CS4054: Bioinformatics

Rushda Muneer

Assignment 3

Exercise 1: Compute **MaxDistance(Data, Centers)** for *Data* shown in the figure 1 and **Centers (2, 4), (6, 7), and (7, 3)**. **(5 points)**
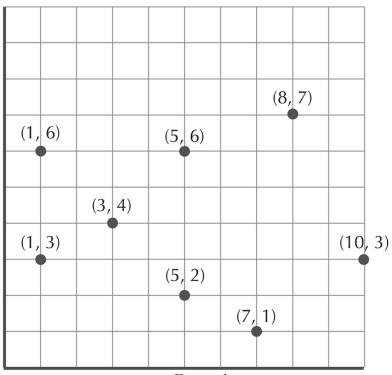

Figure 1

Exercise 2: Compute the values of **MaxDistance(Data, Centers)** and **Distortion(Data, Centers)** for the eight data points in the figure 1 and the **three centers (3, 4.5), (6, 1.5), and (9, 5)**. How do these values differ if the centers are instead **(5/3, 13/3), (6.5, 6.5), and (22/3, 2)**? **(10 points)**

Exercise 3: Run the Lloyd algorithm for the set of four one-dimensional **data points {0, 1, 1.9, 3}** with the two initial **centers {1, 3}**. **(5 points)**

Exercise 4: Apply **Hierarchical Clustering** to the distance matrix using D_{min} and D_{avg} . **(30 points)**

$$D_{avg}(C_1, C_2) = \frac{\sum_{\text{all points } i \text{ in cluster } C_1} \sum_{\text{all points } j \text{ in cluster } C_2} D_{i,j}}{|C_1| \cdot |C_2|}$$

$$D_{min}(C_1,C_2) = \min_{\text{all points } i \text{ in cluster } C_1, \text{ all points } j \text{ in cluster } C_2} D_{i,j} \; .$$

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0