Business Analytics - ETC3250 2018 - Lab 6

Clustering

Souhaib Ben Taieb

10 April 2018

Exercice 1

Read and run the code in Sections 10.5.1 and 10.5.2 of ISLR.

Exercice 2

ISLR Section 10.7, exercises 2(a), 2(b), 2(c) and 2(d).

Exercice 3

1. Let $\{x_1, \dots, x_n\}$ be a set of points where $x_i \in \mathbb{R}^d$, and let x be any point in $\mathbb{R}^d$. Prove that

$$\sum_{i=1}^{n} ||x_i - x||^2 = \sum_{i} ||x_i - \bar{x}||^2 + n \, ||\bar{x} - x||^2, \quad (1) \tag{1}$$

where $||\cdot||$ is the L_2 norm, and $\bar{x}$ is the centroid of the set of points, i.e. $\bar{x} = \frac{1}{n}\sum_i x_i$. (Hint: add and substract $\bar{x}$ in the left-hand side of the previous expression).

2. Which value of x minimizes $\sum_{i=1}^{n} ||x_i - x||^2$? Prove it.

3. Using expression (1), prove that

$$\sum_{i,j} ||x_i - x_j||^2 = 2n \sum_{i} ||x_i - \bar{x}||^2 \tag{2}$$

Data

The crime dataset (https://github.com/bsouhaib/BA2017/blob/master/data/crimes2008.csv) contains FBI crime rate statistics. These are the indices for 9 different types of crimes reported by the states of the USA, for 2008: violent, property, murder, rape, robbery, assault, burglary, ltheft (larceny theft), vtheft (vehicle theft). The values have been population adjusted so that the numers are per million people.

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:reshape':
##
##     rename
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:GGally':
##
##     nasa

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
crime <-  as.data.frame(read_csv(url("https://github.com/bsouhaib/BA2018/raw/master/data/crimes2008.csv
```

```
## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   Population = col_integer(),
##   State = col_character(),
##   index = col_double(),
##   violent = col_double(),
##   property = col_double(),
##   murder = col_double(),
##   rape = col_double(),
##   robbery = col_double(),
##   assault = col_double(),
##   burglary = col_double(),
##   ltheft = col_double(),
##   vtheft = col_double()
## )
```

```
dim(crime)
crime <- crime[,-c(1,2,4)]
head(crime)
#crime[,-1] <- scale(crime[,-1])
base::row.names(crime) <- crime[,1]
crime <- scale(crime[,-1])
```

Exercice 4

Make a scatterplot matrix of the crime indices, with and without Washingto DC. Write a paragraph describing the relationships between the statistics, and about any observations about cluster patterns in the data.

Exercice 5

Cluster the states using hierarchical clustering, with Euclidean distance and wards linkage. Plot the dendrogram. How many clusters would be suggested by the dendrogram?

Exercice 6

Use k-means clustering with k set to several different values, say 2-8. Calculate the ratio of between Sum of Squares (SS) to total SS for each value of k. Tabulate this. What is between SS? total SS? What happens to this value as k ranges from 2 to 8? Why is this? Also, what happens if you change the random seed, which changes the initialization of k-means?

Exercice 7

Use the *fpc* package in R, and the function *cluster.stats* to produce the statistic *wb.ratio* to examine the within group distances to the between group distances for each hierarchical cluster solution. How many clusters would be chosen by this approach?

(The *wb.ratio* statistic reports the ratio between two quantities comparing within to between distances. The average of the distances between points that are in the same cluster, ie within. And the distances between points that are not in the same cluster, ie between. The smaller the value of this the better the result describes clustering as explaining the variation in the data.)

Exercice 8

Decide on an appropriate number of clusters, and report the results. Tabulate the cluster means, standard deviation, and number of points in each cluster. Plot the cluster means using a parallel coordinate plot. List the states in each cluster. Write a paragraph describing the characteristics of each cluster, e.g. cluster 3 is characterized by low larceny and vehicle theft.