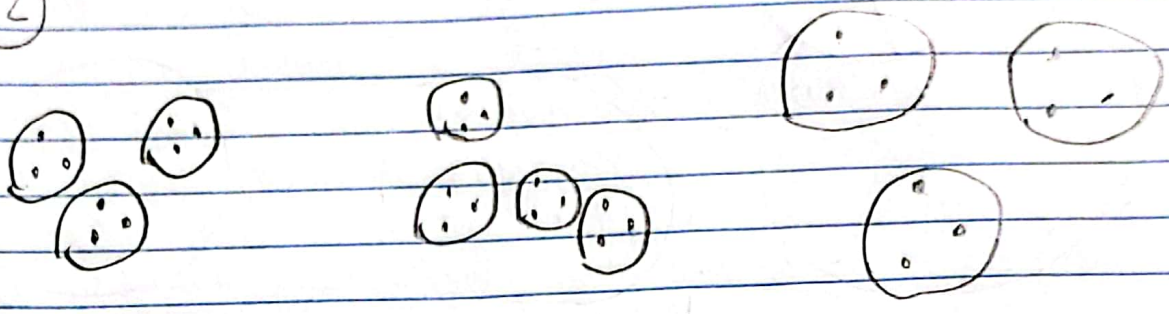
Kamen Petkov

Homework 8

(2)



- Note: The notion of a cluster is not ve
Well defined. It could be the case th
there are other ways of clustering these
points.

---

(6) K-means:

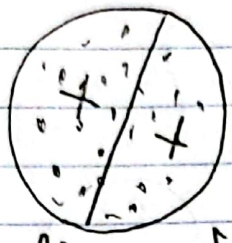
a) K = 2
- points uniformly distributed

K-means algorithm

1. Select K points as initial centroids
2. Form K clusters by assigning each poi
to its closest centroid
3. Recompute the centroid of each clus

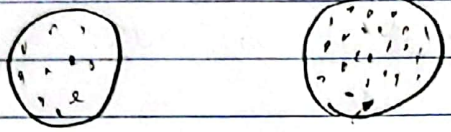
→ Repeat 2 and 3 until centroids do
cha

=> There are infinitely many ways to pa
tion the points into 2 clusters as th
line could be drawn anywhere → the p
are uniformely distributed



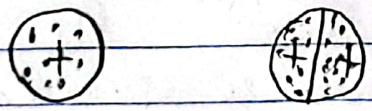
→ The two centroids will be more or less on an equal distance from the line of separation.

b) K=3 → distance between edges is larger than the radii of the circles



=> Since the 2 circles are similar, when we try to do 3 clusters, only one of the two clusters will be split → the one with the greater SSE ; if the circles are exactly the same, then the choice will be "random"

↳ Sketch

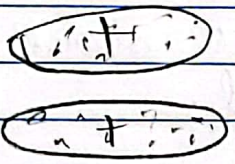


→ The centroid in one of the circles will remain the same and in the other one it will be split like in a).

c) K=3 → distance between the edges is smaller than the radii of the circles.

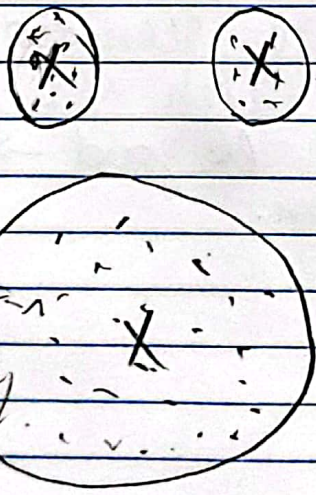
→ The split would affect both circles as the third centroid would be between the circles. The other two centroids would move further away from the middle one. The 3 centroids might happen to be on a similar distance from each other.

d) K=2



→ The points are already clustered in two. I don't think that anything would change. The centroids are indicated in the sketch.

e) K=3



Similar situation to d). The centroids are indicated on the sketch (with +).

(11) - When SSE is low for all clusters, it means that the data is well-clustered and there is no need for additional splits.

- When SSE of one variable is low for only one cluster, it means that this can be our starting point for improving the other clusters during the splits and finding the right positions of the centroids.

- When SSE is high for all clusters, we probably have a lot of noise and the data is not very useful.

- When SSE is high for just one cluster the attribute is not helpful for defining the clusters properly.

- In order to improve clustering we would want to get rid of data with high SSE. Worst case scenario is when all attributes have high SSE because the clusters wouldn't be good -> they will be more or less random.