Accompanying Tutorial for "An Introduction to Some Popular Clustering Methods"

Teck Por Lim

12 Aug 2015

Instructions

The ggplot2 and stats packages are pre-requisites for the tutorial. However, the stats package is available by default, so there is no need to issue the library(stats) function.

The other packages are needed, if you choose to use functions belonging to them. If you do not already have them installed, it is recommended that you install the pre-requisite packages first before trying out this tutorial.

Please examine the 3 examples available, and modify them by substituting k-means (Lloyd's version) with other clustering algorithms:

1. example1.R
2. example2.R
3. example3.R

Suggested Steps

1. Rename each example?.R to example?.kmeans.Lloyd.R
2. Edit each example by hand, and rename the modified example file
3. If unsure about details of the functions, use the R help system. For example, try typing ?kmeans

A suggestion is to base the filenames on a rational format like exampleN.method.argument.R, where

- N is an integer
- "method" refers to the clustering method used
- "argument" is a word which helps one to recall which argument was used

It may be longer to type, but it will make more sense.

Suggested Packages

These packages are available from CRAN. However, if you are aware of good packages and functions which are not listed, please feel free to share them with everyone.

- stats
    - kmeans() (k-means clustering)
    - hclust() (hierarchichal clustering)
    - cutree() (cuts tree into groups)
- EMCluster (Expectation Maximization clustering)
    - simple.init() (initialization)
    - emcluster() (returns object)
- kernlab
    - kkmeans() (kernel k-means)

Challenge Question (Example 4)

Example 4 demonstrates the power of kernel k-means, as this example will cause many clustering algorithms to fail.

1. Innermost group is a circle of randomly generated points
2. Outermost group is another circle of randomly generated points

However, can you please verify that kernel k-means is not magic, by modifying the code in example4.R such that kernel k-means will also fail? Hint: modify the parameters of the functions or modify the generated data.