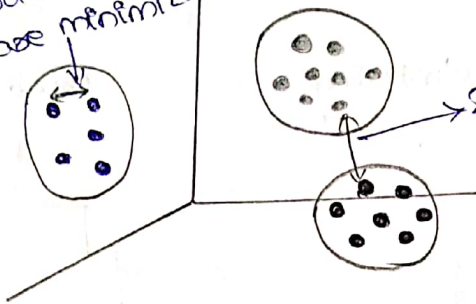


20/2/19

What is 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.

Intra cluster distances are minimized.



Inter-cluster distances are maximized

Applications:-

* Group related documents for browsing, group genes and proteins that have similar functionality (Biology), group stocks with similar price fluctuations (Business).

What is not cluster Analysis

- Supervised classification: Have class label information
- Simple Segmentation: Dividing students into different registration groups alphabetically, by last name.
- Results of a query: Grouping are a result of external specification
- Graph partitioning: some mutual relevance and synergy, but areas are not identical.

** Types of clustering:-

* A clustering is a set of clusters

* partitional clustering

- A division data objects into non-overlapping subsets (clusters)

such that each data objects is in exactly one subset.

eg:-



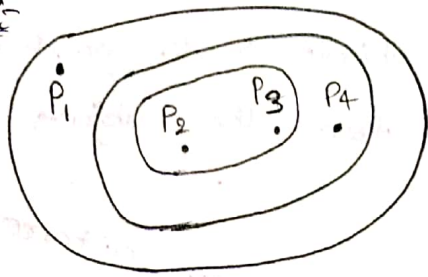
original points

A partitional clustering

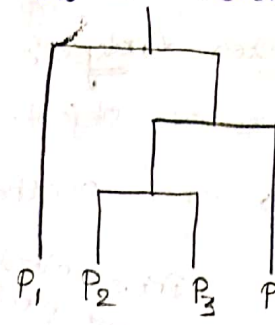
Hierarchical clustering:-

* A set of nested clusters organized as a hierarchical tree.

eg:-

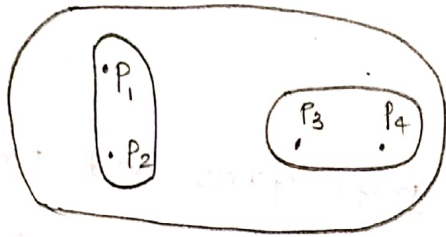


Traditional Hierarchical clustering

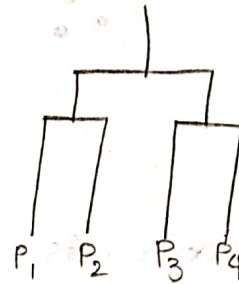


Traditional Dendrogram

collection of
tree structure



Non-traditional Hierarchical clustering



Non-traditional Dendrogram.

Differences between sets of clusters:-

→ Exclusive versus non-exclusive

* In non-exclusive clusterings, points may belong to multiple clusters

* can represent multiple classes (or) 'border' points

→ Fuzzy versus Non-fuzzy

* In fuzzy clustering a point belongs to every cluster with some weight between 0 and 1

* weights must sum to 1.

→ partial versus complete.

* In some cases, we only want to cluster some of the data

→ Heterogeneous versus homogeneous.

* cluster of widely different sizes, shapes and densities.

* types of clusters:-

→ Well-separated clusters: A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 - well separated clusters.

→ Center-based clusters:

* A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster.

* The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster.



21/2/19

4 - center based clusters

Contiguous clusters: (Nearest neighbor or transitive)

A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

eg:-



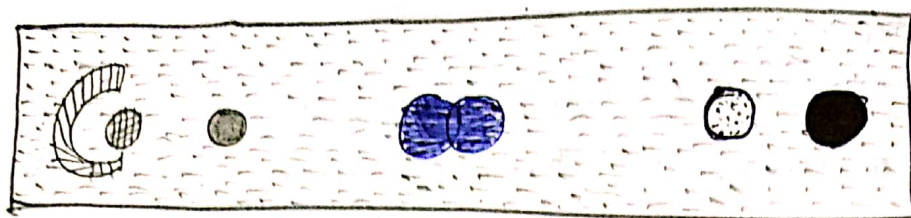
3 - contiguous clusters.

Density-Based clusters:

* A cluster is a density region of points, which is separated by low-density regions, from other regions of high density.

* Used when the clusters are irregular, and when noise and outliers are present.

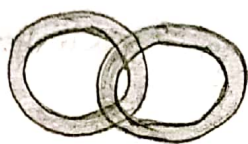
eg:-



6-density based clusters.

Property or Conceptual clusters:-

finds clusters that share some common property or represent a particular concept.



2 overlapping circles

Described by an objective function:-

- * Find clusters that minimize (or) maximize an objective function
- * Enumerate all possible ways of dividing the points into clusters.
- * Evaluate the 'goodness' of each set of clusters by using the given objective function. (NP Hard)
- * Hierarchical clustering algorithms have local objectives.
- * Partitional algorithms have global objectives.
- * Global objective function approach is to fit the data to a model.

Characteristics of the input data:-

- * Type of proximity or density measure
 - This is a desired measure, but is central to clustering.
- * Sparseness
 - Dictates type of similarity
 - Adds to efficiency
- * Attribute type
 - Dictates type of similarity

- * Type of Data
 - Dictates type of Similarity
 - other characteristics, e.g. autocorrelation

- * Dimensionality
 - Noise and outliers
 - Type of Distribution.

Clustering Algorithms:-

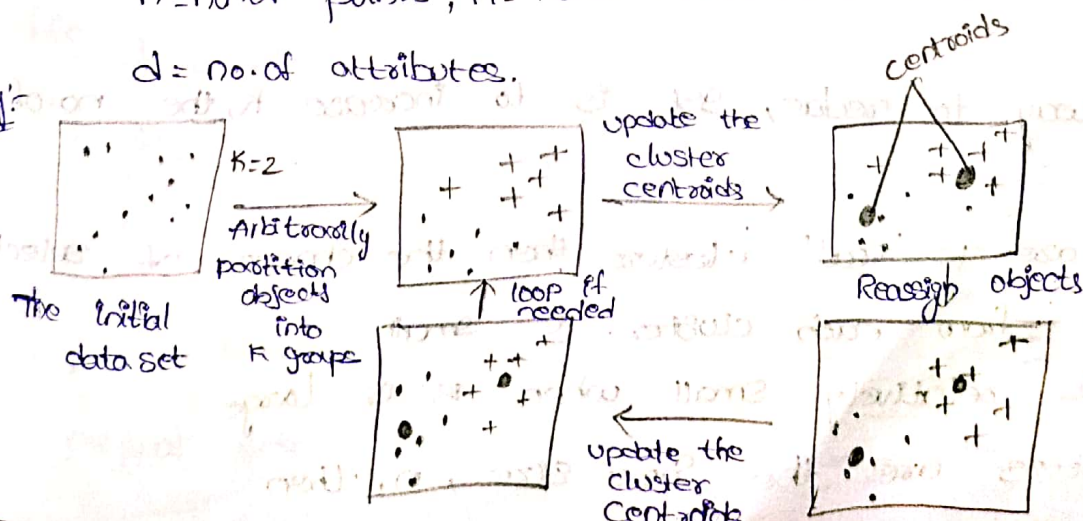
K-means algorithm:-

- * positional clustering approach
- * Each cluster is associated with a centroid (Center point)
- * Each point is assigned to the cluster with the closest centroid
- * Number of clusters, k , must be specified.
- * The basic algorithm is very simple.

Algorithm:-

1. Select k points as the initial centroids
 2. repeat
 3. Form k clusters by assigning all points to the closest centroid.
 4. Recompute the centroid of each cluster
 5. Until the centroids don't change.
- * The centroid is (typically) the mean of the points in the cluster.
 - * 'closeness' is measured by Euclidean distance, cosine similarity, correlation, etc..
 - * The complexity is $O(n * k * I * d)$
 - n = no. of points, k = no. of clusters, I = no. of iterations,
 - d = no. of attributes.

eg:-



* partition objects into k nonempty subsets

* Repeat

→ compute centroid for each partition

→ Assign each object to the cluster of its nearest centroid

* Until no change.

Strengths:- Efficient $O(kn)$ where n is #obj, k is #cluster &

t is #iteration, normally $k, t \ll n$

comparing: PAM: $O(k(n-k)^2)$ CLARA: $O(k^3 + k(n-k))$

comment:- Often terminate at a local optimal

Weakness:- Applicable only to obj in a continuous n -dimensional space

* Using k -modes method for categorical data

* In comparison, k -medoids can be applied to wide range of data

Evaluating k -means clusters:-

* Most common measure is sum of Squared Errors (SSE)

* For each point, the error is the distance to the nearest cluster

* To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} \text{dist}^2(m_i, x)$$

where,

x is a data point in cluster C_i

m_i is the representative point for cluster C_i

can show that m_i corresponds to the center (mean) of cluster.

* Given two clusters, we can choose the one with the smallest error -

* One easy way to reduce, SSE is to increase k , the no. of clusters.

* If there are k 'real' clusters then the chance of selecting one centroid from each cluster is small

- chance is relatively small when k is large

- If clusters are the same size, n , then

$P = \frac{\text{no. of ways to select one centroid from each cluster}}{\text{number of ways to select } k \text{ centroids}}$

$$\frac{k! n^k}{(kn)^k} = \frac{k!}{k^k}$$

* For example, if $k=10$, then probability $= 10! / 10^{10} = 0.00036$.

Bisecting K-means algorithm:-

* It is a variant of K-means that can produce a partitioned (or) hierarchical clustering.

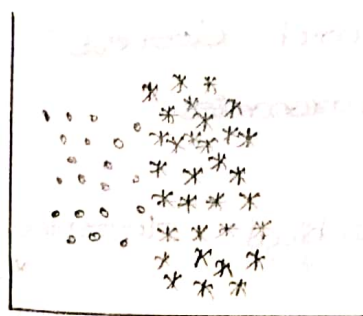
Algorithm:-

- Initialize the list of clusters to containing all points
- repeat
- select a cluster from the list of clusters
- for $i=1$ to no. of iterations do
- Bisect the selected cluster using basic K-means
- end for
- Add the two cluster from the bisection with the lowest SSE to the list of clusters.
- until the list of clusters contain k clusters.

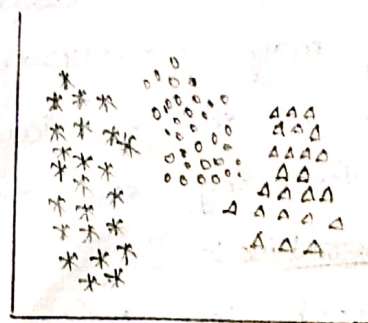
Limitations of K-means:-

- K-means has problem when the clusters are different.
- Sizes
- Densities
- Non-globular shapes

Differing sizes:-

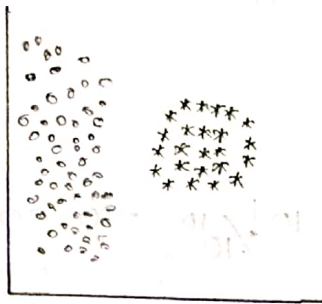


original points

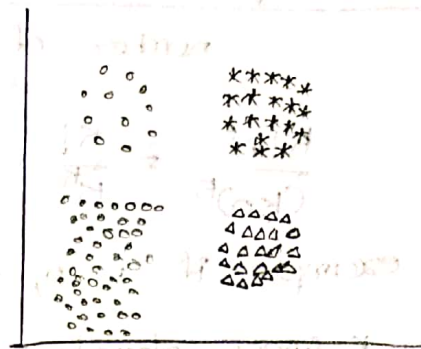


K-mean points (3 clusters)

Differing Densities:-

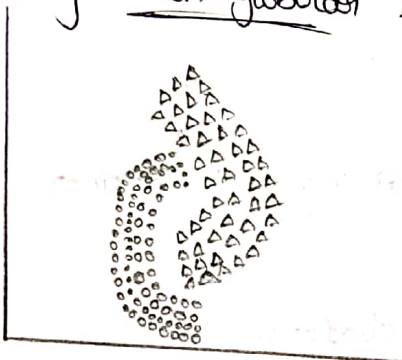


original points

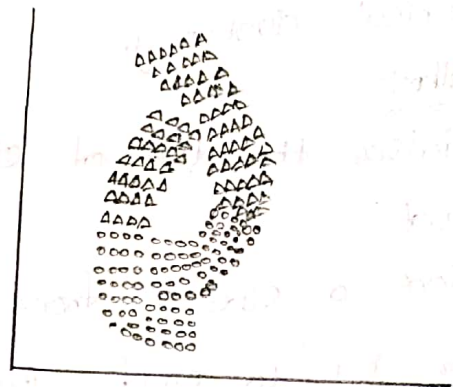


K-means (3 clusters)

Differing Non-globular shapes:-



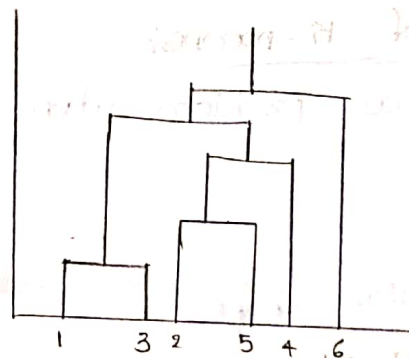
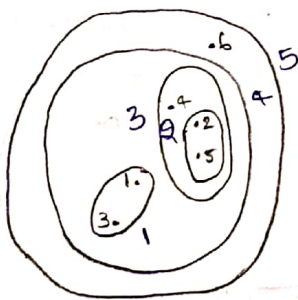
original points



K-Mean points

Hierarchical clustering Algorithms:-

- * produces a set of nested clustering organized as a hierarchical tree.
- * Can be visualized as a dendrogram
- * A tree like diagram that records the sequences of merges or splits.



- * Do not have to assume any particular no. of clusters
- * They may correspond to meaningful taxonomies
- * Example in biological sciences.
- * The following are the types of hierarchical clustering

1) Agglomerative:-

- * Start with the points as individual clusters
- * At each step, merge the closest pair of clusters until only one

clusters (or 13 clusters) left.

2) Divisive:-

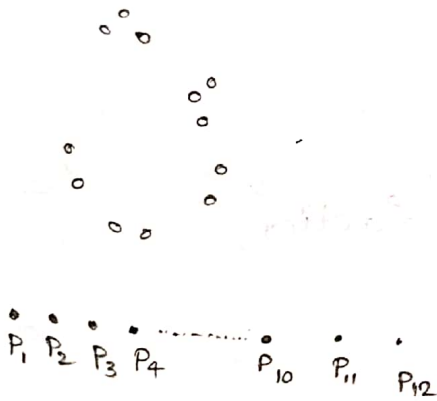
- * Start with one, all inclusive cluster.
- * At each step, split a cluster until each cluster contains a point (or there are 13 clusters)

Agglomerative clustering Algorithm:-

- 1) Compute the proximity matrix
- 2) let each data point be a cluster
- 3) Repeat
- 4) Merge the two closest clusters.
- 5) Update the proximity matrix.
- 6) Until only a single cluster remains.

Starting situation:-

- * start with clusters of individual points and a proximity matrix

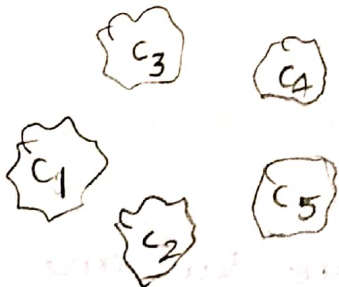


	P ₁	P ₂	P ₃	P ₄	P ₅	...
P ₁						
P ₂						
P ₃						
P ₄						
P ₅						

proximity Matrix

Intermediate Situation:-

- * After some merging steps, we have some clusters.



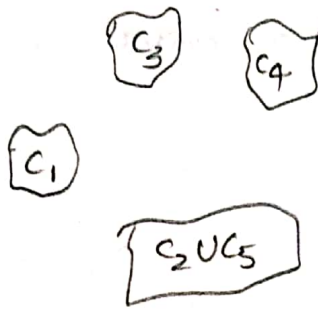
	C ₁	C ₂	C ₃	C ₄	C ₅
C ₁					
C ₂					
C ₃					
C ₄					
C ₅					

proximity Matrix.



5/3/17 After Merging:-

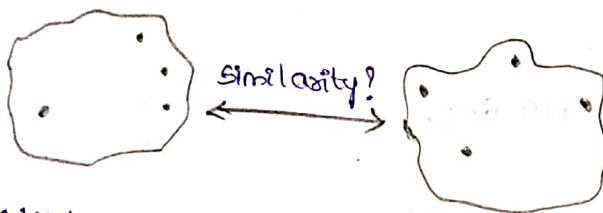
The question is "how do we update the Proximity matrix?"



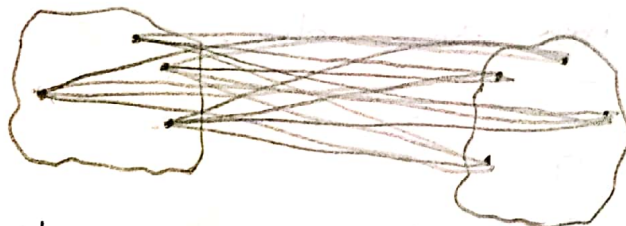
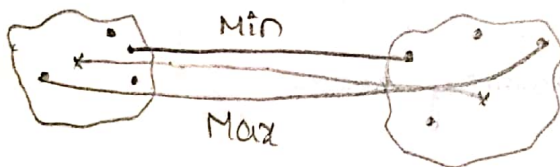
	C_1	$C_2 \cup C_5$	C_3	C_4
C_1		?		
$C_2 \cup C_5$	?	?	?	?
C_3		?		
C_4		?		

Proximity Matrix

How to define Inter-cluster similarity:-



- * MIN
 - * MAX
 - * Group Average
 - * Distance Between Centroids
 - * Other methods driven by an objective function
- Ward's Method uses squared error.



cluster similarity for MIN:-

- * Similarity of two clusters is based on the two most similar (closest) points in the different clusters.
- * Determined by one pair of points, i.e.; by one link in the proximity graph.

cluster similarity for MAX:

* Similarity of two clusters is based on the two least similar (most distant) points in the different clusters.

* Determined by all pairs of points in the two clusters.

cluster similarity for Group Average:

* Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

$$\rightarrow \text{proximity}(\text{cluster}_i, \text{cluster}_j) = \frac{\sum_{\substack{P_i \in \text{cluster}_i \\ P_j \in \text{cluster}_j}} \text{proximity}(P_i, P_j)}{|\text{cluster}_i| * |\text{cluster}_j|}$$

* Need to use average connectivity for scalability since total

proximity
6/3/19
Strengths & Limitations:-

→ Less susceptible to noise and outliers

→ Biased towards globular clusters.

Ward's Method:-

* Similarity of two clusters is based on the increase in squared error when two clusters are merged.

* Similar to group average if distance b/w points is distance squared.

$O(N^2)$ is space, since it uses the proximity matrix.

N is the no. of points.

$O(N^3)$ is time complexity

Complexity can be reduced to $O(N^2 \log(N))$ time for some approximations.

Divisive Hierarchical clustering:-

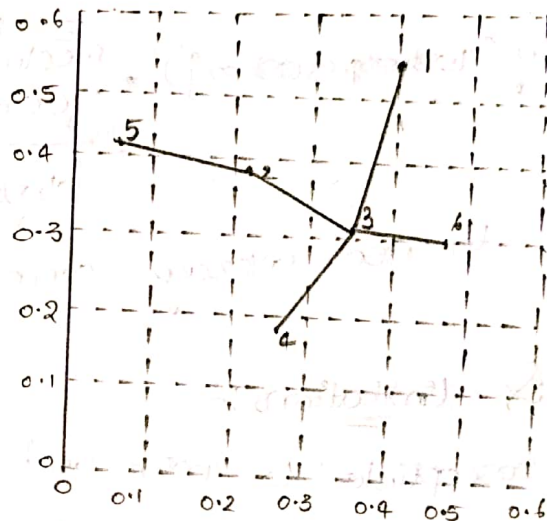
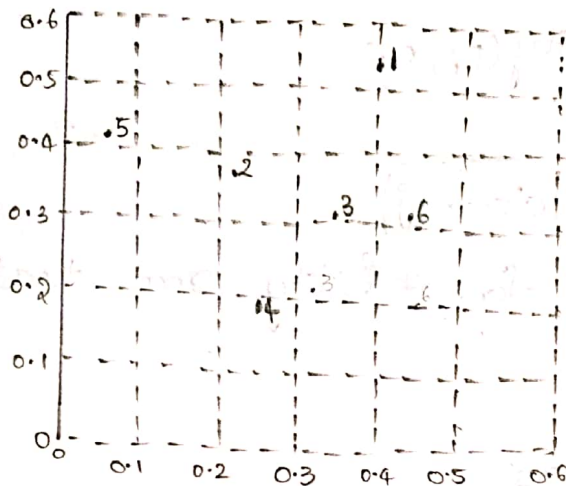
* Build MST (Minimum Spanning Tree)

Types of
K-Means.

Start with a tree that consist of any point.

In successive steps, look for the closest pair of points (p, q) such that one point (p) is in the current tree but the other (q) is not.

Add (q) to the tree and put an edge between p and q .



Algorithm:-

1. Compute a Minimum Spanning Tree for the proximity graph
2. repeat
3. Create a new cluster by breaking the link corresponding to the largest distance (Smallest similarity)
4. until only single cluster remain.

**

DBSCAN Algorithm:-

* DBSCAN is a density-based algorithm.

* Density = no. of points within a specified radius (Eps)

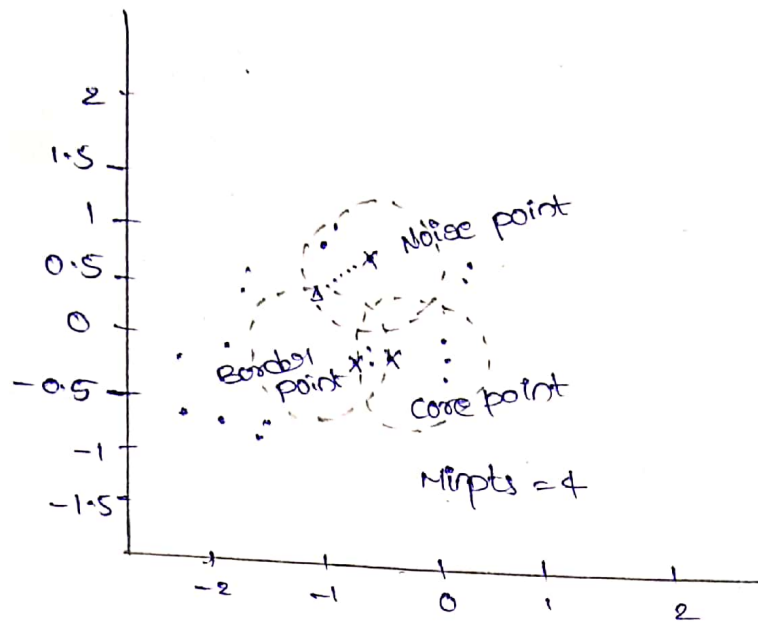
* A point is a core point if it has more than a specified number of points (Minimum points) within Eps

* These are points that are at the interior of a cluster

* A border point has fewer than Minimum points within Eps, but is in the neighborhood of a core point.

* A noise point is any point that is not a core point or a

border point.



Algorithm:-

current_cluster_label $\leftarrow$ 1

for all core points do

if the core point has no cluster label then

current_cluster_label $\leftarrow$ current_cluster_label + 1

Label the current core point with cluster label current_cluster_label

end if

for all points in the Eps-neighborhood, except i^{th} point

if the point does not have a cluster label ^{itself} do

. Label the point with cluster label current_cluster_label

end if

end for

end for