(Answers) Mid\_201.pdf (2020) (11 câu)

[**1. Questions 2**](#_94qvtnmnl639)

[**1. Answers 2**](#_w73ace66hsg5)

[[Tâm] 2](#_tptorm9eoifd)

[**2. Questions 3**](#_y05l5kn7ogcc)

[**2. Answers 4**](#_mcp7mvo4tru7)

[[Tâm] 4](#_rvp9zkyq8tgb)

[[Author 2] 4](#_imdleoikhg1o)

[**3. Questions 4**](#_7qg718bxezer)

[**3. Answers 5**](#_77sjxdx404n6)

[[Tâm] 5](#_h6zk516rpqp)

[[Author 2] 5](#_uhpiut128c4b)

[**4. Questions 5**](#_q6jb85n58196)

[**4. Answers 5**](#_mdpfu9hgssr)

[[Author 1] 5](#_jgk2jccli667)

[[Author 2] 6](#_oybo1d9pvxxe)

[**6. Questions 6**](#_pf1n7pbrggrk)

[**6. Answers 6**](#_n4yuz5i32wvz)

[[Author 1] 6](#_z3cn9oe5u45z)

[[Author 2] 7](#_7x7quvk5uvz2)

[**7. Questions 7**](#_t47s21fs29c0)

[**7. Answers 7**](#_v42t046oh5qq)

[[Tâm] bài có trọng số thì ta dùng Modified k-nn 7](#_go3gzgjmab06)

[[Author 2] 8](#_uqvzukt0nhbp)

[**8. Questions 8**](#_q3f18hpqvs0y)

[**8. Answers 8**](#_u99dcbf62x30)

[[Author 1] 8](#_2wpsc7h686wg)

[[Author 2] 8](#_mbnnizf04imc)

[**9. Questions 8**](#_ic0044fo91eb)

[**9. Answers 9**](#_zfkaansbs7hj)

[[Tâm] 9](#_aea2jx4kxddv)

[[Author 2] 9](#_dymo0rb5fq28)

[**10. Questions 9**](#_rn1z7tg5yszz)

[**10. Answers 9**](#_ar7duyuhcu7u)

[[Tâm] 9](#_py6lo9yn9c3l)

[[Author 2] 10](#_9b8c9hvxss3q)

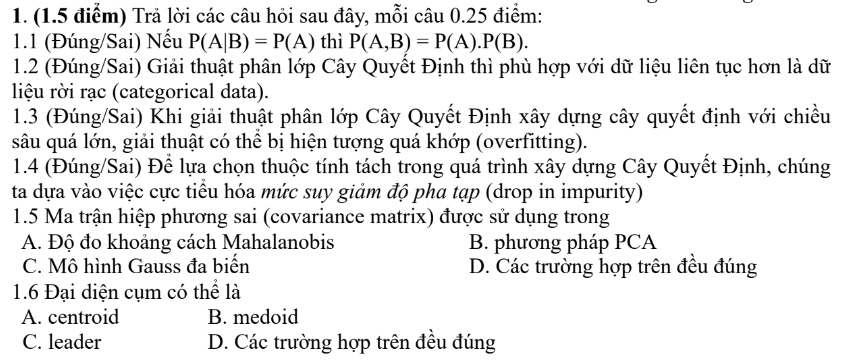
[**11. Questions 10**](#_3f4ul2xipbzl)

[**11. Answers 10**](#_5yrvmex0g39q)

[[Tâm] 10](#_8jan33na78us)

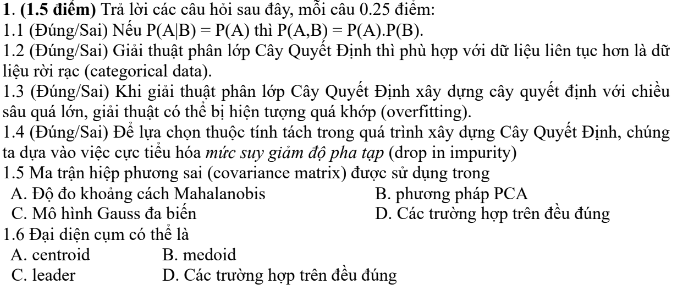
[[Author 2] 10](#_oleslb58leve)

### **1. Questions**

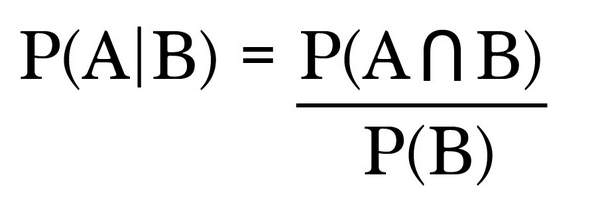


### **1. Answers**

#### [Tâm]

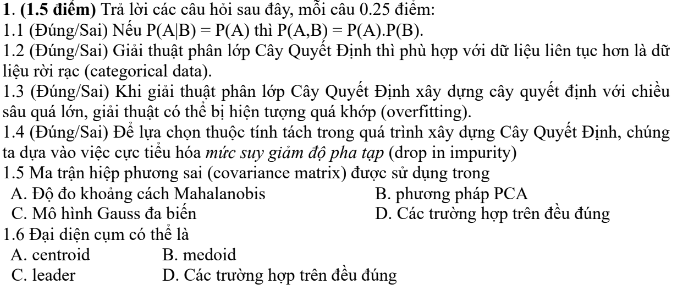


Theo bayes:

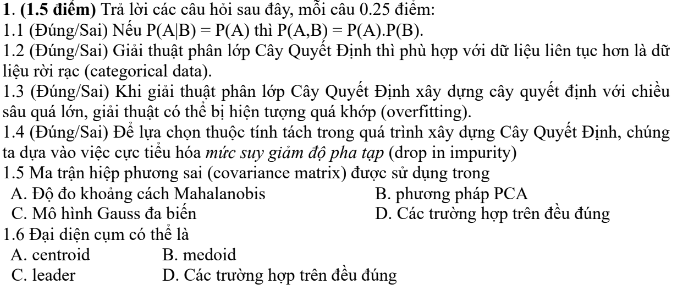


P(A|B) = P(A) thay vào bayes → P(A,B) = P(A).P(B)

### 



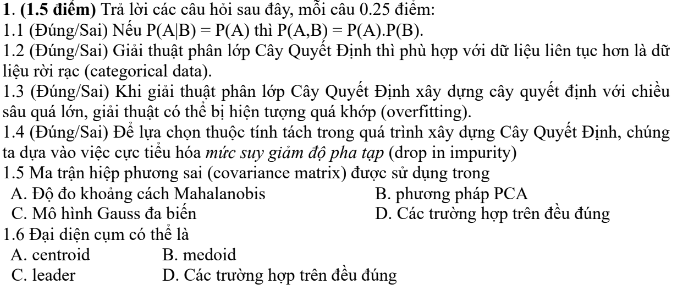
[Sai] Phù hợp hơn với dữ liệu rời rạc



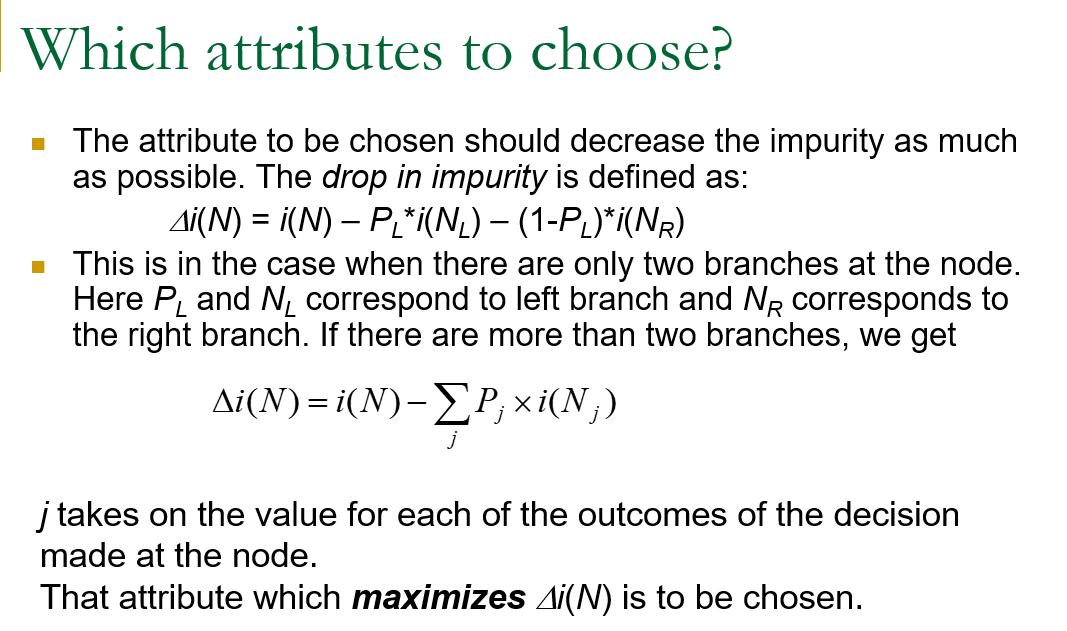
[Đúng]

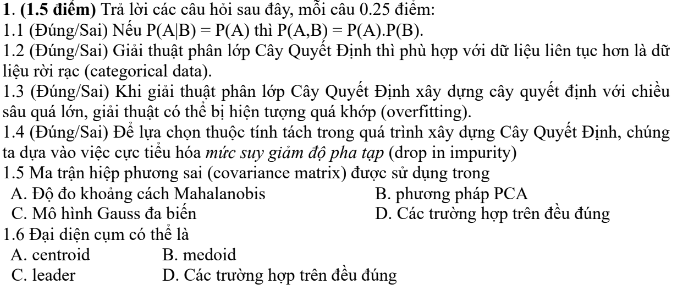
Hiện tượng overfitting (quá khớp) ở cây quyết định xảy ra khi cây quyết định được xây dựng quá phức tạp và quá khớp với tập dữ liệu huấn luyện, mà không thể tổng quát hóa tốt cho các dữ liệu mới.

Điều này thường xảy ra khi cây quyết định được xây dựng quá sâu, quá nhiều lá và quá phức tạp. Khi mô hình quá khớp, nó sẽ cho kết quả tốt trên tập dữ liệu huấn luyện nhưng không cho kết quả tốt trên các tập dữ liệu mới.

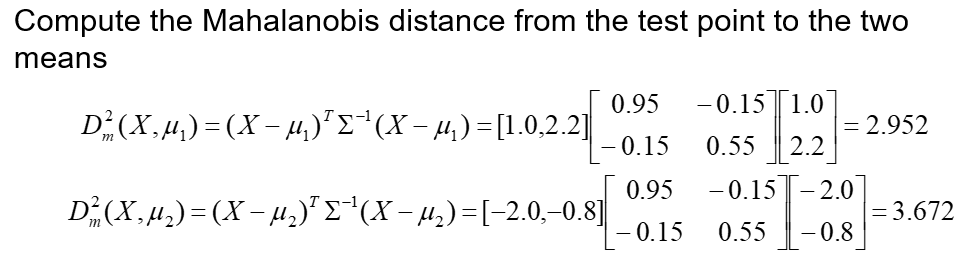


[Đúng]

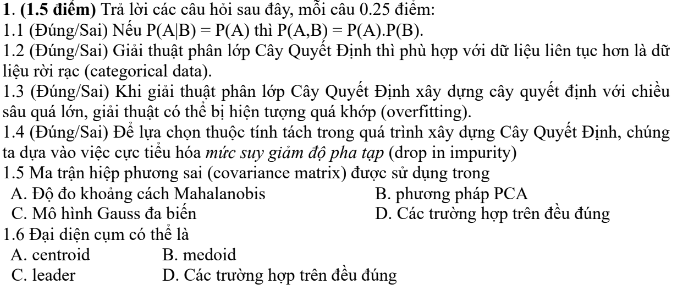




A



C



D

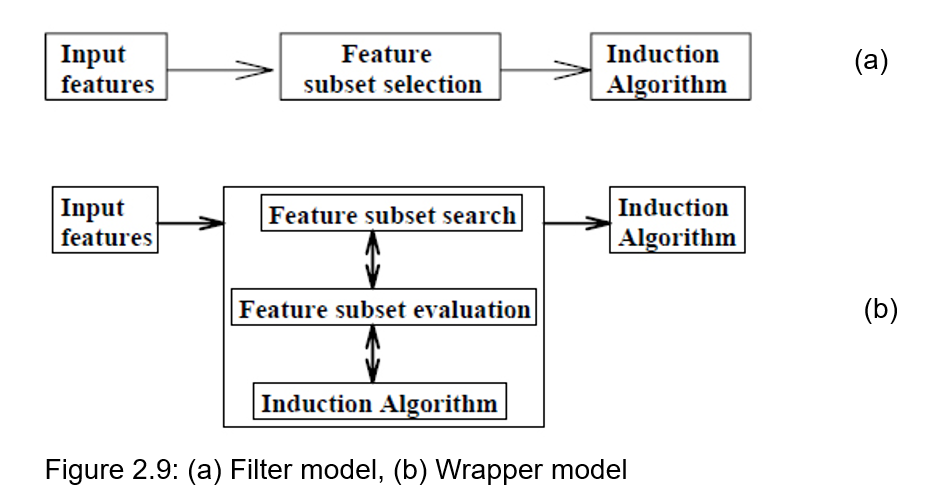
### **2. Questions**

(0.5 điểm) Nêu sự khác biệt giữa mô hình filter và mô hình wrapper trong công tác lựa

chọn đặc trưng (feature selection).

### **2. Answers**

#### [Tâm]



Mô hình filter và mô hình wrapper là hai phương pháp khác nhau trong quá trình lựa chọn đặc trưng (feature selection) của một mô hình.

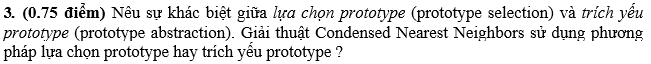
Mô hình filter là một phương pháp lựa chọn đặc trưng độc lập với mô hình dự đoán. Nó thực hiện bằng cách sử dụng các phương pháp thống kê hoặc thông tin để đánh giá mức độ quan trọng của mỗi đặc trưng. Các đặc trưng được lựa chọn dựa trên các giá trị độ quan trọng này, và sau đó được sử dụng để huấn luyện mô hình dự đoán. Mô hình filter có thể giúp giảm chiều dữ liệu và cải thiện hiệu suất của mô hình dự đoán.

Mô hình wrapper là một phương pháp lựa chọn đặc trưng phụ thuộc vào mô hình dự đoán. Nó thực hiện bằng cách sử dụng một mô hình dự đoán để đánh giá mức độ quan trọng của mỗi đặc trưng. Mỗi lần lựa chọn đặc trưng, mô hình dự đoán được huấn luyện lại và đánh giá lại hiệu suất trên tập kiểm tra. Quá trình này được lặp lại cho đến khi tập đặc trưng tốt nhất được tìm thấy. Mô hình wrapper có thể cung cấp kết quả tốt hơn so với mô hình filter vì nó xem xét tương tác giữa các đặc trưng.

Tóm lại, mô hình filter và mô hình wrapper đều là các phương pháp lựa chọn đặc trưng, nhưng có phương pháp tiếp cận khác nhau. Mô hình filter đánh giá mức độ quan trọng của từng đặc trưng độc lập với mô hình dự đoán, trong khi mô hình wrapper sử dụng mô hình dự đoán để đánh giá mức độ quan trọng của từng đặc trưng phụ thuộc vào các tương tác giữa chúng.

#### [Author 2]

### **3. Questions**



### **3. Answers**

#### [Tâm]

Lựa chọn prototype (prototype selection) và trích yếu prototype (prototype abstraction) là hai phương pháp khác nhau để giảm số lượng các điểm dữ liệu trong tập huấn luyện.

Lựa chọn prototype là phương pháp giữ lại một tập con của các điểm dữ liệu gốc, được chọn dựa trên một số tiêu chí nhất định như độ tương đồng với các điểm khác trong tập dữ liệu. Tập con này được sử dụng để đào tạo mô hình.

Trích yếu prototype là phương pháp giảm số chiều của dữ liệu bằng cách biểu diễn dữ liệu bằng một tập hợp các điểm dữ liệu khác. Các điểm này được gọi là prototype và chúng được chọn để đại diện cho nhóm các điểm dữ liệu tương tự. Tập hợp các prototype này được sử dụng để đào tạo mô hình.

Giải thuật Condensed Nearest Neighbors (CNN) sử dụng phương pháp lựa chọn prototype để giảm kích thước của tập dữ liệu ban đầu. Với phương pháp này, CNN sẽ xây dựng một tập dữ liệu con từ tập dữ liệu ban đầu bằng cách chỉ lấy một số mẫu đại diện cho các lớp và loại bỏ các mẫu không cần thiết. Việc loại bỏ các mẫu không cần thiết giúp giảm đáng kể chi phí tính toán khi huấn luyện mô hình.

Phương pháp lựa chọn prototype của CNN là trích yếu prototype. Điều này có nghĩa là CNN sử dụng một thuật toán trích yếu dữ liệu để tìm các mẫu đại diện cho các lớp trong tập dữ liệu ban đầu. Thuật toán trích yếu này sẽ tính toán các vector trọng số đại diện cho các lớp và sử dụng chúng để tìm các mẫu đại diện.

Với phương pháp trích yếu prototype, CNN có thể tìm ra các mẫu đại diện cho các lớp chỉ với một số lượng nhỏ các mẫu, do đó giảm đáng kể kích thước của tập dữ liệu ban đầu. Tuy nhiên, việc trích yếu dữ liệu có thể dẫn đến mất mát thông tin và ảnh hưởng đến hiệu suất của mô hình. Do đó, cần phải chọn một thuật toán trích yếu dữ liệu phù hợp để đảm bảo rằng các mẫu đại diện được chọn thực sự là đại diện cho các lớp và không làm giảm độ chính xác của mô hình.

#### [Author 2]

### **4. Questions**

### 

### **4. Answers**

#### [Author 1]

Cả hai phương pháp Maximum Likelihood (ML) và Maximum A Posteriori (MAP) đều là các phương pháp ước lượng tham số cho các mô hình học máy. Tuy nhiên, cách tiếp cận của chúng khác nhau.

Maximum Likelihood (ML) là một phương pháp ước lượng tham số dựa trên xác suất của dữ liệu quan sát được. ML ước lượng giá trị tham số của mô hình sao cho xác suất quan sát được dữ liệu là lớn nhất. Điều này được thực hiện bằng cách tối đa hóa hàm likelihood, tức là hàm xác suất của dữ liệu dưới giả định rằng các tham số của mô hình đang được ước lượng là chính xác. Các giá trị tham số được ước lượng bằng cách giải phương trình đạo hàm bằng 0 của hàm likelihood.

Maximum A Posteriori (MAP) là một phương pháp ước lượng tham số dựa trên phân bố xác suất hậu nghiệm của tham số dưới điều kiện có sẵn các dữ liệu quan sát được. MAP kết hợp xác suất tiên nghiệm của tham số với xác suất likelihood của dữ liệu quan sát được để ước lượng phân bố xác suất hậu nghiệm của các tham số. Điều này được thực hiện bằng cách tối đa hóa hàm log của phân bố xác suất hậu nghiệm, tức là hàm log của tích giữa xác suất likelihood và xác suất tiên nghiệm.

Về cách tiếp cận, phương pháp ML tập trung vào tối ưu hóa xác suất dữ liệu quan sát được, trong khi phương pháp MAP tập trung vào tối ưu hóa phân bố xác suất hậu nghiệm của tham số dưới điều kiện các dữ liệu quan sát được. Sự khác biệt này là do MAP sử dụng xác suất tiên nghiệm của tham số, tức là kiến thức trước đó về phân bố của các tham số, để ước lượng phân bố xác suất hậu nghiệm của tham số, trong khi ML không sử dụng kiến thức tiên nghiệm nào và chỉ dựa trên dữ liệu quan sát được để ước lượng tham số.

Trong phương pháp Naïve Bayes, MAP được sử dụng để ước lượng các tham số của mô hình, trong đó xác suất tiên nghiệm được xác định bởi kiến thức tiền đề về phân bố xác suất của các lớp và các đặc trưng. Mô hình Naïve Bayes giả định rằng các đặc trưng độc lập với nhau khi biết lớp, do đó xác suất của một đặc trưng cho trước lớp được tính dựa trên số lần xuất hiện của đặc trưng trong tập huấn luyện và số lần xuất hiện của lớp đó trong tập huấn luyện. MAP được sử dụng để tối đa hoá xác suất hậu nghiệm của các tham số đối với một bộ dữ liệu huấn luyện cụ thể, giúp cải thiện tính tổng quát của mô hình và tránh hiện tượng quá khớp.

Tóm lại, ML và MAP đều là các phương pháp ước lượng tham số cho các mô hình học máy, tuy nhiên, cách tiếp cận của chúng khác nhau. ML tập trung vào tối ưu hóa xác suất dữ liệu quan sát được, trong khi MAP tập trung vào tối ưu hóa phân bố xác suất hậu nghiệm của tham số dưới điều kiện các dữ liệu quan sát được và sử dụng xác suất tiên nghiệm để ước lượng phân bố xác suất hậu nghiệm của các tham số. Trong phương pháp Naïve Bayes, MAP được sử dụng để ước lượng các tham số của mô hình và cải thiện tính tổng quát của mô hình.

#### [Author 2]

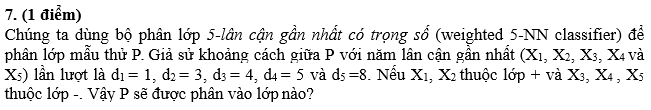
### **6. Questions**

### **6. Answers**

#### [Author 1]

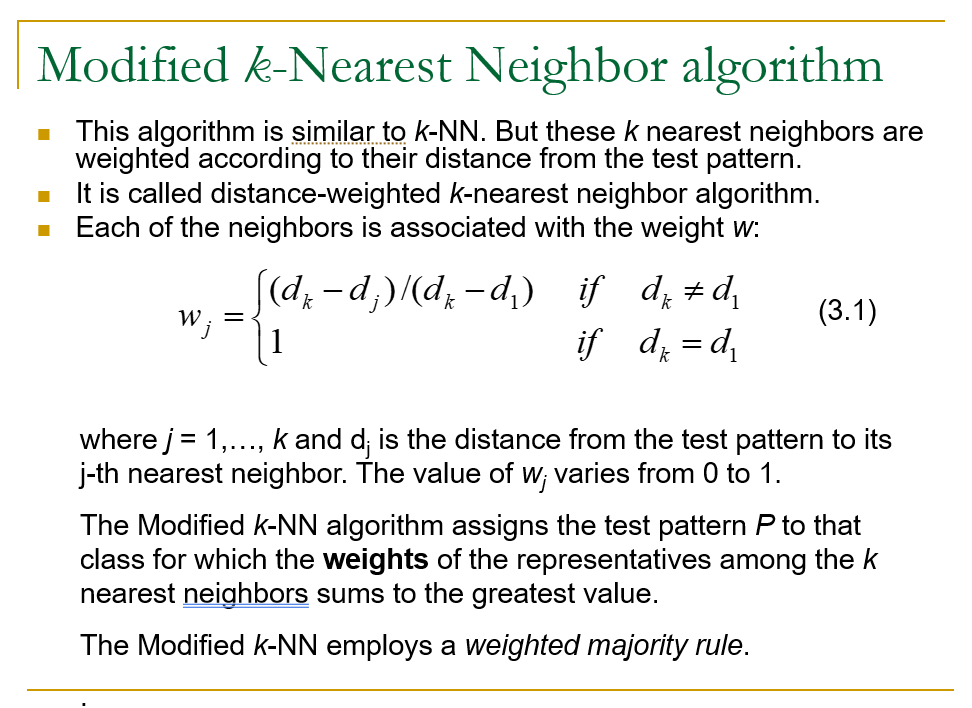
#### [Author 2]

### **7. Questions**



### **7. Answers**

#### [Tâm] bài có trọng số thì ta dùng Modified k-nn



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

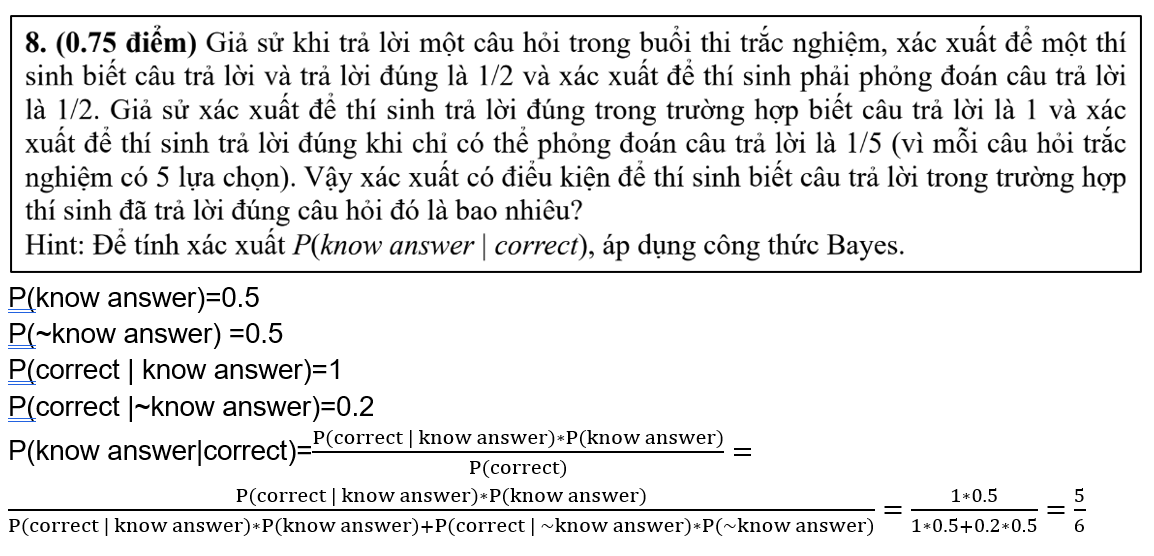
sum + = 1 + 5/7 = 12/7

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

⇒ P thuộc lớp +

#### [Author 2]

### **8. Questions**



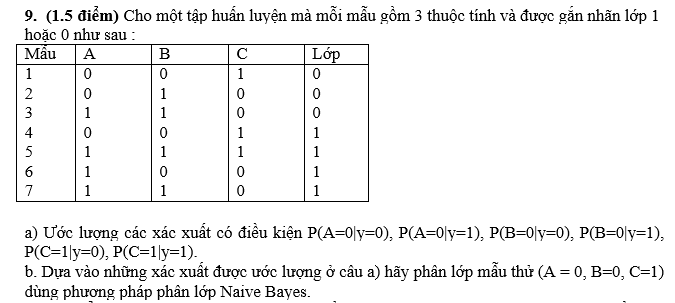
Toàn thấy hơi cấn chỗ “xác suất một thí sinh biết câu trả lời và trả lời đúng là ½” nó trùng với câu hỏi nên mình tạm hiểu nó là P(know answer) = 0.5

### **8. Answers**

#### [Author 1]

#### [Author 2]

### **9. Questions**



### **9. Answers**

#### [Tâm]

P(A=0|y=0) = 2/3

P(A=0|y=1) = 1/4

P(B=0|y=0) = 1/3

P(B=0|y=1) = 2/4

P(C=1|y=0) = 1/3

P(C=1|y=1) = 2/4

phân lớp mẫu thử (A = 0, B=0, C=1) dùng phương pháp phân lớp Naive Bayes

P(y=0| A = 0, B=0, C=1) =P(y=0) P(A=0|y=0).P(B=0|y=0).P(C=1|y=0) =

P(y=1| A = 0, B=0, C=1) =P(y=1) P(A=0|y=1).P(B=0|y=1).P(C=1|y=1) =

#### [Author 2]

### **10. Questions**



### **10. Answers**

#### [Tâm]

Cả hai phương pháp kiểm tra chéo (cross-validation) sau đây đều được sử dụng để đánh giá hiệu quả của các mô hình máy học trên tập dữ liệu huấn luyện.

**Kiểm tra chéo k-phần (k-fold cross-validation):**

Phương pháp này chia tập dữ liệu huấn luyện thành k phần bằng nhau (gọi là fold).

Sau đó, mô hình được huấn luyện trên k-1 phần và được đánh giá trên phần còn lại. Quá trình này được lặp lại k lần để đánh giá hiệu quả của mô hình trên toàn bộ tập dữ liệu huấn luyện.

Kết quả đánh giá cuối cùng được tính trung bình từ k lần lặp lại trên các fold.

Kỹ thuật kiểm tra chéo k-fold thường được sử dụng khi tập dữ liệu huấn luyện có kích thước lớn hoặc khi không đủ dữ liệu để sử dụng phương pháp kiểm tra chéo leave-one-out.

**Kiểm tra chéo bỏ ra một phần tử (leave-one-out cross-validation):**

Phương pháp này đánh giá hiệu quả của mô hình trên từng mẫu dữ liệu trong tập huấn luyện.

Cụ thể, mỗi mẫu dữ liệu sẽ được bỏ ra khỏi tập dữ liệu huấn luyện, mô hình sẽ được huấn luyện trên tập dữ liệu còn lại và đánh giá trên mẫu dữ liệu bị bỏ ra.

Quá trình này được lặp lại cho tất cả các mẫu dữ liệu trong tập huấn luyện để đánh giá hiệu quả của mô hình trên toàn bộ tập huấn luyện.

Kết quả đánh giá cuối cùng được tính trung bình từ số lần lặp lại trên các mẫu dữ liệu.

Kỹ thuật kiểm tra chéo leave-one-out thường được sử dụng khi kích thước tập dữ liệu huấn luyện nhỏ hoặc khi muốn đánh giá chính xác nhất hiệu quả của mô hình. Tuy nhiên, phương pháp này có thể tốn nhiều thời gian tính toán nếu kích thước tập dữ liệu lớn.

#### [Author 2]

### **11. Questions**

### **11. Answers**

#### [Tâm]

Chất lượng gom cụm của giải thuật k-means phụ thuộc vào các yếu tố sau:

**Số lượng cụm (k)**: Nếu số lượng cụm được chọn không phù hợp với số lượng các nhóm thực tế trong dữ liệu, hoặc quá ít hoặc quá nhiều so với số lượng các nhóm thực tế, thì kết quả gom cụm sẽ không chính xác.

**Phương pháp chọn điểm khởi tạo ban đầu:** Việc chọn điểm khởi tạo ban đầu ảnh hưởng lớn đến kết quả của giải thuật k-means. Nếu các điểm khởi tạo ban đầu được chọn ngẫu nhiên hoặc không tốt, thì kết quả gom cụm có thể không chính xác.

**Độ đo khoảng cách:** Việc lựa chọn độ đo khoảng cách phù hợp sẽ giúp tăng hiệu quả của giải thuật k-means. Độ đo khoảng cách Euclidean thường được sử dụng cho dữ liệu liên tục, trong khi độ đo khoảng cách Hamming thường được sử dụng cho dữ liệu rời rạc.

Giải thuật k-means thường được sử dụng để gom cụm dữ liệu liên tục, tuy nhiên nó cũng có thể được áp dụng cho dữ liệu rời rạc. Trong trường hợp dữ liệu rời rạc, độ đo khoảng cách thường được chọn là độ đo Hamming. Nếu dữ liệu rời rạc chứa quá nhiều giá trị khác nhau, giải thuật k-means có thể cho kết quả không chính xác. Trong trường hợp này, nên sử dụng các phương pháp gom cụm khác phù hợp hơn cho dữ liệu rời rạc, chẳng hạn như phương pháp k-prototypes.

#### [Author 2]