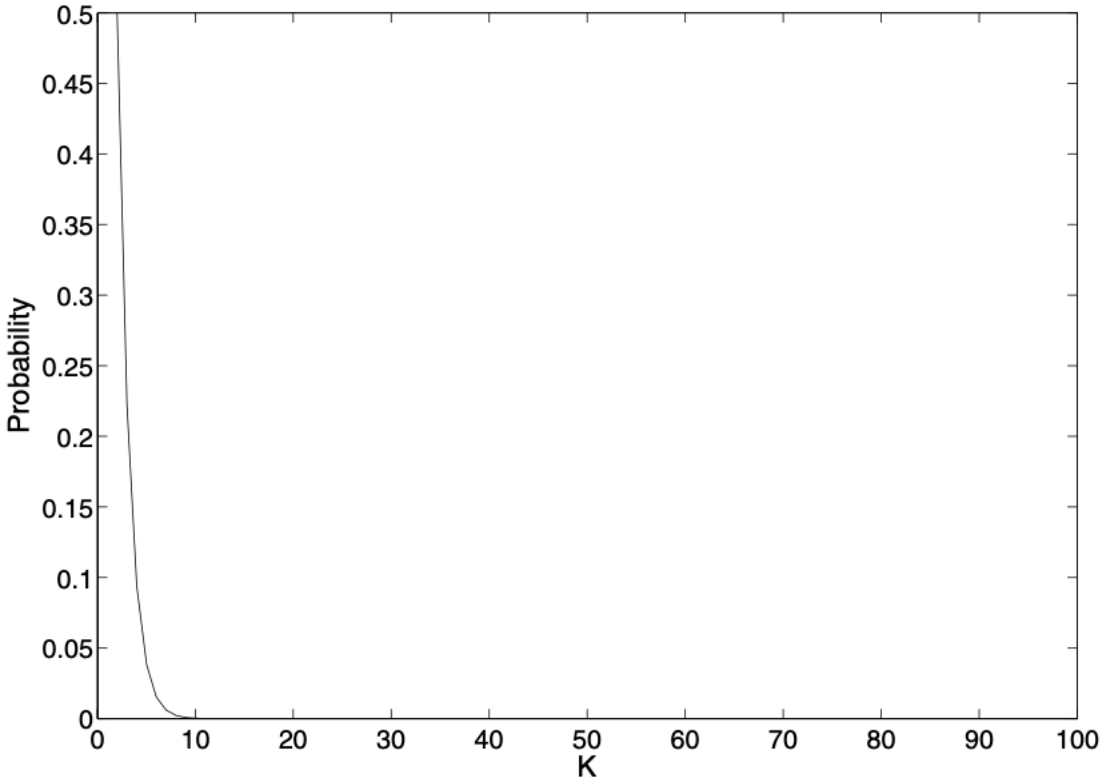
CS422 HW4
SHIQI LIU

1 Recitation Exercises

1.1 Chapter 7
Exercises: 4,7,11,16,17,21,22

4.
(a) Note that the probability is essentially 0 by the time K = 10.



(b) We used simulation to compute it. The probabilities are 0.21 ,<10^(-6) , <10^(-6)

Proceeding analytically, the probability that a point doesn't come from a particular cluster is , 1-(1/k), thus the probability that all 2k points don't come from a particular cluster is (1-(1/k))^(2k). Hence, the probability that last one of the 200 points comes from a particular cluster is 1-(1-(1/k))^(2k).

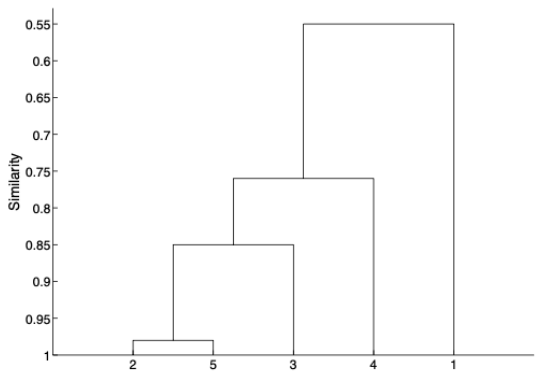
If we assume independence, then an upper bound for the probability that all clusters are represented in the final samples is given by (1-(1-(1/k))^(2k))^(k). The values given by this bound are 0.27, 5.7e-07, 8.3e-64

7
The correct answer is (c). Less dense regions require more centroids if the squared error is to be minimized.

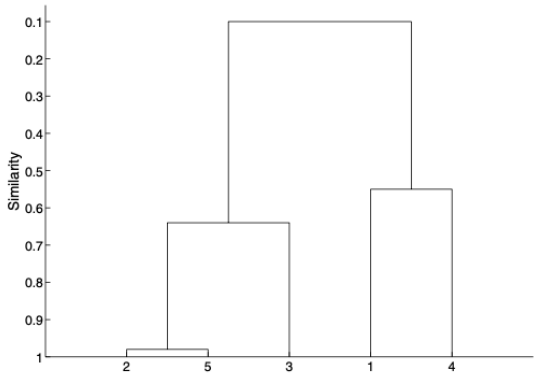
11

a) If the SSE of one attribute is low for all clusters, then the variable is essentially a constant and of little use in dividing the data into groups.
b) if the SSE of one attribute is relatively low for just one cluster, then this attribute helps define the cluster.
c) If the SSE of an attribute is relatively high for all clusters, then it could well mean that the attribute is noise.
d) If the SSE of an attribute is relatively high for one cluster, then it is at odds with the information provided by the attributes with low SSE that define 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.
e) The idea is to eliminate attributes that have poor distinguishing power between clusters, i.e., low or high SSE for all clusters, since they are useless for clustering. Note that attributes with high SSE for all clusters are particularly troublesome if they have a relatively high SSE with respect to other attributes (perhaps because of their scale) since they introduce a lot of noise into the computation of the overall SSE.

16



(a) Single link.



(b) Complete link.

17

a) 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) Yes, both centroids are stable solutions.
c) The two clusters are {6, 12, 18, 24, 30} and {42, 48}
d) MIN produces the most natural clustering.
e) MIN produces contiguous clusters.
f) K-means is not good at finding clusters of different sizes, at least when they are not well separated. The reason for this is that the objective of minimizing squared error causes it to "break" the larger cluster. Thus, in this problem, the low error clustering solution is the "unnatural" one.

21

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Total	Entropy	Purity
#1	1	1	0	11	4	676	693	0.20	0.98
#2	27	89	333	827	253	33	1562	1.84	0.53
#3	326	465	8	105	16	29	949	1.70	0.49
total	354	555	341	943	273	738	3204	1.44	0.61

22

a) Yes. The random points will have regions of lesser or greater density, while the uniformly distributed points will have uniform density throughout the unit square.
b) Random set of points will have a lower SSSE.
c) DBSCAN will merge all points in the uniform data set into one cluster or classify them all as noise, depending on the threshold. There might be some boundary issues for points at the edge of the region. However, DBSCAN can often find clusters in the random data, since it does have some variation in density.

2 Practicum Problems

2.1

If we used origin as a class label, we are able to get the different values. That implies the cluster assignment might do not have any clear relationship with class label

2.2

(1) Provide the Silhouette score to justify which value of k is optimal.
(2) Calculate the mean values for all features in each cluster for the optimal clustering
(3) how do these values differ from the centroid
coordinates?

(1) According to Silhouette score, the k is optimal when the total number of clusters equals 2. That is because 2 is the highest silhouette score, and a value is higher, the object is more matched to its own cluster and more poor matched it neighboring clusters.

(2) The mean value is shown.
(3) According the data, we know the mean of a cluster is the same as the centroid coordinate.
2.3
Homogeneity means all of the observations are the same class label in the same cluster. Completeness means all members which in the same class are in the same cluster.
Both scores range is from 0 to 1, the higher score will be the better result.