

# Unsupervised Learning

**Spoken Tutorial Project**

**<https://spoken-tutorial.org>**

**National Mission on Education through ICT**

**<https://sakshat.ac.in>**

**Tanmay Srinath**

**Madhuri Ganapathi**

**IIT Bombay**

**2 January 2022**



# Learning Objectives

**We will learn about:**



# Learning Objectives

**We will learn about:**

- ▶ **Unsupervised Learning and its applications**



# Learning Objectives

We will learn about:

- ▶ **Unsupervised Learning and its applications**
- ▶ **k-means clustering on iris dataset**



# Learning Objectives

**We will learn about:**

- ▶ **Unsupervised Learning and its applications**
- ▶ **k-means clustering on iris dataset**
- ▶ **Measure the performance using Adjusted RAND Index**



# 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**

**Install 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 R on

<https://spoken-tutorial.org/>



# Unsupervised Learning



# Unsupervised Learning

- It is a technique that is applied on unlabelled datasets



# Unsupervised Learning

- ▶ It is a technique that is applied on unlabelled datasets
- ▶ It uses machine learning algorithms to analyse and cluster unlabelled data



# Unsupervised Learning

- ▶ It is a technique that is applied on unlabelled datasets
- ▶ It uses machine learning algorithms to analyse and cluster unlabelled data
- ▶ It deals with finding groups of data points with similar characteristics



# Types of Unsupervised Learning



# Types of Unsupervised Learning

- **Clustering: used for grouping search engine results**



# Types of Unsupervised Learning

- ▶ **Clustering: used for grouping search engine results**
- ▶ **Anomaly Detection: used to detect fraudulent transactions**



# k-means Clustering

Implement **k-means clustering** on the iris dataset



# Download Files

**We will use:**



# Download Files

We will use:

► **A script file** `Clustering.R`



# Download Files

We will use:

▶ **A script file** `Clustering.R`



# Download Files

We will use:

▶ A script file `Clustering.R`

Download this file from the **Code files** link of this tutorial

Make a copy and then use it for practising



# Posing the Problem

- Can we group data based on sepal and petal dimensions?



# Posing the Problem

- ▶ Can we group data based on sepal and petal dimensions?
- ▶ If so, do the groups represent the original species label accurately?



# Solution

- **The answer to this problem is to use a clustering algorithm**



# Finding Number of Clusters

- For real-life unlabelled data, we should find the optimal number of clusters



# Finding Number of Clusters

- ▶ For real-life unlabelled data, we should find the optimal number of clusters
- ▶ For this we will use the Elbow Method



# Adjusted RAND Index

- This is a measure of similarity between two clusters



# Adjusted RAND Index

- ▶ This is a measure of similarity between two clusters
- ▶ It ranges from  $-1$  to  $+1$ , where  $-1$  is bad and  $+1$  is good



# Summary

## In this tutorial we have learnt:

- ▶ **Unsupervised Learning and its applications**
- ▶ **k-means clustering on iris dataset**
- ▶ **Measure the performance using Adjusted RAND Index**



# Assignment

- **Using inbuilt PlantGrowth dataset, perform k-means clustering**



# Assignment

- ▶ **Using inbuilt** `PlantGrowth` dataset, **perform** `k-means` clustering
- ▶ **Evaluate using** Adjusted RAND index



# About the Spoken Tutorial Project

- ▶ Watch the video available at [https://spoken-tutorial.org/What\\_is\\_a\\_Spoken\\_Tutorial](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](mailto: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

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**

