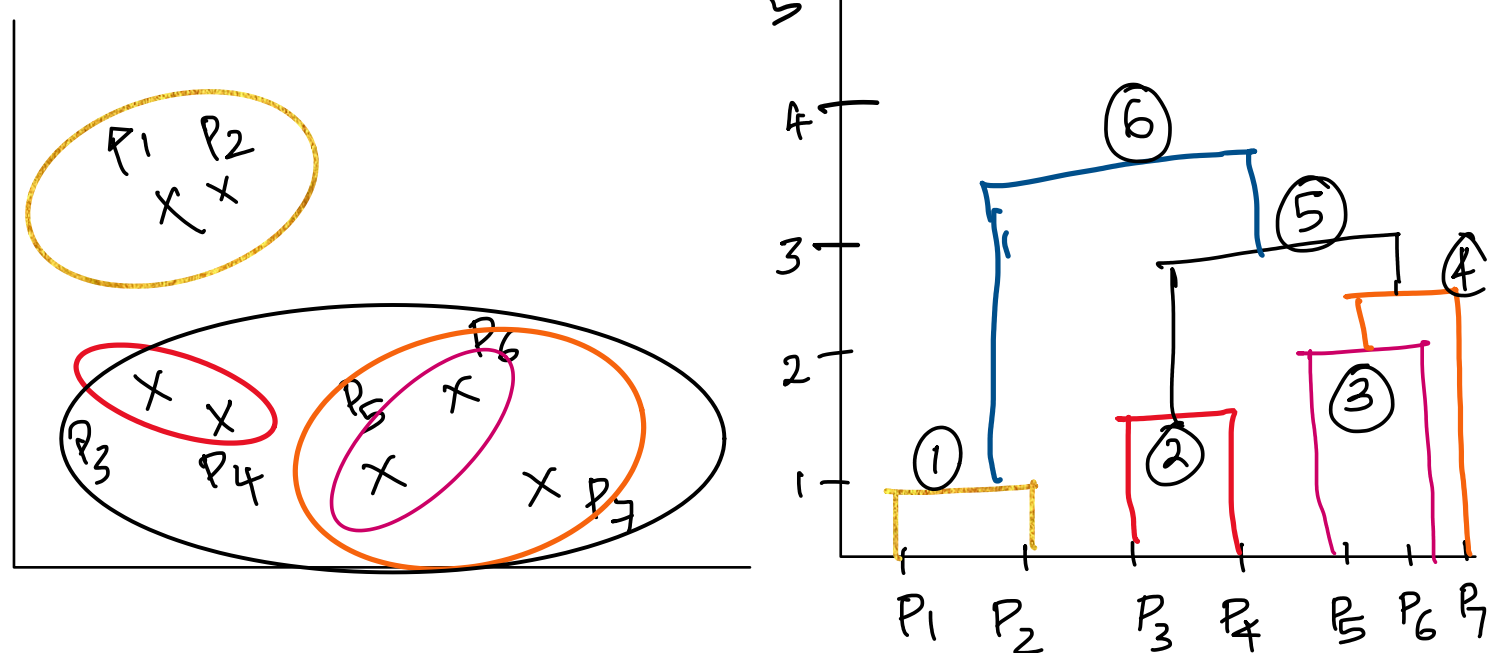


Hierarchical clustering

18 February 2024

15:12



Here, datapoints that are closest to each other group together to form a group. (As shown in the dendrogram)

Once the individual points form a group the groups will also combine together. Finally a completely connected network is formed.

How do we decide on the no. of groups to be formed?

Find the longest vertical line that has no horizontal line passed through it and then based on the no. of intersections so many clusters are considered.

k Means vs Hierarchical clustering:

Which takes more time?

By H.C \rightarrow to create dendrogram.
use for smaller dataset.

How do we validate clustering model?

Silhouette score:

① Choose a cluster C_1 & find the distance between the datapoints ($a(i)$)

② Choose a nearest cluster C_2
Find the dist b/w datapoints in C_1 & datapoints in C_2 .

If $a(i) \gg b(i) \rightarrow$ poor model

If $b(i) \gg a(i) \rightarrow$ better model.

Here we are checking for the distance within the cluster & distance b/w the clusters. Ideally distance within the cluster should be small & distance between clusters should be high.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad \text{will be b/w } -1 \text{ to } +1$$

More the value is towards +1, better the model & vice versa.