CPSC 425 **Segmentation and Classification** **18 (Term 2)**

Practice Questions

Solutions

Multiple Part True/False Questions.

For each question, indicate which of the statements, (A)–(D), are **true** and which are **false**? Note: Questions may have zero, one or multiple statements that are true.

Question 1. A Receiver Operating Characteristic (ROC) curve plots true positive rate (TPR) on the y-axis and false positive rate (FPR) on the x-axis. Which of the following statements about an ROC curve are **true**? Which are **false**?

(A) The diagonal represents random guessing.

(B) A good classifier lies near the upper left.

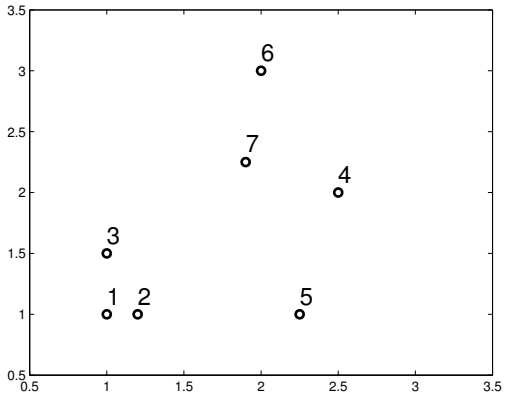
(C) An ROC curve is useful for tuning a given classifier.

(D) ROC curves are useful for comparing 2 classifiers.

Solution : (A) True, (B) True, (C) True, (D) True.

Short Answer Questions.

Question 2. For this question, the task is to construct dendrograms of the seven labelled points shown, as follows:



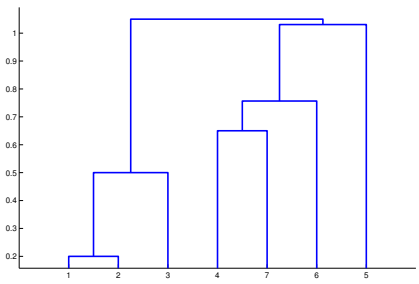
The table below gives the (Euclidean) distance between point i and point j. The table is, of course, symmetric since the distance between point i and point j is the same as the distance between point j and point i.

Pt	1	2	3	4	5	6	7
1	0	0.2000	0.5000	1.8028	1.2500	2.2361	1.5403
2	0.2000	0	0.5385	1.6401	1.0500	2.1541	1.4327
3	0.5000	0.5385	0	1.5811	1.3463	1.8028	1.1715
4	1.8028	1.6401	1.5811	0	1.0308	1.1180	0.6500
5	1.2500	1.0500	1.3463	1.0308	0	2.0156	1.2981
6	2.2361	2.1541	1.8028	1.1180	2.0156	0	0.7566
7	1.5403	1.4327	1.1715	0.6500	1.2981	0.7566	0

We start, as usual, by assigning each point as its own cluster. At each subsequent step, two clusters are merged according to the chosen point-cluster distance measure.

(a) Complete the following table (on the right) and draw the associated dendrogram (on the left) when the point-cluster distance measure is "single-link clustering," (Recall: In single-link clustering the distance is the distance between the two closest elements from each cluster).

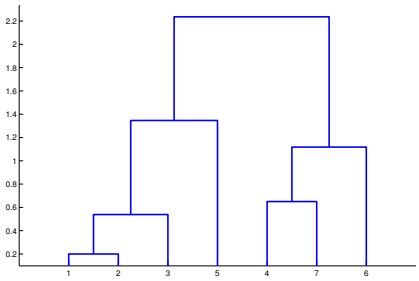
Solution:



Step	Clusters (set notation)
1	$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}$
2	$\{1,2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}$
3	$\{1,2,3\}, \{4\}, \{5\}, \{6\}, \{7\}$
4	$\{1,2,3\}, \{4,7\}, \{5\}, \{6\}$
5	$\{1,2,3\}, \{4,6,7\}, \{5\}$
6	$\{1,2,3\}, \{4,5,6,7\}$
7	$\{1,2,3,4,5,6,7\}$

(b) Complete the following table (on the right) and draw the associated dendrogram (on the left) when the point-cluster distance measure is "complete-link clustering," (Recall: In complete-link clustering the distance is the maximum distance between two elements from each cluster).

Solution:

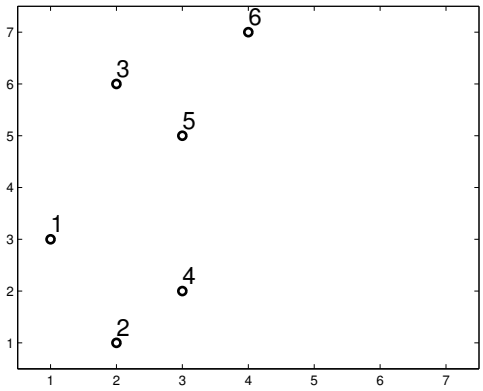


Step	Clusters (set notation)
1	$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}$
2	$\{1, 2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}$
3	$\{1, 2, 3\}, \{4\}, \{5\}, \{6\}, \{7\}$
4	$\{1, 2, 3\}, \{4, 7\}, \{5\}, \{6\}$
5	$\{1, 2, 3\}, \{4, 6, 7\}, \{5\}$
6	$\{1, 2, 3, 5\}, \{4, 6, 7\}$
7	$\{1, 2, 3, 4, 5, 6, 7\}$

Question 3. The K-means algorithm can converge to different solutions depending on which points are selected at random as the initial cluster centers. If we ran the algorithm 10 separate times with different random selections and wanted to select the best solution, how would we determine which solution is the best? Give your answer using one or two sentences.

Solution : The K-means algorithm minimizes the sum of the squared distances of points from the closest centers. We would just select the solution that found the lowest value of this sum.

Question 4. For this question, the task is to perform K-means clustering of the six labelled points at locations (1,3), (2,1), (2,6), (3,2), (3,5), (4,7), shown as follows:



We will attempt to find two clusters, A and B, and will start, as usual, by selecting two points at random as the initial cluster centers. Suppose the two random points initially selected are (1,3) for A and (3,5) for B.

(a) After the first iteration of the algorithm, what points will be assigned to cluster A? What points will be assigned to cluster B? What will be the new, adjusted locations of the cluster centers?

Solution:
Consider the following table of squared distances and the resulting classification (based on shortest distance):

	A:$(1,3)$	B:$(3,5)$	Class
$(1,3)$	0	$2^2+2^2=8$	A
$(2,1)$	$1^2+2^2=5$	$1^2+4^2=17$	A
$(2,6)$	$1^2+3^2=10$	$1^2+1^2=2$	B
$(3,2)$	$2^2+1^2=5$	$0+3^2=9$	A
$(3,5)$	$2^2+2^2=8$	0	B
$(4,7)$	$3^2+4^2=25$	$1^2+2^2=5$	B

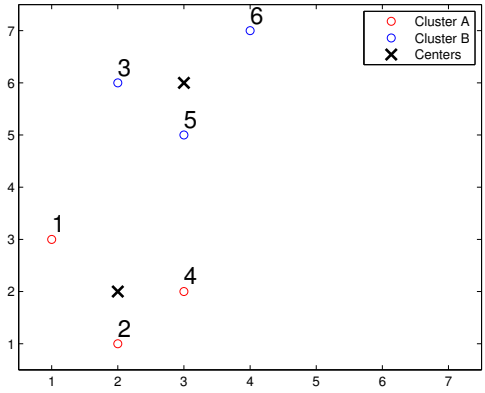
Thus, after the first iteration, points (1,3), (2,1) and (3,2) are assigned to the cluster A. Points (2,6), (3,5) and (4,7) are assigned to the cluster B.

The new cluster center for A is

$$\left(\frac{1+2+3}{3}, \frac{3+1+2}{3}\right) = (2, 2)$$

The new cluster center for B is

$$\left(\frac{2+3+4}{3}, \frac{6+5+7}{3}\right) = (3, 6)$$



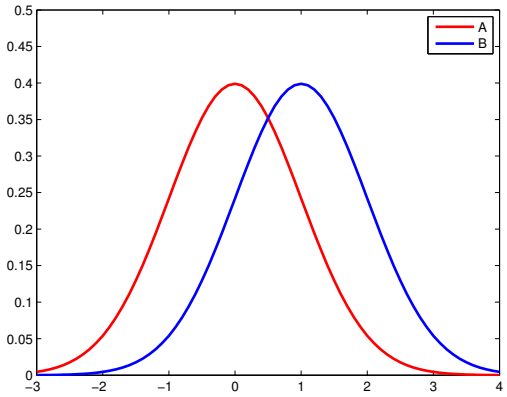
(b) Will subsequent iterations of the algorithm change the assignment of the points to clusters? (Briefly justify your answer).

Solution:
No. By inspection, the algorithm has converged in a single step. Any subsequent iterations of the algorithm will not change the assignment of points to clusters and therefore will not change the cluster centers further.

Question 5. We examine loss in a classifier. Assume we have two probability distributions, $P(A|x)$ and $P(B|x)$, both Gaussians with the same $\sigma = 1$, but A has mean 0 and B has mean 1.

(a) Draw the two distributions so that they are qualitatively correct. You don't have to be precise.

Solution:



(b) Where is the decision boundary determining the classifier deciding whether, given x, you have an A or a B? Draw it and identify the value of x.

Solution:
The decision boundary occurs at $x = 0.5$, the point at which the two curves intersect.

(c) Now, assume that we have a loss function L and that $L(A \rightarrow B) = 10$ and $L(B \rightarrow A) = 1$. The Bayes estimator incorporates the loss function to reflect the cost of errors. In which direction will the decision boundary move to reflect the cost of errors?

Solution:
The decision boundary moves to the right, (i.e., to $x > 0.5$)

(d) Assume we are testing to identify A. What is the term used in classification for the case where we classify something as B when it is in fact A (true positive, false positive, true negative, false negative)?

Solution:
false negative