Children     SmokerSmoker  RegionNorthwest  RegionSoutheast
##      -1.591e-01        5.669e-01       -5.639e-02       -9.178e-02
## RegionSouthwest          Charges
##      -6.587e-02       -6.814e-06
##
## Degrees of Freedom: 935 Total (i.e. Null);  926 Residual
## Null Deviance:       224.6
## Residual Deviance: 123.9     AIC: 785.8
```

```r
summary(reglog)
```

```
##
## Call:
## glm(formula = Claim ~ ., data = df.train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.81050  -0.32147   0.01223   0.28876   1.28009
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.062e-01  7.659e-02  -6.609 6.52e-11 ***
## Age              5.776e-03  9.994e-04   5.780 1.02e-08 ***
## SexMale         -2.262e-02  2.402e-02  -0.942  0.34644
## Bmi              3.539e-02  2.195e-03  16.119  < 2e-16 ***
## Children        -1.591e-01  9.842e-03 -16.162  < 2e-16 ***
## SmokerSmoker     5.669e-01  5.848e-02   9.695  < 2e-16 ***
## RegionNorthwest -5.639e-02  3.482e-02  -1.619  0.10571
## RegionSoutheast -9.178e-02  3.435e-02  -2.672  0.00768 **
## RegionSouthwest -6.587e-02  3.485e-02  -1.890  0.05908 .
## Charges         -6.814e-06  2.076e-06  -3.282  0.00107 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1338352)
##
##     Null deviance: 224.56  on 935  degrees of freedom
## Residual deviance: 123.93  on 926  degrees of freedom
## AIC: 785.77
##
## Number of Fisher Scoring iterations: 2
```

Validasi

```r
# Memprediksi data test
pred.log <- predict(reglog, df.test)

# Menentukan cut off
cutoff <- 0.5
pred.log.class <- ifelse(pred.log > cutoff, 1, 0)

# Menampilkan hasil prediksi
pred.log.class
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##   0   0   1   0   0   0   1   1   0   1   1   0   1   0   1   1   0   1
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##   1   1   1   0   1   1   1   0   1   1   0   0   1   0   0   1   1   1
##  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##   0   1   1   1   0   1   0   0   0   1   0   1   1   1   1   1   0   0
##  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##   0   1   1   0   1   0   0   1   1   1   1   0   0   1   1   1   1   0
##  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
```

```
##   1   0   0   1   0   1   1   0   0   1   0   0   1   1   1   1   1   0
##  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
##   1   1   0   1   0   0   1   1   1   1   1   1   1   0   1   1   1   0
## 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
##   0   1   1   0   1   1   0   0   0   0   1   0   1   0   1   1   0   0
## 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
##   1   0   1   0   0   0   1   1   1   0   0   0   1   1   0   1   1   1
## 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162
##   1   1   0   1   1   1   1   0   1   0   0   1   1   1   0   1   0   1
## 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
##   1   1   1   1   1   0   1   1   1   0   0   0   0   1   0   1   1   0
## 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198
##   0   0   0   1   0   1   1   0   1   0   1   1   0   0   1   1   1   0
## 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216
##   1   1   1   1   0   1   0   0   0   0   1   1   1   1   0   1   1   1
## 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234
##   1   0   0   0   0   1   1   1   1   1   1   0   1   0   0   1   1   1
## 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252
##   0   1   1   0   1   0   1   1   0   1   1   0   0   1   1   0   1   0
## 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270
##   0   1   0   0   0   1   1   0   0   1   0   0   1   1   0   1   1   0
## 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288
##   1   1   0   0   1   1   1   0   0   1   0   0   1   0   0   0   1   1
## 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306
##   1   1   0   0   1   0   0   1   1   1   1   0   0   1   1   0   1   1
## 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324
##   1   1   0   1   0   0   0   0   1   0   0   1   0   1   0   1   0   0
## 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342
##   1   1   0   1   1   0   1   1   1   1   1   0   0   1   0   0   0   0
## 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360
##   0   0   1   1   0   1   0   1   1   0   1   1   0   1   0   1   1   0
## 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378
##   1   0   1   1   1   1   1   0   0   1   1   0   0   0   1   0   1   1
## 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396
##   0   1   1   1   1   1   1   0   0   1   1   0   0   1   1   1   0   0
## 397 398 399 400 401 402
##   1   1   0   0   1   1
```

```
# Melihat prediksi dalam bentuk tabel
df.pred.log <- data.frame(df.test, pred.log.class)
View(df.pred.log)
```

```
# Confussion matrix
conf.log <- table(df.test$Claim, pred.log.class)
conf.log
```

```
##    pred.log.class
##       0   1
##   0 149  32
##   1  24 197
```

```
# Megambil angka TP, FN, FP, TN
TP.log <- conf.log[1, 1]
FN.log <- conf.log[1, 2]
FP.log <- conf.log[2, 1]
TN.log <- conf.log[2, 2]
```

```r
# Menghitung nilai akurasi
acc.log <- (TP.log + TN.log) / (TP.log + FN.log + FP.log + TN.log)
acc.log
```

```
## [1] 0.8606965
```

```r
# Menghitung nilai presisi
prec.log <- TP.log / (TP.log + FP.log)
prec.log
```

```
## [1] 0.8612717
```

```r
# Menghitung Nilai Recall
rec.log <- TP.log / (TP.log + FN.log)
rec.log
```

```
## [1] 0.8232044
```