Relazione dati Netflix

Aldo Solari, matricola 2575

Relazione

L'obiettivo dell'analisi è di prevedere la valutazione (*rating*) di 2931 utenti del test set per il film Miss Detective.

Il modello utilizzato per ottenere le previsioni finali è un semplice modello lineare

$$\hat{y}_i^* = \hat{\beta}_0 + \sum_{j=1}^{99} \hat{\beta}_j x_{ij}^* + \hat{\beta}_{100} z_i^*$$

dove

- x_{ij}^* è la valutazione dell'utente i-simo sul film j-simo del test set (e vale 0 se il dato è mancante);
- z_i^* è una variabile indicatrice che vale 1 se l'utente i-simo del test set assegna mediamente valutazioni più alte ai film di genere romantico o drammatico rispetto ad altri generi di film, altrimento vale 0; questa variabile è stata creata utilizzando un dataset esterno;
- $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{100}$ sono le stime dei coefficienti del modello ottenute sulla base del training set.

Sintesi del processo di modellizzazione

1. *Pre-elaborazione dei dati*

Non è stata effettuata alcuna pre-elaborazione dei dati.

2. *Dati mancanti*

Non è stato effettuato alcun trattamento dei dati mancanti. Visto che nel dataset originale i dati mancanti sono codificati con il valore 0, sono stati trattati come valori numerici nel modello lineare.

3. *Feature engineering*

Utilizzando il dataset esterno messo a disposizione dal gruppo DEFINETTI, è stata creata una variabile indicatrice che vale 1 se l'utente assegna mediamente valutazioni più alte ai film di genere romantico o drammatico rispetto ad altri generi di film, altrimenti vale 0.

4. *Feature selection*

Non è stata effettuata alcuna selezione dei predittori.

5. *Dati esterni*

E' stato utilizzato il dataset messo a disposizione dal gruppo DEFINETTI.

6. *Modelli*

Non sono stati considerati altri modelli oltre al modello lineare utilizzato per la previsione finale (per il quale non erano previsti parametri di *tuning*).

Codice riproducibile

Includere **solo** il codice indispensabile per ottenere la previsione finale, e visualizzare i primi valori previsti con `head(yhat)`.

```
PATH <- "https://raw.githubusercontent.com/aldosolari/DM/master/docs/hw/"
X.tr = read.table(paste(PATH,"Train_ratings_all.dat", sep=""))
y.tr = read.table(paste(PATH,"Train_y_rating.dat", sep=""))
train = data.frame(X.tr, y=y.tr$V1)
X.te = read.table(paste(PATH,"Test_ratings_all.dat", sep=""))
test = data.frame(X.te)
definetti <- read.csv(paste0(PATH,"definetti.csv"))
genre <- definetti$princ_genre2[-100]
rd <- rep(0,length(genre))
rd[(genre=="Romance" | genre=="Drama")] <- 1
m.tr <- apply(X.tr,1,function(x) mean(x[x!=0]))
m1.tr <- apply(X.tr,1,function(x) sum(x[x!=0]*rd[x!=0])/sum(rd[x!=0]))
z.tr <- m1.tr >= m.tr
train$z <- z.tr
m.te <- apply(X.te,1,function(x) mean(x[x!=0]))
m1.te <- apply(X.te,1,function(x) sum(x[x!=0]*rd[x!=0])/sum(rd[x!=0]))
z.te <- m1.te >= m.te
test$z <- z.te
fit = lm(y~.,data=train)
yhat = pmin(predict(fit, newdata=test),5)
head(yhat)
```

```
##        1        2        3        4        5        6
## 3.948703 3.470572 3.676288 3.272309 3.499158 3.772120
```

```
# file .TXT sottomesso
# write.table(file="2575.txt", yhat, row.names = FALSE, col.names = FALSE)
```

```r
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.7
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.0.2  magrittr_1.5    tools_4.0.2     htmltools_0.5.0
##  [5] yaml_2.2.1      stringi_1.5.3   rmarkdown_2.4.2 knitr_1.30
##  [9] stringr_1.4.0   xfun_0.18       digest_0.6.25   rlang_0.4.8
## [13] evaluate_0.14
```