COMPSCI 762 2022 S1 Week 12 Questions

Luke Chang

May 22, 2022

Question 1 – Clustering

Figure 1: A data set has 8 instances, from A1 to A8. The Euclidean distances between every two instances are given as the following:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{45}$	$\sqrt{63}$	$\sqrt{57}$	$\sqrt{41}$	$\sqrt{28}$	$\sqrt{95}$	$\sqrt{6}$
A2		0	$\sqrt{55}$	$\sqrt{49}$	$\sqrt{35}$	$\sqrt{11}$	$\sqrt{5}$	$\sqrt{25}$
A3			0	$\sqrt{11}$	$\sqrt{23}$	$\sqrt{54}$	$\sqrt{47}$	$\sqrt{65}$
A4				0	$\sqrt{2}$	$\sqrt{7}$	$\sqrt{26}$	$\sqrt{5}$
A5					0	$\sqrt{5}$	$\sqrt{21}$	$\sqrt{35}$
A6						0	$\sqrt{13}$	$\sqrt{27}$
A7							0	$\sqrt{53}$
A8								0

Use the data in Figure 1 to answer the questions below:

1. Suppose that the initial seeds (centers of each cluster) are **A1**, **A4** and **A7**. Run the k-means algorithm for 1 epoch only. At the end of this epoch show:
   - The new clusters (i.e. the examples belonging to each cluster)
   - The centers of the new clusters
2. Use single-linkage (MIN) agglomerative clustering to group the data. Show the dendrogram.

Question 2 – Ooutlier/Anomaly Detection

1. What are the three types of anomaly? Give an example for each type.
2. You are given the following list of 2D data points:

   $$[1;1];[1;2];[2;2];[2;1];[3;3];[2;5];[2;3]$$

   If you had to select one point to be anomalous, how to use Manhattan distance to determine the outlier. Explain the anomaly detection technique.
3. Consider a set of points (0,0), (1,0), (0,1), (3,0). Calculate the *Local Outlier Factor* (LOF) score for the points using Manhattan distance and k is 2.