HW #3

Question 1 (10 points):

Many partitional clustering algorithms that automatically determine the number of clusters claim that this is an advantage. List two situations in which this is not the case.

Question 2 (20 points)

You are given a dataset with 100 records and are asked to cluster the data. You use K-means to cluster the data, but for all values for K, $1 <= K <= 100$, the K-means algorithm returns only one non-empty cluster. You then apply an incremental version of K-means, but obtain exactly the same result. How is this possible? How would single link or DBSCAN handle such data?

Question 3 (20 points)
Using the below data, compute the silhouette coefficient for each point, each of the two clusters, and the overall clustering.

Table of cluster labels

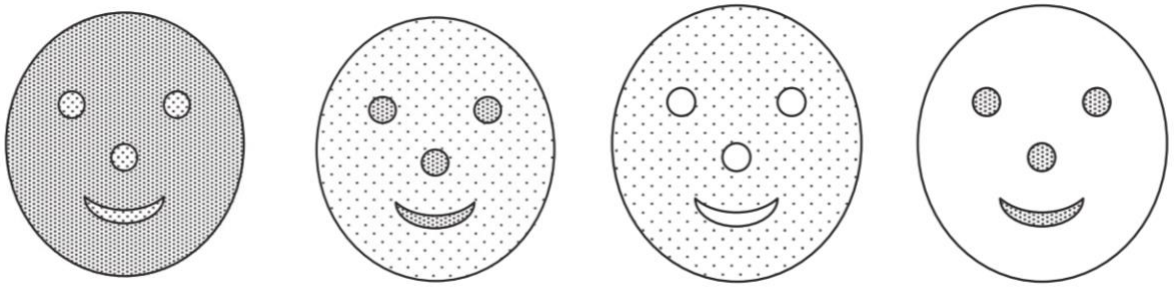
Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Similarity matrix

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

Question 4 (20 points)
Given the below four faces, darkness or number of dots represents density. Lines are used only to distinguish regions and do not represent points.

For each figure, could you use partition, hierarchical, density, and other algorithms we learned in class to find the patterns represented by the nose, eyes, and mouth? Please list at least 3 different types of algorithms and explain the pros and cons of each.



(a) (b) (c) (d)

Question 5 (30 points)
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?
(b). If so, which set of points will typically have a smaller SSE for K = 10 clusters?
(c). What will be the behavior of DBSCAN on the uniform dataset? The random dataset?