Preprocessing Titanic Dataset

Guillem Fernández Pallarès i Miquel Tomé Carreño

20 de maig, 2021

Índex

1 Descripció del dataset

El dataset escollit ha estat *Titanic: Machine Learning from Disaster* disponible a aquest enllaç. Amb aquest joc de dades es pretenen aplicar algoritmes de Machine Learning posteriors al preprocessat de les dades per crear models predictius que permetin construir un model que prevegui, en funció d'unes variables determinades, un determinat passatger sobreviurà o no al conegut accident.

2 Integració i selecció de dades

Les dades utilitzades es troben divididas en dos datasets diferents: un d'ells conté el subconjunt de dades que serà utilitzat al set d'entrenament o *training set* i l'altre conté aquelles que seran utilitzades al test de prova o *testing set* per comprovar l'eficàcia del model construït. Els dos subconjunts s'integraran en un de sol per dur a terme el preprocessat de les dades.

Primerament, es llegiran ambdós fitxers i s'afegirà una columna a cadascun d'ells que indicarà si un determinat registre pertany al subconjunt d'entrenament o de prova. Addicionalment, cal esmentar que al subset de prova s'ha afegit la columna `Survived` amb valors perduts `NA`, que és la variable dicotòmica a predir i que està present a l'altre subset.

```
# Lectura del training set.
r_train <- read.csv("train.csv")
summary(r_train)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
```

```
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

```
# Lectura del testing set i addició de la variable a predir.
r_test <- read.csv("test.csv")
r_test$Survived <- NA
summary(r_test)
```

```
##   PassengerId         Pclass          Name               Sex
##  Min.   : 892.0   Min.   :1.000   Length:418         Length:418
##  1st Qu.: 996.2   1st Qu.:1.000   Class :character   Class :character
##  Median :1100.5   Median :3.000   Mode  :character   Mode  :character
##  Mean   :1100.5   Mean   :2.266
##  3rd Qu.:1204.8   3rd Qu.:3.000
##  Max.   :1309.0   Max.   :3.000
##
##       Age            SibSp            Parch           Ticket
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.0000   Length:418
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :27.00   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   :30.27   Mean   :0.4474   Mean   :0.3923
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :76.00   Max.   :8.0000   Max.   :9.0000
##  NA's   :86
##       Fare            Cabin             Embarked         Survived
##  Min.   :  0.000   Length:418         Length:418         Mode:logical
##  1st Qu.:  7.896   Class :character   Class :character   NA's:418
##  Median : 14.454   Mode  :character   Mode  :character
##  Mean   : 35.627
##  3rd Qu.: 31.500
##  Max.   :512.329
##  NA's   :1
```

```r
# Addició de la columna amb la classificació train-test.
r_train$train_test <- "train"
r_test$train_test <- "test"

# Concatenació dels dos subsets.
data <- rbind(r_train,
              r_test)

# Comprovació que s'ha concatenat correctament.
nrow(r_train) + nrow(r_test) == nrow(data)
```

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

Pel que fa a la selecció de les dades, es trindran en compte la totalitat de registres dels quals es disposa, els quals seran considerats durant la fase de preprocessat de les dades.

3 Neteja de les dades

3.1 Valors perduts

3.2 Valors extrems o *outliers*