Women Techmakers
Algiers

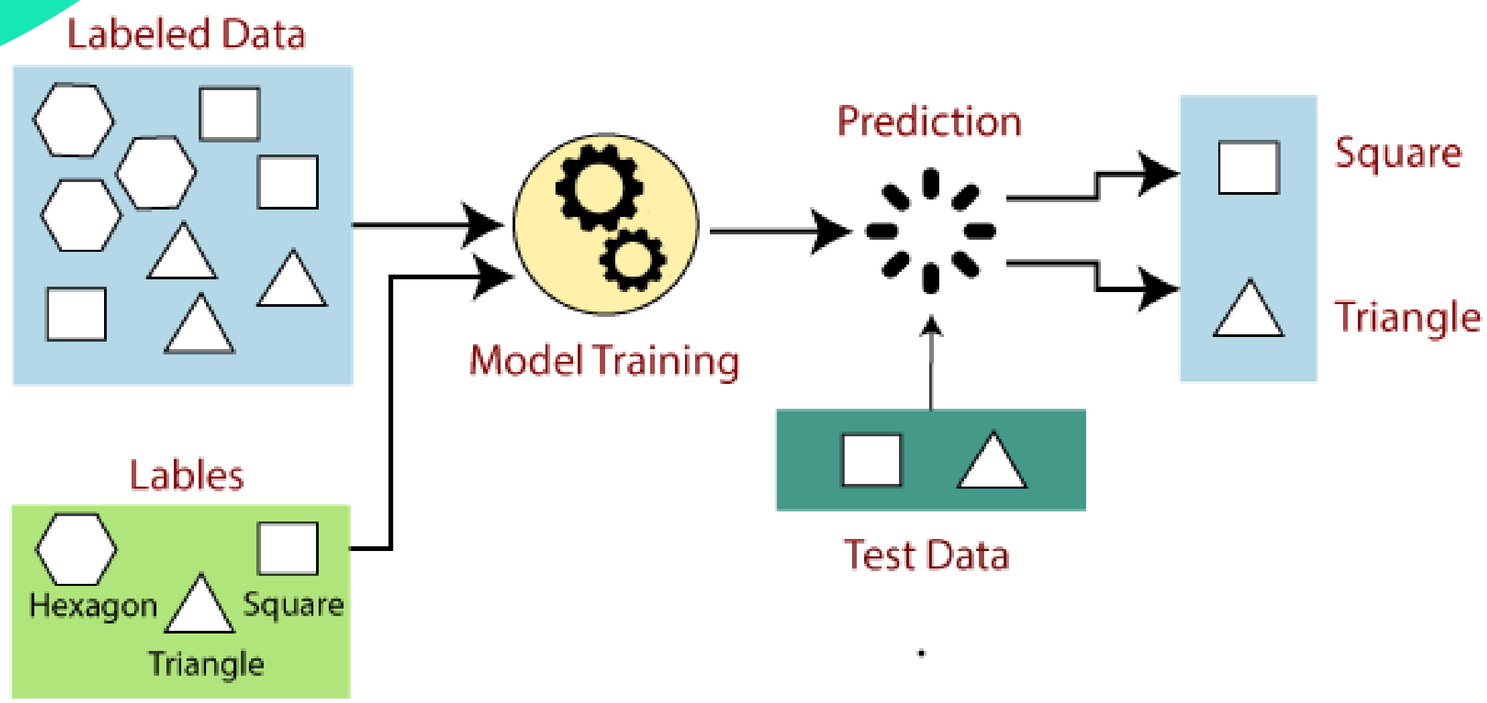
Unsupervised learning

Presented by Abla Hagani

Workshop Plan

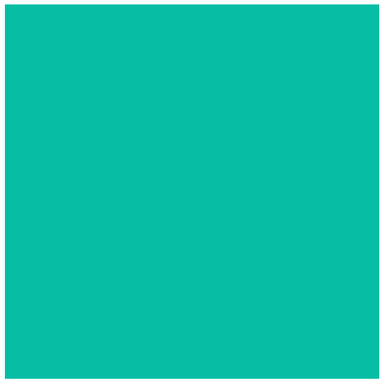
1. Introduction to Unsupervised Learning
2. Types of Unsupervised Learning
3. Key Concepts
4. Common Algorithms
5. Model Evaluation
6. Hands-On Example with Jupyter Notebook

Supervised learning models uses labels to learn



Supervised learning models uses labels to learn

Shape 01

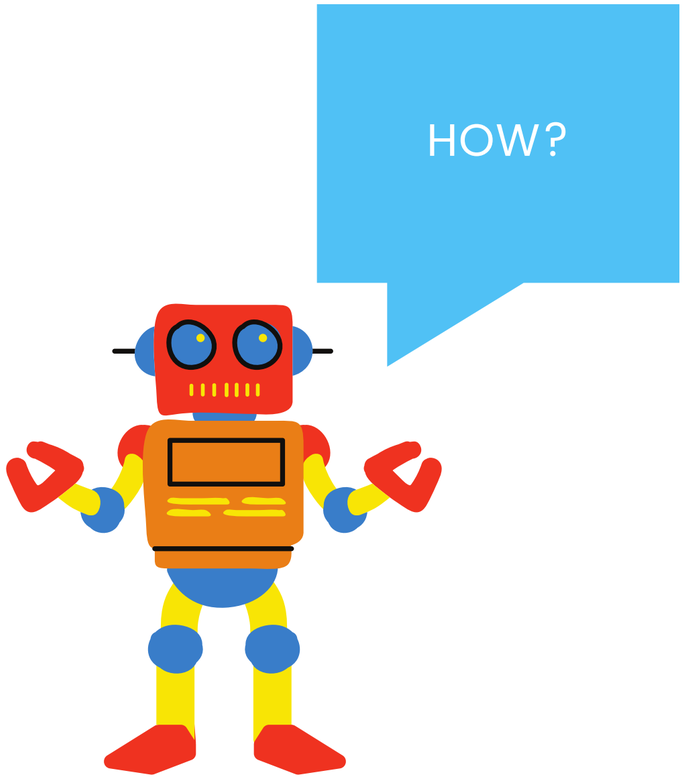


Features

Number of sides: 4
All sides are equal

Class: **Square**

But in Unsupervised learning we don’t have labels



In unsupervised learning, the algorithm explores the data and tries to group similar data points together or identify underlying structures without any guidance on what the outcomes should be.

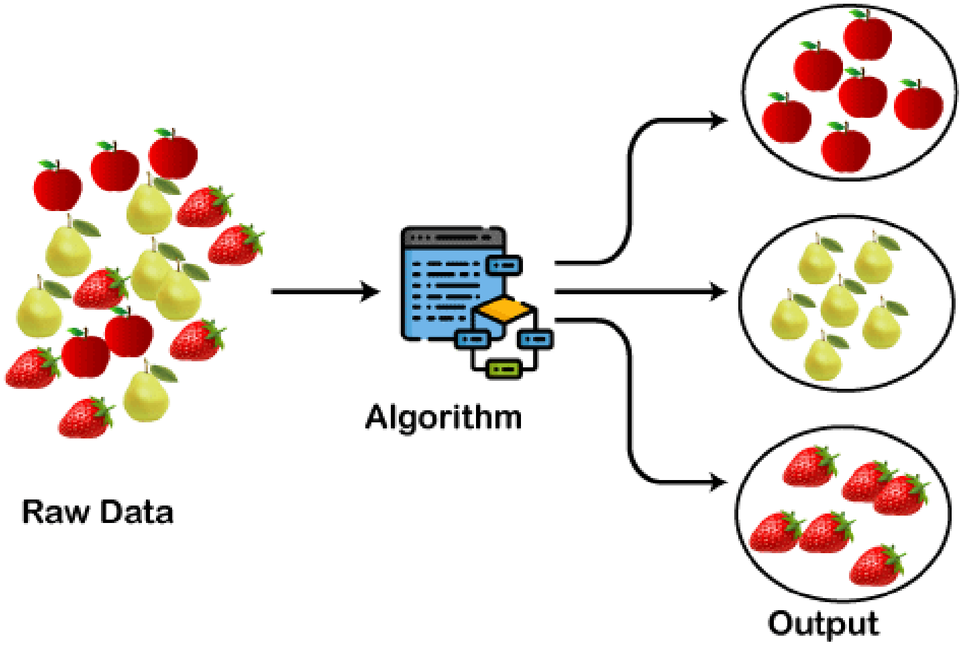
Application domains

- **Market Segmentation:** Clustering is used to identify distinct customer groups based on purchasing behavior and demographics, aiding in targeted marketing strategies.
- **Anomaly Detection:** Algorithms detect unusual patterns or outliers in data, useful for cybersecurity, fraud detection, and equipment maintenance.
- **Recommendation Systems:** Techniques like collaborative filtering and matrix factorization provide personalized recommendations by analyzing user behavior and preferences.
- **Image and Document Clustering:** Algorithms group similar images or documents, enabling tasks like image organization and document categorization.

Types of Unsupervised Learning

Clustering

Clustering is grouping similar data points into clusters based on some similarity measure without prior knowledge of class labels. Common clustering algorithms include K-means clustering, hierarchical clustering, DBSCAN, and Gaussian mixture models (GMM).

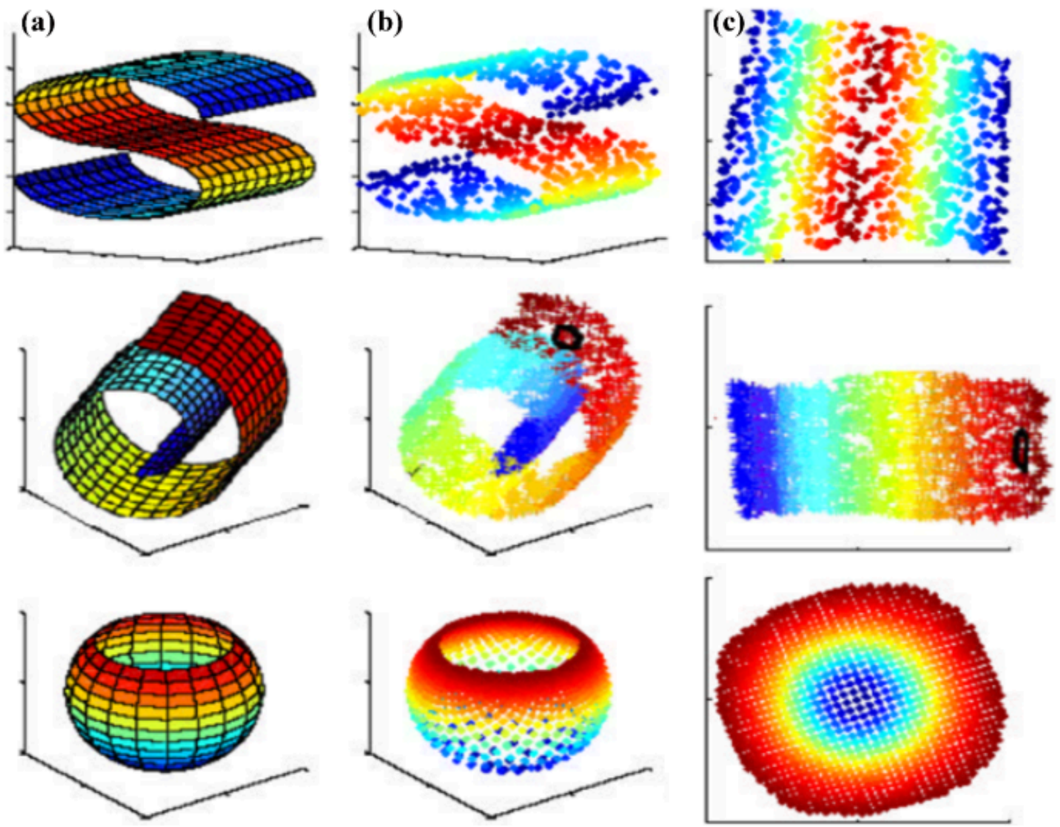


Association

Association rule learning is unsupervised learning where the algorithm discovers interesting relationships or associations between variables in large datasets.

Dimensionality Reduction:

Popular algorithms used for dimensionality reduction include principal component analysis (PCA) and Singular Value Decomposition (SVD). These algorithms seek to transform data from high-dimensional spaces to low-dimensional spaces without compromising meaningful properties in the original data.



Clustering

- Unsupervised learning technique that groups of unlabeled data into clusters.

- The main purpose of clustering is to group similar data points (samples) into clusters while maximizing inter-cluster variance and minimizing the intra-cluster variance.

- This similarity depends on several patterns in the training dataset, such as shape, size, color...

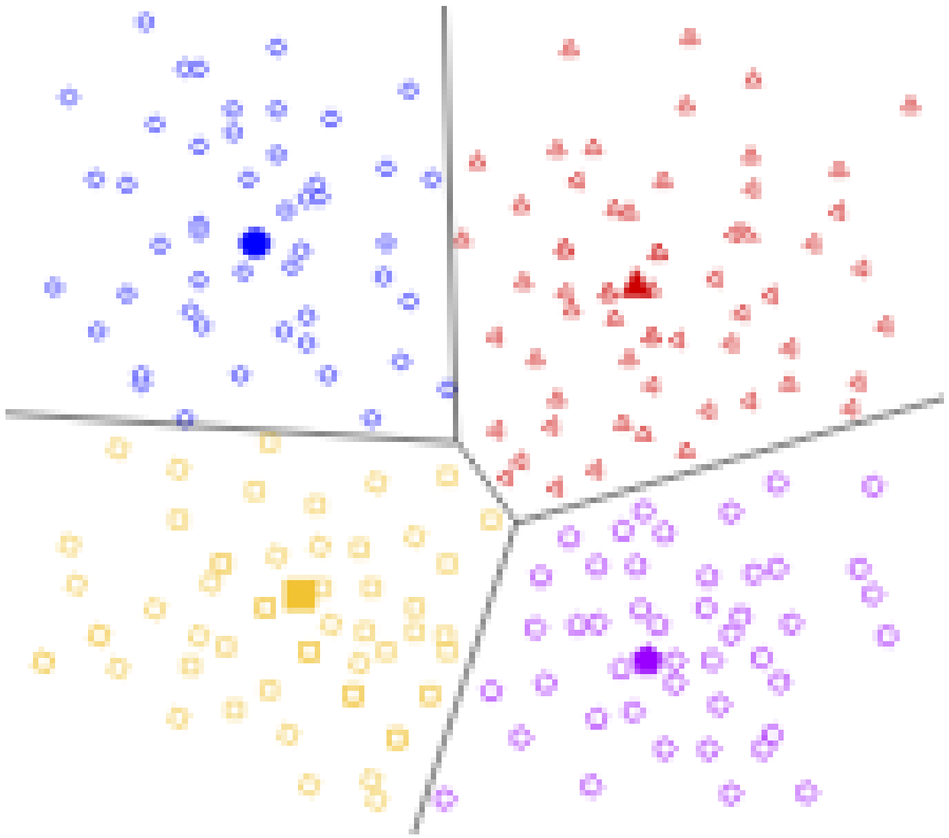
Principle

In clustering, there are several approaches. The choice of the appropriate technique depends mainly on the dataset type and the distribution of data. Among the most commonly used approaches in clustering :

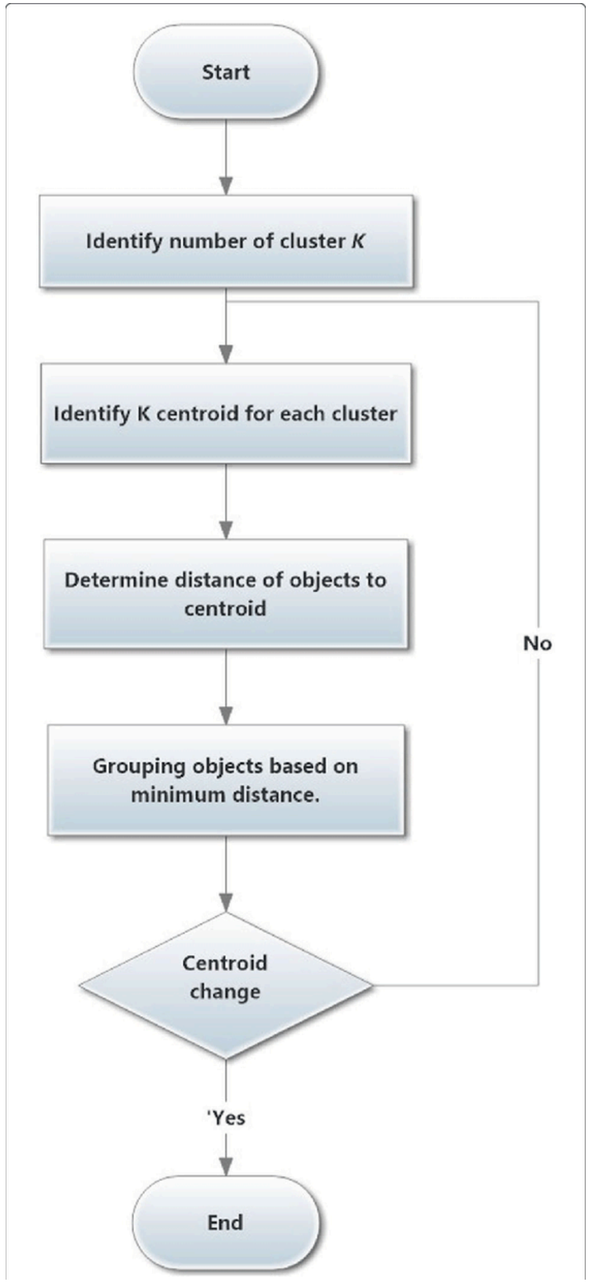
- Hierarchical Clustering.
- Centroid-based Clustering.
- Density-based Clustering.
- Distribution-based Clustering.

Centroid based clustering

Centroid-based clustering is a type of clustering algorithm where clusters are represented by a central point, called the centroid. Each data point is assigned to the cluster with the nearest centroid.



K-means Algorithm



Some Math :

- The distance between a data point X and a centroid C is calculated based on the Euclidean distance as follows:

$$D = \sqrt{\sum_{i=1}^{N}(X_i - C_i)^2}$$,N is the dimensionality of X and C

- The new cluster C' is computed by taking the mean of all data points assigned to that cluster, as follow:

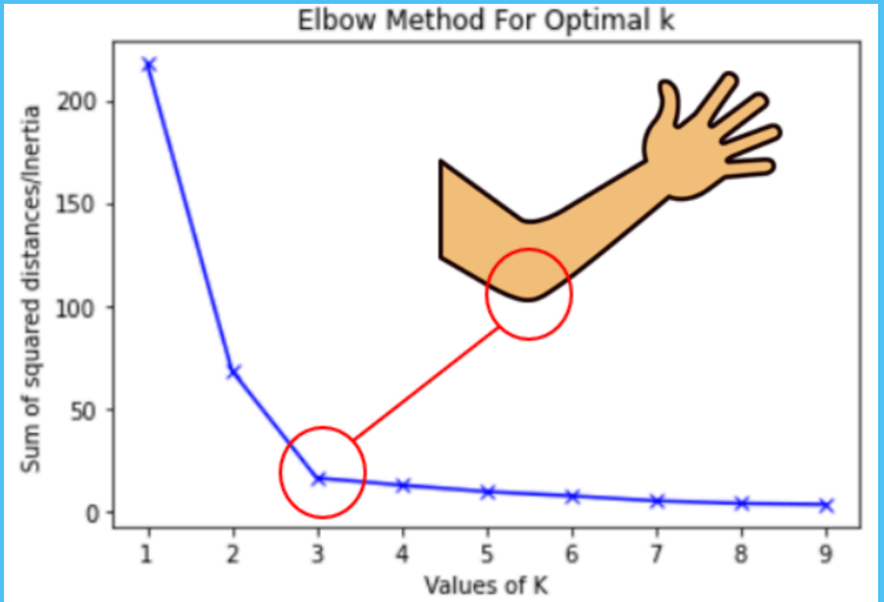
$$C' = \frac{1}{m}\sum_{j=1}^{m} X_j$$

How to choose the K?

The Elbow Method helps determine the optimal number of clusters in K-Means clustering by analyzing the variance explained as a function of the number of clusters.

1. Run K-Means for a range of k values.
2. Calculate the Within-Cluster Sum of Squares (WCSS) for each kkk.
3. Plot WCSS versus the number of clusters.
4. Identify the "elbow" point where the decrease in WCSS slows down.

- Result: The elbow point indicates the optimal kkk, balancing compactness and simplicity.



Line plot between K and inertia

$$\text{WCSS} = \sum_{C_k}^{C_n} \left(\sum_{d_i in\, C_i}^{d_m} distance(d_i, C_k)^2 \right)$$

Where,

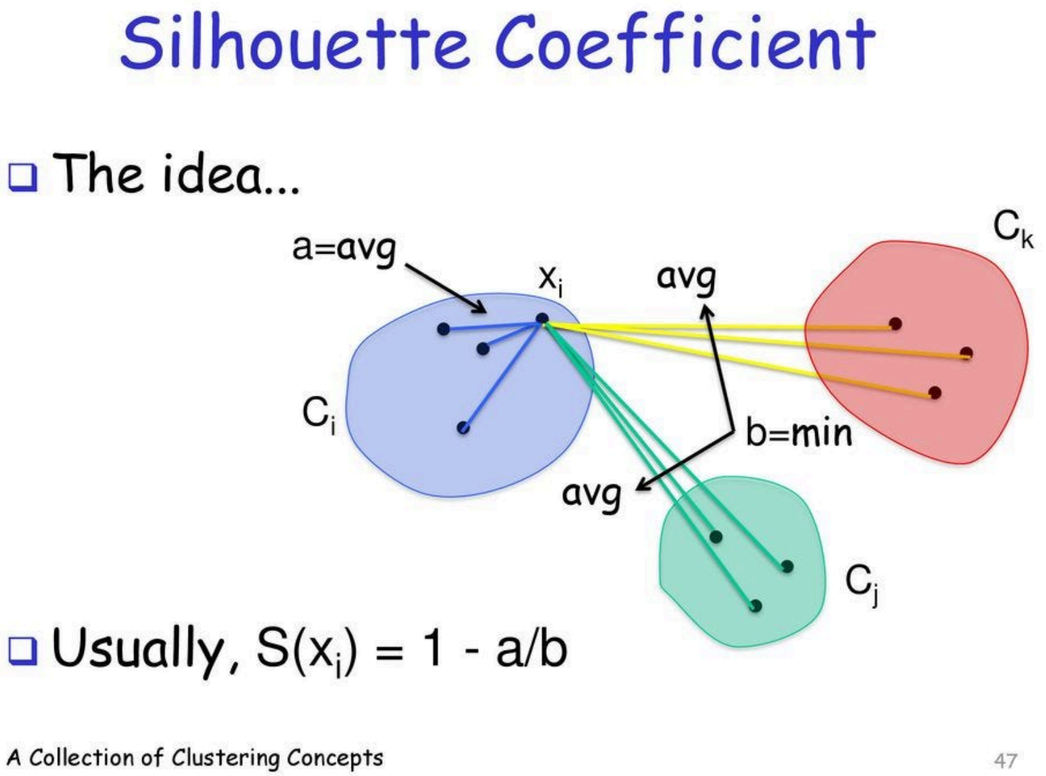
C is the cluster centroids and d is the data point in each Cluster.

Evaluation Methods

Internal validation methods:

1. Sillouette coefficient

evaluates the quality of clusters by measuring how similar each point is to its own cluster (cohesion) compared to other clusters (separation).



Evaluation Methods

Internal validation methods:

1. Sillouette coefficient

b(i): distance between i and its **next nearest cluster centroid**

(cluster labels assigned from ground-truth classes)

a(i): average distance between i and all of the other points in **its own cluster**

For a single point, i

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$-1 \leq s(i) \leq 1$$

Sprawling, overlapped clusters

Tight, well-separated clusters

Evaluation Methods

External validation methods:

Despite internal validation methods, external validation methods require a prior knowledge of the true labels or clusters.

- In these techniques, the set S represents the set of predicted clusters {C1, C2, ……, Cn}, and another set of clusters representing the true labels is defined by P = {P1, P2, ……, Pn}. In this scenario, similar classification metrics can be employed. External validation is not recommended in clustering as it's difficult to know if a cluster Ck defines a class Pk.

Hands on Experience with jupyter notebook

Women Techmakers
Algiers

Thank you!

Together we thrive in tech