

Analisis y Curacion clase 4: Práctico 2

30 de Mayo, 2019, Cohorte Alpha

Análisis exploratorio y curación de datos

- ▶ Gonzalez Nehuen
- ▶ Arja Adel
- ▶ Madoery Pablo

Practico 2: Entregar un Rmd donde se:

- ▶ Elija un dataset clasificado de su preferencia y area (domain expertise), aplique un metodo de clustering y/o mixtura de Gaussianas en el mismo.
- ▶ Investigue los resultados en el meta parametro K numero de cumulos e investigue posibles procesos de seleccion del mismo.
- ▶ Elabore un resumen, y seleccione un mejor valor segun el/los criterios aplicados, discuta el significado de los cumulos encontrados.
- ▶ Comente la influencia de la normalizacion de los datos en los resultados del clustering.

Solución

Utilizamos el método “k-Nearest Neighbour”, con el cual realizamos la clasificación de un data set de distintos tipos de vidrios.

El data set fue descargado de

<https://archive.ics.uci.edu/ml/datasets/glass+identification>.

```
glass_df = read.csv("glass.csv", header = FALSE)
colnames(glass_df) <-
  c("id", "RI", "Na", "Mg", "Al", "Si", "K", "Ca", "Ba", "P")
```

```
glass_df$Type <- as.character(glass_df$Type)
glass_df <- glass_df[sample(nrow(glass_df)),]
```

```
normalize <- function(x) {
  return ((x-min(x))/(max(x)-min(x)))
}
```

```
glass_df_n <- as.data.frame(lapply(glass_df[2:10], normalize))
```

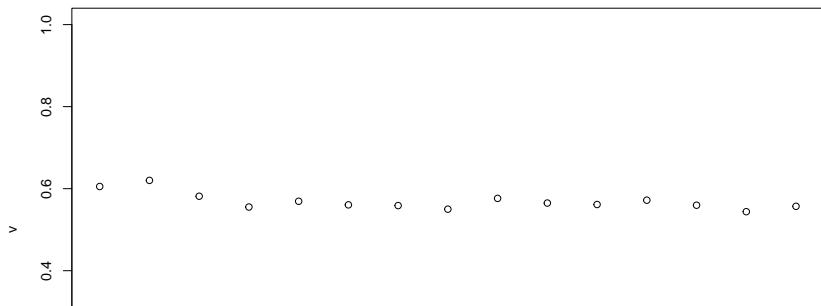
```
glass_df_n_train = glass_df_n[1:100,]
glass_df_n_test = glass_df_n[101:214,]
```

```
train_types = glass_df[1:100, "Type"]
test_types = glass_df[101:214, "Type"]
```

```

library(class)
v = rep(0, 15)
N_reps = 10
for (j in 1:N_reps){
  for (k in 1:15){
    data_test_pred <- knn(train=glass_df_n_train, test=glass_test)
    v[k] <- v[k] + sum(data_test_pred == test_types) / length(data_test_pred)
  }
}
v <- v/N_reps
plot(1:15, v, ylim=c(0,1))

```



```
library(gmodels)
CrossTable(x=test_types, y=data_test_pred, prop.chisq = FALSE)

##
##
##      Cell Contents
## |-----|
## |                                     N |
## |          N / Row Total |
## |          N / Col Total |
## |          N / Table Total |
## |-----|
##
##
## Total Observations in Table:  114
##
##
##           | data_test_pred
## test_types |          1 |          2 |          5 |
## -----|-----|-----|-----|-----|
```

