No.

Date

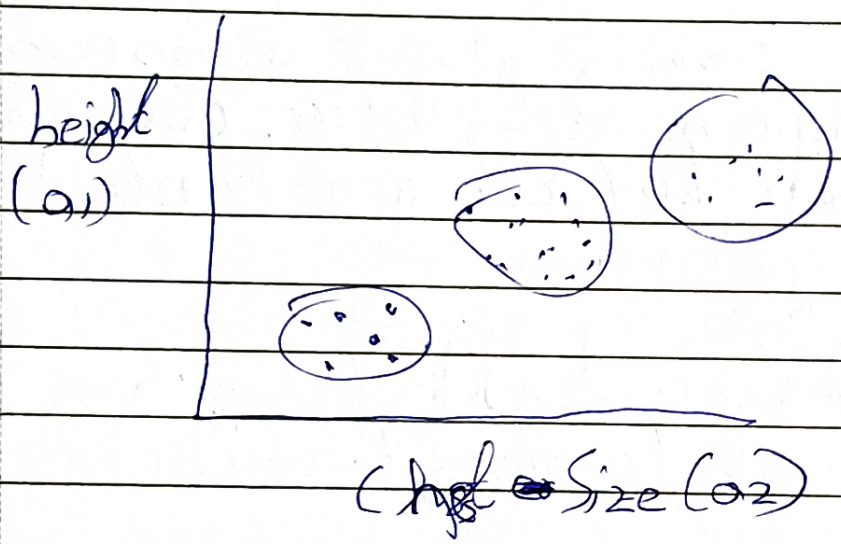
~~Submit Computer Lab Assignment list~~

K-means clustering

↳ greedy algorithm

Input : a dataset

$(x_1, x_2 \cdots x_n) \rightarrow$ where x_i is a vector



Clustering is a process to partition datapoints into multiple groups where each grp contains similar datapoints,

The objective of clustering is to find groups that will maximise inter-cluster distance and minimize intra-cluster distance,

No.

Date

K = # of cluster

Output : K clusters

Step1: Randomly select K data points from X and assign to K diff clusters & consider that those points are the centroids of the cluster.

Step2: For the remaining points assign each data point to the nearest clusters. (where the Euclidean distance between p and centroid of cluster is considered.

Step3: Update ~~the~~ each centroid by taking avg of data points belonging to each cluster

Step4: Repeat the process of assigning data points to cluster until convergence (when centroids remain unchanged.

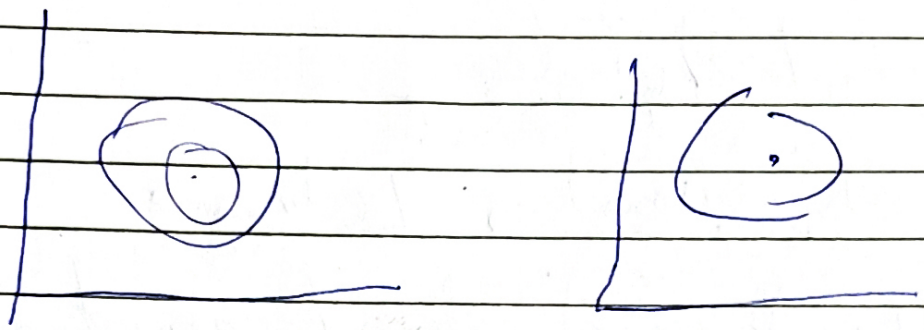
Local greedy choice based on 3 dist from centroids choosing a centroid

Suboptimal but to be acceptable too many iterations,

No.

Date

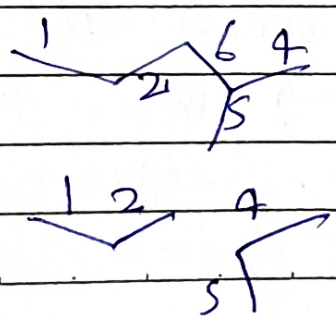
It does not work for arbitary shapes



Same centroid.

MST-based clustering

① Represent data point as weighted graph where each data point is a node of a graph and an edge weight is the Euclidean distance between two data points representing any two nodes
② Construct the MST using prim's algorithm
③ Delete the max weight edge from MST
④ Obtain the subgraphs
⑤ Apply step 3 to each subgraph until recd no of cluster is obtained



May also delete 2nd max weight

No.

Date

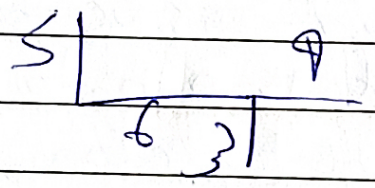
DBSCAN

Density based clustering

Density estimation technique

Density drops sharply on seperation of clusters based on a threshold value.

Another Ratio based edge division



MST-based clustering

1st step $O(n^2) = O(m+n)$

2nd step $O((m+n)\log n)$ → worst case (avr case) $O(n^2 \log n)$

3rd step $O(kn)$ → if directly sort priority queue $O(n \log n)$

best case $O(n \log n)$

$O((m+n)\log n)$