K-Means 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

14 March 2022

Learning Objectives

Learning Objectives

We will learn about:

Learning Objectives

We will learn about:

- **k-means Clustering**

Learning Objectives

We will learn about:

- **k-means Clustering**
- **Benefits of k-means Clustering**

Learning Objectives

We will learn about:

- **k-means Clustering**
- **Benefits of k-means Clustering**
- **Applications of k-means Clustering**

Learning Objectives

We will learn about:

- **k-means Clustering**
- **Benefits of k-means Clustering**
- **Applications of k-means Clustering**
- **k-means++ clustering**

Learning Objectives

We will learn about:

- **k-means Clustering**
- **Benefits of k-means Clustering**
- **Applications of k-means Clustering**
- **k-means++ clustering**
- **Different k-means++ models on iris data**

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/

K-means Clustering

K-means Clustering

- **It partitions `n` observations into `k` clusters**

K-means Clustering

- **It partitions `n` observations into `k` clusters**
- **Observations are homogenous within each cluster**

K-means Clustering

- **It partitions `n` observations into `k` clusters**
- **Observations are homogenous within each cluster**
- **Each observation belongs to a cluster with the nearest cluster mean**

k-means Clustering

- **k-means Clustering is a powerful algorithm**

Benefits of k-means Clustering

Benefits of k-means Clustering

- **k-means Clustering is relatively simple to implement**

Benefits of k-means Clustering

- **k-means Clustering is relatively simple to implement**
- **k-means Clustering scales well on large datasets**

Applications of k-means Clustering

Applications of k-means Clustering

- **Customer Segmentation**
  https://archive.ics.uci.edu/ml/datasets/online+retail

Applications of k-means Clustering

- **Customer Segmentation**
  https://archive.ics.uci.edu/ml/datasets/online+retail
- **Ailment Diagnosis**
  https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

Optimising k-means

Optimising k-means

- **The basic form of k-means clustering is not optimal**

Optimising k-means

- **The basic form of k-means clustering is not optimal**
- **It depends a lot on the initialisation of clusters**

Optimising k-means

- **The basic form of k-means clustering is not optimal**
- **It depends a lot on the initialisation of clusters**
- **To overcome this drawback, we will use an optimised algorithm** `k-means++`

k-means++ algorithm

k-means++ algorithm

- **It is an algorithm for choosing the initial centroid locations**

k-means++ algorithm

- It is an algorithm for choosing the initial centroid locations
- The first center will be chosen at random

k-means++ algorithm

- It is an algorithm for choosing the initial centroid locations
- The first center will be chosen at random
- The next ones will be selected with a certain probability

k-means++ algorithm

- **This probability is proportional to the distance from the closest center**

k-means++ algorithm

- **This probability is proportional to the distance from the closest center**
- **By avoiding random initialisation, it provides faster results**

k-means++ Model

k-means++ Model

We will create 3 k-means++ models and compare their results

k-means++ Model

We will create 3 k-means++ models and compare their results

Let us implement k-means++ on the iris dataset

Download Files

We will use:

Download Files

We will use:

- **A script file K-means.R**

Download Files

We will use:

- **A script file K-means.R**

Download this file from the Code files link of this tutorial

Make a copy and then use it for practising

Summary

In this tutorial we have learnt about:

- k-means Clustering
- Benefits of k-means Clustering
- Applications of k-means Clustering
- k-means++ clustering
- Different k-means++ models on iris data

Assignment

Assignment

- **Apply** `k-means++` **on the** `PimaIndiansDiabetes` **dataset**

Assignment

- **Apply** `k-means++` **on the** `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 the models with different input parameters**

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**