Memo No.

Date / /

Hierarchical Clustering

Disadvantage of K-Means:

- It needs to pre-enter the number of cluster(K).

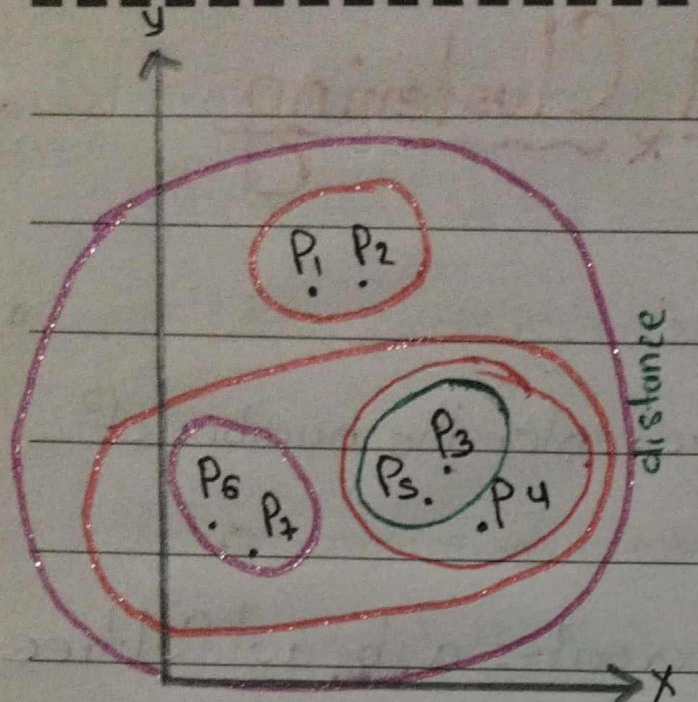
Hierarchical Clustering creates a beautiful tree based structure for visualization.

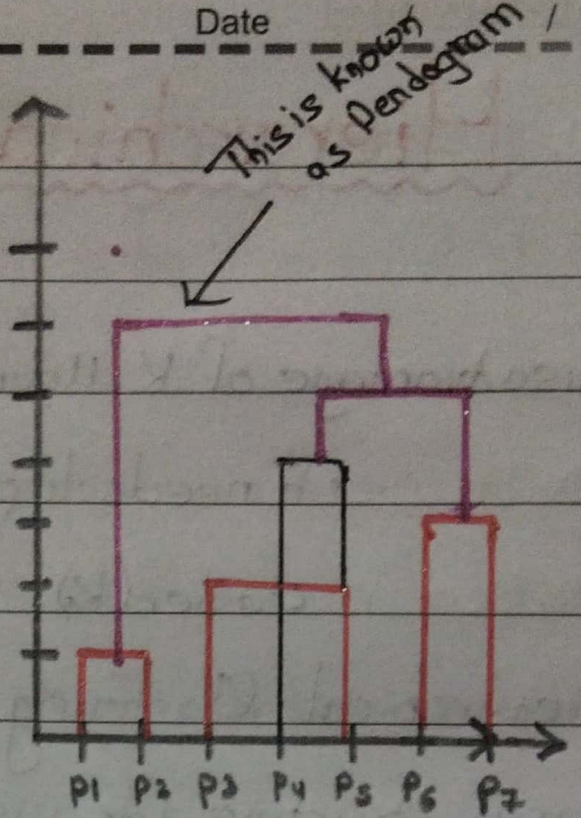
Here, we are going to discuss buttom-up (agglomerative) approach of cluster building. We start by defining any sort of similarity bet^n the datapoints. Generally, we consider the 'Euclidean distances.' The point which are closer to each other are more similar than the point which are farther away. The algorithm starts with considering all points as separate clusters and then grouping pints together to form clusters.

Mo Tu We Th Fr Sa Su

Memo No.

Date /





Algorithm

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

possible pair among n data set

2) Compare all the distances and put the two closest clusters in the same cluster. The dissimilarity between these two clusters

Mo Tu We Th Fr Sa Su

Memo No.

Date / /

indicates the height in the dendogram at which the fusion line should be placed.

3. Compute the new pairwise inter-cluster dissimilarities (or the Euclidean distances) among the remaining clusters.

4. Repeat steps 2 and 3 till we have only one cluster left.

How many group will be formed?

We need to find the longest vertical line that has no horizontal line passed throughIt.

Max time is taken by K-Means or Hierarchal clustering?

→ KMeans Clustering.

Mo | Tu | We | Th | Fr | Sa | Su

Memo No. ______

Date / /

Validating Clustering Method:

The validation of clusters created is a troublesome task. The problem here is

" Clusters are in the eyes of beholder"

A good cluster will have

a) High inter-class similarities

b) low interclass similarities

Memo No. ______

Date / /

DBSCAN

- Density based spatial Clustering of Application with noise.

• It is an unsupervised machine learning algorithm. This algorithm defines clusters as continuous regions of high density.

Some definition first:

Epsilon: This is the distance till which we look for the neighbours point. Radius

Min-points: The minimum number of point specified by the user. cluster [illegible] minimum points

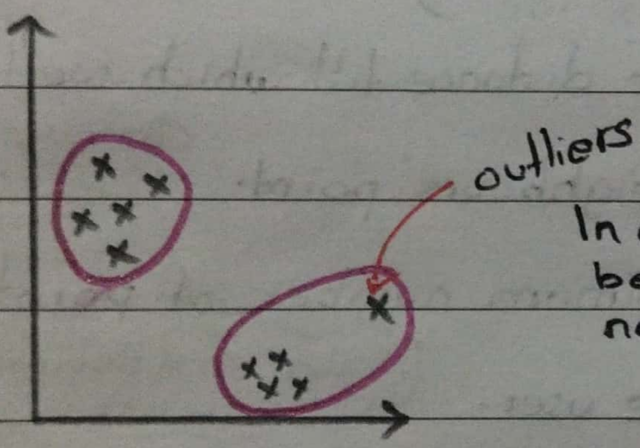
↓ hyperparameter

Core Points: If the number of points inside the epsilon of a point is greater than or equal to the min points then it's called a core point.

Border Points: If the number of points inside the epsilon radius of a point is less than the min points and it lies within the epsilion radius region of a core points, it's called a border point.

Noise: A point which is neither a core nor a border point is a noise point.

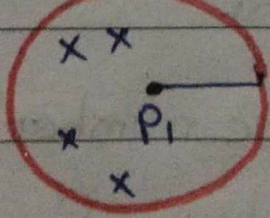
In case of Kmeans



In dB scan, this point will be noise and we will neglect this.

min point = 4

Suppose, P_1 be the point, in Epsilon distance ϵ we draw a circle.



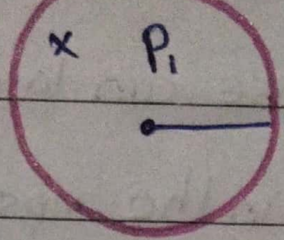
if in that circle, we have min point's number data, then
$P_1 \longrightarrow$ core point.

Mo Tu We Th Fr Sa Su

Memo No. ______

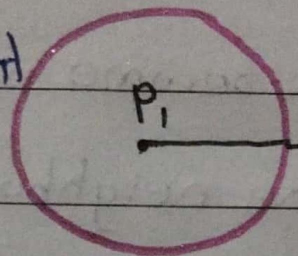
Date / /

If there is less datapoint than min-point then, it is border point



$P_1 \longrightarrow$ border point

If no datapoint exist within epsilion. and it's not a part of core point



$\rightarrow$ it will be outliers and it won't be taking in the model. Neglected.

આપણે cluster બનાવશું
min point હોય તો if
oru core point of epsilion
(જો હોય → border else noise

Algorithm:

1. The algorithm starts with a random point in the dataset which has been visited yet and it's neighboring points are identified based on epsilon value.

2. If the point contains greater than or equal to min points, then cluster formation starts. This point becomes a core point else it's considered as noise. The thing to

Memo No.
Date / /

note here is that point initially classified as noise can later become a border point if it's in the epsilion of a core point.

3. If the point is a core point, then all it's neighbours become a part of cluster. If the points in the neighbours turn out to be core points then their neighbours are also part of the cluster.

4. Repeat the steps above until all points are classified into different clusters or noise