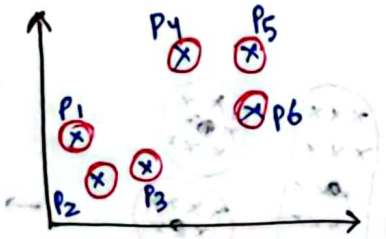
Hierarchical Clustering:

The key difference between hierarchical clustering and KMeans clustering is, Hierarchical clustering don't have any centroids.

There are two techniques in HC clustering:

1) Agglomerative (Combine)
2) Divisive

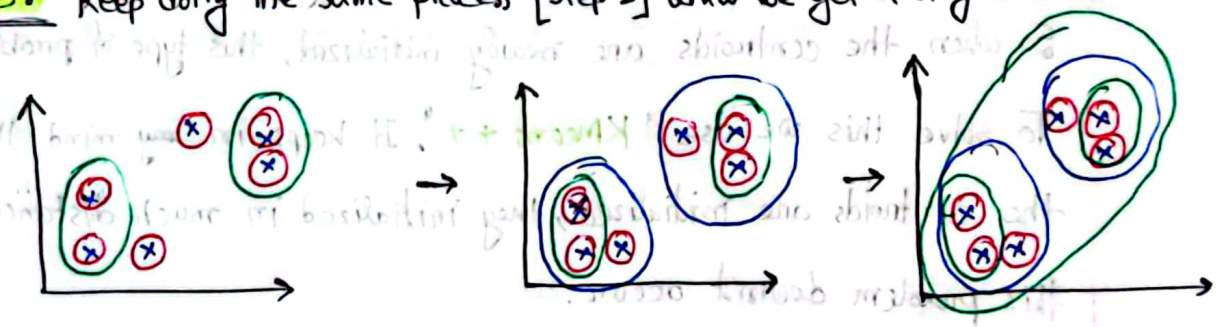
Agglomerative:



Step 01: For each point, we will consider it a seperate cluster.

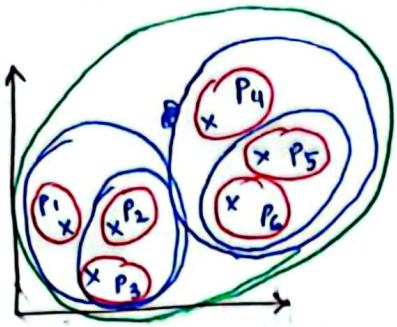
Step 02: Find the nearest point and create a new cluster

Step 03: Keep doing the same process [step 2] until we get a single cluster



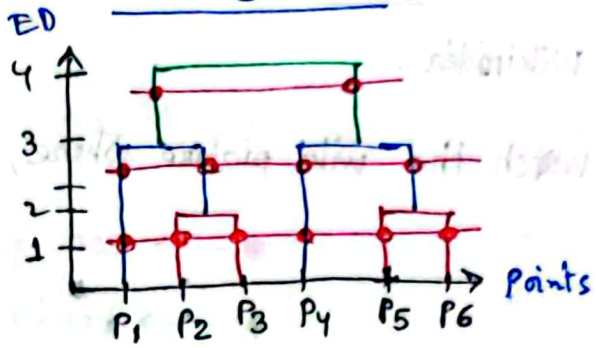
Divisive clustering works in the reverse order.

How many clusters?



Dendogram

ED → Euclidian Distance



To find the number of clusters (K=?) select the longest vertical line in the dendogram such that no horizontal line passes through it.

In our case that is the top most line which has 2 points.

So, K will be = 2. There will be 2 clusters.

Here for theoretical concept purpose, threshold is calculated like this But while implementing, the threshold will get selected automatically.

KMeans Vs Hierarchical Clustering: (Which to use when)

① Dataset
- small → KMeans
- Huge → HC

② Types of Data
- Numerical Data → KMeans or HC
- Variety of Data → HC

DBSCAN Clustering: Diagram is hard to sketch. Source is taken from Wikipedia.

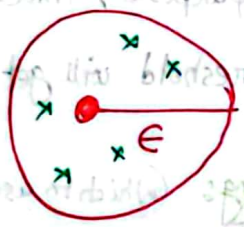
Watch the wiki picture where,

- ● → core point
- ● → border point (yellow)
- ● → Noise/outlier

Hyperparameters →

1) ~~Min~~ minpts (minimum number of points) suppose 4
2) ϵ = radius

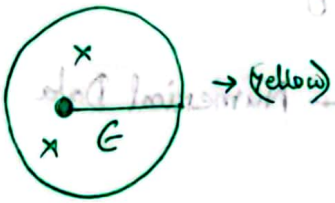
Core point:



The number of points within the ϵ radius should be >= minpts.

We have to consider the red mark as a point also

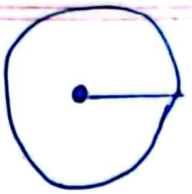
Border point:



The number of points within the ϵ radius will be < minpts

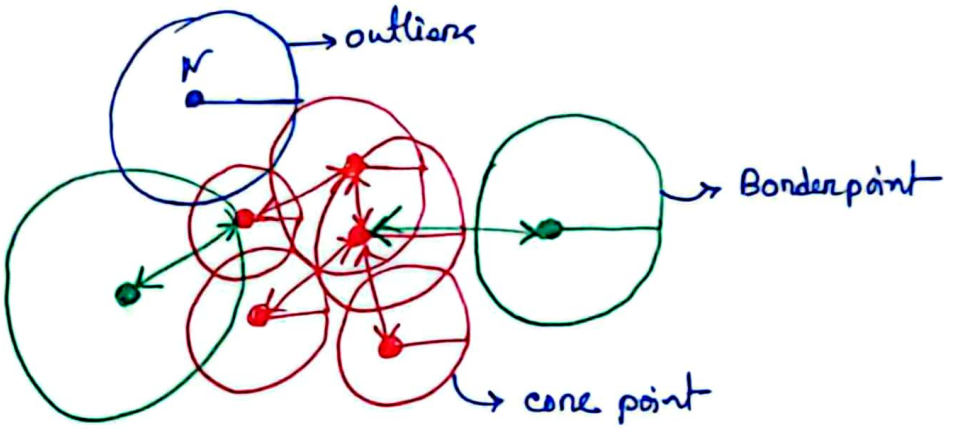
We have to consider the middle mark as a point also.

③ **outliers:**



[DBscan is robust to outliers]

This circle is made for the outliers. If there are no points there, the middle mark considered as outlier



You can check DBSCAN Examples on the internet. It is hard to sketch.

Silhouette Clustering:

When we apply clustering, how to validate that?

In unsupervised ml we check the performance of the models by Silhouette ~~Clustering~~ Scoring.

You can refer to the video of it to learn about the mathematical Intuition. Because pictures are taken from Wikipedia. So complex to note.

value ranges -1 to 1. The more the positive, the better the performance. The more the negative the worse the performance.