CS 4824 / ECE 4424 Homework 4 (Written Portion) Solutions [Total points: 55]

Question 1 (Types of Clustering) [12 points]

a) Partitional, overlapping (most actors have done movies of multiple genres), complete.

b) Hierarchical, overlapping, partial.

c) Partitional, non-overlapping, partial (some students in the CS department wouldn't have taken ML class and thus can't be grouped).

d) Partitional, overlapping, partial.

Question 2 (Computing SSE) [2 + 2 + 4 = 8 points]

i. 4R^2

ii. 4(a^2 + b^2 + R^2)

iii. Subsitute a = R, b = 0, and radius as R/2 in the above expression and multiply by 2 to account for both clusters. The final result would be 10*R^2.

Question 3 (k-means and Bisecting k-means) [6 points]

a) Bisecting K-means would perform better as it is more likely to separate out the bigger cluster from the other two small clusters in the right in the first iteration. And in the second iteration, the cluster with the two small clusters would very likely get bisected into two small globular clusters of radius R/2. Whereas, K-means would face more initialization problems in getting correct initialization of three centroids.

b) K-means would perform better, because bisecting K-means would tend to split the middle cluster in the first iteration, especially if the distance between the two clusters is small compared to the radius of each cluster. Once the middle cluster is split, it can never be rejoined in successive iterations.

Question 4 (Hierarchical Clustering and DBSCAN) [12 points]

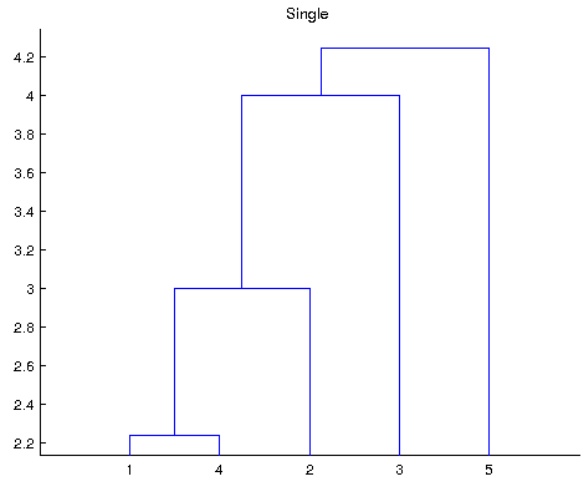
a) Single-link and DBSCAN will work well for this case. Complete-link will have problem due to varying sizes of clusters. b) DBSCAN will perform well but single-link will not perform well, since it is susceptible to noise. Complete link will also not work well due to varying sizes of clusters. c) Single-link and DBSCAN will perform well for this purpose, but complete-link can break the clusters into globular shapes. d) If we assume that the density of the noise is almost same as the left cluster, then none of the three methods will work in this case. Otherwise, DBSCAN will perform well, if noise is less dense than the left cluster.

Question 5 (Hierarchical Clustering) [5 + 4 + 4 points]

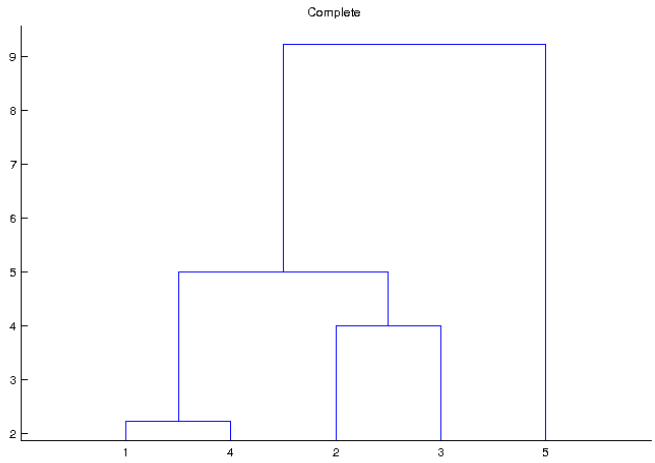
a)

Pair-wise Distances	Point 1	Point 2	Point 3	Point 4	Point 5
Point 1	0	3	5	2.23	9.2
Point 2	3	0	4	4.47	7.6
Point 3	5	4	0	4.47	4.2
Point 4	2.23	4.47	4.47	0	8.6
Point 5	9.2	7.6	4.2	8.6	0

b)



c)



Question 6 (Hierarchical Clustering) [4 points]

MIN will find four clusters as the between-distance of small clusters is larger than distance between points in larger cluster.
MAX will merge three smaller clusters before it finds the large cluster

Practice Questions

Question 7 (Types of Clustering)

a) Partitional, overlapping, complete

b) Partitional, non-overlapping, complete

You can answer incomplete if you took the view that the set of objects was the entire population.

Question 8 (k-means and DBSCAN)

(i) In (a), DBSCAN will either place all the points in one cluster, or will produce no clusters, depending on its parameter settings. On (b), DBSCAN will produce the dense regions, if any, as clusters.

(ii) K-means will find K globular clusters even if the points in them are just randomly distributed. DBSCAN will produce the dense regions, if any, as clusters.

Question 9 (Comparing clustering methods)

(a)
center-based 2 clusters. The rectangular region will be split in half. Note that the noise is included in the two clusters.
contiguity-based 1 cluster because the two circular regions will be joined by noise. Ideally, we would have wanted to obtain three clusters: one for each circle and noise points as the third cluster. However, as we proceed with merging points in the MIN approach, we would first merge points in the two circles to form two clusters, but all noise points will also be left as singleton clusters. These noise points would then start to get merged with points from the two circles and we would never have a clean separation between the two circles and the noise points at any level of the dendrogram.
density-based 2 clusters, one for each circular region. Noise will be eliminated.

(b)
center-based 1 cluster that includes both rings.
contiguity-based 2 clusters, one for each ring.
density-based 2 clusters, one for each ring.

(c)
center-based 3 clusters, one for each triangular region. One cluster is also an acceptable answer.
contiguity-based 1 cluster. The three triangular regions will be joined together because they touch.
density-based 3 clusters, one for each triangular region. Even though the three triangles touch, the density in the region where they touch is lower than throughout the interior of the triangles.

(d)

center-based 2 clusters. The two groups of lines will be split in two.
contiguity-based 5 clusters. Each set of lines that intertwines becomes a cluster.
density-based 2 clusters. The two groups of lines define two regions of high density separated by a region of low density.

Question 10 (Hierarchical Clustering)

a) MIN will perform well but MAX may break the larger cluster and merge it with the smaller cluster. b) MIN is sensitive to noise, MAX is good to remove the noise but can break the elliptical cluster. c) MIN will perform well, MAX will break also the clusters into globular shape. d) MIN will perform well if the distance between the two circles is greater than the distance between nearest points in the sparser cluster. MAX will tend to break the larger cluster and merge it with the smaller cluster.

Question 11 (k-means)

a) **True**. The clusters are too far away for one centroid to attract points from another.

b) **False**. The final clusters will have points from both of the two shaded regions since they are close to each other and not of circular shape.

c) **True**. The centroid at 12.5 is farther away from all points than any other clusters and will become empty.