**CITY UNIVERSITY OF HONG KONG**

---

Course code & title : EE4146 Data Engineering and Learning System

Session : Semester A 2021/22

Time allowed : 2 hours

- This paper has **Nine** pages (including this cove page)

---

**Instructions**:

Please make sure you follow all instructions from the University, ARRO, and EE. Please note the following:

1. This paper consists of **Eight** questions (A-H). The questions are ALL compulsory. Make sure that you attempt all of them. The total score is 100.
2. This is an **open-book open-notes exam.** Students can read the lecture notes and/or other materials available.

Answering this exam paper implies your acknowledgment of the Pledge for following the Rules on Academic Honesty:

"I pledge that the answers in this examination are my own and that I will not seek or obtain an unfair advantage in producing these answers. Specifically,

1. I will not plagiarize (copy without citation) from any source;
2. I will not communicate or attempt to communicate with any other person during the examination; neither will I give or attempt to give assistance to another student taking the examination; and
3. I will use only approved devices (e.g., calculators) and/or approved device models.
4. I understand that any act of academic dishonesty can lead to disciplinary action."

*I pledge to follow the Rules on Academic Honesty and understand that violations may lead to severe penalties.*

Name: _____

Signature: _____

Date : _____

Student ID: _____

**Question A Multiple choice questions (20%)**

1. Which of the following descriptions is/are right? (**BD**)

A. The larger the probability of an outcome, the more information it provides and vice-versa
B. Small entropy means "very predictable"
C. For a coin with probability P, the maximum entropy is 2.3219
D. For a coin with probability P, the maximum entropy is 1

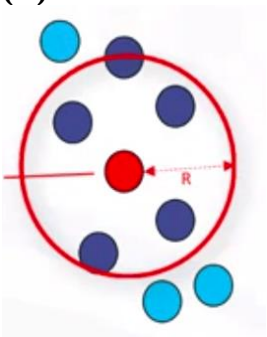
2. Which of the following statements about KMedoids algorithm are true? **(ABC)**

A. K-Medoids algorithm can determine spherical shaped clusters
B. Number of clusters to be determined must be specified
C. Less sensitive to noise data than KMeans
D. Suitable for large volume of data (Scalable)

3. Assume you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations: C1: {(2,2), (4,4), (6,6)}; C2: {(0,4), (4,0)}; C3: {(5,5), (9,9)}. What will be the Manhattan distance for observation (9, 9) from cluster centroid C1 in second iteration. **(A )**

A. 10
B. 5*sqrt(2)
C. 13*sqrt(2)
D. None of these

4. Given the Radius (R) as marked in the figure and the minpts = 6. Then the red point in the figure is a _____. **(A)**




A. Core point
B. Border point
C. Neither core nor border point
D. Can't say

5. Consider a dataset containing six one-dimensional points: {2, 4, 7, 9, 13, 14}. After three iterations of Hierarchical Agglomerative Clustering using Euclidean distance between points, we get the 3 clusters: C1 = {2, 4}, C2 = {7, 9} and C3 = {13, 14}. What is the distance between clusters C1 and C2 using Complete Linkage? **(B)**

A. 3
B. 7
C. 4
D. 6

6. Consider a dataset containing six one-dimensional points: {2, 4, 7, 9, 13, 14}. After three iterations of Hierarchical Agglomerative Clustering using Euclidean distance between points, we get the 3 clusters: C1 = {2, 4}, C2 = {7, 9} and C3 = {13, 14}. What is the distance between clusters C1 and C2 using average Linkage? **(D)**

A. 3
B. 7
C. 4
D. 5

7. You want to cluster this data into 2 clusters. Which of these algorithms would work well? **(AC)**

A. DBSCAN
B. K-means
C. Density based model
D. K-medoid

8. Which of the following algorithm is/are sensitive to outliers? **(AD)**

A. Kmeans
B. Kmedoids
C. CLARA
D. Fuzzy kmeans

9. Which of the following algorithm is/are data-preprocessing steps? **(ABCD)**

A. Aggregation
B. Sampling
C. PCA
D. Binarization

10. Assume you want to cluster 7 observations into 3 clusters using K-Medoid clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations: C1: {(2,3), (4,5), (6,10)}; C2: {(0,4), (4,0)}; C3: {(5,5), (9,9)}. What will be the Manhattan distance for observation (9, 9) from cluster centroid C1 in second iteration. **(C)**

A. 10
B. 7
C. 9
D. 8

**<u>Question B</u> True/False (10%) (randomly choose 10 of them)**

1. Linear Discriminant Analysis (LDA) finds a space of lower dimensionality by choosing the directions where the data varies most.
2. The single link agglomerative clustering algorithm groups two clusters based on the maximum distance between points in the two clusters.
3. K-Means will always give the same results regardless of the initialization of the centroids.
4. In the fuzzy clustering method, every data object is assigned to exactly one cluster.
5. The calculation of correlation is invariant to scaling and translation.
6. Different attributes can be mapped to the same set of values.
7. Sampling is the main technique employed for data aggregation.
8. K-medoids is a kind of divisive clustering.
9. Discretization is the process of converting a continuous attribute into an ordinal attribute
10. The more certain an outcome, the less information that it contains and vice-versa

**Solutions: F F F F T T F F T T**

**<u>Question C</u> (15%)**

(a) Give one advantage of hierarchical clustering over K-means clustering, and one advantage of K-means clustering over hierarchical clustering. (5%)

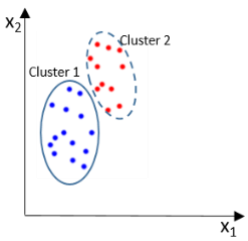
**Solutions**:

Some advantages of hierarchical clustering:
1. Don't need to know how many clusters you're after
2. Can cut hierarchy at any level to get any number of clusters
3. Easy to interpret hierarchy for particular applications
4. Can deal with long stringy data

Some advantages of K-means clustering:
1. Can be much faster than hierarchical clustering, depending on data
2. Nice theoretical framework
3. Can incorporate new data and reform clusters easily

(b) Please illustrate one of the data quality issues and illustrate the corresponding strategies to deal with that issue. (5%)

(c) Illustrate the method of LDA and summarize the corresponding steps. (5%)



**Question D (12%)**

For the following data, We aim to reduce the data into a single dimension representation. The first principal component (0.694, 0.720).

| data # | x | y |
|---|---|---|
| 1 | 5.51 | 5.35 |
| 2 | 20.82 | 24.03 |
| 3 | -0.77 | -0.57 |
| 4 | 19.30 | 19.38 |
| 5 | 14.24 | 12.77 |
| 6 | 9.74 | 9.68 |
| 7 | 11.59 | 12.06 |
| 8 | -6.08 | -5.22 |

(1).What is the representation (projected coordinate) for data #1 (x=5.51, y=5.35) in the first principal space?

(2).What are the xy coordinates in the original space reconstructed using this first principal representation for data #1 (x=5.51, y=5.35)?

(3).What is the representation (projected coordinate) for data #1 (x=5.51, y=5.35) in the second principal space?

(4).What is the reconstruction error if you use two principal components to represent original data?

(1) $\bar{x} = \frac{1}{8}(5.51+20.82+(-0.77)+19.30+14.24+9.74+11.59+(-6.08)) = 9.29$

$\bar{y} = \frac{1}{8}(5.35+24.03+(-0.57)+19.38+12.77+9.68+12.06+(-5.22)) = 9.69$

The representation for data #1 in the first principal space is: $0.694 \times (5.51-9.29) + 0.720 \times (5.35-9.69)$

$= -2.62 - 3.12$

$= -5.74$

(2) $newC = (0.694 \quad 0.720) \times \begin{matrix} 5.51 \\ 5.35 \end{matrix} = 0.694 \times 5.51 + 0.720 \times 5.35$

$= 3.82 + 3.85$

$= 7.67$

The xy coordinates in the original space are

$(0.694 \quad 0.720)^T newC = (0.694 \times 7.67 \quad 0.720 \times 7.67)$

$= (5.32 \quad 5.52)$

(3) The second principal component is $(0.720, -0.694)/(-0.720, 0.694)$

Take $(0.720, -0.694)$ for example.

The representation for data #1 in the 2nd principal space is $0.720 \times (5.51-9.29) + (-0.694) \times (5.35-9.69)$

$= -2.72 + 3.01 = -0.29$

Hence. the answer is $\pm 0.29$

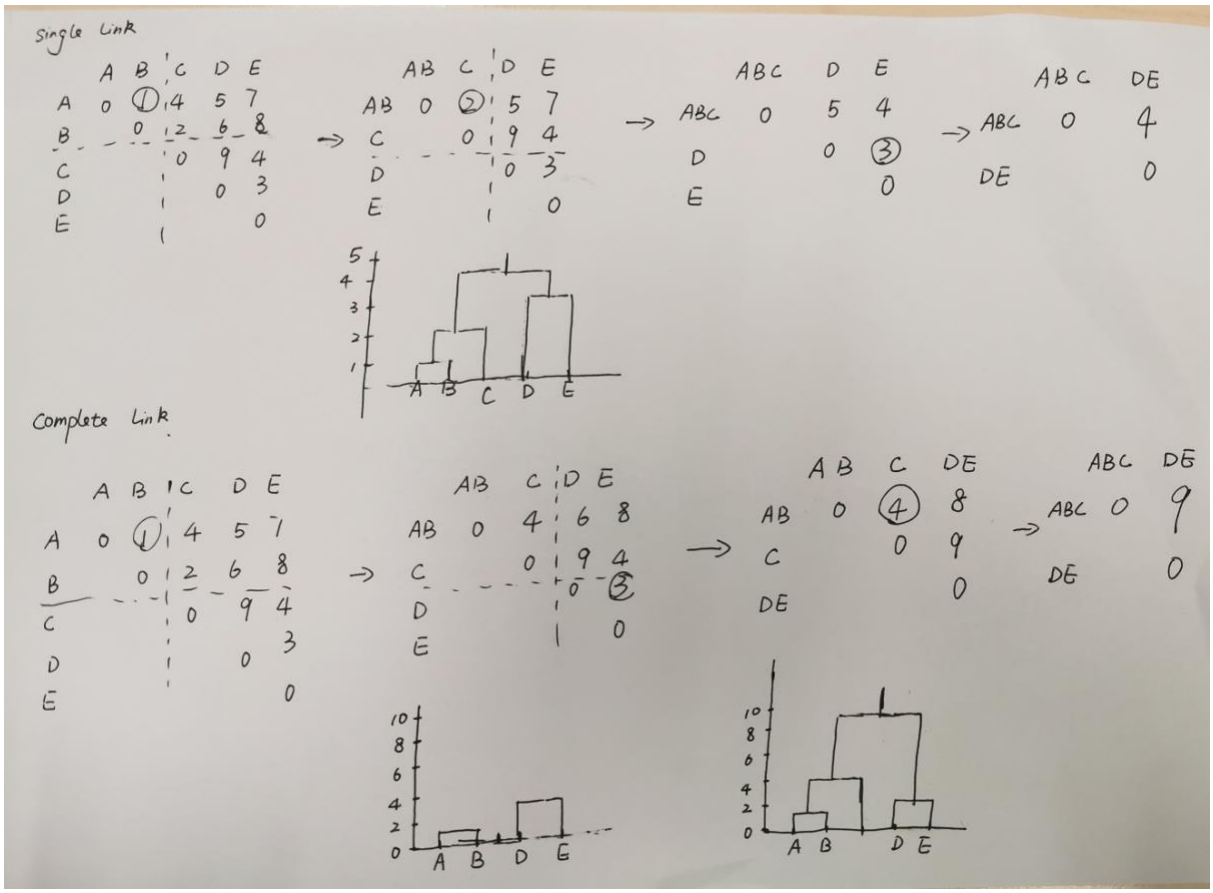
(4) 0.

**Question E (12%)**

Use complete link and single link hierarchical clustering to group the data described by the following distance matrix. Show the dendrograms.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | 1 | 4 | 5 | 7 |
| B | | 0 | 2 | 6 | 8 |
| C | | | 0 | 9 | 4 |
| D | | | | 0 | 3 |
| E | | | | | 0 |

**Solutions**



**Question F (8%)**

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). 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:

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|---|---|---|---|---|---|---|---|---|
| A1 | 0 | $\sqrt{25}$ | $\sqrt{36}$ | $\sqrt{13}$ | $\sqrt{50}$ | $\sqrt{52}$ | $\sqrt{65}$ | $\sqrt{5}$ |
| A2 | | 0 | $\sqrt{37}$ | $\sqrt{18}$ | $\sqrt{25}$ | $\sqrt{17}$ | $\sqrt{10}$ | $\sqrt{20}$ |
| A3 | | | 0 | $\sqrt{25}$ | $\sqrt{2}$ | $\sqrt{2}$ | $\sqrt{53}$ | $\sqrt{41}$ |
| A4 | | | | 0 | $\sqrt{13}$ | $\sqrt{17}$ | $\sqrt{52}$ | $\sqrt{2}$ |
| A5 | | | | | 0 | $\sqrt{2}$ | $\sqrt{45}$ | $\sqrt{25}$ |
| A6 | | | | | | 0 | $\sqrt{29}$ | $\sqrt{29}$ |
| A7 | | | | | | | 0 | $\sqrt{58}$ |
| A8 | | | | | | | | 0 |

a) The new clusters (i.e. the examples belonging to each cluster)
b) The centers of the new clusters

**Solutions:**

a)

d(a,b) denotes the Eucledian distance between a and b. It is obtained directly from the distance matrix or calculated as follows: $d(a,b)=\text{sqrt}((x_b-x_a)^2+(y_b-y_a)^2))$
seed1=A1=(2,10), seed2=A4=(5,8), seed3=A7=(1,2)

epoch1 – start:

A1:
d(A1, seed1)=0 as A1 is seed1
d(A1, seed2)= $\sqrt{13}$ >0
d(A1, seed3)= $\sqrt{65}$ >0
➔A1 ∈ cluster1

A2:
d(A2,seed1)= $\sqrt{25}$ = 5
d(A2, seed2)= $\sqrt{18}$ = 4.24
d(A2, seed3)= $\sqrt{10}$ = 3.16 ← smaller
➔ A2 ∈ cluster3

A3:
d(A3, seed1)= $\sqrt{36}$ = 6
d(A3, seed2)= $\sqrt{25}$ = 5 ← smaller
d(A3, seed3)= $\sqrt{53}$ = 7.28
➔ A3 ∈ cluster2

A4:
d(A4, seed1)= $\sqrt{13}$
d(A4, seed2)=0 as A4 is seed2
d(A4, seed3)= $\sqrt{52}$ >0
➔ A4 ∈ cluster2

A5:
d(A5, seed1)= $\sqrt{50}$ = 7.07
d(A5, seed2)= $\sqrt{13}$ = 3.60 ← smaller
d(A5, seed3)= $\sqrt{45}$ = 6.70
➔ A5 ∈ cluster2

A6:
d(A6, seed1)= $\sqrt{52}$ = 7.21
d(A6, seed2)= $\sqrt{17}$ = 4.12 ← smaller
d(A6, seed3)= $\sqrt{29}$ = 5.38
➔ A6 ∈ cluster2

A7:
d(A7, seed1)= $\sqrt{65}$ >0
d(A7, seed2)= $\sqrt{52}$ >0
d(A7, seed3)=0 as A7 is seed3
➔ A7 ∈ cluster3

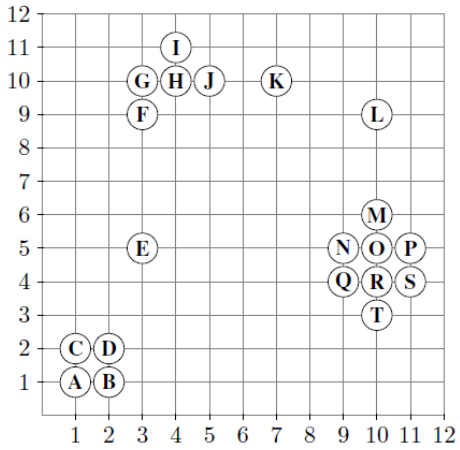
A8:
d(A8, seed1)= $\sqrt{5}$
d(A8, seed2)= $\sqrt{2}$ ← smaller
d(A8, seed3)= $\sqrt{58}$
➔ A8 ∈ cluster2

end of epoch1

new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

b) centers of the new clusters:
C1= (2, 10), C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

**<u>Question G</u> (15%)**
Given the following data set:




As distance function, use Manhattan Distance. Compute DBSCAN and indicate which points are core points, border points and noise points with the following parameter settings:

1) Radius Epsilon = 1.1 and minPts = 2
2) Radius Epsilon = 1.1 and minPts = 3
3) Radius Epsilon = 2.1 and minPts = 4

**Solutions:**

1) All points are core points, no border points, noise points{ EKL}
2) Core points{ABCD,GH, NOPQRS}, Border points{IJF, MT}, noise points{ EKL}
3) Core points{ABCD,GHIJ,MNOPQRST}, Border points{KF}, noise points{EL}

**Question H (8%)**

For given dataset x=[2,3,4,8,10], we choose the initial cluster center, c1=4, c2=10, please illustrate the first step with fuzzing clustering models.

*For node* 1

$$w_{11} = \frac{(2-10)^2}{(2-10)^2+(2-4)^2} = \frac{64}{68} = 0.9412$$

$$w_{12} = \frac{(2-4)^2}{(2-10)^2+(2-4)^2} = \frac{4}{68} = 0.0588$$

*or* $w_{12} = 1 - w_{11}$

*For node* 2

$$w_{21} = \frac{(3-10)^2}{(3-10)^2+(3-4)^2} = \frac{49}{50} = 0.98$$

$$w_{22} = \frac{(3-4)^2}{(3-10)^2+(3-4)^2} = \frac{1}{50} = 0.02$$

*For node* 3

$$w_{31} = \frac{(4-10)^2}{(4-10)^2+(4-4)^2} = 1$$

$$w_{32} = \frac{(4-4)^2}{(4-10)^2+(4-4)^2} = 0$$

*For node* 4

$$w_{41} = \frac{(8-10)^2}{(8-10)^2+(8-4)^2} = \frac{4}{20} = 0.2$$

$$w_{42} = \frac{(8-4)^2}{(8-10)^2+(8-4)^2} = \frac{16}{20} = 0.8$$

*For node* 5

$$w_{51} = \frac{(10-10)^2}{(4-10)^2+(4-4)^2} = 0$$

$$w_{52} = 1$$

$$c1 = \frac{0.9421^2 \times 2 + 0.98^2 \times 3 + 1^2 \times 4 + 0.2^2 \times 8 + 0^2 \times 10}{(0.9421)^2 + 0.98^2 + 1^2 + 0.2^2 + 0^2}$$

$$= 2.875$$

$$c2 = \frac{0.0588^2 \times 2 + 0.02^2 \times 3 + 0^2 \times 4 + 0.8^2 \times 8 + 1^2 \times 10}{0.0588^2 + 0.02^2 + 0^2 + 0.8^2 + 1^2}$$

$$= 9.187$$

- END -