RECITATION PROBLEMS

Q4. Given K equally sized clusters, the probability that a randomly chosen initial centroid will come from any given cluster is lf K, but the probability that each cluster will have exactly one initial centroid is much lower. (It should be clear that having one initial centroid in each cluster is a good starting situation for K-means.) In general, if there are K clusters and each cluster has n points, then the probability, p, of selecting in a sample of size K one initial centroid from each cluster is given by Equation 8.20. (This assumes sampling with replacement.)

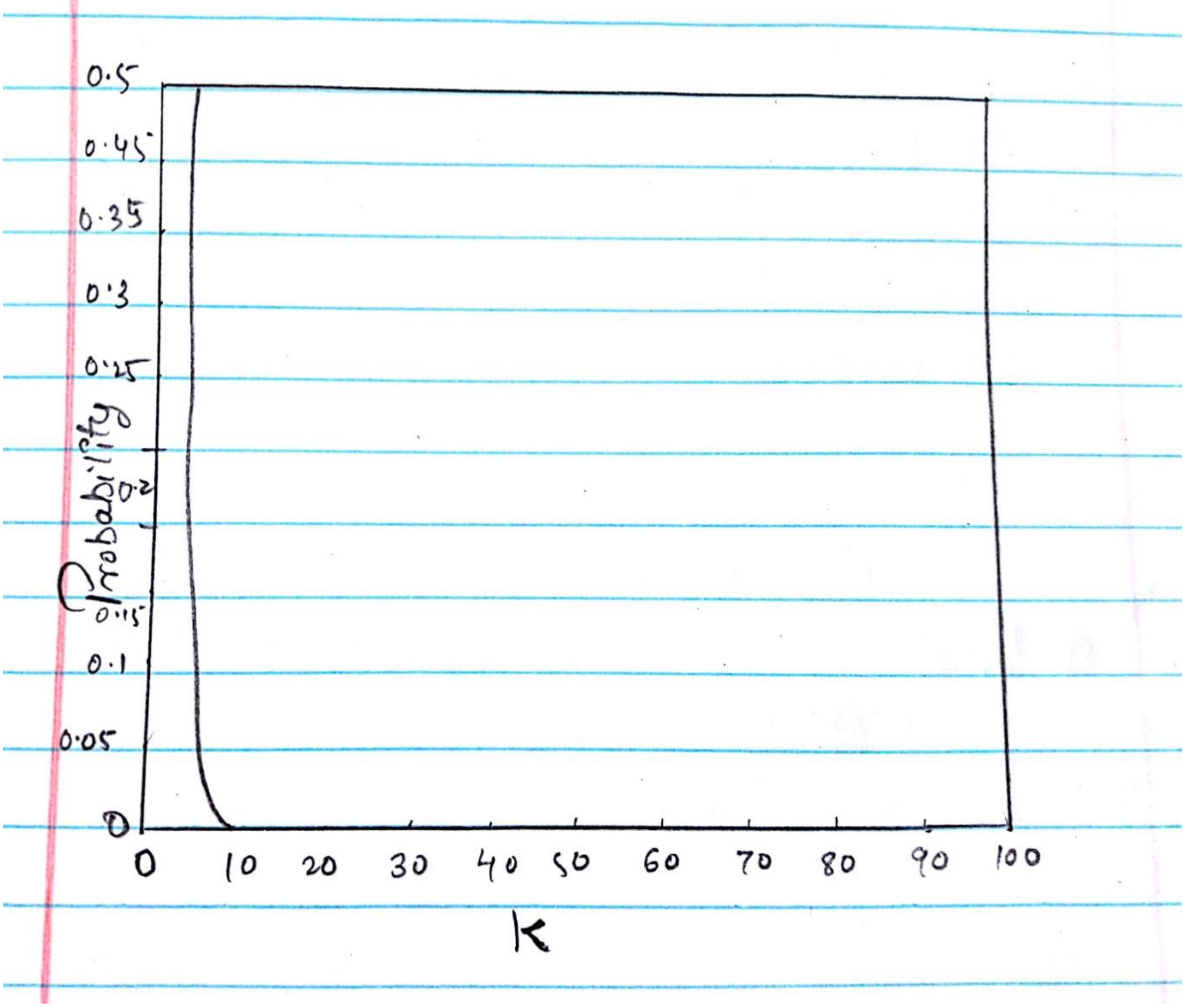
From this formula we can calculate, for example, that the chance of having one initial centroid from each of four clusters is $4!/4^4 = 0.0938$.

P=no. of ways to select one centroid from each cluster/no. of ways to select k centroid $=K!/K^K$

(a) Plot the probability of obtaining one point from each cluster in a sample of size K for values of K between 2 and 100.

Answer:

Here, we can observe that probability value reaches 0 by the time K is 10.



(b) For K clusters, K : 10,100, and 1000, find the probability that a sample of size 2K(contains at least one point from each cluster. You can use either mathematical methods or statistical simulation to determine the answer.

Answer:

By simulation, the probabilities are 0.21, < 10−6, and < 10−6.

Proceeding analytically, the probability that a point doesn't come from a particular cluster is, $1 – 1/K$, and thus, there is a probability that all $2K$ points don't come from a particular cluster is $(1 – 1/K\ )^{2K}$. Hence, there is a probability that at least one of the 200 points comes from a particular cluster is $1 - (1 – 1/K\ )^{2K}$. So, if we assume independence, then an upper bound for the probability that all clusters are represented in the final sample is given by $(1-(1- 1/K\ )^{2K})^{K}$. The values given by this bound are 0.27, 5.7e-07, and 8.2e-64.

Q7. Suppose that for a data set

- there are m points and K clusters,
- half the points and clusters are in "more dense" regions,
- half the points and clusters are in "less dense" regions, and
- the two regions are well-separated from each other.

For the given data set, which of the following should occur in order to minimize the squared error when finding if clusters:

(a) Centroids should be equally distributed between more dense and less dense regions.

(b) More centroids should be allocated to the less dense region.

(c) More centroids should be allocated to the denser region.

Note: Do not get distracted by special cases or bring in factors other than density. However, if you feel the true answer is different from any given above, justify your response.

Answer:

(c). More centroids should be allocated to the denser region because the less dense regions require more centroids if the squared error is to be minimized.

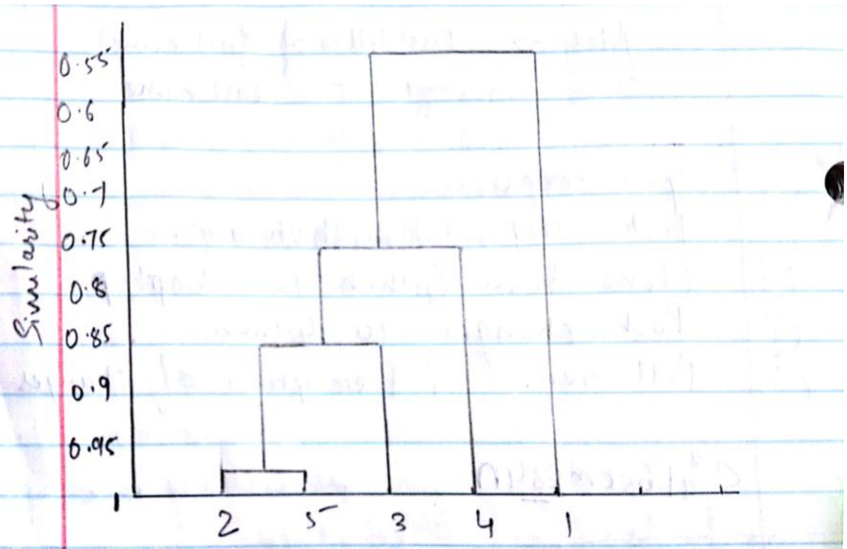
Q11. Total SSE is the sum of the SSE for each separate attribute. What does it mean if the SSE for one variable is low for all clusters? Low for just one cluster? High for all clusters? High for just one cluster? How could you use the per variable SSE information to improve your clustering?

Answer:
If the SSE of one attribute is low for all clusters, then the variable basically must be a constant and cannot be use to separate data into groups, if the SSE of one attribute is relatively low for just one cluster, then this attribute helps in identifying that cluster. If the SSE of an attribute is relatively high for all clusters, then it could mean that the attribute is noise. If the SSE of an attribute is relatively high for one cluster, then there is a likelihood that the information provided by the attributes with low SSE differs from that which defines the cluster. It could merely be the case that the clusters defined by this attribute are different from those defined by the other attributes; but in any case, it means that this attribute does not help define the cluster. Here, the idea is to eliminate attributes that have poor distinguishing power between clusters for all clusters. To improve our clustering the attributes with high SSE for all clusters which a have relatively high SSE with respect to other attributes should be eliminated.

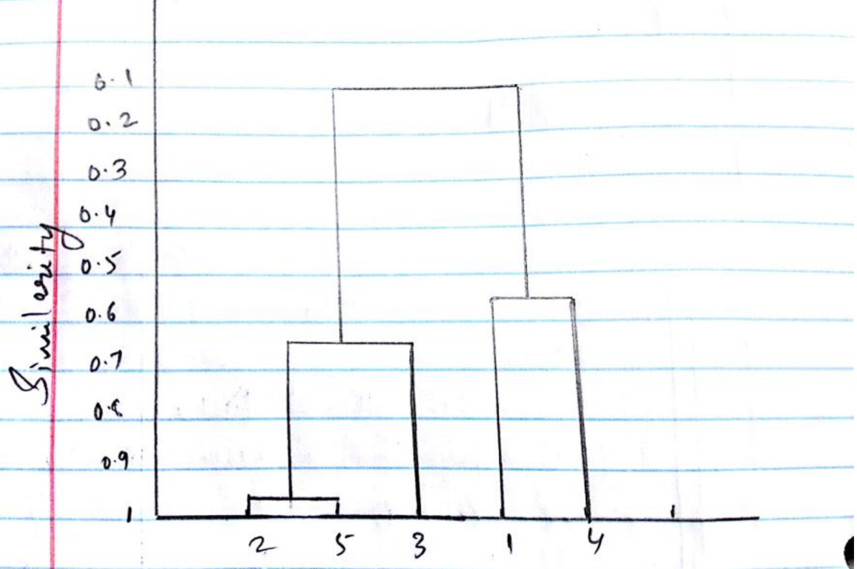
Q17. Hierarchical clustering is sometimes used to generate K clusters, K > I by taking the clusters atl,he Kth level of the dendrogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, and also compare hierarchical approaches to K-means.
The following is a set of one-dimensionapl oints: {6,12,18,24,30,42,48}.

(a) For each of the following sets of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroids.
Answer:
Similarity Matrix :

	P1	P2	P3	P4	P5
P1	1.0	0.10	0.41	0.55	0.35
P2	0.10	1.0	0.64	0.47	0.98
P3	0.41	0.64	1.0	0.44	0.85
P4	0.55	0.47	0.44	1.0	0.76
P5	0.35	0.98	0.85	0.76	1.0



Single link



Complete Link

i. {18, 45}
First cluster is 6, 12, 18, 24, 30.
Error = 360.
Second cluster is 42, 48.
Error = 18.
Total Error = 378

ii. {15,40}
First cluster is 6, 12, 18, 24 .
Error = 180.
Second cluster is 30, 42, 48.

Error = 168.
Total Error = 348.

(b) Do both sets of centroids represent stable solutions; i.e., if the K-means algorithm was run on this set of points using the given centroids as the starting centroids, would there be any change in the clusters generated?
Answer: Yes, both centroids are stable solutions.

(c) What are the two clusters produced by single link?
Answer: The two clusters are {6, 12, 18, 24, 30} and {42, 48}.

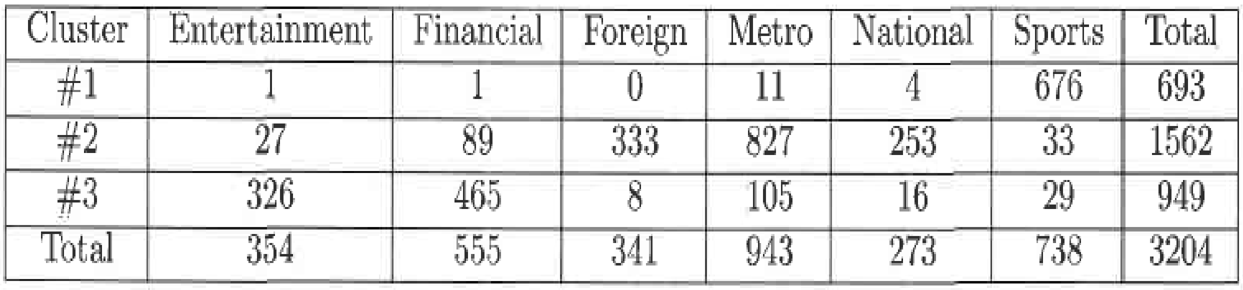
(d) Which technique, K-means or single link, seems to produce the "most natural" clustering in this situation? (For K-means, take the clustering with the lowest squared error.)
Answer: The most natural clustering is produced by Min technique.

(e) What definition(s) of clustering does this natural clustering correspond to? (Well-separated, center-based, contiguous, or density.)
Answer: MIN produces contiguous clusters. Also, density can be considered. Even center-based can be thought of as correct.

(f) What well-known characteristic of the K-means algorithm explains the previous behavior?
Answer:
K-Means works well if the data set has values which creates well separated clusters. It works poor otherwise for finding different shapes. The idea of minimizing squared error creates small clusters. Thus, in this problem, the low error clustering solution is unnatural.

Q21. Compute the entropy and purity for the confusion matrix in Table 8.14

Table 8.14. Confusion matrix for Exercise 21.



Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Total
#1	1	1	0	11	4	676	693
#2	27	89	333	827	253	33	1562
#3	326	465	8	105	16	29	949
Total	354	555	341	943	273	738	3204

Answer:
Entropy -
Cluster 1 - 0.20
Cluster 2 - 1.84
Cluster 3 – 1.70
Total – 1.44
Purity –
Cluster 1 - 0.98
Cluster 2 – 0.53
Cluster 3 – 0.49
Total – 0.61

Q22. You are given two sets of 100 points that fall within the unit square. One set of points is arranged so that the points are uniformly spaced. The other set of points is generated from a uniform distribution over the unit square.

(a) Is there a difference between the two sets of points?
Answer:
Yes. The random points will have regions of lesser or greater density, while the uniformly distributed points will, of course, have uniform density throughout the unit square.

(b) If so, which set of points will typically have a smaller SSE for K:10 clusters?
Answer:
For K:10, the random set of points will have a lower SSE.

(c) What will be the behavior of DBSCAN on the uniform data set? The random data set?
Answer:
For the uniform dataset, it will be treated as a single cluster and all the members will be classified into it accordingly. However, it might vary according to the threshold and can also be classified as noise. In random data set, DBSCAN will find new shapes or patterns and classify them into clusters as density varies.

Q23. Using the data in Exercise 24, compute the silhouette coefficient for each point, each of the two clusters, and the overall clustering.
Answer:

	P1	**P2**	**P3**	**P4**
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Let *a* indicate the average distance of a point to other points in its cluster.
Let *b* indicate the minimum of the average distance of a point to points in another cluster.
Point P1: SC = 1- a/b = 1 - 0.1/((0.65+0.55)/2)= 5/6 = 0.833
Point P2: SC = 1- a/b = 1 - 0.1/((0.7+0.6)/2) = 0.846
Point P2: SC = 1- a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556
Point P2: SC = 1- a/b = 1 - 0.3/((0.55+0.6)/2) = 0.478
Cluster 1 Average SC = (0.833+0.846)/2 = 0.84
Cluster 2 Average SC = (0.556+0.478)/2 = 0.52
Overall Average SC = (0.840+0.517)/2 = 0.68

Q24. Given the set of cluster labels and similarity matrix shown in Tables 8.15 and 8.16, respectively, compute the correlation between the similarity matrix and the ideal similarity matrix, i.e., the matrix whose ijth entry is 1 if two objects belongs to the same cluster, and 0 otherwise.

Answer:
Here, for Point P1, cluster label is 1, P2 is 1, P3 is 2, P4 is 2.
Similarity matrix can be given as,

Point	P1	P2	P3	P4
P1	1	0.8	0.65	0.55
P2	0.8	1	0.7	0.6
P3	0.65	0.7	1	0.9
P4	0.55	0.6	0.9	1

We need to compute the correlation between the off-diagonal elements of the distance matrix and the ideal similarity matrix.
We get:
Standard deviation of the vector **x** : $\sigma x = 0.5164$
Standard deviation of the vector **y** : $\sigma y = 0.1703$
Covariance of **x** and **y**: $\mathrm{cov}(\mathbf{x}, \mathbf{y}) = -0.200$
Therefore, $\mathrm{corr}(\mathbf{x}, \mathbf{y}) = \mathrm{cov}(\mathbf{x}, \mathbf{y})/\sigma x\sigma y = -0.227$