HW7: Clustering (All bonus problems)

Update Nov. 11: HW 7 is all bonus problems.

Due Friday Nov. 20 2:15PM. Upload your answers as a PDF file to Canvas.

Help sessions: Thursday Nov. 19, 3-4pm, and 8-9pm. Use class Zoom meeting link.

How to see this in nice PDF

For precompiled PDF, click here. To compile, `pandoc hw7.md -o hw7.pdf`

Part 1: K-means

1. [1pts] Suppose we have a set of data represented in the matrix below:

$$X = \begin{pmatrix} -0.57 & 0.87 & -0.89 \\ 0.04 & -0.76 & 0.41 \\ 0.55 & -0.38 & 0.56 \\ -0.65 & -1.66 & 0.35 \\ -0.28 & 0.25 & -1.54 \\ -1.18 & 1.26 & -0.33 \end{pmatrix}$$

   where each row is a sample (denoted as X_i, and i **starts from 1**) and each column is a feature.

   We know that a sample should be assigned to a cluster if the sample is closer to the centroid of the cluster than to that of any other cluster. Suppose we have two clusters: cluster 1 and cluster 2, whose centroids are C_1 and C_2. What is the condition that a sample X_i belongs to cluster 1? Write it as a comparison expression in terms of X_i, C_1, and C_2. The distance definition is Euclidean distance.

2. [2pts] Continuing from Problem 1 above, if we use sample 0 and sample 1 as the initial centroids for two clusters, what samples will be assigned to cluster 1 and what will be assigned to cluster 2? Samples 1 and 2 obviously belong to clusters 1 and 2.

   Show your steps. You may use a computer program instead of keying lengthy instructions on a calculator. The function `kmeans` in the class demo file `kmeans.py` may be used.

3. [2pts] Continuing from Problems 2 and 3 above, what are the centroids of new clusters 1 and 2?

   Show your steps. You may use a computer program instead of keying lengthy instructions on a calculator. The function `kmeans` in the class demo file `kmeans.py` may be used.

Part II: Single-linkage clustering

4. [3pts] Using the same data in Problem 1, what is distance matrix initially? Populate your results in the table below by replacing all dashes with proper values:

	(1)	(2)	(3)	(4)	(5)	(6)
(1)	0	-	-	-	-	-
(2)	-	0	-	-	-	-
(3)	-	-	0	-	-	-
(4)	-	-	-	0	-	-
(5)	-	-	-	-	0	-
(6)	-	-	-	-	-	0

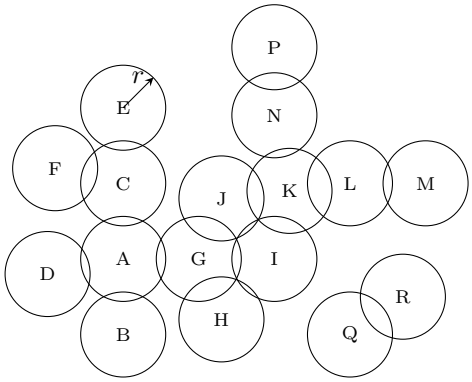
   Each column or row corresponds to a sample which is one row in the matrix X in Problem 1. Row and column indexes are already populated.

You may write a computer program to help you.

5. [1pt] Based on the distance matrix obtained in Problem 4, what two clusters should be merged?
6. [4pts] If we want 3 clusters, what are they per single-linkage clustering? You must show distance matrix D obtained in each iteration, and the dendrogram with proper node values (δ's, divided by 2 or not).

Part III: DBSCAN

7. [1pt] Given an illustration of samples in the figure below, what are the neighbors of sample A if $T = 3$? A sample is a neighbor of another sample if the circles representing the two samples intersect.



8. [2pts] Continuing from Problem 7 above, which neighbors of sample A are core points? Show your steps.
9. [2pts] Show all members of the cluster starting from node A. The order to visit samples is alphabetical.
10. [2pts] The pseudocode for DBSCAN in slides is very long. However, lines 5–7 do something very similar to what lines 10, 12, 13 do. Can you somehow merge them so the pseudocode can be shorter?