Machine Learning 1
Data Science
Summer Semester 2018

Prof. Tim Downie

Workshop 3 – 25 April 2018
Robust Clustering with PAM, Hierarchical Clustering

Preliminaries

Open RStudio. Clear any old objects in your workspace using Session > clear workspace.
Open a new script file, set the working directory and save it with the name `Workshop3.R`.

In this workshop you need the packages `cluster` and `ISLR`. First check to see if `cluster` is already installed using

```
> find.package("cluster")
```

If a directory path is returned, then the package is installed. If you get an error, then it is not installed and you need to install it using

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

If installing is unsuccessful try

```
> update.packages()
```

answering `y` to each question. Load the package into R with: `library(cluster)`
Repeat the above with the `ISLR` library.

Exercise 1 PAM Clustering

Use the following R code in order to run the Partitioning Around Medoids PAM algorithm on the `USArrests` dataset from Workshop 2. You will need to replace ??? with the appropriate code.

The `USArrests` data set has the state names as row names. When plotting the names it is bette to use the state abbreviations.

```
row.names(USArrests)<-state.abb
pairs(USArrests)
```

Have an initial look at the the PAM clustering with 4 clusters.

```
pam.out<-pam(USArrests,k=???)
clusplot(pam.out,labels=3)
```

The cluster plot automatically plots the first two principal components. Each cluster is represented by an ellipse that encircles all the points in that cluster.

Another plot uses to assess the appropriateness of a clustering is using a *silhouette plot*. The elements are sorted into their clusters and its *silhouette width* is plotted. The silhouette width measures the similarity of

each point to its cluster. A rough guide is a silhouette width over 0.4 is good. Negative silhouette widths suggest that maybe that element would be better assigned to the neighbour cluster. See the `silhouette` help page (or Wikipedia) for more details.

```
sp<-silhouette(pam.out)
plot(sp,col=1:4)
mean(sp[,"sil_width"])
abline(v=mean(sp[,"sil_width"]))
```

The mean silhouette width can be used to assess which is the best number of clusters.

```
avesw.vec<-rep(NA,7)
for(k in 2:7)
  avesw.vec[??]<-mean(silhouette(pam(USArrests,k=???))[,"sil_width"])
plot(1:7,avesw.vec,type="b",ylim=c(0,0.6))
```

Which number of clusters gives the largest mean silhouette width?

Rerun the PAM algorithm with the optimal number of clusters

```
pam.out<-pam(???,k=???)
clusplot(???,labels=3)
sp<-silhouette(???)
plot(sp,col=1:???)
abline(v=mean(sp[,"sil_width"]))
```

Compare the optimal PAM clustering with the K-Means result.

```
km.out<-kmeans(USArrests,centers=??,nstart=20)
table(km.out$???,pam.out$???)
```

There is no difference between the two.

If you want to use PAM clustering using the Manhattan distance metric there is an argument to `pam` called `metric="manhattan"`. Alternatively you can specify a distance matrix rather than the data frame for advanced options such as the correlation metric mentioned in James et al. pages 396-399.

Silhouette plot for K-means

The function `sihouette` was written for objects created by functions in the `cluster` package. `kmeans` is a function in the `stats` package loaded with the "base R" installation. As a result `silhouette(km.out)` doesn't work, but with a little work we can coerce the data into the correct format. The first argument to silhouette should be the cluster vector `km.out$cluster`. The second argument should be a distance matrix `dist(USArrests)`, a matrix containing the pairwise euclidean distance between each pair of points, e.g the Euclidean distance in the USArrests dataset between the elements Alabama and Alaska is 37.17.

The following code gives a silhouette plot for your last K-means clustering.

```
sp<-silhouette(km.out$cluster, dist(USArrests))
plot(sp,col=1:???)
```

Exercise 2

In Workshop 2 Exercise 3 you ran K-means and PCA on the Hamburg Decathlon data. Load in these data as last week. Carry out a similar analysis to Exercise 1 using PAM to see if outliers were influencing the results. Hint: in `clustplot` use `labels=0` as the row names are the row numbers which are not informative.

Exercise 3 Hierarchical Clustering

Work through 10.6 Lab 3 on pp. 407–413 of James et al. Just concentrate on the `"complete"` linkage method.

Homework: Hierarchical Clustering

Below are 6 points in 2-dimensions.

Element	X1	X2
1	4	2
2	5	5
3	6	11
4	10	2
5	14	7
6	15	9

(a) Calculate the distances between each pair of points rounded to 1 decimal place. You can use the R function `dist()` to do this.

(b) Use the distance matrix to find the first cluster join. Write down the two element numbers and the distance between them.

(c) Using the complete linkage rule repeat step (b) to obtain the hierarchical clusters. At each step write down the element numbers for each element in the new cluster and the distance used to determine the cluster.

(d) Draw the scatter plot of the points and the dendrogram of the clustering.