Hierarchical Clustering in R

Spoken Tutorial Project
https://spoken-tutorial.org
National Mission on Education through ICT
https://sakshat.ac.in

Tanmay Srinath
Madhuri Ganapathi
IIT Bombay

22 March 2022

Learning Objectives

Learning Objectives

We will learn about:

Learning Objectives

We will learn about:

- **Hierarchical Clustering**

Learning Objectives

We will learn about:

- **Hierarchical Clustering**
- **Types of Hierarchical Clustering**

Learning Objectives

We will learn about:

- **Hierarchical Clustering**
- **Types of Hierarchical Clustering**
- **Advantages of Hierarchical Clustering**

Learning Objectives

- **Linkage and its types**

Learning Objectives

- **Linkage and its types**
- **Applications of Hierarchical Clustering**

Learning Objectives

- Linkage and its types
- Applications of Hierarchical Clustering
- Hierarchical Clustering on iris dataset

System Specifications

System Specifications

- **Ubuntu Linux OS version 20.04**

System Specifications

- **Ubuntu Linux OS version 20.04**
- **R version 4.1.2**

System Specifications

- **Ubuntu Linux OS version 20.04**
- **R version 4.1.2**
- **RStudio version 1.4.1717**

System Specifications

- **Ubuntu Linux OS version 20.04**
- **R version 4.1.2**
- **RStudio version 1.4.1717**

System Specifications

- **Ubuntu Linux OS version 20.04**
- **R version 4.1.2**
- **RStudio version 1.4.1717**

R version 4.1.0 or higher

Pre-requisites

Pre-requisites

- **Basics of R Programming**

Pre-requisites

- **Basics of R Programming**
- **Basics of Machine Learning**

Pre-requisites

- **Basics of R Programming**
- **Basics of Machine Learning**

Pre-requisites

- **Basics of R Programming**
- **Basics of Machine Learning**

If not, please access the relevant tutorials on

https://spoken-tutorial.org/

Hierarchical Clustering

Hierarchical Clustering

- **It is a method that works by grouping data into a tree of clusters**

Hierarchical Clustering

- It is a method that works by grouping data into a tree of clusters
- It begins by treating every data point as a separate cluster

Hierarchical Clustering

- **It is a method that works by grouping data into a tree of clusters**
- **It begins by treating every data point as a separate cluster**
- **Then it forms a `dendrogram` by combining the two closest clusters**

Dendrogram

Dendrogram

- **It is a tree-like diagram**

Dendrogram

- It is a tree-like diagram
- It represents the clusters formed by hierarchical clustering

Types of Hierarchical Clustering

Types of Hierarchical Clustering

There are two types of hierarchical clustering:

Types of Hierarchical Clustering

There are two types of hierarchical clustering:

- **Agglomerative clustering**

Types of Hierarchical Clustering

There are two types of hierarchical clustering:

- **Agglomerative clustering**
- **Divisive clustering**

Agglomerative Clustering

Agglomerative Clustering

- **It uses a bottom-up approach**

Agglomerative Clustering

- It uses a bottom-up approach
- Each data point starts in its own cluster

Divisive Clustering

Divisive Clustering

- **It uses a top-down approach**

Divisive Clustering

- It uses a top-down approach
- All data points start in the same cluster

Agglomerative Clustering

In this tutorial, we will focus on agglomerative clustering

Agglomerative Clustering

Agglomerative Clustering

- **This is the most common type of hierarchical clustering**

Agglomerative Clustering

- **This is the most common type of hierarchical clustering**
- **In this type, the `dendrogram` is built by starting from the leaves**

Agglomerative Clustering

- **This is the most common type of hierarchical clustering**
- **In this type, the `dendrogram` is built by starting from the leaves**
- **Clusters are then combined up the trunk**

Linkage

Linkage

- **Linkage defines the dissimilarity between two groups of observations**

Linkage

- **Linkage defines the dissimilarity between two groups of observations**
- **There are primarily 4 types of linkages**

Types of Linkage

Types of Linkage

- **Complete: It is the maximum pairwise dissimilarity between observations in two different clusters**

Types of Linkage

- **Complete: It is the maximum pairwise dissimilarity between observations in two different clusters**
- **Single: It is the minimum pairwise dissimilarity between observations in two different clusters**

Types of Linkage

Types of Linkage

- **Average: It is the mean pairwise dissimilarity between observations in two different clusters**

Types of Linkage

- **Average: It is the mean pairwise dissimilarity between observations in two different clusters**
- **Centroid: It is the dissimilarity between centroids of two different clusters**

Advantages of Hierarchical Clustering

Advantages of Hierarchical Clustering

- **It gives homogeneous clusters**

Advantages of Hierarchical Clustering

- **It gives homogeneous clusters**
- **One can decide the number of clusters based on dendrograms**

Advantages of Hierarchical Clustering

- It gives homogeneous clusters
- One can decide the number of clusters based on dendrograms
- It is mathematically very easy to understand

Applications of Hierarchical Clustering

Applications of Hierarchical Clustering

- **It can be used to cluster shoppers based on past shopping history**

  https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

Applications of Hierarchical Clustering

- **It can be used to build phylogenetic trees that show evolutionary relationships**

  https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

Hierarchical Clustering

Let's implement Hierarchical Clustering on the iris dataset

Download Files

We will use:

Download Files

We will use:

- **A script file HClust.R**

Download Files

We will use:

- **A script file HClust.R**

Download Files

We will use:

- A script file HClust.R

Download this file from the Code files link of this tutorial

Make a copy and then use it for practising

Summary

We have learnt about:

- **Hierarchical Clustering**
- **Types of Hierarchical Clustering**
- **Advantages of Hierarchical Clustering**

Summary

- **Linkage and its types**
- **Applications of Hierarchical Clustering**
- **Hierarchical Clustering on iris dataset**

Assignment

Assignment

- **Apply hierarchical clustering on** `PimaIndiansDiabetes` **dataset**

Assignment

- **Apply hierarchical clustering on `PimaIndiansDiabetes` dataset**
- **Install and import the `mlbench` package**

Assignment

- **Run the**
  `data(PimaIndiansDiabetes2)`
  command to load the dataset

Assignment

- **Run the** `data(PimaIndiansDiabetes2)` **command to load the dataset**
- **Compare between various linkage methods**

About the Spoken Tutorial Project

- Watch the video available at https://spoken-tutorial.org/What_is_a_Spoken_Tutorial
- It summarises the Spoken Tutorial project
- If you do not have good bandwidth, you can download and watch it

Spoken Tutorial Workshops

The Spoken Tutorial Project Team

- **Conducts workshops using spoken tutorials**
- **Gives certificates to those who pass an online test**
- **For more details, please write to contact@spoken-tutorial.org**

Answers for THIS Spoken Tutorial

- Questions in THIS Spoken Tutorial?
- Visit https://forums.spoken-tutorial.org
- Choose the minute and second where you have the question
- Explain your question briefly
- The FOSSEE project will ensure an answer

fossee better education

You will have to register to ask questions

Forum to answer questions

- Questions not related to the Spoken Tutorial?
- Do you have general/technical questions on the Software?
- Please visit the FOSSEE Forum

  https://forums.fossee.in/
- Choose the Software and post your question

Textbook Companion Project

- The FOSSEE team coordinates the coding of solved examples of popular books and case study projects
- We give certificates to those who do this

For more details, please visit these sites:

https://r.fossee.in/

https://fossee.in/

Acknowledgements

- **The** Spoken Tutorial **and** FOSSEE **projects are funded by the** Ministry of Education**, Govt. of India**

About the Contributors

- **This tutorial is contributed by Tanmay Srinath and Madhuri Ganapathi, IIT Bombay**