CS-37 Machine Learning with Python

Unit– 03 Unsupervised Learning

- Clustering :
- Data using k-means clustering
- compressing image using vector quantization
- mean shift clustering model
- agglomerative clustering, case study implementation using Python.

Clustering in Machine Learning

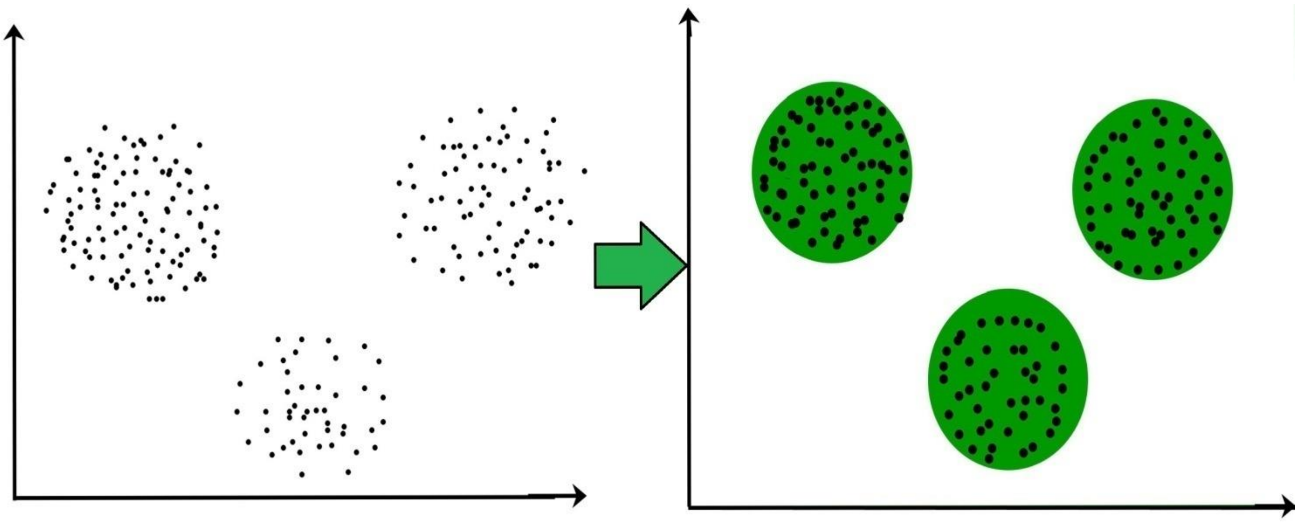
- ✓ In real world, not every **data we work upon has a target variable**.
- ✓ Have you ever wondered how Netflix groups similar movies together or how Amazon organizes its vast product catalog? These are **real-world applications of clustering**.
- ✓ This kind of data cannot be analyzed using supervised learning algorithms.
- ✓ When the goal is to group similar data points in a dataset, then we use cluster analysis.
- ✓ In this guide, we'll learn understand concept of clustering, its applications, and some popular clustering algorithms.

What is Clustering?

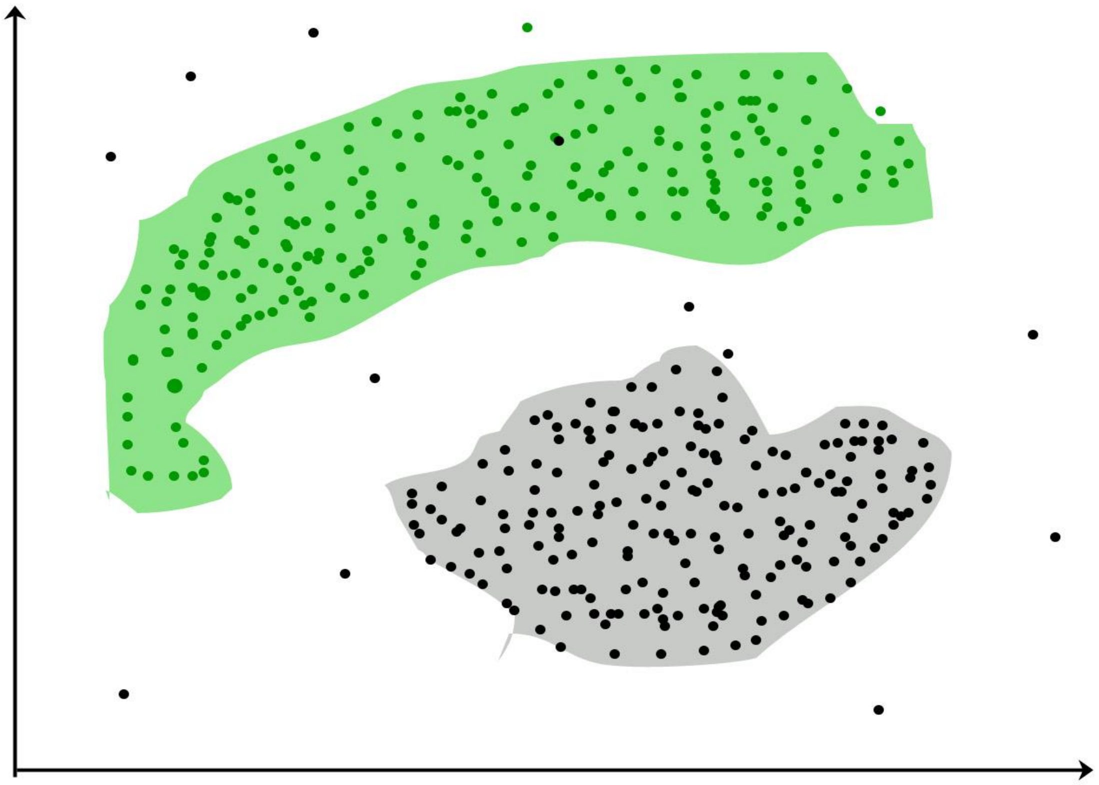
- ✓ The task of **grouping data points based on their similarity with each other is called Clustering or Cluster Analysis**.
- ✓ This method is defined under the branch of unsupervised learning, which aims at gaining insights from unlabelled data points.
- ✓ Think of it as you have a dataset of customers shopping habits.
- ✓ **Clustering can help you group customers with similar purchasing behaviors, which can then be used for**

targeted marketing, product recommendations, or customer segmentation

- ✓ For Example, In the graph given below, we can clearly see that there are 3 circular clusters forming on the basis of distance.



- ✓ Now it is not necessary that the clusters formed **must be circular in shape**.
- ✓ The shape of clusters can be arbitrary.
- ✓ There are many algorithms that work well with detecting arbitrary shaped clusters.
- ✓ For example, In the below given graph we can see that the clusters formed are not circular in shape.



• Types of Clustering

There are 2 types of clustering that can be performed to group similar data points:

- **Hard Clustering:** In this type of clustering, each data point belongs to a cluster completely or not. For example, Let's say there are 4 data point and we have to cluster them into 2 clusters. So each data point will either belong to cluster 1 or cluster 2.

Data Points	Clusters
A	C1
B	C2
C	C2
D	C1

- **Soft Clustering:** In this type of clustering, instead of assigning each data point into a separate cluster, a probability or likelihood of that point being that cluster is evaluated. For example, Let’s say there are 4 data point and we have to cluster them into 2 clusters. So we will be evaluating a probability of a data point belonging to both clusters. This probability is calculated for all data points.

Data Points	Probability of C1	Probability of C2
A	0.91	0.09
B	0.3	0.7
C	0.17	0.83
D	1	0

Uses of Clustering

Now before we begin with types of clustering algorithms, we will go through the use cases of Clustering algorithms. Clustering algorithms are majorly used for:

- **Market Segmentation:** Businesses use clustering to group their customers and use targeted advertisements to attract more audience.
- **Market Basket Analysis:** Shop owners analyze their sales and figure out which items are majorly bought together by the customers. For example, In USA, according to a study diapers and beers were usually bought together by fathers.
- **Social Network Analysis:** Social media sites use your data to understand your browsing behavior and provide you

with targeted friend recommendations or content recommendations.

- **Medical Imaging:** Doctors use Clustering to find out diseased areas in diagnostic images like X-rays.
- **Anomaly Detection:** To find outliers in a stream of real-time dataset or forecasting fraudulent transactions we can use clustering to identify them.
- **Simplify working with large datasets:** Each cluster is given a cluster ID after clustering is complete. Now, you may reduce a feature set's whole feature set into its cluster ID. Clustering is effective when it can represent a complicated case with a straightforward cluster ID. Using the same principle, clustering data can make complex datasets simpler.

Types of Clustering Methods

- ✓ At the surface level, **clustering helps in the analysis of unstructured data.**
- ✓ **Graphing, the shortest distance, and the density of the data points are a few of the elements that influence cluster formation**.
- ✓ Clustering is the process of determining how related the objects are **based on a metric called the similarity measure**.
- ✓ Similarity metrics **are easier to locate in smaller sets of features and harder as the number of features increases**.
- ✓ Depending on the type of clustering algorithm being utilized, several techniques are employed to group the data from the datasets.
- ✓ In this part, the clustering techniques are described.

Various types of clustering algorithms are:

1. **Centroid-based Clustering (Partitioning methods):** Groups data points around central centroids, requiring a predefined number of clusters.
   e.g. (K-means ,K-medoids).

2. **Density-based Clustering (Model-based methods):** Identifies clusters as dense regions in data space, automatically determining the number of clusters.
   e.g.(*DBSCAN* (Density-Based Clustering Algorithm), *OPTICS* (Ordering Points To Identify the Clustering Structure)).

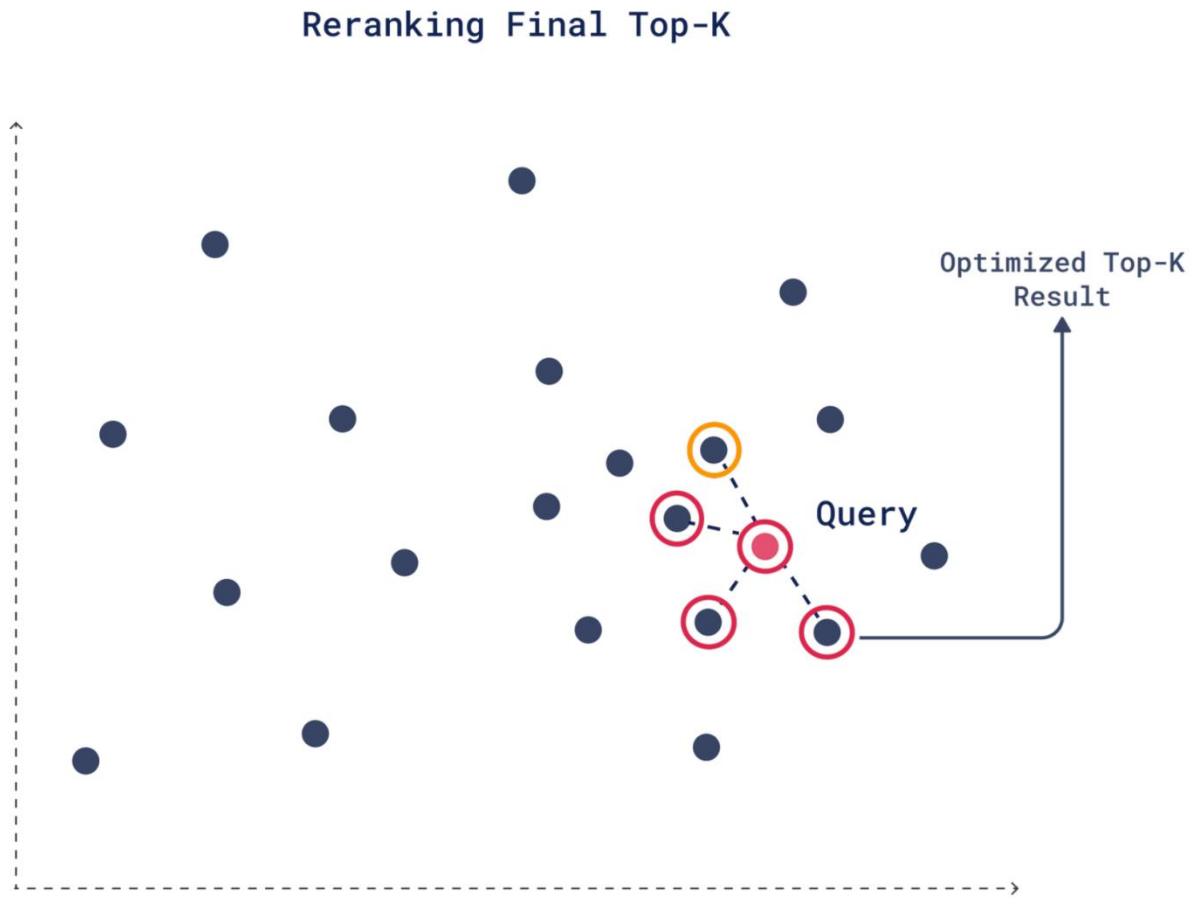
3. **Connectivity-based Clustering(** Hierarchical clustering:**):** Builds a hierarchy of clusters based on object similarity, represented in a dendrogram (e.g.: Agglomerative clustering and Divisive clustering).

- Two of the most popular soft clustering techniques are:

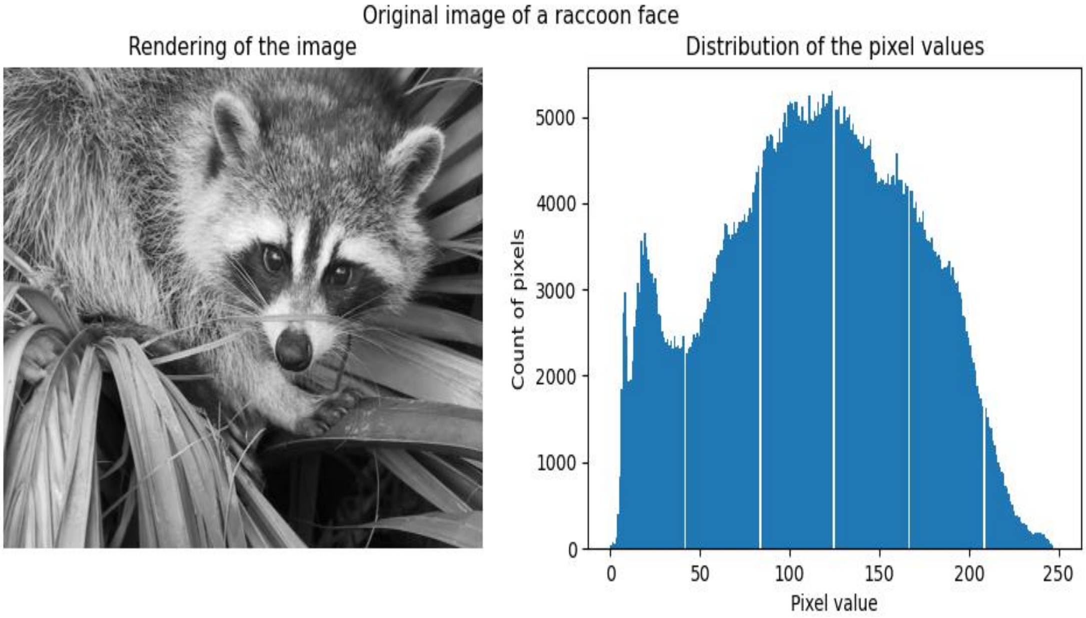
4. **Distribution-based Clustering**: Assumes data is generated from probability distributions and assigns points based on likelihood (e.g. Gaussian Mixture Model).
5. **Fuzzy Clustering**: Allows data points to belong to multiple clusters with varying degrees of membership (e.g. fuzzy clustering methods).

Vector Quantization

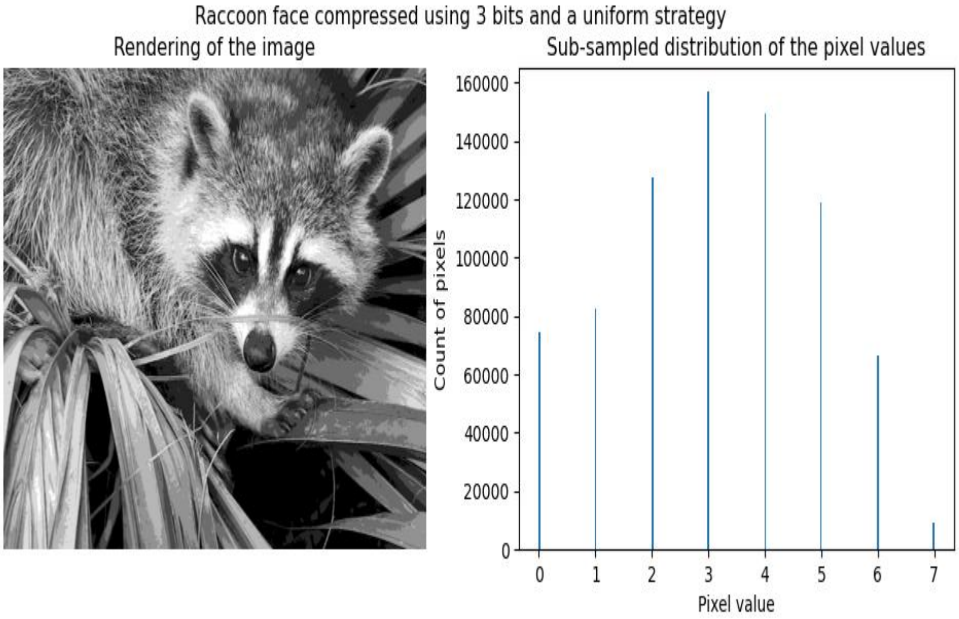
- ✓ Vector quantization is a data compression technique used to reduce the size of high-dimensional data.
- ✓ Compressing vectors reduces memory usage while maintaining nearly all of the essential information.
- ✓ This method allows for more efficient storage and faster search operations, particularly in large datasets.



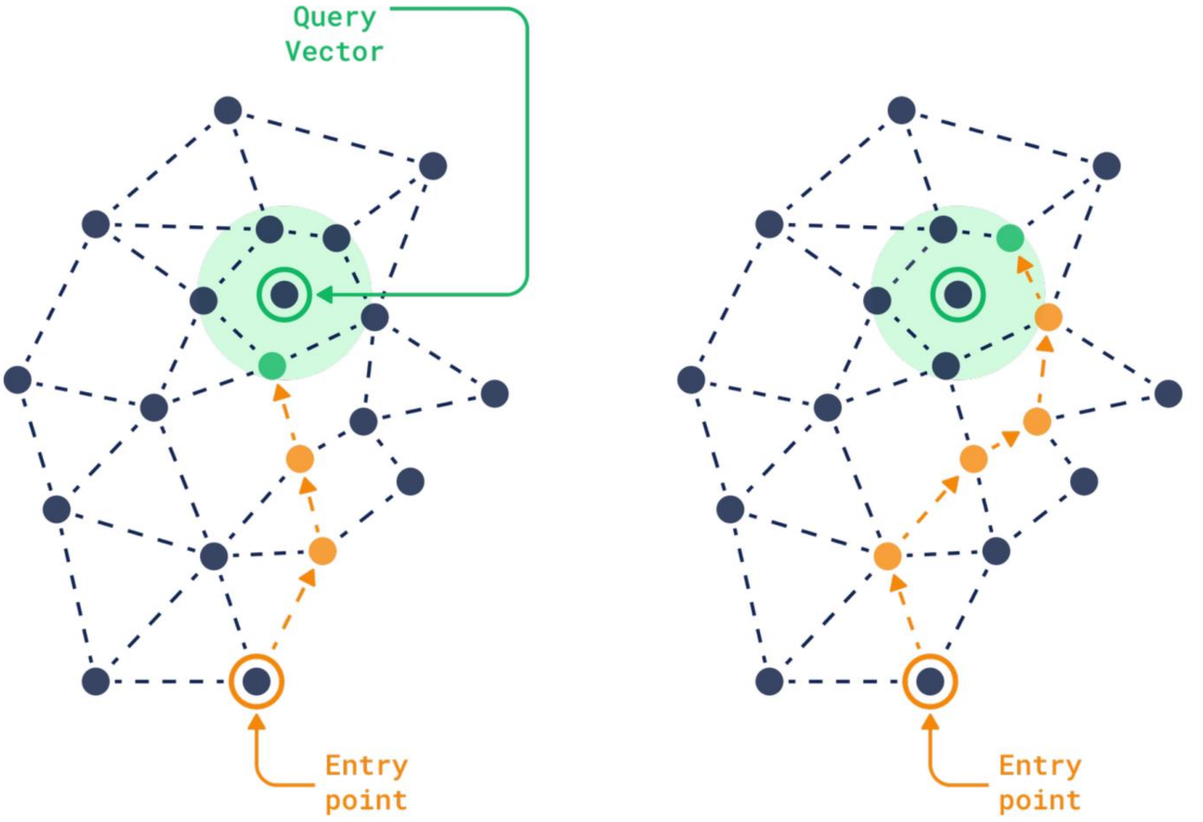
- ## Original image



- ## Compression via vector quantization



- **How vector quantization are Work**



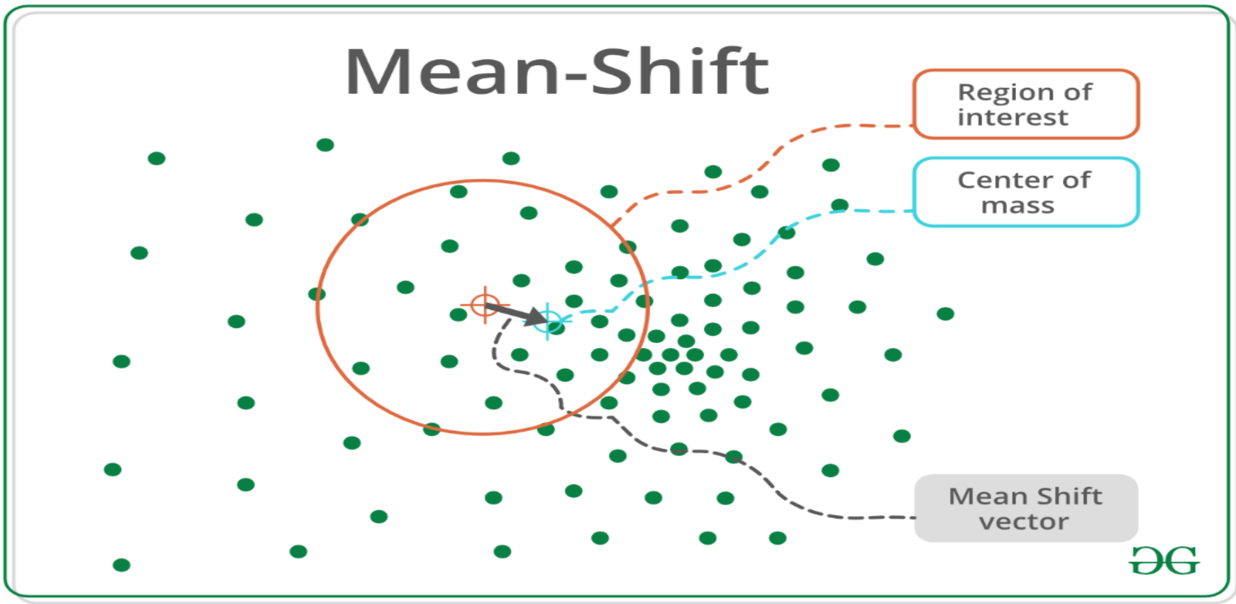
• Mean-Shift Clustering

✓ **Meanshift** is falling under the category of a clustering algorithm in contrast of Unsupervised learning that assigns the data points to the clusters iteratively by shifting points towards the mode (mode is the highest density of data points in the region, in the context of the Meanshift).

- ✓ As such, it is also known as the **Mode-seeking algorithm**. Mean-shift algorithm has applications in the field of image processing and computer vision.
- ✓ Unlike the popular K-Means cluster algorithm, mean-shift does not require specifying the number of clusters in advance.
- ✓ The number of clusters is determined by the algorithm with respect to the data.
- ✓ Mean-shift clustering is a non-parametric, density-based clustering algorithm that can be used to identify clusters in a dataset.
- ✓ It is particularly useful for datasets where the clusters have arbitrary shapes and are not well-separated by linear boundaries.
- ✓ The basic idea behind mean-shift clustering is to shift each data point towards the mode (i.e., the highest density) of the distribution of points within a certain radius.
- ✓ The algorithm iteratively performs these shifts until the points converge to a local maximum of the density function. These local maxima represent the clusters in the data.

- **Iterative Mode Search –**

  1. Initialize random seed and window W.
  2. Calculate the center of gravity (mean) of W.
  3. Shift the search window to the mean.
  4.Repeat Step 2 until convergence



Pros:

- Finds variable number of modes
- Robust to outliers
- General, application-independent tool
- Model-free, doesn't assume any prior shape like spherical, elliptical, etc. on data clusters
- Just a single parameter (window size h) where h has a physical meaning (unlike k-means)

Cons:

- Output depends on window size
- Window size (bandwidth) selecHon is not trivial
- Computationally (relatively) expensive (approx 2s/image)
- Doesn't scale well with dimension of feature space.

• Agglomerative Clustering in Machine Learning

- ✓ Agglomerative clustering is a hierarchical clustering algorithm that starts with each data point as its own cluster and iteratively merges the closest clusters until a stopping criterion is reached.
- ✓ It is a bottom-up approach that produces a dendrogram, which is a tree-like diagram that shows the hierarchical relationship between the clusters.
- ✓ The algorithm can be implemented using the scikit-learn library in Python.

• Agglomerative Clustering Algorithm

- ✓ Agglomerative Clustering is a hierarchical algorithm that creates a nested hierarchy of clusters by merging clusters in a bottom-up approach.
- ✓ This algorithm includes the following steps –

- ▪ Treat each data point as a single cluster
- ▪ Compute the proximity matrix using a distance metric
- ▪ Merge clusters based on a linkage criterion
- ▪ Update the distance matrix
- ▪ Repeat steps 3 and 4 until a single cluster remains

Why use Agglomerative Clustering?

- ✓ The Agglomerative clustering allows easy interpretation of relationships between data points.
- ✓ Unlike k-means clustering, we do not need to specify the number of clusters. It is very efficient and can identify small clusters.

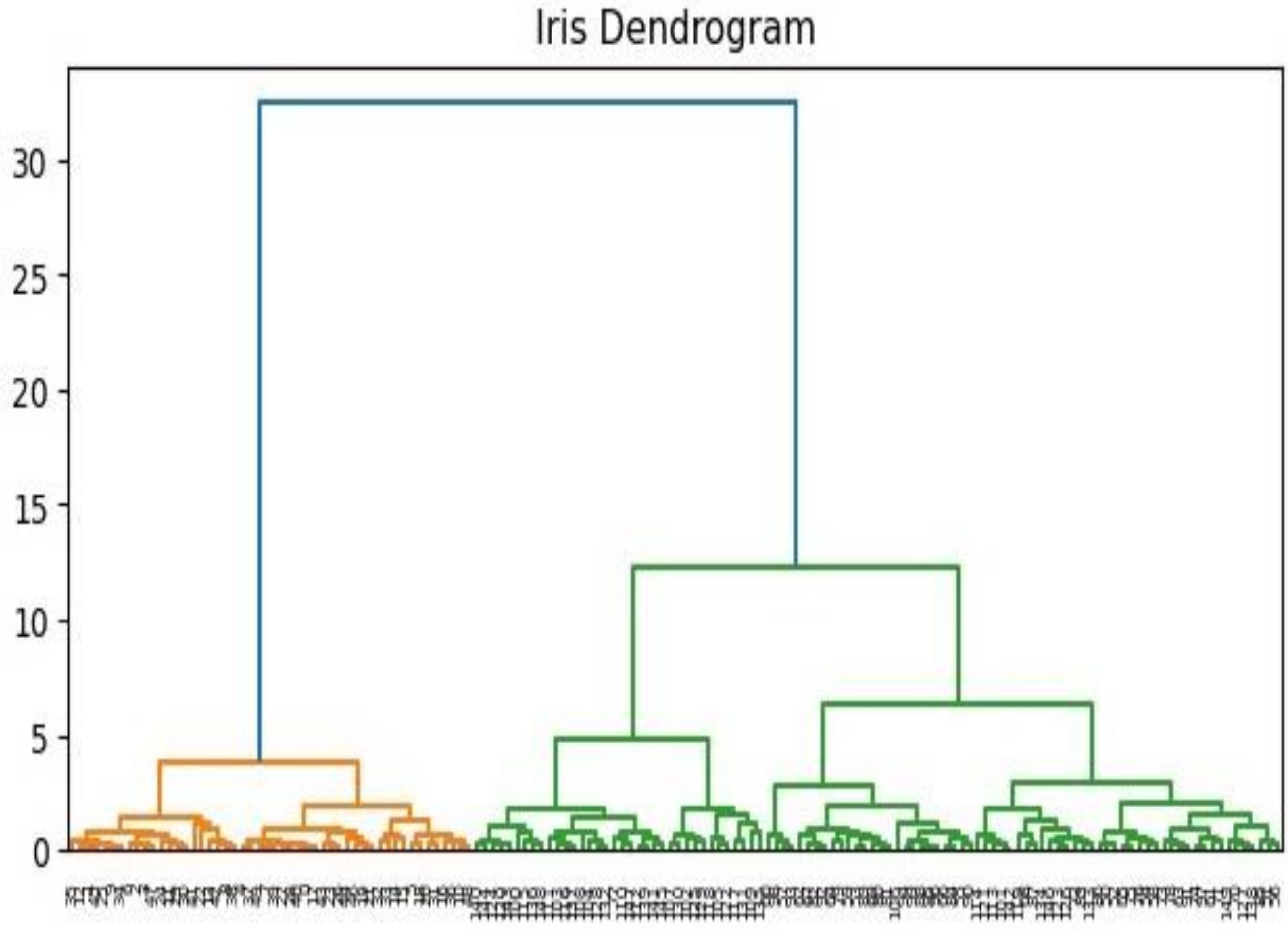
Implementation of Agglomerative Clustering in Python

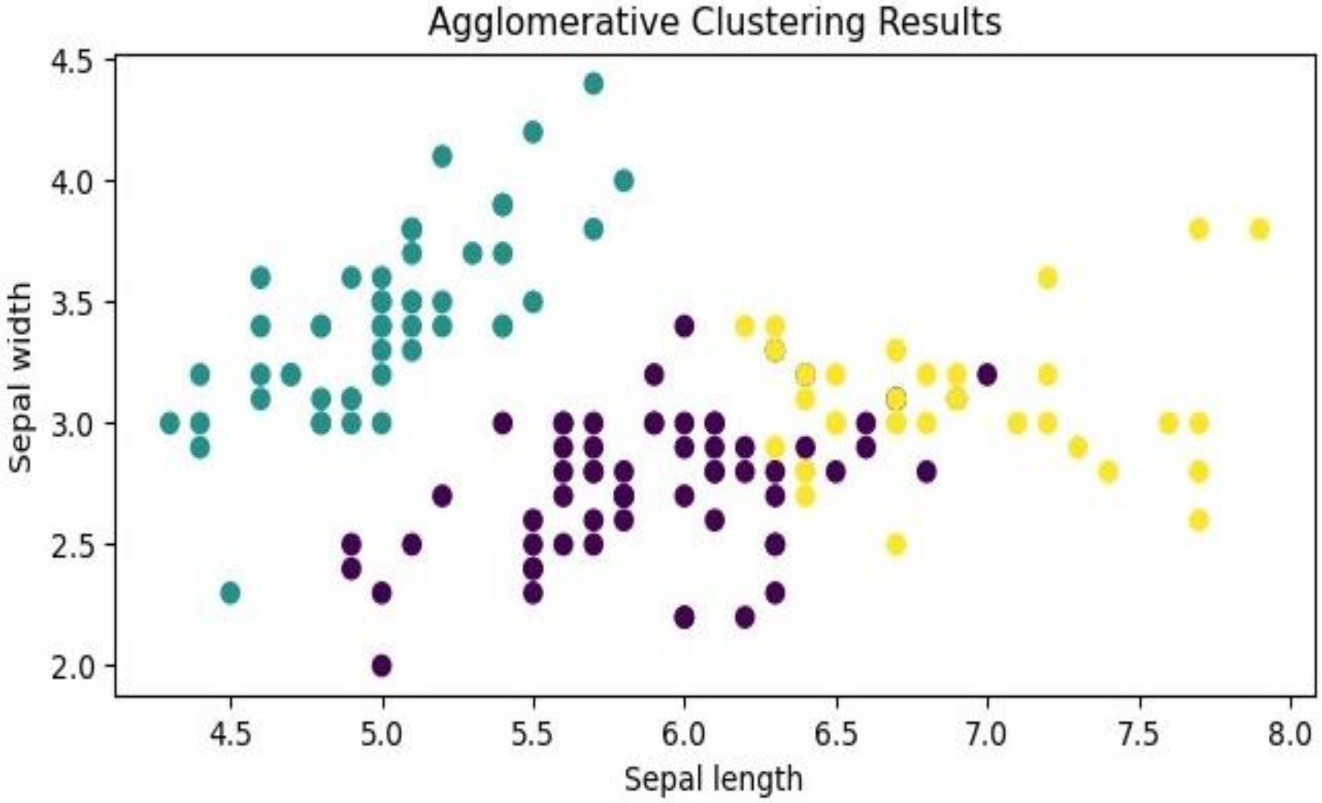
- ✓ We will use the iris dataset for demonstration.
- ✓ The first step is to import the necessary libraries and load the dataset.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
from sklearn.cluster import AgglomerativeClustering
from scipy.cluster.hierarchy import dendrogram, linkage

# Load the Iris dataset
iris = load_iris()
X = iris.data
y = iris.target
Z = linkage(X, 'ward')
```

```
# Plot the dendogram
plt.figure(figsize=(7.5, 3.5))
plt.title("Iris Dendrogram")
dendrogram(Z)
plt.show()

# create an instance of the AgglomerativeClustering class
model = AgglomerativeClustering(n_clusters=3)

# fit the model to the dataset
model.fit(X)
labels = model.labels_

# Plot the results
plt.figure(figsize=(7.5, 3.5))
plt.scatter(X[:, 0], X[:, 1], c=labels)
plt.xlabel("Sepal length")
plt.ylabel("Sepal width")
plt.title("Agglomerative Clustering Results")
plt.show()
```

Output:


Iris Dendrogram
30
25
20
15
10
5
0


Agglomerative Clustering Results
4.5
4.0
3.5
3.0
2.5
2.0
Sepal width
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
Sepal length

Advantages of Agglomerative Clustering

- Produces a dendrogram that shows the hierarchical relationship between the clusters.
- Can handle different types of distance metrics and linkage methods.
- Allows for a flexible number of clusters to be extracted from the data.
- Can handle large datasets with efficient implementations.

Disadvantages of Agglomerative Clustering

- Can be computationally expensive for large datasets.
- Can produce imbalanced clusters if the distance metric or linkage method is not appropriate for the data.
- The final result may be sensitive to the choice of distance metric and linkage method used.
- The dendrogram may be difficult to interpret for large datasets with many clusters.

Applications of Agglomerative Clustering

- Image Segmentation
- Document Clustering
- Customer Behaviour Analysis (Customer Segmentation)
- Market Segmentation
- Social Network Analysis

EXAM IMP QUESTIONS (UNIT 3)

One Mark Questions

1. Definition-based Questions:

1. **What is clustering in machine learning?**
   Ans: Clustering is the process of grouping similar data points based on their characteristics in an unsupervised manner.

2. **What is the main objective of clustering?**
   Ans: To group similar data points together without prior labels.

3. **Which type of machine learning does clustering belong to?**
   Ans: Unsupervised learning.

4. **What is the difference between hard clustering and soft clustering?**
   Ans: Hard clustering assigns each data point to exactly one cluster, whereas soft clustering assigns probabilities to data points for belonging to multiple clusters.

2. True/False Questions:

5. **K-Means clustering is an example of hierarchical clustering.** *(True/False)*
   Ans: False

6. **DBSCAN is a density-based clustering algorithm.** *(True/False)*
   Ans: True

7. **Agglomerative clustering follows a bottom-up approach.** *(True/False)*
   Ans: True

8. **Mean-Shift clustering requires the number of clusters to be specified beforehand.** *(True/False)*
   Ans: False

3. Fill in the Blanks:

9. **________ clustering is used to find clusters in datasets without specifying the number of clusters beforehand.**
   Ans: Density-based

10. **In Agglomerative Clustering, the clustering process starts with each data point as a _________.**
    Ans: Separate cluster

11. **The technique used in clustering to measure similarity is called _________.**
    Ans: Similarity measure

12. **Market segmentation is a real-world application of _________.**
    Ans: Clustering

4. Importance-Based Questions:

13. **What is an advantage of hierarchical clustering over K-means?**
    Ans: No need to specify the number of clusters in advance.

14. **Which clustering method is best for detecting arbitrary shaped clusters?**
    Ans: Density-based clustering (e.g., DBSCAN).

Two Mark Questions

1. Explain the difference between centroid-based and density-based clustering.
2. What are the advantages and disadvantages of Agglomerative Clustering?
3. How does Mean-Shift clustering work, and what is its major advantage?
4. Differentiate between hierarchical and partition-based clustering.
5. What are some real-world applications of clustering algorithms?

Three Mark Questions

1. Explain the concept of soft clustering with an example.
2. Describe the steps involved in the Agglomerative Clustering algorithm.
3. What are the key differences between K-Means and DBSCAN clustering algorithms?

4. How does hierarchical clustering work, and what are the two main types of hierarchical clustering?
5. What is vector quantization in clustering, and how does it help in data compression?

Five Mark Questions

1. Explain the concept of clustering in machine learning and its importance in real-world applications.
2. Discuss the different types of clustering methods with suitable examples.
3. Compare and contrast centroid-based clustering, density-based clustering, and hierarchical clustering.
4. Describe the Mean-Shift clustering algorithm, including its working, advantages, and disadvantages.
5. Explain Agglomerative Clustering in detail, including its working steps, advantages, disadvantages, and applications.