Machine Learning & Machine Learning (extended)

Practice Exercise Sheet – Clustering

Question 1: Use K-means algorithm and squared Euclidean distance measure to cluster the following 2-dimensional objects in to 3 clusters. The points are described as: O1 (2,10), O2 (2,5), O3 (8,4), O4 (5,8), O5 (7,5), O6 (6,4), O7 (1,2), O8 (4,9). Suppose that initial cluster centres are O1, O4, and O7. Draw a 10x10 grid with all the 8 objects with initialized cluster centres marked.

a) Run the K-means algorithm steps for 1 iteration and show: (i) the clusters (i.e. objects belonging to each cluster), (ii) the centres of new clusters, and (iii) draw a 10x10 grid with all the 8 objects and show the clusters and centres after the first iteration.
b) Using graphical drawing, illustrate the algorithm iterations until the algorithm converges. How many iterations it takes to converge?

Question 2: Repeat the above with Manhattan distance measure.

Question 3: Use min/single link to perform agglomerative clustering by showing the dendrogram for the data described by the distance matrix below:

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Note: the height of each “junction” in the dendrogram represents the distance between the pair of clusters.

Question 4: Repeat the above with max/complete link to perform agglomerative clustering.

Question 5: Use single link agglomerative clustering to cluster the following 8 objects by showing the dendrograms: O1 (2,10), O2 (2,5), O3 (8,4), O4 (5,8), O5 (7,5), O6 (6,4), O7 (1,2), O8 (4,9).

Question 6: Use complete link agglomerative clustering to cluster the following 8 objects by showing the dendrograms: O1 (2,10), O2 (2,5), O3 (8,4), O4 (5,8), O5 (7,5), O6 (6,4), O7 (1,2), O8 (4,9).

Question 7: What is the goal of clustering, and how it differs from classification?

Question 8: Describe in what situation the conventional k-means algorithm would fail to cluster the data. Can you suggest a modification to overcome the problem?

Question 9: Suppose you have run k-means clustering on an available data set. Later you get more data points which are observed over similar attributes/features. Can we cluster the new data points using the results of first run of k-means algorithm?