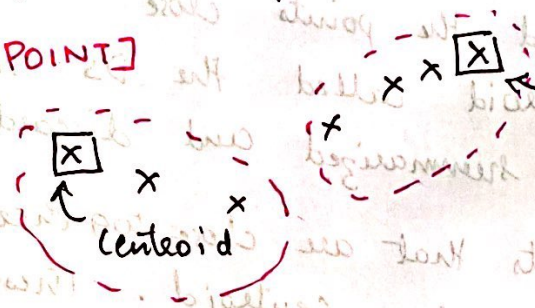
QUIZ - 11

1) K-MEANS [1 POINT]

→ We initialize k points far away from each other
→ for each point assign to a cluster whose centroid is the closest.
→ Update the new centroid based on points in the cluster
→ Reassign points to the closest centroid.
→ Repeat previous 2 steps until the centroids don't change.

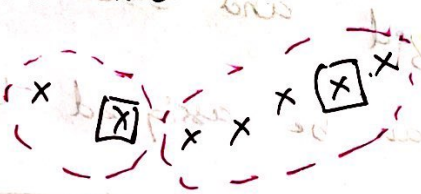
eg:- [1 POINT]

1)



if k = 2.

2) Calculate new centroid

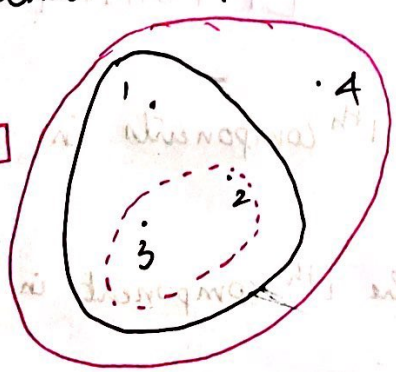


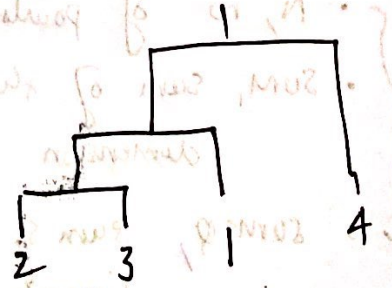
No more changes so algorithms stops.

2) Hierarchical clustering [1 POINT]

→ Start with each point as cluster
→ Repeatedly combine the two nearest clusters into one.
→ We stop when either 1 big cluster is formed or when a particular k value is reached.

eg:-
[1 POINT]





Can they work in non Euclidean Space? [1 POINT]

Both the algorithms can only work in Euclidean space, however they can work in non Euclidean space where we can use clustroids instead of centroids and measure nearness with intercluster distance or by picking a notion of cohesion.

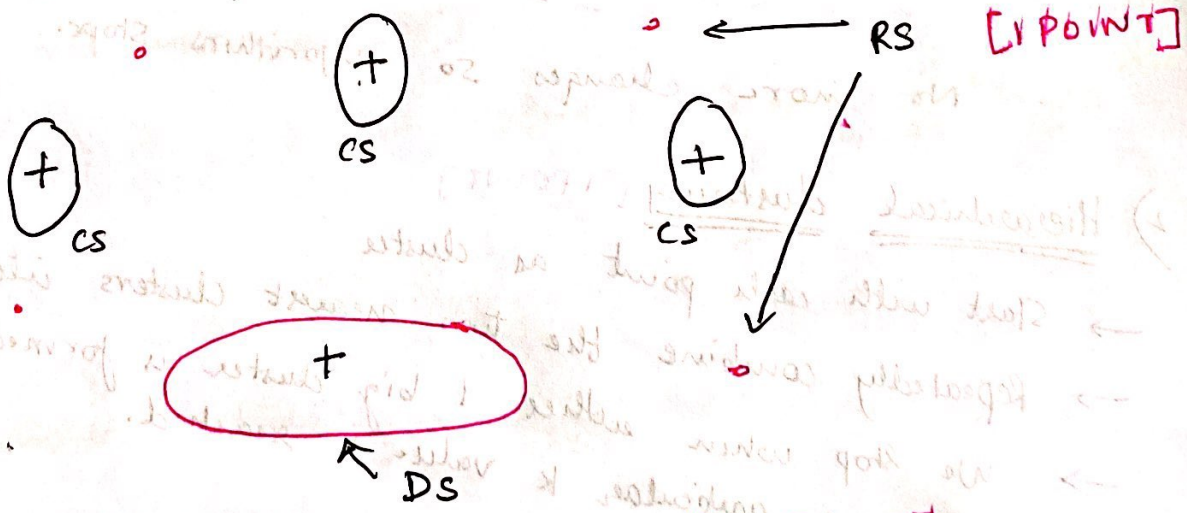
2) BFR

[1 POINT] → load a sample of entire data into memory and Take k random points

→ Use kmeans to find the points close enough to a centroid called the DS set these points can be summarized and discarded.

→ Similarly, group points that are close together but not close to any existing centroid. These points too can be summarized and discarded. (CS set)

→ The isolated points can be assigned to RS set



→ Summarize like the following :- [1 POINT]

2d+1 values.
- N, no of points
- SUM, sum of the respective i^{th} components in the i^{th} dimension
- SUMSQ, sum of squares of the i^{th} component in the i^{th} dimension.

→ load 2 :- [1 POINT]

- Add new points ~~that are closest~~ to DS or CS whichever is the closest.
- Update the DS or CS statistics accordingly

→ consider merging compressed sets in the CS.

→ In last round, merge all CS compressed sets in the CS and all RS points into the nearest DS cluster.

We can use Mahalanobis distance to measure the nearness

LIMITATIONS [1 POINT]

→ Assumes the data is Normally distributed

→ axis are fixed - ellipses at an angle are not ok.