Aprendizaje Automático I

Ejercicio: Aprendizaje No Supervisado

DSLab

octubre, 2023

En este ejercicio vamos a trabajar con los datos del *"Dow Jones Index Data Set"* que podéis descargar aquí: DOW JONES INDEX. Se trata de datos semanales del Dow Jone Industrial Index.

Datos

En primer lugar descargamos y leemos los datos

```
djidata = read.table("./dow_jones_index/dow_jones_index.data",
    header = TRUE, sep = ",")
djidata = as.data.frame(djidata)
head(djidata)
```

```
##   quarter stock      date   open   high    low  close    volume
## 1       1    AA  1/7/2011 $15.82 $16.72 $15.78 $16.42 239655616
## 2       1    AA 1/14/2011 $16.71 $16.71 $15.64 $15.97 242963398
## 3       1    AA 1/21/2011 $16.19 $16.38 $15.60 $15.79 138428495
## 4       1    AA 1/28/2011 $15.87 $16.63 $15.82 $16.13 151379173
## 5       1    AA  2/4/2011 $16.18 $17.39 $16.18 $17.14 154387761
## 6       1    AA 2/11/2011 $17.33 $17.48 $16.97 $17.37 114691279
##   percent_change_price percent_change_volume_over_last_wk previous_weeks_volume
## 1             3.792670                                NA                    NA
## 2            -4.428490                          1.380223             239655616
## 3            -2.470660                        -43.024959             242963398
```

```
## 4             1.638310                          9.355500            138428495
## 5             5.933250                          1.987452            151379173
## 6             0.230814                        -25.712195            154387761
##   next_weeks_open next_weeks_close percent_change_next_weeks_price
## 1          $16.71           $15.97                       -4.428490
## 2          $16.19           $15.79                       -2.470660
## 3          $15.87           $16.13                        1.638310
## 4          $16.18           $17.14                        5.933250
## 5          $17.33           $17.37                        0.230814
## 6          $17.39           $17.28                       -0.632547
##   days_to_next_dividend percent_return_next_dividend
## 1                    26                     0.182704
## 2                    19                     0.187852
## 3                    12                     0.189994
## 4                     5                     0.185989
## 5                    97                     0.175029
## 6                    90                     0.172712
```

```
table(djidata$stock)
```

```
##
##   AA  AXP   BA  BAC  CAT CSCO  CVX   DD  DIS   GE   HD  HPQ  IBM INTC  JNJ  JPM
##   25   25   25   25   25   25   25   25   25   25   25   25   25   25   25   25
##   KO KRFT  MCD  MMM  MRK MSFT  PFE   PG    T  TRV  UTX   VZ  WMT  XOM
##   25   25   25   25   25   25   25   25   25   25   25   25   25   25
```

Cada fila corresponde a datos semanales de un valor bursatil. En este ejercicio vamos a trabajar con los datos correspondientes a la variable *close*, esto es, el valor de cada stock al cierre de la semana.

Necesitamos transformar la variable de interés como sigue:

```
djidata$close = as.numeric(sub("\\$", "", djidata$close))
```

1. Construir la matriz de series temporales

En primer lugar hemos de construir la matriz con las series que necesitamos. Necesitamos una matriz de series con las series por columnas para cada uno de los valores bursátiles.

```
stocks = unique(djidata[, "stock"])
n = dim(djidata[stocks == "AA", ])[1]
stocksdata = matrix(0, n, length(stocks))
for (i in 1:length(stocks)) stocksdata[, i] = djidata[djidata$stock ==
```

```
    stocks[i], "close"]
colnames(stocksdata) = stocks
stocksts1 = as.ts(stocksdata[1:12, ])
stocksts2 = as.ts(stocksdata[13:25, ])
stocksts = as.ts(stocksdata)
```

2. Representar las series con las que vamos a trabajar

3. Realizar un análisis cluster usando como variables de interés la media y desviación estándar de cada serie

¿Pueden identificarse valores atípicos?

¿Existe relacion entre las dos variables consideradas en el análisis? ¿Como interpretas este resultado?

4. Representar las series escaladas

5. Realizar un análisis Cluster para cada uno de los cuatrimestres

¿En cuantos grupos podemos dividir la muestra?

Representar graficamente la media de cada cluster para tratar de identificar el comportamiento medio de los valores en cada cluster.

6. Análisis Cluster para todo el periodo

Elegir una tecnica para determinar el mejor numero de clusters.

7. Representar graficamente la media de cada cluster

8. Localizar atípicos en los clusters

9. Repetir el análisis, para todo el periodo, empleando la distancia DTW.

10. Identificar las diferencias entre los dos análisis