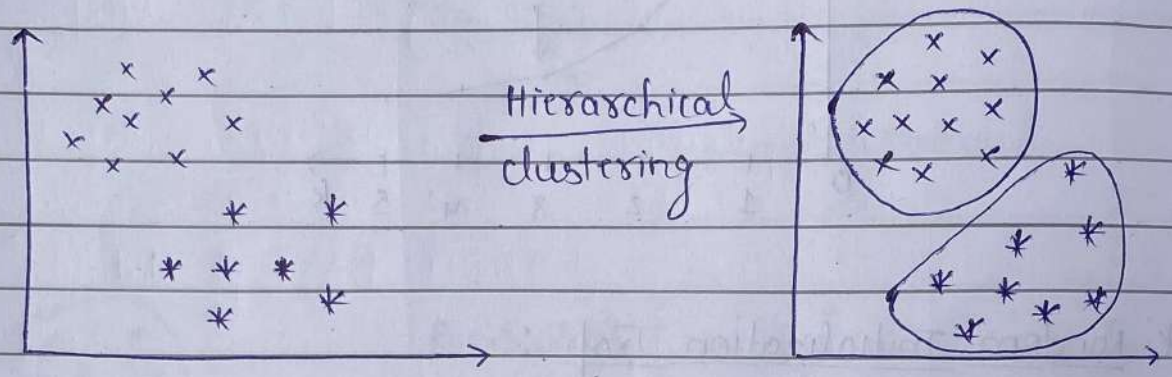
Heirarchical Clustering

→ It is an unsupervised machine learning algorithm which is used to solve clustering problem by grouping the unlabeled dataset into different clusters.



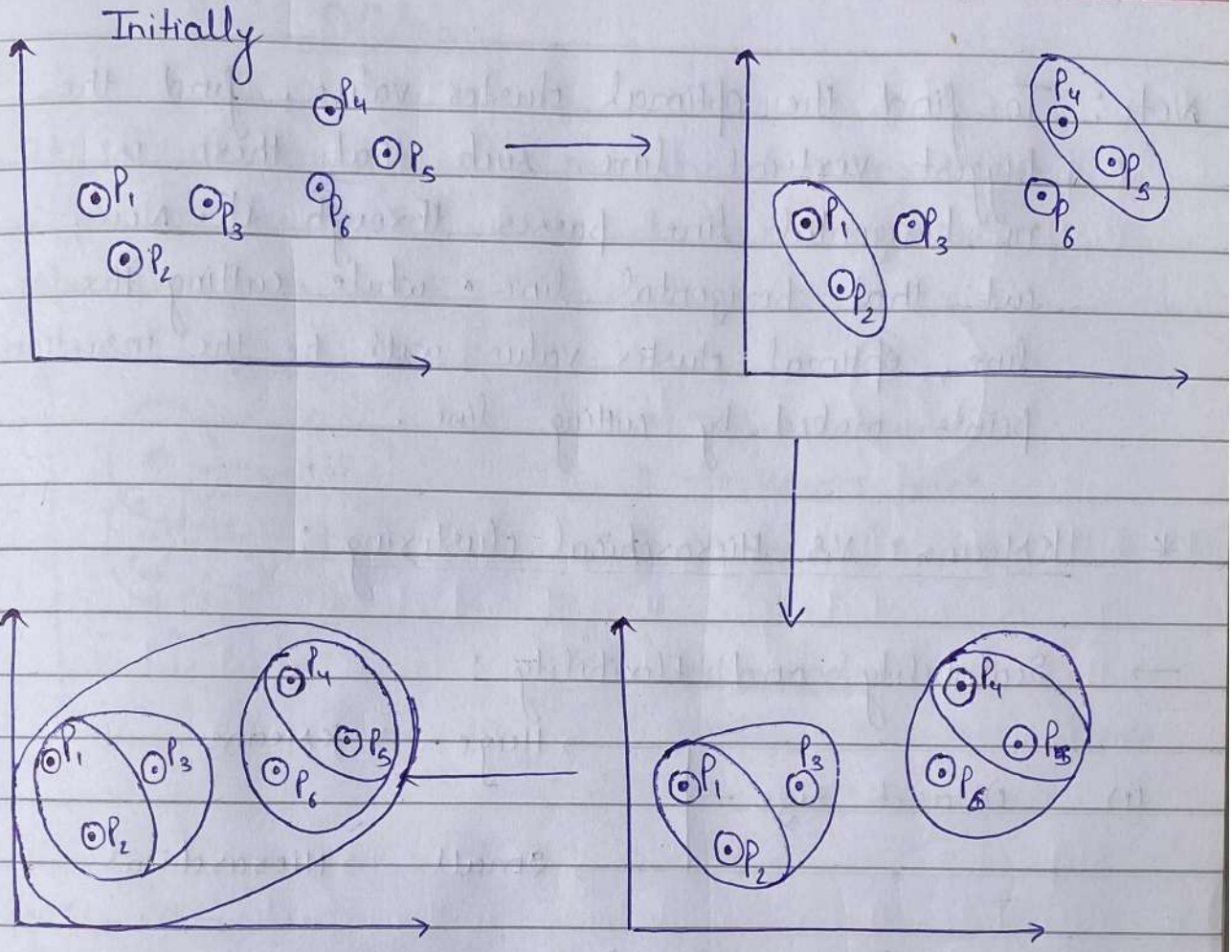
Note: There are two types of hierarchical clustering:
(i) Agglomerative clustering
(ii) Divisive clustering

* Agglomerative Clustering:

Steps:
(i) Initially every datapoint will be considered as a separate cluster.
(ii) Find the nearest cluster, then merge them to make a new cluster.
(iii) Repeat 2nd step until whole dataset becomes part of new/one cluster.

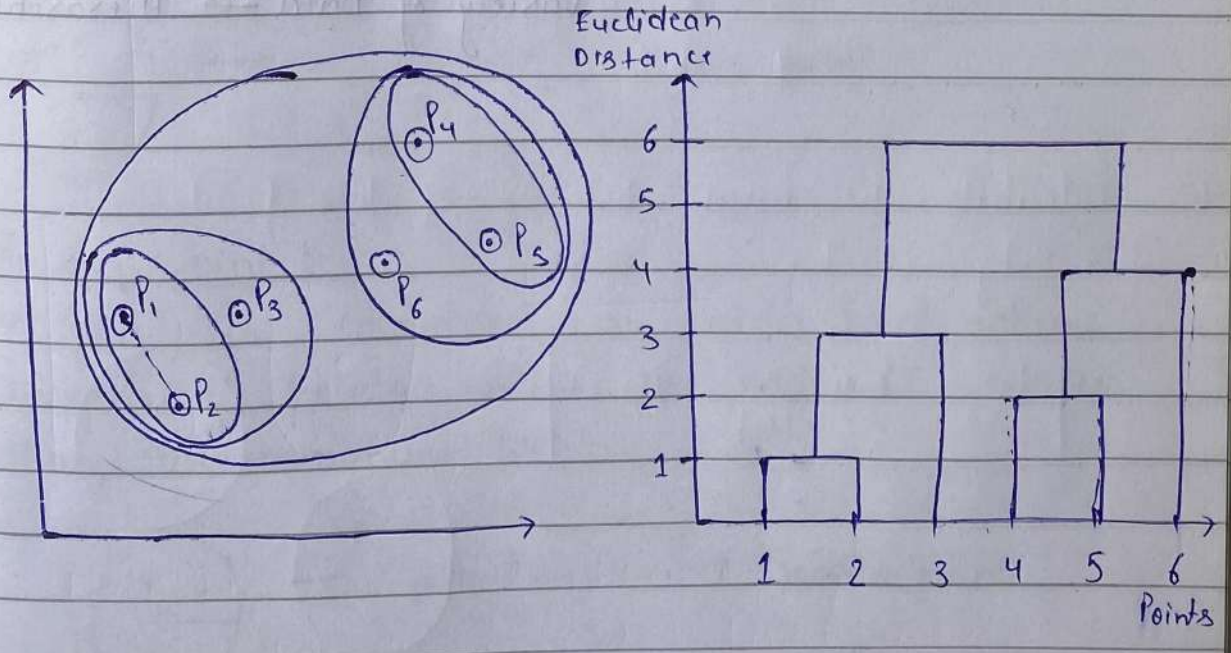
- Distance calculation: → Euclidean distance

$$ED = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}$$



Q How to select optimal number of clusters value?

→ By dendogram visualization.



Note: To find the optimal clusters value, find the biggest vertical line such that there is no horizontal line passes through it. Now, cut that horizontal line. while cutting the line, optimal clusters value will be the insection points maked by cutting line.

* KMeans VS Hierarchical clustering:

→ Scalability and Flexibility:

(i) Dataset size
- Huge → KMeans
- Small → Hierarchical

(ii) Types of Data
- Numerical → KMeans, Hierarchical
- Variety of Data → Hierarchical