Machine learning 1
Master Data Science
Summer Semester 2018

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 1 – 11 April 2018
K-Means Clustering

1 Mathematical Exercises

Exercise 1

In this problem, you will perform K-means clustering manually, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ variables. The observations are as follows.

Obs.	X1	X2	Init
1	1	4	1
2	1	3	2
3	0	4	1
4	5	1	2
5	6	2	1
6	4	0	2

(a) The initial clustering is given in the table. Represent the the observations in a drawn scatter plot.

(b) Compute the centroid for each cluster.

(c) Assign each observation to the centroid to which it is closest, in terms of squared Euclidean distance. Report the new cluster labels for each observation and calculate the squared Euclidean distance.

(d) Repeat (b) and (b) until the answers obtained stop changing.

(e) In your plot from (a), colour the observations according to the cluster labels obtained in (d).

Exercise 2 Homework

In James et al. p. 388 there is a formula relating the within cluster sum of squared differences from the centroid with the within cluster sum of squared differences between all points.

$$\frac{1}{|C_k|}\sum_{i,j\in C_k}\|\boldsymbol{x_i}-\boldsymbol{x_j}\|^2 = 2\sum_{i\in C_k}\|\boldsymbol{x_i}-\boldsymbol{\mu_k}\|^2,$$

It is possible to prove this algebraically, but we will only consider the one dimensional ($p = 1$) case for one cluster. In practice you do not do clustering on one variable, but it gives an insight to the proof for p-dimensional clustering.

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2 = 2\sum_{i=1}^{n}(x_i-\overline{x})^2. \tag{1}$$

(a) Show that this is the case for the data $x_1 = 1, \quad x_2 = 5, \quad x_3 = 10, \quad x_4 = 17, \quad x_5 = 22.$

(b) Show that equation 1 above is true. Hint: Add and subtract $\overline{x}$ to the left hand side of the equation, giving:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}((x_i-\overline{x})-(x_j-\overline{x}))^2$$

2 K-means in R

Preliminaries
Start RStudio in the usual way:

- Open RStudio
- Strg+Shift+N opens a new R script.
- Create a new directory for this course called ML1
- Open a new script file set the working directory and save it with the name `Workshop1.R`.
- Clear any old objects in your workspace using
  Session > clear workspace.
- As with all your work in RStudio you should type you commands into the script file and then execute them using CTRL+Enter

Exercise 3 K-Means in R: Simulated Data

Simulate a matrix with two columns representing the x and y coordinates of 5 points. The first 25 points will be shifted in a South-East direction to create two clusters.

```
set.seed(2)
x=matrix(rnorm(50*2),ncol=2)
x[1:25,1]<-x[1:25,1]+2
x[1:25,2]<-x[1:25,2]-2
plot(x,pch=16)
```

Execute the following code one line at a time:

```
km.out<-kmeans(x,centers=2,nstart=1)  #run with two clusters
km.out$cluster           #a vector specifying which cluster each row belongs to
names(km.out)            #all the different outputs from kmeans
km.out$totss             #the sum of squares without clustering
km.out$tot.withinss      #the sum of squares with this clustering
km.out$withinss          #the sum of squares within each cluster
km.out$centers           #Matrix with the center coordinates
plot(x,col=km.out$cluster+1,pch=16)   #plot the poiunts colorfed by cluster
points(km.out$centers,col=2:3,pch=3)  #add the cluster centers
```

The `kmeans` command has an argument called `nstart`. This indicates how many times the algorithm is to be repeated using different random starting configurations. `nstart=1` is too small, you might be unlucky and have found a poor local minimum.

Repeat this using 20 repeats

```
km.out<-kmeans(x,centers=2,nstart=20)
plot(x,col=km.out$cluster+1,pch=16)
points(km.out$centers,col=2:3,pch=3)
km.out$tot.withinss
```

Ok it seems that the first attempt was a good (or the best) solution.

We can now see what happens when we try to force more than two clusters to the data.

```
set.seed(4)
km.out<-kmeans(x,centers=3,nstart=20)
plot(x,col=km.out$cluster+1,pch=16)
km.out<-kmeans(x,centers=4,nstart=20)
plot(x,col=km.out$cluster+1,pch=16)
```

```
km.out$tot.withinss
```

The within cluster sum of squares W is now much lower even though there are only really two clusters in our data matrix. W almost always decreases with increasing number of clusters $\Rightarrow$ The best number of clusters to fit cannot be found by simply by minimising W.

We will now work through a similar example where the data has 3 clusters.

```
x<-matrix(rnorm(50*3),ncol=2)
x[1:25,1]<-x[1:25,1]+2
x[1:25,2]<-x[1:25,2]-2
x[50+1:25,1]<-x[50+1:25,1]+2
x[50+1:25,2]<-x[50+1:25,2]+2
km.out<-kmeans(x,3,nstart=20)
plot(x,col=km.out$cluster+1,pch=16)
points(km.out$centers,col=2:4,pch=3)
```

Exercise 4 Clustering City Locations

Load the `mdsr` library

```
> library(dplyr)
```

If you get an error at this stage then you need to install the package using

```
> install.packages("dplyr")
```

Load the `mdsr` library

```
> library(mdsr)
```

Again you might need to install the package.

Lets take a first look at the structure of the data frame.

```
> names(WorldCities)
> dim(WorldCities)
```

How many cities are listed in this data set?

We will restrict the data to just those with a population of at least 100 000 inhabitants using the `filter` command in the `dplyr` package, and then keep just two variables `longitude` and `lattitude`.

```
> BigCities<-filter(WorldCities,population >= 100000)
> BigCities<-select(BigCities,longitude, latitude)
```

How many cities have at least one hundred thousand inhabitants?

We will now run K-Means using 6 clusters, and plot the clusters.

```
> set.seed(15)
> city.km<-kmeans(BigCities,centers = 6)
> with(BigCities,plot(longitude,latitude,col=city.km$cluster,pch=16,cex=0.6))
```

What do you notice?

It is important to realise that there is nothing in the data defining continent, but the clustering method can easily and surprisingly effectively identify the continents.

Try using different numbers of clusters from 2 upwards. Also notice that the borders of each cluster can differ depending on the seed.