Initially, All points as seperate clusters"

(11) Hierarchial Clustering

Bottom-up/Agglomerative approach

→ one main disadvantage of K-means is that it needs us to pre-enter the no. of clusters (K).

→ Hierarchial clustering is an alternative approach which does not need us to give the value of K beforehand and also, it creates a beautiful 'tree-based structure" for visualization.

→ we start by defining any sort of similarity between the datapoints. Generally, we consider Eucledian distance. the points which are closer to each are more similar than the points which are farther away.

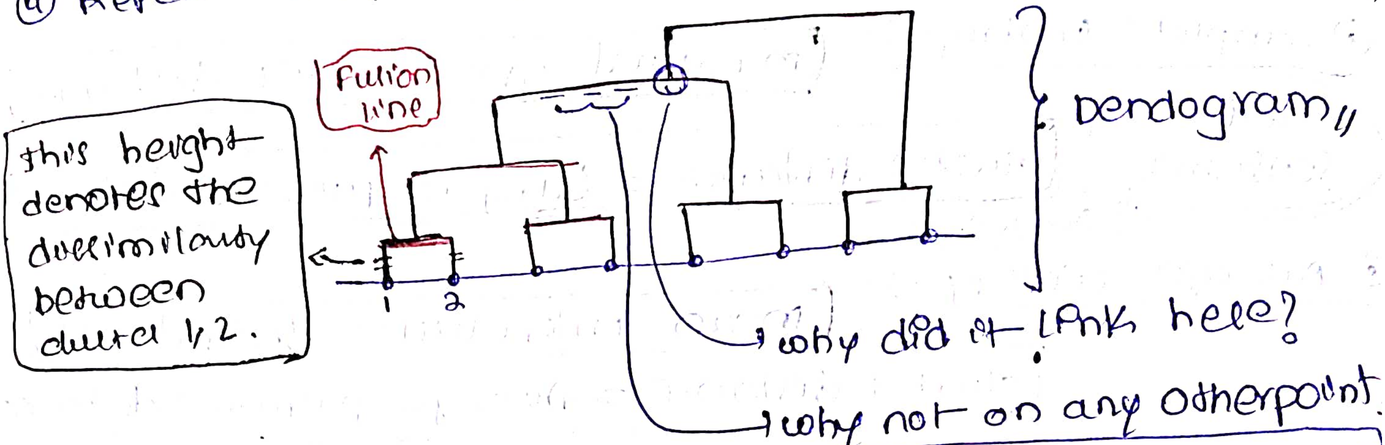
Algorithm

① Begin with n observations and a measure (such as Eucledian distance) of all the $n(n-1)/2$ pairwise dissimilarities (distances). Treat each observation as its own cluster. so, initially we have n clusters.

② compare all the distances and put two closest points/clusters in the same cluster. the dissimilarity between these two clusters indicates the height in the dendogram at which "fusion line" should be placed.

③ compute the new pairwise inter-cluster dissimilarities among the remaining clusters.

④ Repeat steps 2 & 3 till we have only one cluster left,,



this height denotes the dissimilarity between cluster 1,2.

why did it link here?

why not on any otherpoint.

this is where linkage methods comes to place.

Linkage methods

→ Based on pairwise distances, we can now compute a linkage matrix.
the linkage matrix is simply a table listing which pairs or points are merged at what step & what distance.

→ we can cut dendogram to form flat clusters.

⊛ we know, initially HC starts with clusters consisting of individual points.

→ later it compares the clusters with each other and merges the two "closest clusters".

→ since clusters are pair or points, there are many different kinds of linkage methods,,

Notes

⊙ Single Linkage:-

cluster distance = smallest pairwise distance

⊙ complete Linkage:- (maximal intercluster distance)

(less sensitive to outliers)

cluster distance = Largest pairwise distance

⊙ Average Linkage:- (mean intercluster dissimilarity)

cluster distance = Average pairwise distance

⊙ centroid Linkage:- (can result in undesirable inversions)

cluster distance = distance between the centroids of clusters

⊙ ward's Linkage:- (Before & after merging)

cluster distance = minimize the variance in cluster.

Simple Linkage:-

(minimal intercluster dissimilarity.)

→ Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.

→ cluster distance is the smallest distance b/w any point in cluster ① & any point in cluster②.

→ 'High sensitive' to outliers when forming flat clusters.

→ works well for low-noise data with unusual structures