Lab12 - Clustering

Machine Learning usando o R - Análise Macro

Thalles Quinaglia Liduares

2022-08-25

Upload base de dados

```
setwd("C:\\Program Files\\R\\Dados\\ML")

data<-read.csv("snsdata.csv")

attach(data)
```

Upload pacotes

```
library(tidyverse)
library(gridExtra)
```

Análise dos dados

```
str(data)
```

```
## 'data.frame':    30000 obs. of  40 variables:
##  $ gradyear    : int  2006 2006 2006 2006 2006 2006 2006 2006 2006 2006 ...
##  $ gender      : chr  "M" "F" "M" "F" ...
##  $ age         : num  19 18.8 18.3 18.9 19 ...
##  $ friends     : int  7 0 69 0 10 142 72 17 52 39 ...
##  $ basketball  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ football    : int  0 1 1 0 0 0 0 0 0 0 ...
##  $ soccer      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ softball    : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ volleyball  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ swimming    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cheerleading: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ baseball    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ tennis      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ sports      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cute        : int  0 1 0 1 0 0 0 0 0 1 ...
##  $ sex         : int  0 0 0 0 1 1 0 2 0 0 ...
##  $ sexy        : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ hot         : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ kissed      : int  0 0 0 0 5 0 0 0 0 0 ...
##  $ dance       : int  1 0 0 0 1 0 0 0 0 0 ...
##  $ band        : int  0 0 2 0 1 0 1 0 0 0 ...
##  $ marching    : int  0 0 0 0 0 1 1 0 0 0 ...
##  $ music       : int  0 2 1 0 3 2 0 1 0 1 ...
##  $ rock        : int  0 2 0 1 0 0 0 1 0 1 ...
##  $ god         : int  0 1 0 0 1 0 0 0 0 6 ...
##  $ church      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ jesus       : int  0 0 0 0 0 0 0 0 0 2 ...
##  $ bible       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hair        : int  0 6 0 0 1 0 0 0 0 1 ...
##  $ dress       : int  0 4 0 0 0 1 0 0 0 0 ...
##  $ blonde      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mall        : int  0 1 0 0 0 0 2 0 0 0 ...
##  $ shopping    : int  0 0 0 0 2 1 0 0 0 1 ...
##  $ clothes     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hollister   : int  0 0 0 0 0 0 2 0 0 0 ...
##  $ abercrombie : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ die         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ death       : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ drunk       : int  0 0 0 0 1 1 0 0 0 0 ...
##  $ drugs       : int  0 0 0 0 1 0 0 0 0 0 ...
```

```
summary(data)
```

```
##     gradyear       gender               age             friends
##  Min.   :2006   Length:30000       Min.   :  3.086   Min.   :  0.00
##  1st Qu.:2007   Class :character   1st Qu.: 16.312   1st Qu.:  3.00
##  Median :2008   Mode  :character   Median : 17.287   Median : 20.00
##  Mean   :2008                      Mean   : 17.994   Mean   : 30.18
##  3rd Qu.:2008                      3rd Qu.: 18.259   3rd Qu.: 44.00
##  Max.   :2009                      Max.   :106.927   Max.   :830.00
##                                    NA's   :5086
##    basketball         football           soccer           softball
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   : 0.2673   Mean   : 0.2523   Mean   : 0.2228   Mean   : 0.1612
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :24.0000   Max.   :15.0000   Max.   :27.0000   Max.   :17.0000
##
##    volleyball         swimming        cheerleading       baseball
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median : 0.0000   Median :0.0000   Median : 0.0000
##  Mean   : 0.1431   Mean   : 0.1344   Mean   :0.1066   Mean   : 0.1049
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 0.0000
##  Max.   :14.0000   Max.   :31.0000   Max.   :9.0000   Max.   :16.0000
##
##      tennis             sports           cute              sex
##  Min.   : 0.00000   Min.   : 0.00   Min.   : 0.0000   Min.   :  0.0000
##  1st Qu.: 0.00000   1st Qu.: 0.00   1st Qu.: 0.0000   1st Qu.:  0.0000
##  Median : 0.00000   Median : 0.00   Median : 0.0000   Median :  0.0000
##  Mean   : 0.08733   Mean   : 0.14   Mean   : 0.3229   Mean   :  0.2094
##  3rd Qu.: 0.00000   3rd Qu.: 0.00   3rd Qu.: 0.0000   3rd Qu.:  0.0000
##  Max.   :15.00000   Max.   :12.00   Max.   :18.0000   Max.   :114.0000
##
##       sexy              hot              kissed            dance
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   : 0.1412   Mean   : 0.1266   Mean   : 0.1032   Mean   : 0.4252
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##  Max.   :18.0000   Max.   :10.0000   Max.   :26.0000   Max.   :30.0000
##
##       band            marching           music              rock
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   : 0.2996   Mean   : 0.0406   Mean   : 0.7378   Mean   : 0.2433
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 1.0000   3rd Qu.: 0.0000
##  Max.   :66.0000   Max.   :11.0000   Max.   :64.0000   Max.   :21.0000
##
##       god              church            jesus             bible
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.00000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.00000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.00000
##  Mean   : 0.4653   Mean   : 0.2482   Mean   : 0.1121   Mean   : 0.02133
##  3rd Qu.: 1.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.00000
##  Max.   :79.0000   Max.   :44.0000   Max.   :30.0000   Max.   :11.00000
```

```
##
##       hair             dress           blonde              mall
##  Min.   : 0.0000   Min.   :0.000   Min.   :  0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.:0.000   1st Qu.:  0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median :0.000   Median :  0.0000   Median : 0.0000
##  Mean   : 0.4226   Mean   :0.111   Mean   :  0.0989   Mean   : 0.2574
##  3rd Qu.: 0.0000   3rd Qu.:0.000   3rd Qu.:  0.0000   3rd Qu.: 0.0000
##  Max.   :37.0000   Max.   :9.000   Max.   :327.0000   Max.   :12.0000
##
##     shopping         clothes         hollister        abercrombie
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median : 0.000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   : 0.353   Mean   :0.1485   Mean   :0.06987   Mean   :0.05117
##  3rd Qu.: 1.000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :11.000   Max.   :8.0000   Max.   :9.00000   Max.   :8.00000
##
##       die              death             drunk              drugs
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :0.00000   Min.   : 0.00000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.: 0.00000
##  Median : 0.0000   Median : 0.0000   Median :0.00000   Median : 0.00000
##  Mean   : 0.1841   Mean   : 0.1142   Mean   :0.08797   Mean   : 0.06043
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:0.00000   3rd Qu.: 0.00000
##  Max.   :22.0000   Max.   :14.0000   Max.   :8.00000   Max.   :16.00000
##
```

Tratamento dos valores faltantes

```
val.medio = ave(data$age, data$gradyear,
FUN = function(x) mean(x, na.rm = TRUE))
data$age = ifelse(is.na(data$age), val.medio, data$age)
```

```
interesses = data[5:40]
interesses1= as.data.frame(lapply(interesses, scale))
```

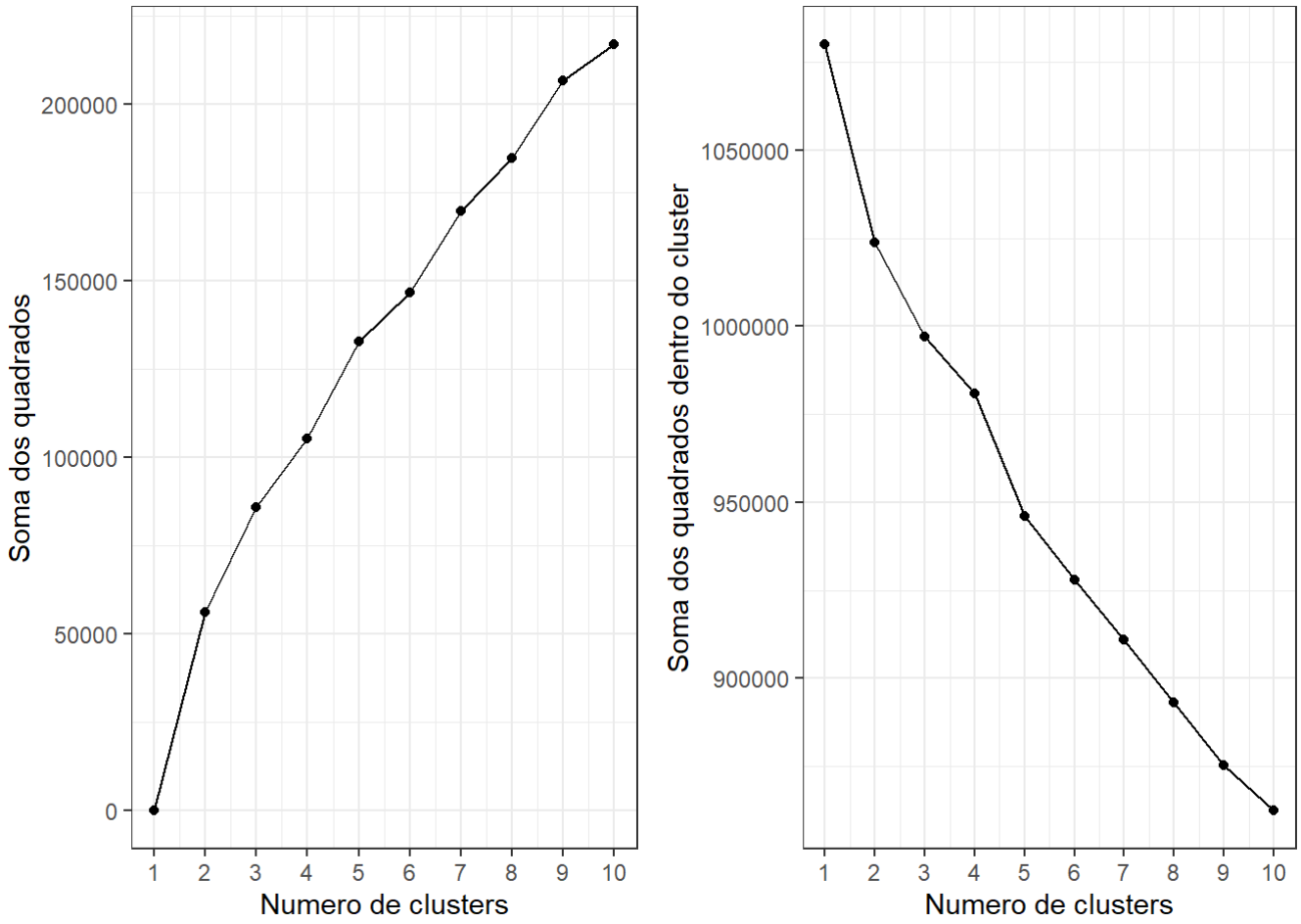
Calculo do k ótimo

```
btw=numeric()
tws=numeric()

set.seed(2708)

for(i in 1:10){
  btw[i]=kmeans(interesses1, centers = i, iter.max = 30)$betweenss
  tws[i]=kmeans(interesses1, centers = i, iter.max = 30)$tot.withinss
}
```

```
k=qplot(1:10, btw, geom=c("point","line"),
        xlab= "Numero de clusters",
        ylab = "Soma dos quadrados")+
        scale_x_continuous(breaks = seq(0,10,1))+
        theme_bw()

kk=qplot(1:10, tws, geom=c("point","line"),
         xlab= "Numero de clusters",
         ylab = "Soma dos quadrados dentro do cluster")+
         scale_x_continuous(breaks = seq(0,10,1))+
         theme_bw()
```

Analise gráfica

```
grid.arrange(k,kk,ncol=2)
```



Com base no plot acima, percebe-se que a partir do cluster k=5, os valores passam a decrescer de forma mais parcimoniosa. Portanto, a escolha ótima é igual a k=5.

Calculo do k-means com k=5

```
mod.k5<-kmeans(interesses1, centers = 5)
```

Previsão do modelo com os valores médios para cada cluster

```
prev<-mod.k5$cluster

aggregate(interesses1,by=list(prev),mean)
```

```
##   Group.1  basketball   football      soccer    softball    volleyball
## 1       1  0.06206280  0.0379521  0.01610821 -0.04051709  0.0073556572
## 2       2  0.01981887  0.1163459  0.09072890 -0.04004432  0.0004899175
## 3       3 -0.18685453 -0.1846446 -0.07765125 -0.13901400 -0.1353200182
## 4       4  0.37038453  0.3797970  0.14290864  0.12524713  0.0918672172
## 5       5  1.33772192  1.1814896  0.44276422  1.15267582  1.0659653371
##      swimming cheerleading    baseball      tennis      sports        cute
## 1  0.08061997  -0.02000226 -0.02919604  0.11383082 -0.04323113 -0.02766834
## 2  0.33189323   0.52403366 -0.04832416  0.05913741 -0.05504704  0.84577076
## 3 -0.08355887  -0.10792168 -0.14083719 -0.03686695 -0.16172539 -0.17053280
## 4  0.26079946   0.20583304  0.25777847  0.10820143  0.76427154  0.46878424
## 5  0.08086519   0.04507462  1.12544297  0.14730979  1.11171495 -0.01119974
##            sex        sexy          hot      kissed        dance        band
## 1 -0.013393673 -0.08491799 -0.018073438 -0.05259385 -0.009930898  0.18357705
## 2 -0.004279147  0.30708572  0.661921419 -0.01519770  0.683524283  0.05136743
## 3 -0.092513233 -0.07946642 -0.131489174 -0.12996663 -0.143384757 -0.03963083
## 4  2.055917516  0.54456839  0.295670790  3.03877172  0.459711560  0.63720869
## 5 -0.035871365  0.01032110  0.002028647 -0.08803313  0.002101166 -0.04040469
##       marching       music        rock         god     church        jesus
## 1  0.049521901  0.23055937  0.09111477  2.44476318  1.8853454  2.594415921
## 2 -0.018301823  0.23323405  0.11383254  0.06117483  0.2512165 -0.005486482
## 3 -0.002307144 -0.11640700 -0.10017770 -0.09847364 -0.1171163 -0.071025272
## 4  0.210252524  1.25504995  1.26260882  0.35726848  0.1382386  0.051986494
## 5 -0.047594626  0.07132671  0.14051488  0.02513704  0.1128803 -0.017077657
##         bible        hair       dress        blonde        mall     shopping
## 1  4.63615459  0.03724049  0.03030028 -0.0004771042 -0.11444476  0.003425952
## 2 -0.08407417  0.37176907  0.61072537  0.0346067376  0.87380788  1.139270041
## 3 -0.10424551 -0.19050692 -0.12651524 -0.0277298361 -0.17977790 -0.217972249
## 4  0.03201289  2.56116327  0.53185123  0.3649034128  0.62098705  0.262661078
## 5 -0.08413642  0.01038533 -0.06710311  0.0341982598 -0.01609944  0.023462836
##         clothes  hollister abercrombie         die       death       drunk
## 1  0.0267213296 -0.0912992 -0.07607198  0.19994701  0.25187787  0.06677431
## 2  0.6639986658  0.9293809  0.85614148  0.03983618  0.09227983  0.03493274
## 3 -0.1773060830 -0.1635246 -0.15662150 -0.08524457 -0.06480263 -0.08437753
## 4  1.2299165207  0.2920795  0.39481183  1.70522216  0.94040925  1.78880548
## 5 -0.0007444395 -0.1065370 -0.10005799 -0.07074648 -0.02770380 -0.07323818
##         drugs
## 1  0.03662975
## 2 -0.05468738
## 3 -0.10921543
## 4  2.71092242
## 5 -0.09261611
```

Clustering do modelo

```
### O plot da erro com todas as colunas, pois fica muito grande. Com 20 colunas



plot(interesses1[,1:20],col=prev)
```