**Chương 6 - Clustering**

[**6.0 DANH SÁCH CÁC CÂU CHƯA LÀM 3**](#_iv17sv5qb71v)

[**6.1 Question 3**](#_pgb6iasvnx2r)

[**6.1 Answer 3**](#_bdh6qqrowm7s)

[[Trâm] 3](#_5x16k1qtzsm7)

[[Author 2] 4](#_a6fx7s1stmmy)

[[Author 3] 4](#_5wqdm5i4gfls)

[**6.2 Question 4**](#_ujnjhupx8s9z)

[**6.2 Answer 4**](#_ipy1tbstvype)

[[HX Long] 4](#_48bo76723njy)

[[Author 2] 4](#_95x3ejoi1exq)

[**6.3 Question 4**](#_t8tjbpfz3g2y)

[**6.3 Answer 5**](#_vkxlt2ktya51)

[[Nghia Pham] 5](#_6t9m4xhb7ka9)

[[Author 2] 5](#_shapp1reiokg)

[**6.4. Question 5**](#_9k222vlqzx30)

[**6.4 Answer 5**](#_3t4c4iq9ywh)

[[HX Long] 5](#_zigtpptwc3td)

[[Author 2] 5](#_mythugkseza2)

[**6.5 Question ⇒ Chầy chưa dạy tới, sẵn làm luôn 5**](#_1kpygz1kepqv)

[**6.5 Answer 5**](#_76xlr6x6v1sb)

[[Thành] 5](#_a7ome8bdg7wp)

[[Author 2] 6](#_nj7t6nfprkm)

[**6.6 Question ⇒ Chưa học 6**](#_dbgj89eed27t)

[**6.6 Answer 6**](#_sp5ynzxebwb8)

[[Author 1] 6](#_9zjcios2ewnm)

[[Author 2] 6](#_grp7er1q4vx5)

[**6.7 Question 6**](#_tlldabj1lcl9)

[**6.7 Answer 6**](#_c7tyglp4rho7)

[[Thành]: 6](#_icylbatxowi)

[[Toàn] checked - ok 6](#_maozseqli933)

[**6.8 Question 6**](#_ivdq23vwq8w7)

[**6.8 Answer 6**](#_7gxu77stqoc2)

[[Thành]: 6](#_4x707q1orgse)

[[Author 2] 7](#_tlwrg5s1bj9s)

[**6.9. Question 7**](#_vxi3hqqn9uf)

[**6.9 Answer 7**](#_2psc877e2w1i)

[[Thành]: 7](#_9nquekwlafdc)

[[Author 2] 7](#_v7kt34hifych)

[**6.10 Question 7**](#_h6wzz6rfg93k)

[**6.10 Answer 7**](#_535jtndw6qk3)

[[Thành] ⇒ dựa vào giải thuật k-mean nên kq sẽ phụ thuộc vào k điểm chọn ban đầu của người giải nên kq sẽ ó thể khác nhau 7](#_8jy3xuul1jpy)

[[Author 2] 8](#_jhsuoyhyp4qq)

[**6.11 Question⇒ chưa học tới 8**](#_ssrj4odellsx)

[**6.11 Answer 9**](#_3zjf6frw8yuu)

[[Author 1] 9](#_9iab326r8roo)

[[Author 2] 9](#_1lffxsr30ciq)

[**6.12 Question ⇒ chưa học tới 9**](#_fexykpj80vfh)

[**6.12 Answer 9**](#_t4mrzdktbdf4)

[[Author 1] 9](#_y5w4amci8pxn)

[[Author 2] 9](#_e6owclr45qq9)

[**6.13 Question⇒ chưa học tới 9**](#_9znvct88l4e)

[**6.13 Answer 9**](#_b2tttg62yhqi)

[[Author 1] 9](#_2qmhyw6oc6u3)

[[Author 2] 9](#_9dgt4fsd5cmn)

[**6.14 Question ⇒ chưa học tới 9**](#_z2im7jiibqx5)

[**6.14 Answer 10**](#_y4cfoace5bws)

[[Author 1] 10](#_iu14x013azzo)

[[Author 2] 10](#_10zjig6lonk6)

[**6.15. Question 10**](#_muc0rpjoxwm6)

[**6.15 Answer 10**](#_hty2jm63jxyg)

[[Thành]: 10](#_mu1wcgs0da71)

[[Toàn] - checked: ok 10](#_ghtqybzfxun9)

[**6.16. Question 10**](#_4ymohbhowm6i)

[**6.16 Answer 10**](#_5zgouu86prgz)

[[Thành]: 10](#_ur0t6q9tkelj)

[[Toàn] - checked: ok 10](#_biavgm7900o7)

[**6.17. Question 10**](#_nqtudy7ymrfg)

[**6.17 Answer 11**](#_kk0c32cmkjwt)

[[Thành]: 11](#_mxpzs5hhlszm)

[[Toàn] checked: ok 11](#_2wwptttf02f4)

[**6.18 Question 11**](#_4z1uqzphpl9k)

[**6.18 Answer 11**](#_dxbisltlywwy)

[[Thành]: 11](#_ia82f1z05ch4)

[[Author 2] 11](#_ee3iilvuogub)

[**6.19 Question⇒ chưa học tới 11**](#_qduh9oa9n7k4)

[**6.19 Answer 11**](#_pux6lc1smdb8)

[[Author 1] 11](#_i4vdq38a2mtr)

[[Author 2] 11](#_nxbokawfjv0l)

[**6.20 Question⇒ chưa học tới 11**](#_a9lltphw187v)

[**6.20 Answer 12**](#_fck0yfa6iq9g)

[[Author 1] 12](#_98hm5mtvme1a)

[[Author 2] 12](#_wpfrzlfzdm3k)

[**6.21 Question 12**](#_ftwz6xa90lkw)

[**6.21 Answer 12**](#_s3mnm9nbfgp)

[[Thành] 12](#_q1ym7ohwlkek)

[[Author 2] 12](#_fcsimz4o085a)

—------------------------------------------

### 6.0 DANH SÁCH CÁC CÂU CHƯA LÀM

[**6.5 Question ⇒ Chầy chưa dạy tới, sẵn làm luôn**](#_1kpygz1kepqv)

[**6.6 Question ⇒ Chưa học**](#_dbgj89eed27t)

[**6.11 Question⇒ chưa học tới**](#_ssrj4odellsx)

[**6.12 Question ⇒ chưa học tới**](#_fexykpj80vfh)

[**6.13 Question⇒ chưa học tới**](#_9znvct88l4e)

[**6.14 Question ⇒ chưa học tới**](#_z2im7jiibqx5)

[**6.19 Question⇒ chưa học tới**](#_qduh9oa9n7k4)

[**6.20 Question⇒ chưa học tới**](#_a9lltphw187v)

—------------------------------------------

### **6.1 Question**

State the difference between supervised learning (classification) and unsupervised learning (clustering).

*Nêu sự khác biệt giữa học có giám sát (phân loại) và học không giám sát (phân cụm).*

### 6.1 Answer

##### **[Trâm]**

**Học có giám sát và học không giám sát** là hai loại kỹ thuật học máy khác nhau và chúng khác nhau về bản chất của dữ liệu được sử dụng trong đào tạo và mục đích của quá trình học.

**Học có giám sát** là một loại máy học trong đó mô hình được đào tạo trên dữ liệu được gắn nhãn, có nghĩa là mỗi điểm dữ liệu trong tập huấn luyện có một nhãn hoặc biến đầu ra tương ứng. Mục tiêu của học có giám sát là học một chức năng ánh xạ đầu vào thành đầu ra một cách chính xác. Hai loại học tập có giám sát chính là phân loại, trong đó mô hình được đào tạo để dự đoán nhãn phân loại và hồi quy, trong đó mô hình được đào tạo để dự đoán đầu ra liên tục.

**Học không giám sát** là một loại học máy trong đó mô hình được đào tạo trên dữ liệu không được gắn nhãn, có nghĩa là tập huấn luyện không có bất kỳ biến đầu ra tương ứng nào. Mục tiêu của học không giám sát là tìm các mẫu hoặc cấu trúc ẩn trong dữ liệu và loại học không giám sát chính là phân cụm, trong đó mô hình được đào tạo để nhóm các điểm dữ liệu tương tự lại với nhau thành các cụm dựa trên điểm giống hoặc khác nhau của chúng.

**Tóm lại, sự khác biệt chính** là sự hiện diện hay vắng mặt của dữ liệu được gán nhãn trong tập huấn luyện và bản chất của mục tiêu học tập. Học có giám sát được sử dụng cho các nhiệm vụ dự đoán trong đó mô hình được đào tạo để dự đoán một biến đầu ra cụ thể, trong khi học không giám sát được sử dụng để phân tích dữ liệu khám phá và khám phá các mẫu trong dữ liệu mà không cần lưu ý đến một biến kết quả cụ thể.

##### **[Author 2]**

##### **[Author 3]**

### 6.2 Question

Consider the following 10 patterns:

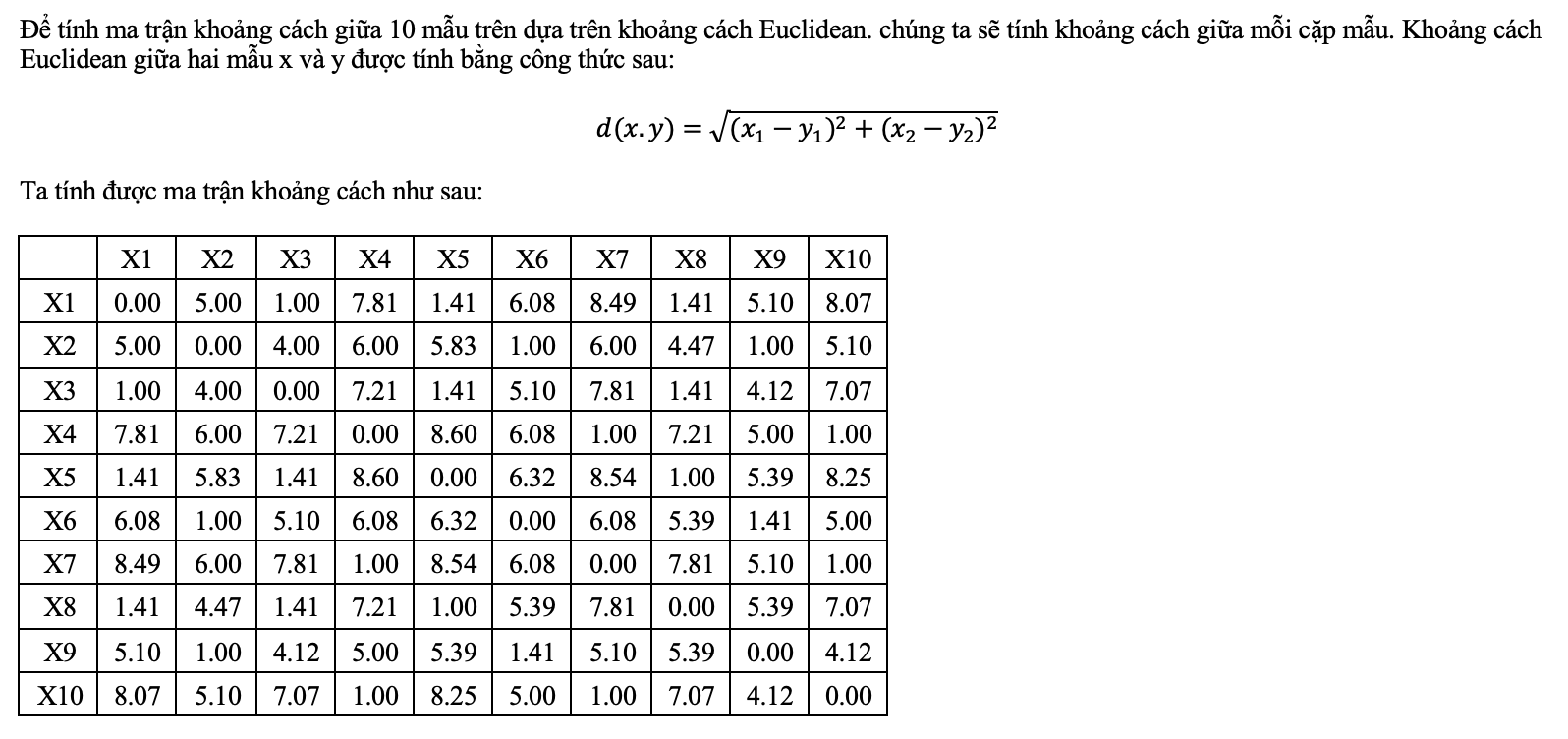
X1 = (1, 1), X2 = (6, 1), X3 = (2, 1), X4 = (6, 7), X5 = (1, 2), X6 = (7, 1), X7 = (7, 7), X8 = (2, 2), X1 = (6, 2), X10 = (7, 6)

Obtain the distance matrix using the Euclidean distance as the distance between two patterns.

*Lấy ma trận khoảng cách sử dụng khoảng cách Euclide là khoảng cách giữa hai mẫu.*

### 6.2 Answer

##### **[HX Long]**



##### **[Author 2]**

### 6.3 Question

If there is a set of *n* patterns and it is required to cluster these patterns to form two clusters, how many such partitions will there be?

*Nếu có một tập hợp n mẫu và cần phải phân cụm các mẫu này để tạo thành hai cụm, thì sẽ có bao nhiêu phân vùng như vậy?*

### 6.3 Answer

##### **[Nghia Pham]**

2^n

Vì mỗi pattern có thể được gán vào 1 trong 2 cluster.

Ref: <https://math.stackexchange.com/a/3671782>

##### **[Author 2]**

1 patterns có thể có 1 trong 2 lựa chọn vậy nên : 2^n

### 6.4. Question

Given the cluster of 5 patterns:

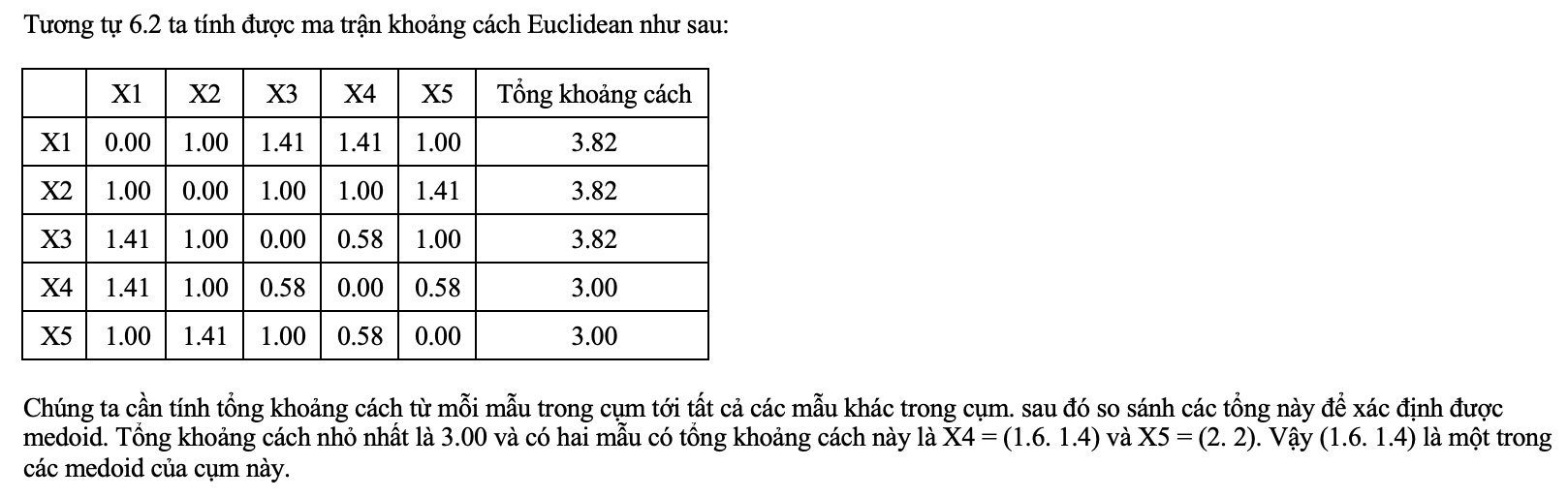
X1 = (1, 1), X2 = (1, 2), X3 = (2, 1), X4 = (1.6, 1.4), X5 = (2, 2)

Show that the medoid of the cluster is (1.6, 1.4).

*Chứng tỏ rằng trung điểm của cụm là (1.6, 1.4).*

### 6.4 Answer

##### **[HX Long]**

****

##### **[Author 2]**

### **6.5 Question** ⇒ Chầy chưa dạy tới, sẵn làm luôn

In agglomerative hierarchical clustering, among the current clusters, how to select the most suitable pair of clusters to be merged?

*Trong phân cụm theo thứ bậc tập hợp, trong số các cụm hiện tại, làm thế nào để chọn ra cặp cụm phù hợp nhất để hợp nhất?*

### 6.5 Answer

##### **[Thành]**

* In the single-link algorithm: **the minimum** of the distances d(X, Y), where X C1 and Y C2.   
  *Trong thuật toán liên kết đơn: khoảng cách nhỏ nhất d(X, Y), trong đó X C1 và Y C2.*
* In the complete-link algorithm: **the maximum** of the distances d(X, Y), where X C1 and Y C2.  
  *Trong thuật toán liên kết đầy đủ: khoảng cách lớn nhất d(X, Y), trong đó X C1 và Y C2.*

**Ref: Slide 18 Chap 6**

##### **[Author 2]**

### **6.6 Question** ⇒ Chưa học

In divisive hierarchical clustering,

a. How to find the best way to split a cluster into two clusters.

b. How to select among the current clusters the most suitable cluster to be split.

*Trong phân cụm phân cấp phân chia,*

*a. Cách tìm cách tốt nhất để tách một cụm thành hai cụm.*

*b. Làm thế nào để chọn trong số các cụm hiện tại cụm phù hợp nhất để tách.*

### 6.6 Answer

##### **[Author 1]**

##### **[Author 2]**

### **6.7 Question**

State the computational complexity of k-means algorithm.

*Nêu độ phức tạp tính toán của thuật toán k-means.*

### 6.7 Answer

##### **[Thành]:**

**The time complexity** of the algorithm is **O(nkdl)**, where n is the number of points in data set, k is number of clusters, l is the number of iterations and d is the dimensionality.

**The space requirement** is **O(kd)**. These features make the algorithm very attractive.

**Ref: Slide 13 Chap 6**

##### **[Toàn] checked - ok**

### **6.8 Question**

State the strong points and weak points of k-means algorithm.

*Nêu điểm mạnh và điểm yếu của thuật toán k-means.*

### 6.8 Answer

##### **[Thành]:**

*Strong points*

1. it’s fast, simple, and surprisingly flexible if you pre-process your data and engineer useful features.
2. It is easy to implement and identify unknown groups of data from complex data sets.

*Weak points*

1. Users have to determine k, the number of clusters and initial centers.
2. K-Means can not handle non-spherical clusters.
3. K-Means is sensitive with data that contains outliers.
4. K-Means is restricted to data for which there is a notion of centroid.

##### **[Author 2]**

### **6.9**. **Question**

State the computational complexity of agglomerative hierarchical clustering and divisive hierarchical clustering.

*Nêu độ phức tạp tính toán của phân cụm theo thứ bậc tổng hợp và phân cụm theo thứ bậc phân chia.*

### 6.9 Answer

##### **[Thành]:**

* agglomerative algorithms **O(n2)** in time and space complexity
* divisive algorithms require exponential time: **O(n!)**

**Ref: Slide 30 Chap 6**

##### **[Author 2]**

### **6.10 Question**

Consider the two dimensional data set given below:

(1, 1), (1, 2), (2, 1), (2, 1.5), (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5, 2), (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5)

Use the k-means algorithm to cluster these patterns with k = 3.

*Hãy xem xét tập dữ liệu hai chiều được đưa ra dưới đây:*

*(1, 1), (1, 2), (2, 1), (2, 1.5), (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5 , 2), (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5)*

*Sử dụng thuật toán k-means để nhóm các mẫu này với k = 3.*

### 6.10 Answer

##### **[Thành] ⇒** *dựa vào giải thuật k-mean nên kq sẽ phụ thuộc vào k điểm chọn ban đầu của người giải nên kq sẽ ó thể khác nhau*

**Step 1**: Randomly select 3 points as the initial centroids: (1, 1), (2, 1), and (4, 1.5).

**Step 2**: Assign each observation to the nearest centroid:

Cluster 1: (1, 1), (1, 2), (2, 1), (2, 1.5)

Cluster 2: (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5, 2), (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5)

Cluster 3: None

The distance between each observation and each centroid is calculated using the Euclidean distance formula:

d = sqrt((x2 - x1)^2 + (y2 - y1)^2)

For example, the distance between (1, 1) and (1, 1) is:

d = sqrt((1 - 1)^2 + (1 - 1)^2) = 0

Similarly, the distance between (1, 1) and (2, 1) is:

d = sqrt((2 - 1)^2 + (1 - 1)^2) = 1

We can use this formula to calculate the distance between each observation and each centroid. Then, we assign each observation to the nearest centroid.

**Step 3**: Recalculate the centroids of each cluster:

Cluster 1: (1.5, 1.5)

Cluster 2: (4.1, 3.4)

Cluster 3: None

The new centroids are calculated by taking the mean of all the observations in each cluster.

**Step 4**: Repeat steps 2 and 3 until the centroids no longer move.

Cluster 1: (1.5, 1.5)

Cluster 2: (4.1, 3.4)

Cluster 3: None

Cluster 1: (1.5, 1.5)

Cluster 2: (4.1, 3.4)

Cluster 3: None

Since the centroids no longer move, we can stop the algorithm. Therefore, the dataset can be clustered into 2 clusters: Cluster 1 with (1, 1), (1, 2), (2, 1), (2, 1.5) and Cluster 2 with (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5, 2), (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5).

##### **[Author 2]**

### **6.11 Question**⇒ chưa học tới

Given a set of 2-dimensional patterns: X1 = (1, 3), X2 = (1.5, 3.2), X3 = (1.3, 2.8), X4 =(3,1). Let apply k-means with k=2 to cluster this dataset. Assume that at a certain iteration, the dataset is grouped into 2 clusters as follows : the first cluster consists of X1, and the second cluster consists of X2, X3, X4. Let perform the next iteration which consists of two steps : recalculating the centroids and assigning the patterns to the clussters.

Note: Euclidean distance is used in the k-means.

*Cho một tập hợp các mẫu 2 chiều: X1 = (1, 3), X2 = (1,5, 3,2), X3 = (1,3, 2,8), X4 =(3,1). Hãy áp dụng phương tiện k với k=2 để phân cụm tập dữ liệu này. Giả sử rằng tại một lần lặp nhất định, tập dữ liệu được nhóm thành 2 cụm như sau: cụm thứ nhất gồm X1 và cụm thứ hai gồm X2, X3, X4. Hãy thực hiện lần lặp tiếp theo bao gồm hai các bước: tính toán lại các trọng tâm và gán các mẫu cho các cụm.*

***Lưu ý:*** *Khoảng cách Euclide được sử dụng trong k-mean.*

### 6.11 Answer

##### **[Author 1]**

##### **[Author 2]**

### **6.12 Question** ⇒ chưa học tới

Consider the two dimensional data set given below:

(1, 1), (1, 2), (2, 1), (2, 1.5), (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5, 2), (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5)

Use agglomerative hierarchical clustering with single-link and agglomerative hierarchical clustering with complete-link to cluster the dataset into 4 clusters.

*Hãy xem xét tập dữ liệu hai chiều được đưa ra dưới đây:*

*(1, 1), (1, 2), (2, 1), (2, 1.5), (3, 2), (4, 1.5), (4, 2), (5, 1.5), (4.5 , 2), (4, 4), (4.5, 4), (4.5, 5), (4, 5), (5, 5)*

*Sử dụng phân cụm theo cấp bậc tổng hợp với liên kết đơn và phân cụm theo cấp bậc tổng hợp với liên kết đầy đủ để phân cụm tập dữ liệu thành 4 cụm.*

### 6.12 Answer

##### **[Author 1]**

##### **[Author 2]**

### **6.13 Question**⇒ chưa học tới

State the similarity between two clustering algorithms: k-means and fuzzy-c-means.

*Nêu sự giống nhau giữa hai thuật toán phân cụm: k-mean và Fuzzy-c-mean.*

### 6.13 Answer

##### **[Author 1]**

##### **[Author 2]**

### 6.14 Question ⇒ chưa học tới

Given a set of 2-dimensional patterns: X1 = (1, 6), X2 = (2,5), X3 = (3, 8), X4 =(4,4). X5 = (5, 7), X6 =(6,9). Let apply fuzzy-c-means with k = 2 to cluster this dataset. Assume that at a certain iteration, the dataset is grouped into 2 clusters with the membership weights as follows.

|  | X1 | X2 | X3 | X4 | X5 | X6 |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster c1 | 0.8 | 0.9 | 0.7 | 0.3 | 0.5 | 0.2 |
| Cluster c2 | 0.2 | 0.1 | 0.3 | 0.7 | 0.5 | 0.8 |

Let perform the next iteration which consists of two steps : recalculating the centroids and assigning the membership weights for each pattern.

Note: Euclidean distance is used in the fuzzy-c-means.

*Cho một tập các mẫu 2 chiều: X1 = (1, 6), X2 = (2,5), X3 = (3, 8), X4 =(4,4). X5 = (5, 7), X6 =(6,9). Hãy áp dụng Fuzzy-c-means với k = 2 để phân cụm tập dữ liệu này. Giả sử rằng tại một lần lặp nhất định, tập dữ liệu được nhóm thành 2 cụm với trọng số thành viên như sau.*

*…*

*Hãy thực hiện bước lặp tiếp theo bao gồm hai bước: tính toán lại trọng tâm và gán trọng số thành viên cho mỗi mẫu.*

*Lưu ý: Khoảng cách Euclide được sử dụng trong phương pháp fuzzy-c-means.*

### 6.14 Answer

##### **[Author 1]**

##### **[Author 2]**

### **6.15**. **Question**

(True/False) K-Means can generate clusters with arbitrary shapes.

*K-Means có thể tạo các cụm có hình dạng tùy ý.*

### 6.15 Answer

##### **[Thành]:**

False. K-means is a centroid-based clustering algorithm that partitions data into K clusters. It requires the number of clusters to be specified beforehand and can only generate clusters with spherical shapes.

*SAI. K-means là một thuật toán phân cụm dựa trên trọng tâm phân vùng dữ liệu thành các cụm K. Nó yêu cầu số lượng cụm phải được chỉ định trước và chỉ có thể tạo các cụm có hình cầu.*

##### **[Toàn] - checked: ok**

### **6.16**. **Question**

(True/False) DBSCAN can generate clusters with arbitrary shapes.

*DBSCAN có thể tạo các cụm có hình dạng tùy ý*

### 6.16 Answer

##### **[Thành]:**

True. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that can generate clusters with arbitrary shapes. Unlike K-means, DBSCAN does not require the number of clusters to be specified beforehand and can identify noise points as well.

*ĐÚNG VẬY. DBSCAN (Phân cụm không gian dựa trên mật độ của các ứng dụng có tiếng ồn) là một thuật toán phân cụm dựa trên mật độ có thể tạo các cụm có hình dạng tùy ý. Không giống như K-mean, DBSCAN không yêu cầu chỉ định trước số lượng cụm và cũng có thể xác định các điểm nhiễu.*

##### **[Toàn] - checked: ok**

### **6.17**. **Question**

(True/False) K-Means can generate clusters only with spherical shapes.

*K-Means chỉ có thể tạo các cụm có hình cầu.*

### 6.17 Answer

##### **[Thành]:**

True. K-means is a centroid-based clustering algorithm that partitions data into K clusters. It requires the number of clusters to be specified beforehand and can only generate clusters with spherical shapes.

##### **[Toàn] checked: ok**

### **6.18 Question**

Give an example in which clustering can be used as a preprocessing step for an another data classification task.

*Cho một ví dụ trong đó phân cụm có thể được sử dụng như một bước tiền xử lý cho một nhiệm vụ phân loại dữ liệu khác.*

### 6.18 Answer

##### **[Thành]:**

An example of clustering as a preprocessing step for document classification is to group similar news articles together. For example, news articles can be clustered based on their content, such as the topic, keywords, and sentiment. Then, a classifier can be trained on each cluster to classify the news articles in that cluster. This can help improve the accuracy of the classifier as it can focus on a smaller subset of the data that is more similar to each other[1](https://www.sciencedirect.com/topics/computer-science/document-clustering).

*Một ví dụ về phân cụm như một bước tiền xử lý để phân loại tài liệu là nhóm các bài báo tương tự lại với nhau. Ví dụ: các bài báo có thể được nhóm lại dựa trên nội dung của chúng, chẳng hạn như chủ đề, từ khóa và tình cảm. Sau đó, một bộ phân loại có thể được đào tạo trên mỗi cụm để phân loại các bài báo trong cụm đó. Điều này có thể giúp cải thiện độ chính xác của bộ phân loại vì nó có thể tập trung vào một tập hợp con nhỏ hơn của dữ liệu giống nhau hơn1.*

##### **[Author 2]**

### 6.19 Question⇒ chưa học tới

Explain the term *incremental clustering*. State the weak point of the Leader algorithm for incremental clustering.

*Giải thích thuật ngữ phân cụm gia tăng. Nêu điểm yếu của thuật toán Leader để phân cụm tăng dần.*

### 6.19 Answer

##### **[Author 1]**

##### **[Author 2]**

### 6.20 Question⇒ chưa học tới

Given a set of 2-dimensional patterns :

A = (1, 1), B = (1, 2), C = (2, 2), D = (6, 2), E = (7, 2), F = (6, 6), G = (7, 6)

Let apply Leader algorithm to cluster the dataset. Assume that the data will be processed in the order A, B, C, D, E, F and G, and the user specified threshold *T* be 3.

*Cho một tập hợp các mẫu 2 chiều:*

*A = (1, 1), B = (1, 2), C = (2, 2), D = (6, 2), E = (7, 2), F = (6, 6), G = (7, 6)*

*Hãy áp dụng thuật toán Leader để phân cụm tập dữ liệu. Giả sử rằng dữ liệu sẽ được xử lý theo thứ tự A, B, C, D, E, F và G và ngưỡng T do người dùng chỉ định là 3.*

### 6.20 Answer

##### **[Author 1]**

##### **[Author 2]**

### 6.21 Question

How to evaluate clustering quality based on objective function.

*Cách đánh giá chất lượng phân cụm dựa trên hàm mục tiêu.*

### 6.21 Answer

##### **[Thành]**

Typical objective functions in clustering formalize the goal of attaining high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents from different clusters are dissimilar). This is an internal criterion for the quality of a clustering. But good scores on an objective function do not necessarily mean that the clustering is useful for a particular application. Therefore, it is important to evaluate the quality of a clustering based on both internal and external criteria.

Internal criteria are based on the data alone and include measures such as the sum of squared errors (SSE), the sum of squared deviations from the cluster mean (SSM), and the ratio of the between-cluster variance to the within-cluster variance (F-ratio). These measures are used to evaluate the quality of a clustering based on how well the data points fit into the clusters.

External criteria, on the other hand, are based on external information such as class labels or expert knowledge. These criteria include measures such as the adjusted Rand index (ARI), the normalized mutual information (NMI), and the F-measure. These measures are used to evaluate the quality of a clustering based on how well the clusters correspond to the external information.

*Các hàm mục tiêu điển hình trong phân cụm chính thức hóa mục tiêu đạt được độ tương tự cao trong nội bộ cụm (các tài liệu trong một cụm là tương tự nhau) và độ tương đồng giữa các cụm thấp (các tài liệu từ các cụm khác nhau là không giống nhau). Đây là một tiêu chí nội bộ cho chất lượng của một cụm. Nhưng điểm tốt của một hàm mục tiêu không nhất thiết có nghĩa là việc phân cụm hữu ích cho một ứng dụng cụ thể. Do đó, điều quan trọng là phải đánh giá chất lượng của một cụm dựa trên cả tiêu chí bên trong và bên ngoài.*

*Tiêu chí nội bộ chỉ dựa trên dữ liệu và bao gồm các thước đo như tổng sai số bình phương (SSE), tổng bình phương độ lệch so với trung bình cụm (SSM) và tỷ lệ phương sai giữa các cụm với phương sai bên trong cụm (F-tỷ lệ). Các biện pháp này được sử dụng để đánh giá chất lượng của một cụm dựa trên mức độ phù hợp của các điểm dữ liệu với các cụm.*

*Mặt khác, các tiêu chí bên ngoài dựa trên thông tin bên ngoài như nhãn lớp hoặc kiến ​​thức chuyên môn. Các tiêu chí này bao gồm các thước đo như chỉ số Rand được điều chỉnh (ARI), thông tin chung được chuẩn hóa (NMI) và thước đo F. Các biện pháp này được sử dụng để đánh giá chất lượng của một cụm dựa trên mức độ tương ứng của các cụm với thông tin bên ngoài.*

##### **[Author 2]**