CS 4824 / ECE 4424 **Homework 4 (Written Portion)** **[Total points: 30]**

Question 1. Ensemble Learning True/False [2 + 2 = 4 points]

For a binary classification problem, you are given a collection of base classifiers where every base classifier has an error rate of ϵ. The ensemble prediction is simply the majority vote of the predictions of the base classifiers. State whether the following statements are "true" or "false." No explanation needed.

a) If $\epsilon = 0.3$ and the predictions of all base classifiers are independent, the error rate of the ensemble classifier will be smaller than ϵ. **[2 points]**

True.

b) If $\epsilon = 0.3$ and all base classifiers are identical, the error rate of the ensemble classifier will be smaller than ϵ. **[2 points]**

False.

Question 2. Clustering True/False [3 + 3 + 3 = 9 points]

For the following questions, give an answer and a short (1 or 2 sentences) explanation. For the rest of this question, "agglomerative hierarchical clustering" refers to procedures such as single link, complete link, and group average, while "k-means clustering" refers to k-means with random initialization of centroids and Euclidean distance.

a) Agglomerative hierarchical clustering procedures are better able to handle outliers than k-means. **[3 points]**

True.

K-means are sensitive to outliers.

b) For any given data set, different runs of k-means can produce different clusterings, but agglomerative hierarchical clustering procedures will always produce the same clustering. **[3 points]**

True.

In agglomerative hierarchical clustering, we calculate the distance between random mean centroid and each point and then determine the cluster belongings. So it is always the same.

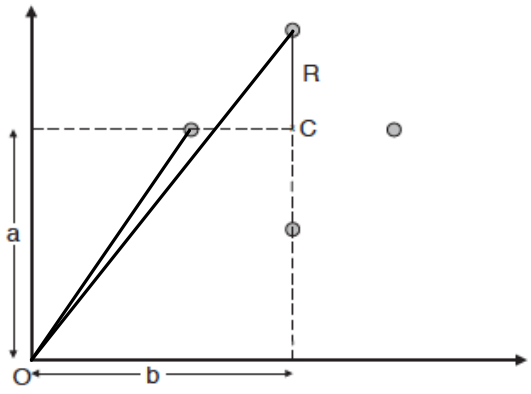
c) When clustering a dataset using K-means, SSE is guaranteed to monotonically decrease as the number of clusters increases. **[3 points]**

True.

As the number of cluster increases, the mean square error decreases and the algorithom becomes more stable as less iteractions.

Question 3 (Computing SSE) [2 + 2 + 2 = 6 points]

Consider the four data points shown in Figure below. The distance between each point to the center C is R.



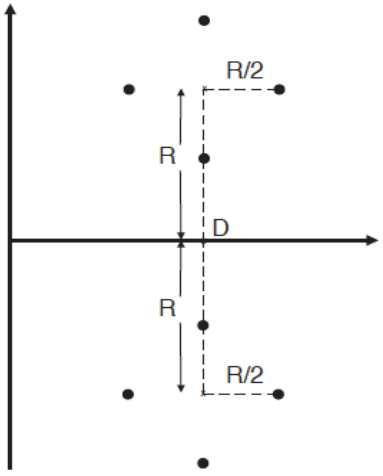
(i) Compute the total SSE of the data points from the centroid, C. **[2 points]**

$4R^2$

(ii) Compute the total SSE of the data points to the origin, O. **[2 points]**

$4(a^2 + b^2 + R^2)$

(iii) Use your approach in (ii) to compute the SSE for the 8 data points shown below with respect to the centroid, D. **[2 points]**



$a \to R$

$b \to 0$

$R \to \frac{R}{2}$

$4(a^2 + b^2 + R^2)$

$= 4(R^2 + (\frac{R}{2})^2)$

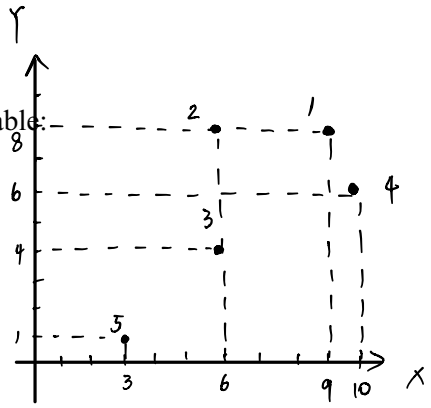
$= 5R^2$

$5R^2 \cdot 2 = 10R^2$

Question 4 (Hierarchical Clustering) [5 + 3 + 3 = 11 points]

Consider a set of 5 points in two-dimensional space, shown in the following table:

Point ID	X	Y
1	9	8
2	6	8
3	6	4
4	10	6
5	3	1

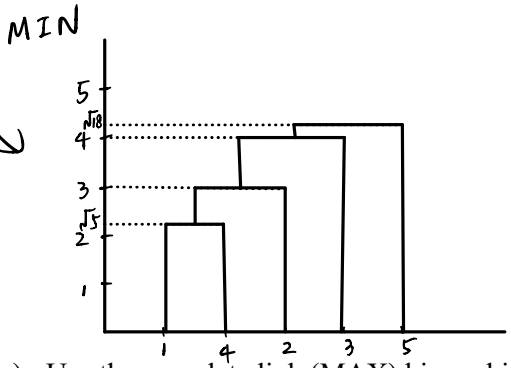


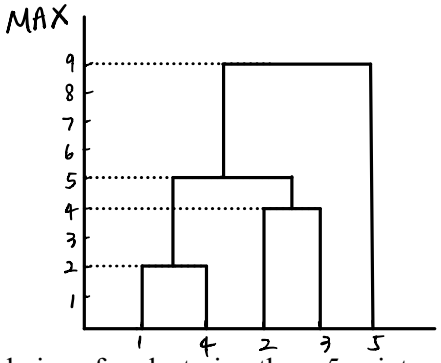
Assuming Euclidean distance as the distance measure, answer the following questions:

a) Compute the matrix of pair-wise distances between the 5 points, where the (i, j)th entry in the matrix corresponds to the distance between point i and point j. **[5 points]**

Pair-wise Distances	Point 1	Point 2	Point 3	Point 4	Point 5
Point 1	0	3	5	$\sqrt{5}$	$\sqrt{85}$
Point 2	3	0	4	$\sqrt{20}$	$\sqrt{58}$
Point 3	5	4	0	$\sqrt{20}$	$\sqrt{18}$
Point 4	$\sqrt{5}=2.24$	$\sqrt{20}=4.47$	$\sqrt{20}$	0	$\sqrt{74}$
Point 5	$\sqrt{85}=9.22$	$\sqrt{58}=7.62$	$\sqrt{18}=4.24$	$\sqrt{74}=8.60$	0

b) Use the single link (MIN) hierarchical clustering technique for clustering these 5 points, and show the dendrogram of the clustering. The Y-axis of the dendrogram (height at which two clusters are merged) can be chosen to be the pair-wise distance between the two clusters. **[3 points]**





c) Use the complete link (MAX) hierarchical clustering technique for clustering these 5 points, and show the dendrogram of the clustering. The Y-axis of the dendrogram (height at which two clusters are merged) can be chosen to be the pair-wise distance between the two clusters. **[3 points]**