Unsupervised Learning

Nipun Batra

IIT Gandhinagar

August 30, 2025

The need for Unsupervised Learning

- Aids the search of patterns in data.
- Find features for categorization.
- Easier to collect unlabeled data.

The need for Unsupervised Learning

- Aids the search of patterns in data.
- Find features for categorization.
- Easier to collect unlabeled data.

Places where you will see unsupervised learning

- It can be used to segment the market based on customer preferences.
- A data science team reduces the number of dimensions in a large dataset to simplify modeling and reduce file size.

Clustering

Clustering

AIM: To find groups/subgroups in a dataset.

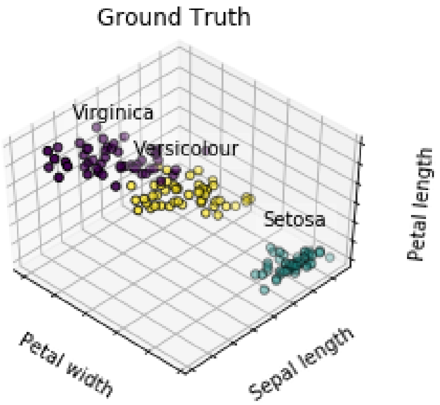
Clustering

AIM: To find groups/subgroups in a dataset.
REQUIREMENTS: A predefined notion of similarity/dissimilarity.

Clustering

AIM: To find groups/subgroups in a dataset.
REQUIREMENTS: A predefined notion of similarity/dissimilarity. **Examples:**
Market Segmentation: Customers with similar preferences in the same groups. This would aid in targeted marketing.

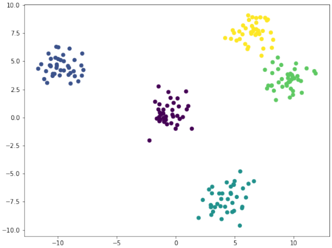
Clustering



Iris Data Set with ground truth

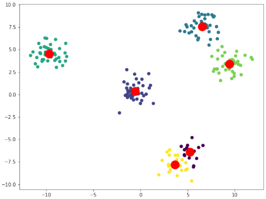
K-Means Clustering

- N points in a R^d space.
- C_i: set of points in the i^{th} cluster.
- $C_1 \cup C_2 \cup \ldots C_k = \{1, \ldots, n\}$
- $C_i \cap C_j = \{\phi\}$ for $i \neq j$

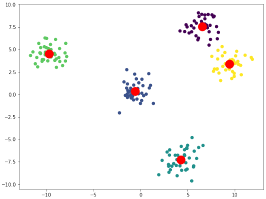


Dataset with 5 clusters

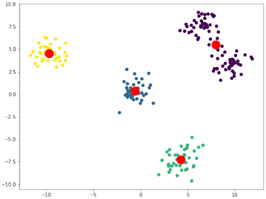
K-Means Clustering



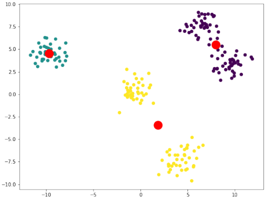
K=6



K=5



K=4



K=3

K-Means Intuition

- Good Clustering: Within the cluster the variation (WCV) is small.

K-Means Intuition

- Good Clustering: Within the cluster the variation (WCV) is small.
- Objective:

$$\min_{C_1,\ldots,C_k} \left(\sum_{i=1}^{k} WCV(C_i) \right)$$

K-Means Intuition

- Good Clustering: Within the cluster the variation (WCV) is small.
- Objective:

$$\min_{C_1,\ldots,C_k}\left(\sum_{i=1}^{k} WCV(C_i)\right)$$

Minimize the WCV as much as possible

K-Means Intuition

Objective:

$$\min_{C_1,\ldots,C_k} \left(\sum_{i=1}^{k} WCV(C_i) \right)$$

K-Means Intuition

Objective:

$$\min_{C_1,\ldots,C_k} \left(\sum_{i=1}^{k} WCV(C_i) \right)$$

$$WCV(C_i) = \frac{1}{|C_i|} \;\texttt{(Distance between all points)}$$

$$WCV(C_i) = \frac{1}{|C_i|} \sum_{a \in C_i} \sum_{b \in C_i} ||x_a - x_b||_2^2$$

where $|C_i|$ is the number of points in C_i

K-Means Algorithm

1. Randomly assign a cluster number i to every point (where $i \in \{1, \ldots n\}$)

K-Means Algorithm

1. Randomly assign a cluster number i to every point (where $i \in \{1, \ldots n\}$)

   1) For each cluster C_i compute the centroid (mean of all points in C_i over d dimensions)

K-Means Algorithm

1. Randomly assign a cluster number i to every point (where $i \in \{1, \ldots n\}$)

   1) For each cluster C_i compute the centroid (mean of all points in C_i over d dimensions)
   2) Assign each observation to the cluster which is the closest.

K-Means Algorithm

1. Randomly assign a cluster number i to every point (where $i \in \{1, \ldots n\}$)
2. Iterate until convergence:
   1) For each cluster C_i compute the centroid (mean of all points in C_i over d dimensions)
   2) Assign each observation to the cluster which is the closest.

Working of K-Means Algorithm

Why does K-Means work?

$$\text{Let, } x_i \in R^d = \text{Centroid for} i^{th} \text{cluster}$$
$$= \frac{1}{|C_i|} \sum_{a \in C_i} x_a$$

Why does K-Means work?

$$\text{Let, } x_i \in R^d = \text{Centroid for} i^{th} \text{cluster}$$
$$= \frac{1}{|C_i|} \sum_{a \in C_i} x_a$$

Then,

$$WCV(C_i) = \frac{1}{|C_i|} \sum_{a \in C_i} \sum_{b \in C_i} ||x_a - x_b||_2^2$$
$$= 2 \sum_{a \in C_i} ||x_a - x_i||_2^2$$

Why does K-Means work?

$$\text{Let, } x_i \in R^d = \text{Centroid for} i^{th} \text{cluster}$$
$$= \frac{1}{|C_i|} \sum_{a \in C_i} x_a$$

Then,

$$WCV(C_i) = \frac{1}{|C_i|} \sum_{a \in C_i} \sum_{b \in C_i} ||x_a - x_b||_2^2$$
$$= 2 \sum_{a \in C_i} ||x_a - x_i||_2^2$$

This shows that K-Means gives the **local minima**.

Hierarchal Clustering

Hierarchal Clustering

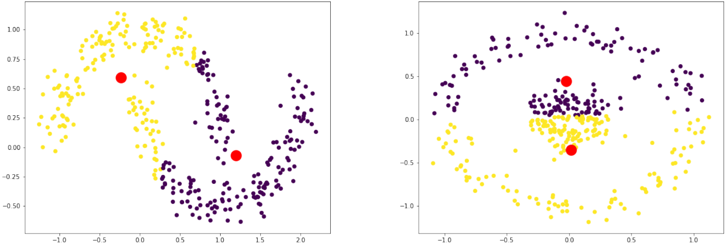
Gives a clustering of all the clusters

Hierarchal Clustering

Gives a clustering of all the clusters
There is no need to specify K at the start

Hierarchal Clustering

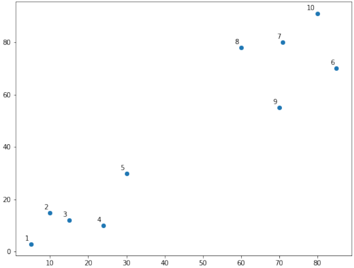
Gives a clustering of all the clusters
There is no need to specify K at the start



Examples where K-Means fails

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster

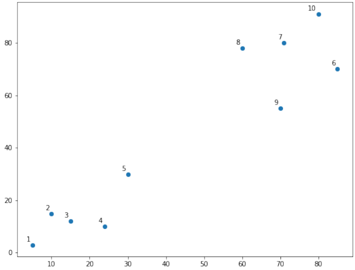


Example Dataset

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster

    1) Identify the 2 closest points

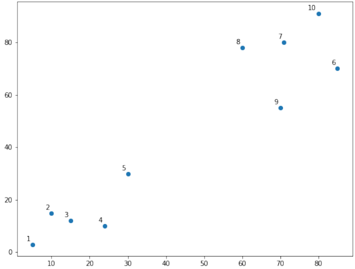


Example Dataset

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster

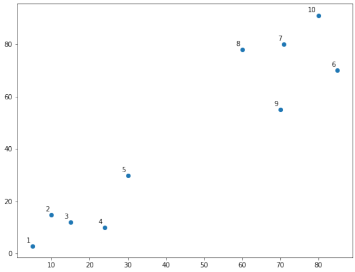
    1) Identify the 2 closest points
    2) Merge them



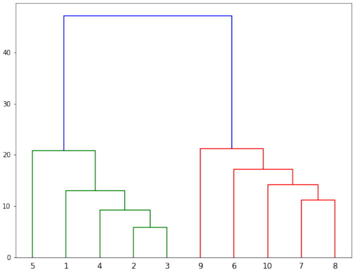
Example Dataset

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster
2. Repeat until all points are in a single cluster
   1) Identify the 2 closest points
   2) Merge them



Example Dataset



Final Clustering

Joining Clusters/Linkages

Complete
Max inter-cluster similarity

Single
Min inter-cluster similarity

Centroid
Dissimilarity between cluster centroids

More Code

Google Colab Link