Clustering

CIS400/600 Fundamentals of Data and Knowledge Mining

Package installations and troubleshooting issues occurring during installations are documented on the last page

Problem 1 - Clustering Tendency

Clustering algorithms cannot differentiate between data that is completely random and data obtained empirically. They tend to cluster data irrespectively. To address this issue, it is important to determine if a dataset is actually clusterable i.e. it is important to find the clustering tendency of data. An available statistical approach to check clustering tendency of data is **Hopkins Statistic**

Read about **Hopkins Statistic** from
Introduction to Data Mining, Steinbach and Kumar, Chapter 8, Section - 8.5.6 Clustering Tendency

1. Install `clustertend` package and load the same
2. Read `WINE` dataset provided along with the exercise
3. Apply **Hopkins Statistic** on `WINE` data that is read. Choose 30% of your data as input to `hopkins()` function. From the value that is output, what do do you infer about data. Is the data highly random or otherwise. Provide your answer in comments

   Read about `clustertend` package from
   https : *//cran.r−project.org/web/packages/clustertend/clustertend.pdf*

Problem 2 - KMeans Clustering

1. Install package `Rattle` and load the same
2. Shuffle `WINE` dataset. Remove column-1 from dataset and **Standardize** the data
3. How to determine the value of `K` i.e. how to determine the number of clusters to form. There is a statistical approach known as `Elbow Method` to determine the number of clusters.
   Refer **figure 8.32** from **Introduction to Data Mining, Steinbach and Kumar, Chapter 8, Section - 8.5.5 Determining the Correct Number of Clusters**
   (a) Set `K = 2, 3, 4, 5, 6, 7` and run `K-Means`. Store the sum of within sum of squares of clusters. The attribute available in `kmeans` method is `tot.withinss`.
   (b) Plot a graph with values of `K` on X-axis and `tot.withinss` on Y-axis. Read `Section - 8.5.5` and deduce what is the best value of `K` from the graph
4. With the `K` value deduced from (3), apply `K-Means` to the data and print cluster centers and cluster sizes of formed clusters
5. Print `Confusion Matrix` based on clusters formed

Problem 3 - Hierarchical Clustering

Use **Euclidean Distance** as the distance metric and perform

1. Hierarchical clustering using **Single Link** proximity measure. Print the dendogram and cut it to `K` clusters. Use `K` value from problem - 2, part 3
2. Perform all operations in (1) using **Complete Link** proximity measure.
3. Perform all operations in (1) using **Ward's** proximity measure.
4. Print confusion matrices for (1), (2) and (3)

Package Installation & Troubleshooting

Students might encounter issues when installing `Rattle` package. `Rattle` is dependent on `RGtk2` library.

1. Mac OSX users can use the below links to fix any issues
   `http://marcoghislanzoni.com/blog/2014/08/29/solved-installing-rattle-r-3-1-ma`
   `https://gist.github.com/sebkopf/9405675#troubleshooting-for-gtk-224-issues`

2. Windows users can use the following link
   `http://www.learnanalytics.in/blog/?p=31`

3. Ubuntu users can follow the below steps
   Open terminal and enter

   ```
   ~$ sudo apt-get install wajig
   ~$ wajig install libgtk2.0-dev
   ```

   Open R and install

   ```
   install.packages("RGtk2", depen=T)
   install.packages("rattle")
   ```

Students are requested to explore more options on troubleshooting any installation issues. Some of the above mentioned fixes are not tested.