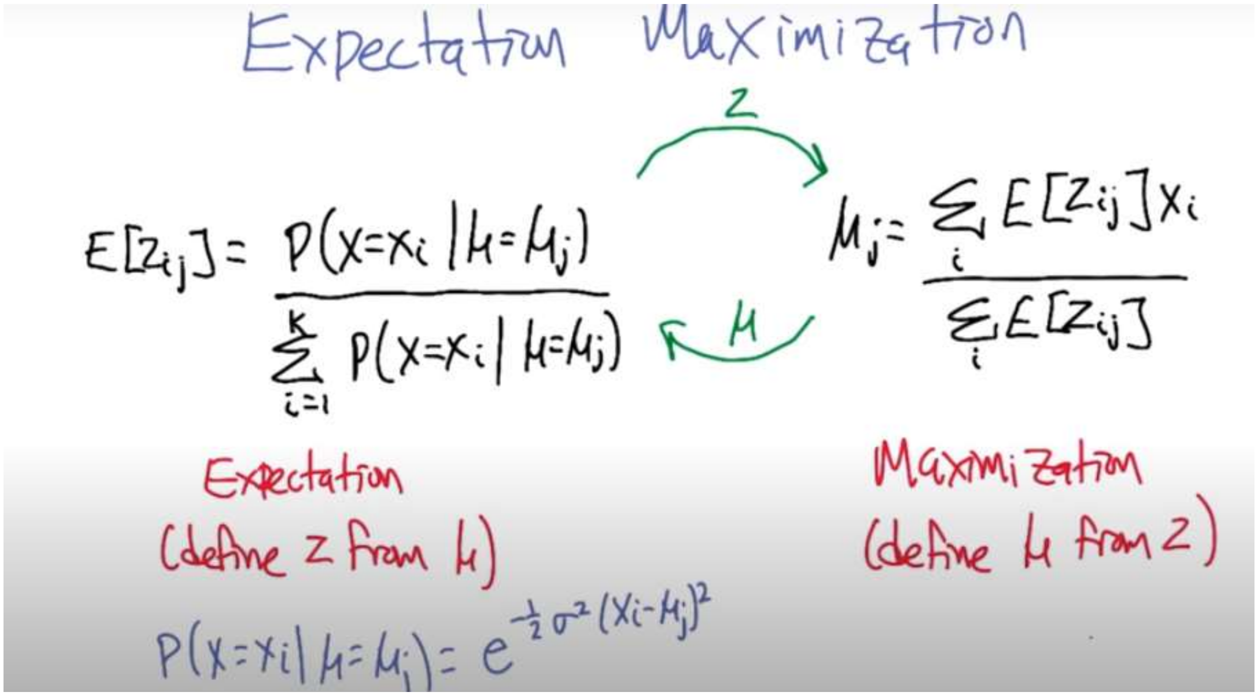
Problem Set 2-CS7641
Author:-Sahil Soni
Emai:-ssoni41@gatech.edu

These are the answers for problem set 2 as part of CS7641 .

Q:-2Show that the K-means procedure can be viewed as a special case of the EM algorithm applied to an appropriate mixture of Gaussian densities model.

Ans:2)
K mean is special case of EM algorithm if we treat it as Gaussian mixture model where we treat EM as hard cluster assignment problem . This is only by theory and special case where they are related but that does not mean we can used Gaussian mixture model to computer kmeans. I find below references where they have clearly defined how they are related.

As explained in the below on the Lecture by Professor Isabel , if all the probability were 0 and 1 then it will be Exactly like K-Mean in special case of Hidden Values. Professor Litman further explain how if one of the Values of z is 0 and One it will be Special case of K mean .
More References:-
https://ttic.uchicago.edu/~dmcallester/ttic101-07/lectures/em/em.pdf



Expectation Maximization

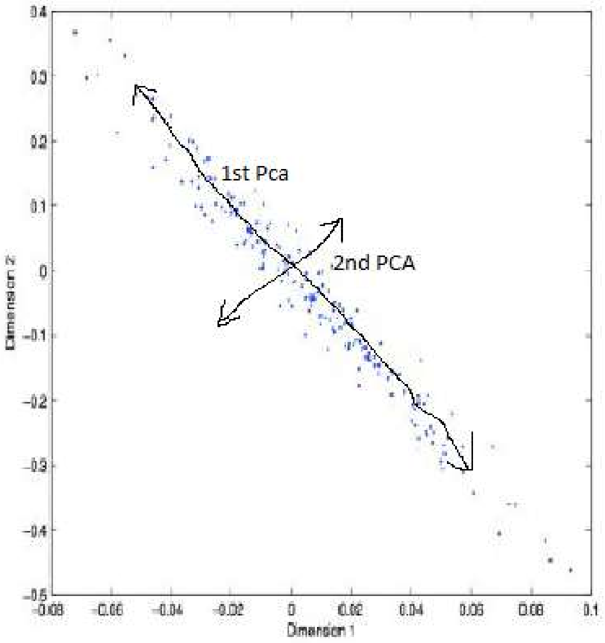
$$E[z_{ij}] = \frac{P(x=x_i \mid \mu=\mu_j)}{\sum_{i=1}^{k} P(x=x_i \mid \mu=\mu_j)} \quad \xrightarrow{z} \quad \mu_j = \frac{\sum_i E[z_{ij}]\, x_i}{\sum_i E[z_{ij}]}$$

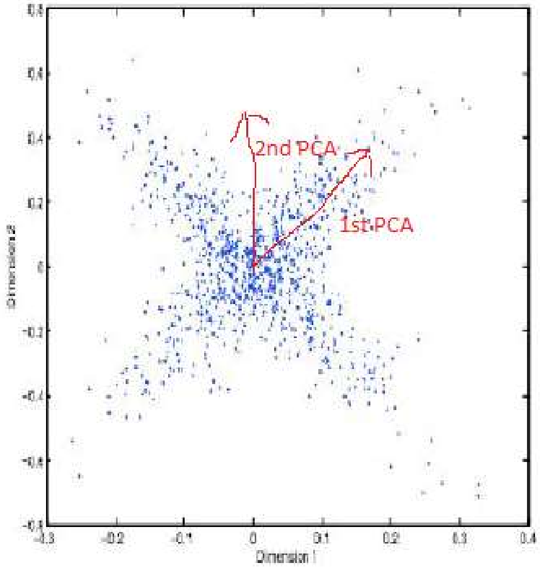
Expectation
(define z from μ)

Maximization
(define μ from z)

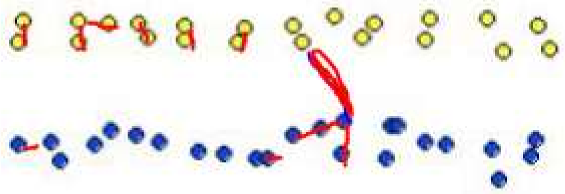
$$P(x=x_i \mid \mu=\mu_j) = e^{-\frac{1}{2}\sigma^2 (x_i-\mu_j)^2}$$

Ans3:-



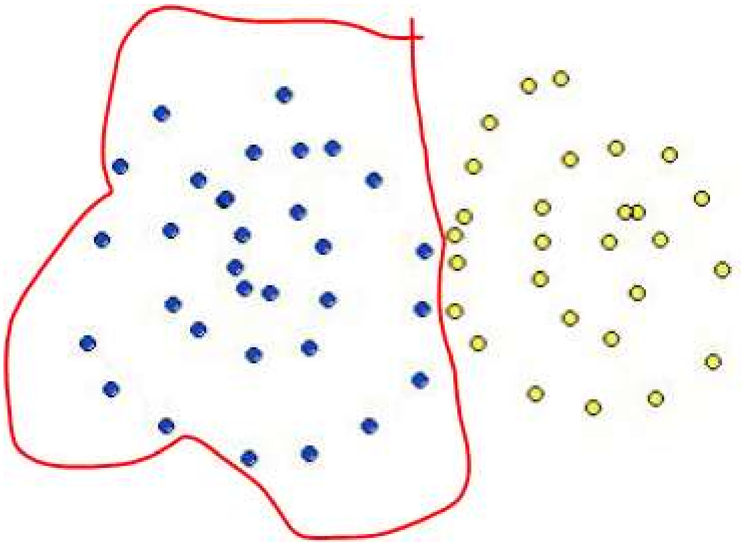


Ans:-4)

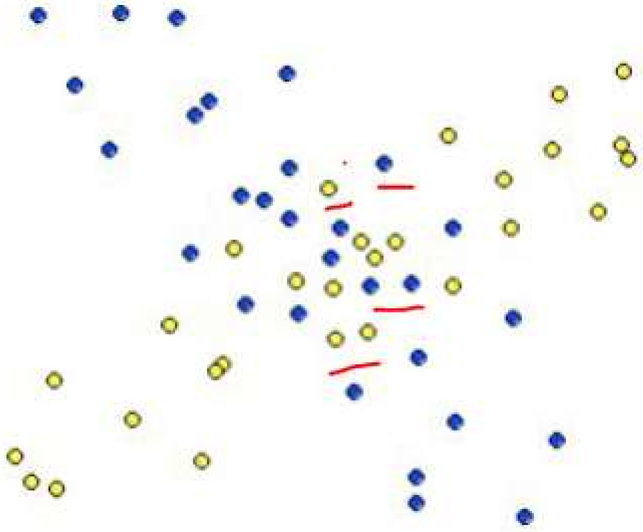


Ans:-

I think Hierarchical clustering with a single link works here because distance between two clusters is shortest as shown on red(so single link) and boundary can be easily separated and large enough and distance between data points are more. So kmeans and EM will not work here. It will formed a hireachy and can be defined on a paper by Parent Child Relationship.



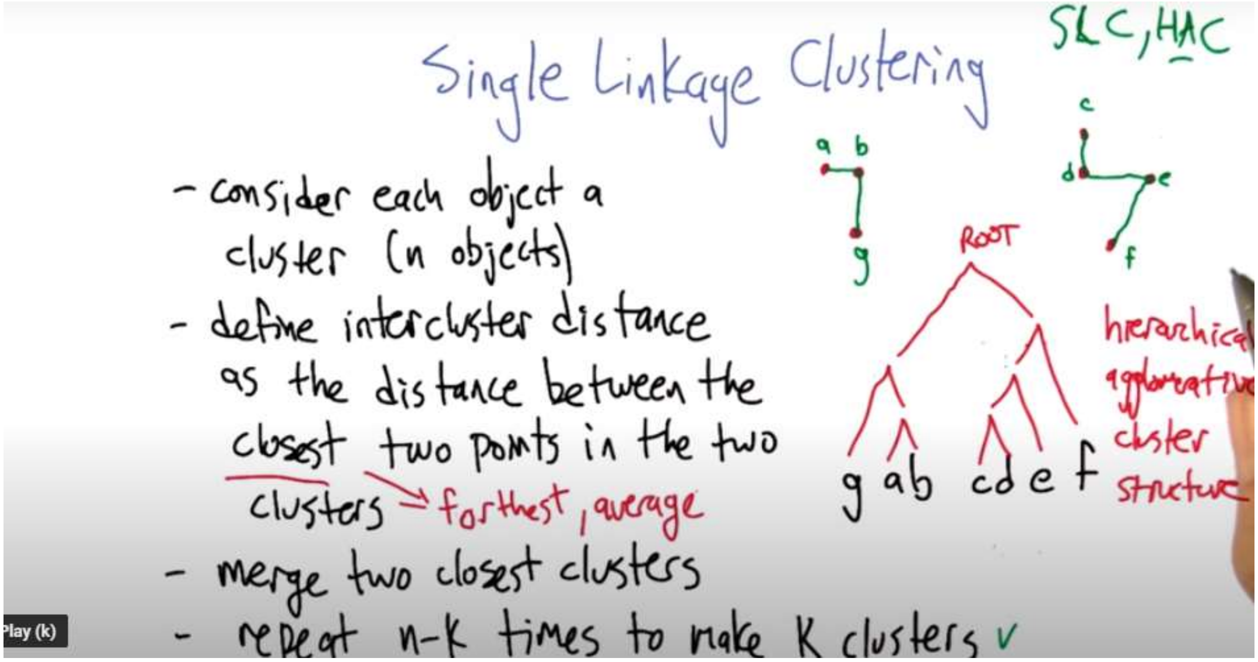
Ans:-
Kmeans and EM will work here simply because we have a very thin boundary to separate the cluster and data points which separate them are very close.The Bridge between clusters are small between data points. When we apply Hierarchical clustering it will fail again because the boundary between data points are too close .That's the issue with SLC as explained on lecture .



Ans:-

Em will work here because it can only work where data points are mixed to each other and we have to separate them , as clusters are overlapping to each other . So only EM can work here.

References:-

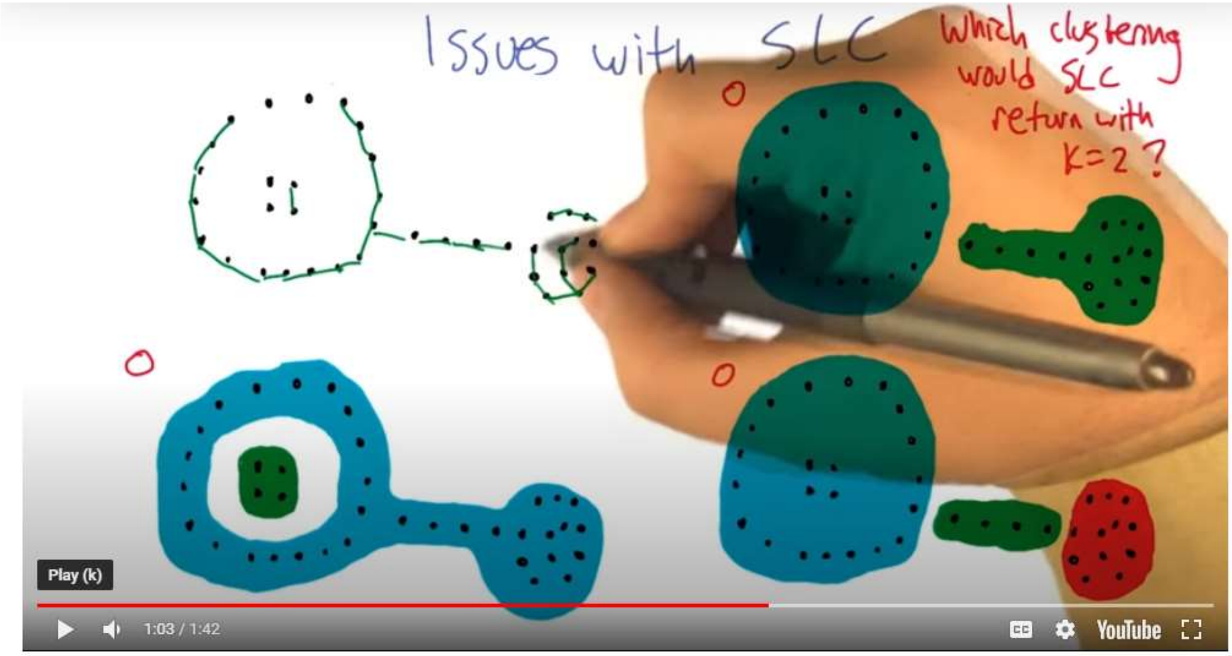


Single Linkage Clustering

SLC, HAC

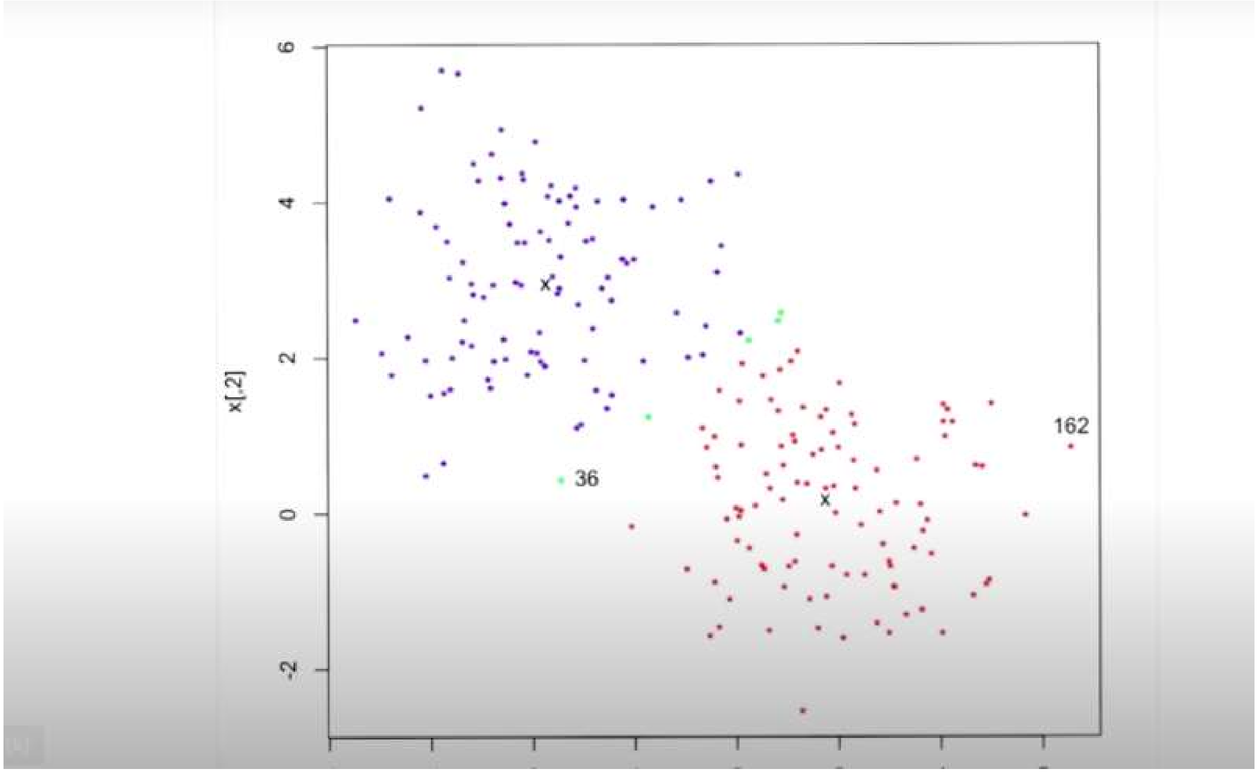
- consider each object a cluster (n objects)
- define intercluster distance as the distance between the closest two points in the two clusters → forthest, average
- merge two closest clusters
- repeat n-k times to make K clusters ✓

hierarchical agglomerative cluster structure



Issues with SLC

which clustering would SLC return with K=2 ?


6
4
2
0
-2
x[,2]
x
x
36
162