UNIT-4

CLUSTER ANALYSIS

- Cluster Analysis.
- Types of Clustering
- Types of clusters.
- Algorithm for cluster analysis.
  (i) K-means algorithm.
  (ii) Agglomerative hierarchial clustering.
  (iii) DBSCAN
  - Agglomerative
  - divisive

Cluster Analysis

Finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups.

Types of ~~Cluster Analysis~~ Clustering

1. Heirarchial versus Partitional.
2. Exclusive versus Overlapping (Non-Exclusive)
3. Complete versus Partial

Types of clusters.

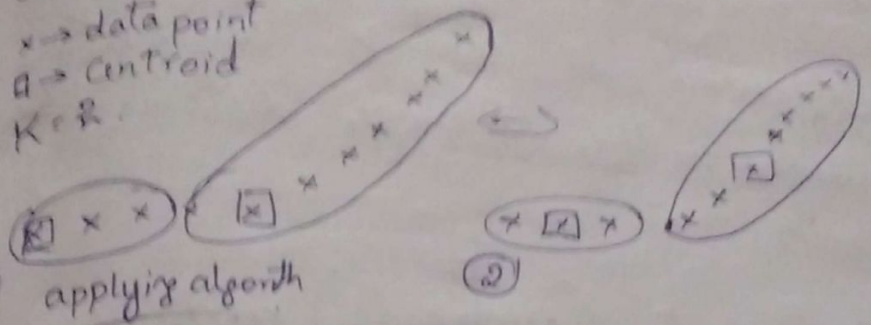
1. Well defined cluster
2. prototype based (centre based)
3. Graph based cluster
4. Density based
5. shared cluster (conceptual clusters)

K-means algorithm

K - no of clusters (family)

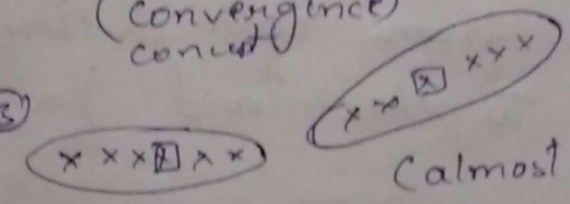
original data

x → data point
a → centroid
K = 2



① applying algorithm ②

1. Select K points as initial centroid $\{c_1, c_2, \ldots c_k\}$
2. Repeat
3. From K clusters by assigning each point to the closest centroid
4. Recompute the centroid for each cluster
5. Until centroids do not change (convergence) (almost all centroids are stabilized)



③

(Selected random points (centroids) the distance b/w them should be maximum)

We are terminating the algorithm because all centroids stabilised and points don't move b/w the clusters.

- for each point place it in the cluster whose current centroid is nearest.
- After all points are assigned, update the location of centroids of K clusters.
- Reassign all points to closest centroid
- Repeat 2 and 3 until convergence.

(Partitional clustering)

What is right value of k? or How to pick k-value

Try different k-values. by looking at the change in the average distance to centroid.

Initial k-points (centroids) picking.

approach 1:-

Sampling

- Cluster a sample of data using hierarchical clustering to obtain k clusters.
- Pick a point from each cluster (point nearest to the centroid)

approach 2:

- Pick "dispersed" set of points.

pick first ~~pick~~ point at random.

- Pick the next point to be the one whose minimum distance from the selected points is as large as possible.
- Repeat until complete k-points

final clustering depends on inilial k-points (centroid -> how you pick)

Complexity.

1.
2.
3.
4.
5.

Complexity:
In each round we have to examine each input point exactly once to find the closest centroid.
Each round is $O(KN)$
Until convergence (centroid is stabilised)
The no of rounds to reach the convergence can be very large

dules

Agglomerative Hierarchical clustering

A set of nested cluster organised as a hierarchical tree is called as Hierarchical clustering.

Two types of algorithm.
1. Agglomerative (bottom-up)
2. Devisive (Top down)

Agglomerative
- Initially each point in a cluster
- Repeatedly combine the two nearest clusters to one

Devisive
- Starts with one cluster and recursively split it.

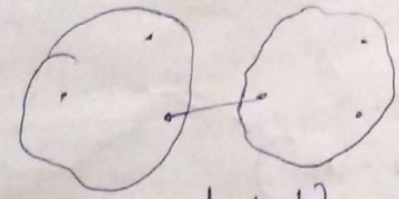
Algorithm Agglomerative
1. Compute the proximity matrix, if necessary.
2. Repeat
3. Manage the closest two clusters.
4. Update the proximity matrix to reflect the proximity b/w the new cluster and the original clusters.
5. Until only one cluster remains.

Dendrogram

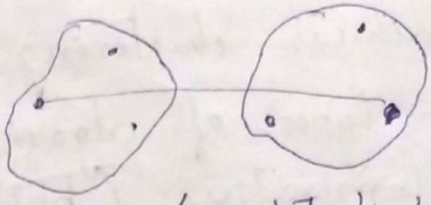
The hierachical clustering is often displayed graphically using a tree like diagram called Dendrogram, which displays both the cluster - sub cluster relationships and the order in which clusters are merged (agglomerative view) / divisive view)

11/4/18

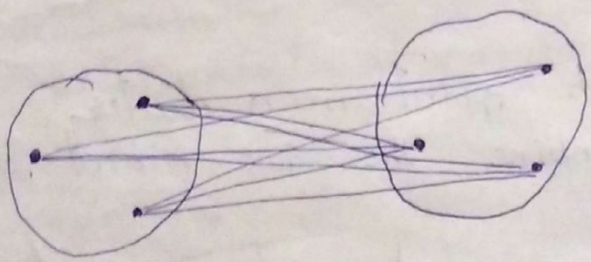
Define Proximity



Min (single list)

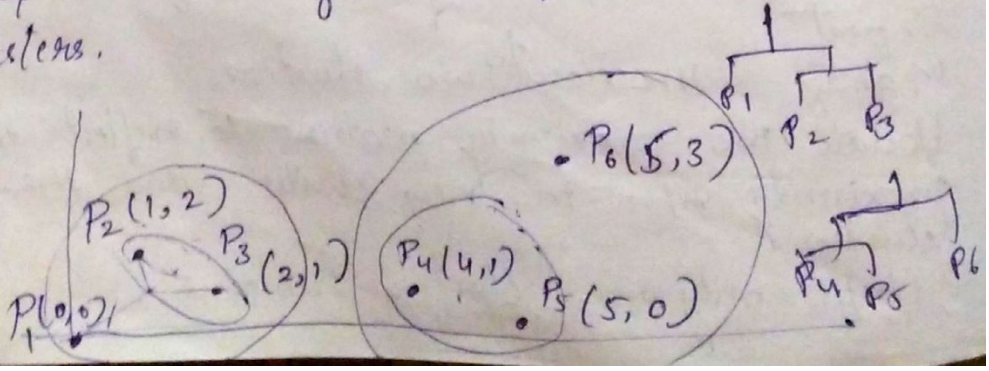


MAX (complete list)



Group average.

Graph band definitions for proximity clusters.



Agglomeration hierarchical clustering

1. How do you represent a cluster of more than one point?
2. How do you determine the measures of the clusters?
3. When to stop combining clusters?

Euclidean Space

Represent a each cluster by its centroid
= average of its point.

Fermi Condition.

Approach 1:- Pick a k number and stop when we reach k clusters.
(Naturally falls into k clusters)

Approach 2:- Stop when the next merge would create a cluster with low 'cohesion' centroid (clustroid)

Cohesion

Diameter - Maximum distance b/w the points in a cluster

Radius - Maximum distance of point from cluster

Density - No of points per unit volume

P6

Merits of DBSCAN

→ Unlike K-mean, DBSCAN doesn't require user to specify the no of clusters to be generated.

→ DBSCAN can find any shape of cluster need not be an circular in shape

→ It produces efficient and accurate results

→ We can easily diminate the noise points

DBSCAN algorithm is able to find arbitarary size and arbitrarily shaped clusters.

DRAWBACKS :-

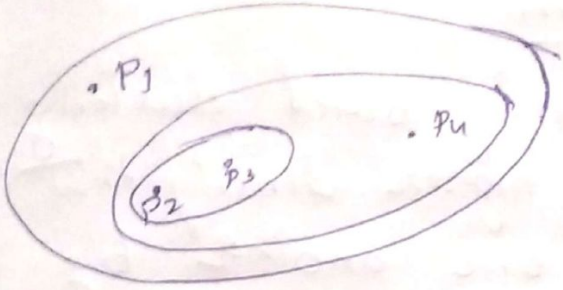
→ DBSCAN algorithm fails in case of variable density clusters.

→ Selecting area (EPS, ε) and minimum points (parameters for DBSCAN) are tricky.

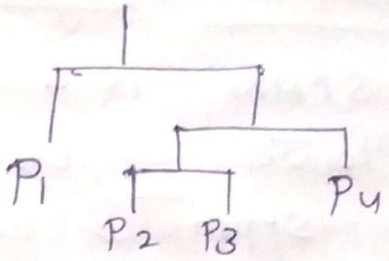
→ Doesnot work well in high dimensionality of data.

Strength & Weakness of K-means & Agglomerative

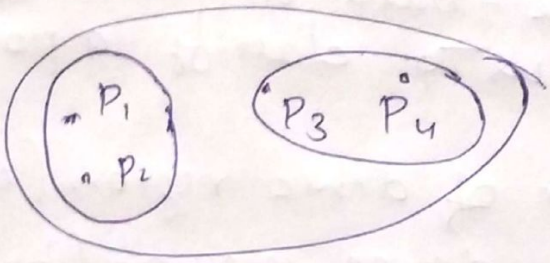
Traditional

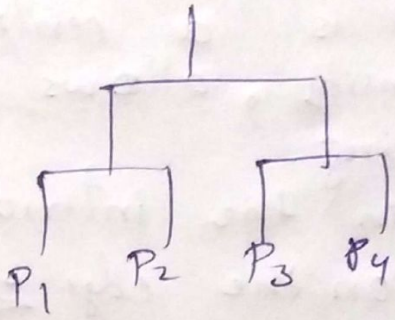


Dendrogram



Non tradition





→ Min — defines cluster proximity b/w the 2 progressive points or wrt (graph terms, the shortest edge b/w 2 nodes in different subset of nodes).

→ Max:- takes the proximity b/w the farthest 2 points in different cluster to be the cluster proximity. longest edge b/w 2 nodes.

→ Group avg - is the technique used to define the cluster proximity to be the average pair wise proximities of all the pairs of points from different clusters

DBSCAN Algorithm

DBSCAN is a density based clustering
It locates region of higher density that
are separated from one another by
regions of low density.

→ It is a centre-base approach to
density, allows us to classify a point
being.

1. In the interior of dense region (core)
2. on the edge of dense region (border)
3. in a sparsely occupied region (Noise)

Density = No of points per unit volume
Density based clustering that produces
a partitional clustering.

. Min points Essentially given you some
kind of threshold or how many you
consider as being dense.

↳ EPS (area which you with perform the
count

Algorithm:-

1. label all the points as core, border and noise points.
2. Eliminate the noise points.
3. Put an edge b/w all core points that are within Eps of each other.
4. Make each group of connected group core points into separate clusters.
5. Assign each border point to any no of clustering it associated core point.

DBSCAN (Density Based Spatial Cluste of Application with noise)

Graph based clustering