(Answers) MID\_10\_2021.pdf (10.2021-2022) (11 câu)

[**1. Questions (0.75 điểm) 2**](#_5h7cs3thsunn)

[**1. Answers 2**](#_f4wvmlwag6pt)

[[Thành] 2](#_e1kdtlz0j7mp)

[**2. Questions(1 điểm) 2**](#_adk02lomt2al)

[**2. Answers 2**](#_eixd3mr72lnp)

[[Thành] 2](#_bs2judy3x2wc)

[**3. Answers 3**](#_rw8sh5ld60ix)

[[Thành] 3](#_ahaneetxb5e4)

[**4. Questions (1.5 điểm) 3**](#_rhijdb1423ee)

[**4. Answers 3**](#_n5z514r6ikmf)

[[Thành] 3](#_86ho4kvwx0th)

[**5. Questions (0.75 điểm) 4**](#_226g1fvr4cwn)

[**5. Answers 4**](#_7r5hyfgt0vl9)

[[Nghia] 4](#_u8aj5uenhmfz)

[**6. Questions (1.5 điểm) 4**](#_mywvxv2e4r3q)

[**6. Answers 5**](#_snavz35etc7)

[[Nghia] 5](#_57oezj4tardd)

[**7. Answers 5**](#_fki39z3kiojg)

[[Nghia] 5](#_2m6gjujd2rgp)

[**8. Questions (0.5 điểm) 6**](#_mmbdxngfrbl)

[**8. Answers 6**](#_hxoq9qx455jr)

[[Nghia] 6](#_28s959mtzuf7)

[**9. Questions (0.5 điểm) 7**](#_fm94yur5tekl)

[**9. Answers 7**](#_s82mnmj8m24r)

[[Nghia] 7](#_ewptx91bosbb)

[**10. Questions (1 điểm) 7**](#_bonhwb27rhhe)

[**10. Answers 7**](#_kxsa3okxdlvl)

[[Nghia] 7](#_266ehoq3g83i)

[**11. Answers 8**](#_khrsi4vq7p7l)

[[Nghia] 8](#_rfi7xz4otwt)

### **1. Questions (0.75 điểm)**

Nêu sự khác biệt giữa *lựa chọn prototype* (prototype selection) và *trích yếu prototype* (prototype abstraction) khi thu giảm tập dữ liệu. Giải thuật Condensed Nearest Neighbors sử dụng phương pháp lựa chọn prototype hay trích yếu prototype ?

### **1. Answers**

#### **[Thành]**

1. Lựa chọn nguyên mẫu và trừu tượng hóa nguyên mẫu là hai phương pháp khác nhau được sử dụng trong việc giảm dữ liệu. Lựa chọn nguyên mẫu là phương pháp chọn một số lượng nhỏ mẫu từ tập dữ liệu ban đầu để tạo tập dữ liệu mới với kích thước nhỏ hơn. Trừu tượng hóa nguyên mẫu là một phương pháp sử dụng các mẫu đã chọn để tạo các mẫu mới bằng cách trích xuất các tính năng quan trọng của chúng.

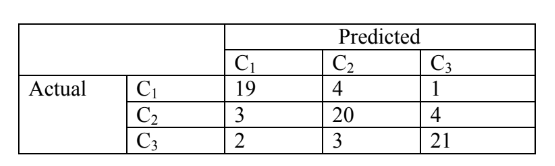
Về cơ bản, lựa chọn nguyên mẫu làm giảm kích thước của tập dữ liệu bằng cách loại bỏ các mẫu trùng lặp hoặc không quan trọng. Trừu tượng hóa nguyên mẫu làm giảm kích thước của tập dữ liệu bằng cách sử dụng các tính năng quan trọng của mẫu để tạo mẫu mới.

1. Condensed Nearest Neighbors (CNN) đang sử dụng phương pháp lựa chọn nguyên mẫu để giảm tập dữ liệu cho phân loại k-NN. Nó chọn một tập hợp con các ví dụ có thể phân loại chính xác tập dữ liệu gốc bằng thuật toán 1-NN.

### **2. Questions(1 điểm)**

a. Cho một bộ phân lớp với 3 lớp C1, C2 và C3. Hãy lập công thức tính độ đo accuracy của bộ phân lớp này. (*0.5 điểm*)

b. Confusion matrix của một bộ phân lớp với 3 lớp được cho như sau:



Hãy tính độ đo accuracy của bộ phân lớp này. (0.5 điểm)

### **2. Answers**

#### **[Thành]**

Accuracy = (TP\_C1 + TP\_C2 + TP\_C3) / (Total number of samples)

where TP\_Ci is the number of true positive samples for class Ci.

b. Confusion matrix của một bộ phân lớp với 3 lớp được cho như sau:

|  | | Predicted | | |
| --- | --- | --- | --- | --- |
| C1 | C2 | C3 |
| Actual | C1 | 19 | 4 | 1 |
| C2 | 3 | 20 | 4 |
| C3 | 2 | 3 | 21 |

Hãy tính độ đo accuracy của bộ phân lớp này. (*0.5 điểm*)

Accuracy = (TP\_C1 + TP\_C2 + TP\_C3) / (Total number of samples)

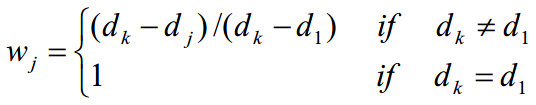
= (19 + 20 + 21) / (19 + 4 + 1 + 3 + 20 + 4 + 2 + 3 + 21) = 60 / 77 = 0.78 = 78%

**3. Questions (1 điểm)**

Chúng ta dùng bộ phân lớp *5-lân cận gần nhất có trọng số* (weighted 5-NN classifier) để phân lớp mẫu thử P. Giả sử khoảng cách giữa P với năm lân cận gần nhất (X1, X2, X3, X4 và X5) lần lượt là d1 = 1, d2 = 3, d3 = 4, d4 = 5 và d5 =8. Nếu X1, X2 thuộc lớp + và X3, X4 , X5 thuộc lớp -. Vậy P sẽ được phân vào lớp nào?

### **3. Answers**

#### **[Thành]**



w1 = (8-1)/(8-1) =1

w2 = (8-3)/(8-1) = 5/7

w3 = (8-4)/(8-1) = 4/7

w4 = (8-5)/(8-1) = 3/7

w5 = (8-8)/(8-1) = 0

sum + = 1 + 5/7 = 12/7

sum - = 4/7+3/7 = 1 < sum +

=> P thuộc về lớp +

### **4. Questions** (1.5 điểm)

Cho một tập các mẫu hai chiều sau đây:

(1, 1, 1), (1, 2, 1), (1, 3, 1), (2, 1, 1), (2, 2, 1), (2, 3, 1), (2, 3.5, 1),

(2.5, 2, 1), (3.5, 1, 1), (3.5, 2, 1), (3.5, 3, 2), (3.5, 4, 2), (4.5, 1, 2)

(4.5, 2, 2), (4.5, 3, 2), (5, 4, 2), (5, 5, 2), (6, 3, 2), (6, 4, 2), (6, 5, 2)

trong đó mỗi mẫu gồm đặc trưng thứ nhất, đặc trưng thứ hai và nhãn lớp. Tìm hai điểm centroid của hai lớp 1 và 2. Sử dụng *minimum-distance classifier* để tìm lớp của mẫu thử *P* là (3.8, 3.1).

### **4. Answers**

#### [Thành]

C1 = centroid lớp 1: (2.05, 2.05)

C2 = centroid lớp 2: (4.85, 3.4)

Tính khoảng cách Euclid từ P với centroid lớp 1 và 2

d(P, C1) = sqrt((2.05-3.8)^2 + (2.05-3.1)^2) = 2.040

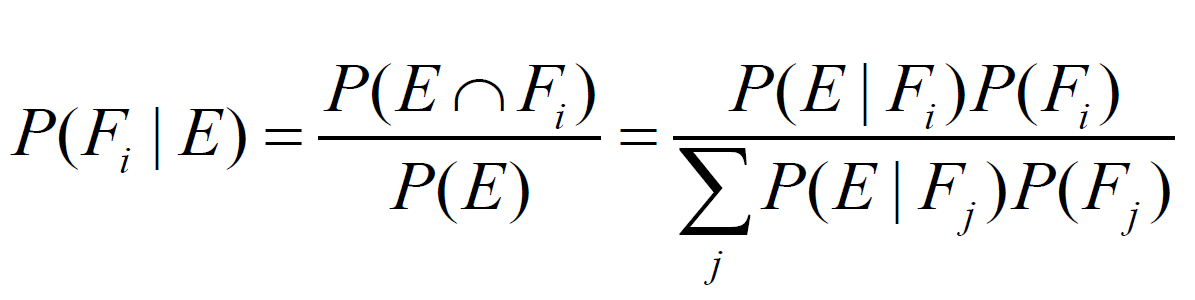
d(P, C2) = sqrt((4.85-3.8)^2 + (3.4-3.1)^2) = 1.092

Do d(P, C2) < d(P, C1) nên ta chọn lớp C2 là lớp của mẫu thử P

### **5. Questions (0.75 điểm)**

a. Cho các xác suất: P(A|B) = 2/3, P(A|~B) = 1/3, P(B) = 1/3. Hãy tính xác xuất có điều kiện P(B|A).

Hint : Áp dụng công thức

**

*b. Để có thể áp dụng bộ phân lớp Naive Bayes, chúng ta cần một giả định gì về tính chất dữ*

*liệu.*

### **5. Answers**

#### [Nghia]

a.

Công thức Bayes: P(B|A) = P(A|B) \* P(B) / P(A).

Ở đây, ta có P(A|B) = 2/3, P(A|~B) = 1/3, P(B) = 1/3.

Ta cần tính P(A) để có thể tính được P(B|A).

Theo công thức xác suất có điều kiện, ta có: P(A) = P(A|B) \* P(B) + P(A|~B) \* P(~B).

Vì P(~B) = 1 - P(B) = 2/3, nên ta có: P(A) = (2/3) \* (1/3) + (1/3) \* (2/3) = 4/9.

Sau đó, ta có thể tính được P(B|A) = P(A|B) \* P(B) / P(A) = (2/3) \* (1/3) / (4/9) = 1/2.

Vậy xác suất có điều kiện P(B|A) là 1/2.

b. Để có thể áp dụng bộ phân lớp Naive Bayes, chúng ta cần một giả định gì về tính chất dữ liệu.

Cần giả định về tính độc lập giữa các đặc trưng của dữ liệu. Cụ thể, giả định này là các đặc trưng của dữ liệu là độc lập với nhau, tức là giá trị của một đặc trưng không phụ thuộc vào giá trị của các đặc trưng khác.

### **6. Questions (1.5 điểm)**

Cho tập mẫu như sau:

| Đặc trưng 1 | Đặc trưng 2 | Đặc trưng 3 | Lớp |
| --- | --- | --- | --- |
| 0  1  1  1  0  0 | 0  0  0  1  1  1 | 0  1  0  1  1  1 | 0  1  0  1  1  0 |

trong đó mỗi mẫu gồm 3 đặc trưng và nhãn lớp.

Nếu mẫu thử *P* với đặc trưng 1 là 0 và đặc trưng 2 là 0 và đặc trưng 3 là 1, hãy phân lớp mẫu thử này dùng giải thuật phân lớp Naïve Bayes.

### **6. Answers**

#### [Nghia]

Tính P(class = 0 | F(0, 0, 1)) & P(class = 1 | F(0, 0, 1)) so sánh P nào giá trị lớn hơn để phân vào class tương ứng

P(class = 0) = P(class = 1) = 3/6 = ½

P(class = 0 | F(0, 0, 1)) = P(class = 0) \* P(F1 = 0 | class = 0) \* P(F2 = 0 | class = 0) \* P(F3 = 1 | class = 0) = 1/2 \* ⅔ \* ⅔ \* ⅓ = 2/27

P(class = 1 | F(0, 0, 1)) = P(class = 1) \* P(F1 = 0 | class = 1) \* P(F2 = 0 | class = 1) \* P(F3 = 1 | class = 1) = 1/2 \* ⅓ \* ⅓ \* 3/3 = 1/18

Ta có: P(class = 1 | F(0, 0, 1)) > P(class = 0 | F(0, 0, 1)) => F(0,0,1) thuộc lớp 1

**7**. Questions **(1 điểm)**

Tìm centroid và medoid của tập mẫu sau đây:

(1, 1), (1, 3), (1, 4), (2, 2), (2, 3), (3,1)

### **7. Answers**

#### **[Nghia]**

**Tìm centroid**

Trung bình cộng của các giá trị của chiều thứ nhất là: (1 + 1 + 1 + 2 + 2 + 3) / 6 = 1.67.

Trung bình cộng của các giá trị của chiều thứ hai là: (1 + 3 + 4 + 2 + 3 + 1) / 6 = 2.33..

Vậy centroid của tập mẫu là (1.67, 2).

**Tìm medoid**

Bản khoảng cách Euclid giữa các điểm

|  | (1, 1) | (1, 3) | (1, 4) | (2, 2) | (2, 3) | (3, 1) |
| --- | --- | --- | --- | --- | --- | --- |
| (1, 1) | 0 | 2 | 3 | 1.41 | 2.24 | 2.24 |
| (1, 3) | 2 | 0 | 1 | 1 | 1.41 | 2.83 |
| (1, 4) | 3 | 1 | 0 | 2.24 | 2.83 | 3.61 |
| (2, 2) | 1.41 | 1 | 2.24 | 0 | 1 | 2.24 |
| (2, 3) | 2.24 | 1.41 | 2.83 | 1 | 0 | 3.16 |
| (3, 1) | 2.24 | 2.83 | 3.61 | 2.24 | 3.16 | 0 |

| Điểm | Tổng khoảng cách |
| --- | --- |
| (1, 1) | 10.93 |
| (1, 3) | 8.25 |
| (1, 4) | 11.68 |
| (2, 2) | 6.65 |
| (2, 3) | 11.64 |
| (3, 1) | 10.27 |

Điểm (2, 2) có tổng khoảng cách Euclid tới các điểm khác nhỏ nhất -> là medoid

### **8. Questions** (**0.5 điểm**)

Nếu tập mẫu có số mẫu là 10 và cần thiết phải tách tập mẫu này thành hai tập con phân ly. Như vậy có tổng cọng bao nhiêu cách tách có thể có?

### **8. Answers**

#### **[Nghia]**

Số cách chọn một tập con từ tập mẫu ban đầu là 2^N, trong đó N là số phần tử trong tập mẫu ban đầu.

Trong trường hợp này, tập mẫu có 10 phần tử, vì vậy số cách tách tập mẫu thành hai tập con phân ly là 2^10 = 1024.

### **9. Questions (0.5 điểm)**

Nêu sự khác biệt giữa gom cụm cứng (hard clustering) và gom cụm mềm (soft clustering). Nêu điểm tương đồng giữa giải thuật gom cụm k-means và giải thuật gom cụm fuzzy-c-means.

### **9. Answers**

#### [Nghia]

| Gom cụm cứng | Gom cụm mềm |
| --- | --- |
| mỗi điểm dữ liệu chỉ thuộc về một cụm duy nhất | mỗi điểm dữ liệu có thể thuộc về nhiều cụm khác nhau với một mức độ xác suất nhất định |
| mỗi điểm dữ liệu được gán vào cụm gần nhất dựa trên khoảng cách Euclid giữa các điểm dữ liệu và các trung tâm cụm. | mỗi điểm dữ liệu được gán một vector xác suất, trong đó mỗi phần tử của vector biểu thị mức độ xác suất của điểm dữ liệu đó thuộc về mỗi cụm |

Điểm tương đồng giữa giải thuật gom cụm k-means và giải thuật gom cụm fuzzy-c-means:

- đều là các phương pháp gom cụm cứng

- đều sử dụng một số lượng cụm cố định để gom các điểm dữ liệu

- đều sử dụng khoảng cách Euclid để tính toán khoảng cách giữa các điểm dữ liệu và các trung tâm cụm

- đều sử dụng một quá trình lặp để tối ưu hóa vị trí của các trung tâm cụm

### **10. Questions (1 điểm)**

a. Cho hai mẫu (mỗi mẫu gồm 3 thuộc tính) X = (7, 4, 3), Y = (4, 1, 8), hãy tính khoảng cách

Manhattan giữa hai mẫu X và Y. (0.5 điểm)

b. Cho hai mẫu (mỗi mẫu gồm 8 thuộc tính nhị phân) X = (1, 0, 1, 1, 1, 0, 1, 1), Y = (0, 1, 1,

0, 0, 1, 0, 0), hãy tính khoảng cách giữa hai mẫu X và Y. (0.5 điểm)

### **10. Answers**

#### [Nghia]

***a. Cho hai mẫu (mỗi mẫu gồm 3 thuộc tính) X = (7, 4, 3), Y = (4, 1, 8), hãy tính khoảng cách Manhattan giữa hai mẫu X và Y. (0.5 điểm)***

|X - Y| = |(7-4)| + |(4-1)| + |(3-8)| = 3 + 3 + 5 = 11

***b. Cho hai mẫu (mỗi mẫu gồm 8 thuộc tính nhị phân) X = (1, 0, 1, 1, 1, 0, 1, 1), Y = (0, 1, 1, 0, 0, 1, 0, 0), hãy tính khoảng cách giữa hai mẫu X và Y. (0.5 điểm)***

Khoảng cách giữa hai mẫu X và Y được tính bằng số lượng thuộc tính khác nhau giữa hai mẫu => |X - Y| = 5

**11. Questions (0.5 điểm)**

Gom cụm gia tăng là gì? Nêu một nhược điểm của giải thuật gom cụm gia tăng Leader.

### **11. Answers**

#### **[Nghia]**

Gom cụm gia tăng (incremental clustering) là một phương pháp gom cụm dữ liệu trong đó các điểm dữ liệu được thêm vào một cách tuần tự và các cụm được cập nhật sau mỗi lần thêm điểm dữ liệu mới.

Nhược điểm của giải thuật gom cụm gia tăng Leader là nó có thể dẫn đến việc tạo ra các cụm không tối ưu. Điều này xảy ra khi một điểm dữ liệu mới được thêm vào và nó không thuộc cụm nào, giải thuật sẽ tạo ra một cụm mới cho điểm dữ liệu này thay vì thêm nó vào cụm gần nhất. Điều này có thể dẫn đến việc tạo ra các cụm không tối ưu và ảnh hưởng đến kết quả của giải thuật.