Lab 9 Clustering

Problem Statement

We are going to study how to cluster a dataset using k-means and hierarchical clustering approaches.

Dataset 1 for K-means Clustering

Chatterjee-Price Attitude Data `attitude` from the `datasets` package. The dataset is a survey of clerical employees of a large financial organization. The data are aggregated from questionnaires of approximately 35 employees for each of 30 (randomly selected) departments. The numbers give the percent proportion of favourable responses to seven questions in each department. For more details, see `?attitude`.

When performing clustering if data contains multiple (or more than 2) variables, one option would be to perform Principal Component Analysis (PCA) and then plot the first two vectors and maybe additionally apply K-Means.

In this exercise, we'll take a subset of the attitude dataset and consider only two variables `privileges` and `learning`, that is we would like to cluster the attitude dataset with the responses from all 30 departments when it comes to `privileges` and `learning`. The subset is defined as follows:

```
dat = attitude[,c(3,4)]
```

Questions to Dataset 1

1) Plot the dataset `dat`.
2) Let `k = 2` and `nstart = 1`. Set a seed and then perform the k-means clustering based on the two parameters.
3) Report the `tot.withinss`.
4) Plot the two clusters with two different colours.
5) Let `nstart = 100` and repeat 2)-4). Compare the two `tot.withinss`.
6) Write a for-loop to record the `tot.withinss` when `k` is 1 to 15. Plot the result.
7) Use the elbow method to identify the best `k`.

With the elbow method, the solution criterion value (within groups sum of squares) will tend to decrease substantially with each successive increase in the number of clusters.

8) Plot the k clusters with the best `k` you get in 7).

Dataset 2 for Hierarchical Clustering

On the book website, http://www-bcf.usc.edu/~gareth/ISL/data.html, there is a gene expression data set (Ch10Ex11.csv) that consists of 40 tissue samples (40 columns) with measurements on 1,000 genes (1000 rows). The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

- Save the dataset to your local hard drive.

- Load in the data using `read.csv()`. You will need to select `header = F`.

```
DatasetName = read.csv("yourLocalDirectory\\Ch10Ex11.csv",header = FALSE)
```

Note you need to replace \ with \\ in your directory. For instance, your directory is

`C:\myDocument\RPrograms\Session9`. Then

```
DatasetName = read.csv("C:\\myDocument\\RPrograms\\Session9\\Ch10Ex11.csv", header =
FALSE)
```

Questions to Dataset 2

9) Read the description of the dataset again. Do you think the current layout of the dataset is a natural way to present the relationship between tissue samples (as columns) and genes (as rows)? Note each tissue may contain hundreds of genes. If not, transform the dataset in a more natural way.

10) Calculate the dissimilarity metric.

    Hint: We will take as our dissimilarity metric between the ith and jth samples to be $1 - rij$, where rij is the correlation between the two samples.
    Notice that this function will have its smallest value (of zero) if $rij = 1$ i.e. the two samples are perfectly correlated.
    This function will have its largest value (of two) if $rij = -1$ i.e. the two samples are perfectly anti-correlated.

11) Apply hierarchical clustering to the samples using correlation based distance for

    a. Complete linkage
    b. Average linkage
    c. Single linkage
    d. Centroid linkage

12) Plot the four dendrograms in the same plot by using `par(mfrow = c(i,j))`, where `i` is the number of rows and `j` is the number of columns in the plot.

13) Do the genes separate the samples into the two groups? To answer this question, we need to generate a confusion matrix on the predicted and true number of healthy/diseased patients.

14) Do your results depend on the type of linkage used?