### **Chater 4 (Bayes Classifier)**

**Outline**

[**Chater 4 (Bayes Classifier) 1**](#_67emb7pvcjs7)

[4.1 Question 2](#_vnoem8y0sggx)

[4.1 Answer 2](#_lhcsbnx4cm5b)

[[Tuấn] Áp dụng ct bayes 2](#_5x16k1qtzsm7)

[[Phim]Giống kết quả 2](#_a6fx7s1stmmy)

[[Lâm] Giống kết quả 2](#_u84zosug1wcc)

[4.2 Question 2](#_mmu30nsu6osy)

[4.2 Answer ⇒ cần xem xét lại 2](#_vmo5ib7u6zjf)

[[Trâm +Tuấn ] ⇒ False 2](#_na74ujc7r2x7)

[[Phim] ⇒ 3](#_whgxs2tjlvmd)

[[HX Long] 3](#_ar771m4n35ex)

[4.3 Question 3](#_6021fr5d2hoj)

[4.3 Answer 3](#_t710xn6bjjpz)

[[Tuấn] 3](#_69i1d6e9sw1q)

[[Phim] (bổ sung thêm của a Tuấn) 4](#_qvytgdcju55z)

[K-NN classifier is a supervised lazy classifier which has local heuristics. 4](#_mj7764tna9v2)

[4.4 Question 4](#_cq74sr8l63bo)

[4.4 Answer 4](#_vcqgbyhz7oic)

[[Tuấn] 4](#_n04c1d1jsjr9)

[[Author 2] 5](#_jeo84yubgpzu)

[4.5 Question 5](#_k4je1c6u85kd)

[4.5 Answer 5](#_8j2h0kqj5d3d)

[[Tuấn] 5](#_73vf4bx4tuce)

[[Phim] ⇒ Bổ sung bài a Tuấn: Chỗ tính xác xuất nên cần tính 2 TH thôi, với lại xem lại bảng có sai sót gì ko 7](#_3g3snmq7l5kn)

[4.6. Question 8](#_7yd2iw670xvj)

[4.6 Answer 8](#_who74fnvcj)

[[Tuấn] 8](#_llpunidc13gx)

[[Phim]giống kết quá 9](#_5pc98lfmtiwx)

#### 4.1 Question

A study of a university found out that 15% of undergraduate students who smoke and 23% of graduate students who smoke. If 1/5 students in the university are graduate students and the rest are undergraduate students, what is the probability that a graduate student who smokes?

*Một nghiên cứu của một trường đại học cho thấy 15% sinh viên đại học hút thuốc và 23% sinh viên sau đại học hút thuốc. Nếu 1/5 sinh viên trong trường đại học là sinh viên cao học và số còn lại là sinh viên đại học, xác suất để một sinh viên cao học hút thuốc là bao nhiêu?*

#### 4.1 Answer

##### **[Tuấn] Áp dụng ct bayes**

P(Graduate) = ⅕ = 0.2

P(Undergraduate) = ⅘ = 0.8

P(Smoke|Undergraduate) = 0.15

P(Smoke|Graduate) = 0.23

Áp dụng công thức Bayes:

P(Graduate|Smoke) = P(Smoke|Graduate) \* P(Graduate) / P(Smoke) (1)

**Trong đó: P(Smoke) = P(Smoke|Undergraduate) \* P(Undergraduate) + P(Smoke|Graduate) \* P(Graduate) = 0.15 \* 4/5 + 0.23 \* 1/5 = 0.174 *(chú ý dòng này)***

Thay vào (1):

P(Graduate|Smoke) = 0.15\*0.2/0.174 = 0.2644

##### **[Phim]Giống kết quả**

##### [Lâm] Giống kết quả

P(smoke|graduate)=0.15

P(smoke|undergraduate)=0.23

P(graduate)=0.2

P(undergraduate)=0.8

P(graduate|smoke)=P(graduate).P(smoke|graduate)/P(smoke)

=P(graduate).P(smoke|graduate)/(P(smoke|gradudate).P(graduate) + P(smoke|undergraduate). P(undergraduate))

#### 

#### 4.2 Question

(True/false) If P(A|B) = P(A) then P(A,B) = P(A).P(B).

#### 4.2 Answer ⇒ cần xem xét lại

##### **[Trâm +Tuấn ] ⇒ False**

Nếu P(A|B) = P(A), điều đó có nghĩa là sự kiện B không xảy ra sự kiện A ⇒ A và B là các sự kiện độc lập.

**Tuy nhiên, P(A,B) = P(A).P(B). Phương trình đúng là P(A,B) = P(A|B).P(B) = P(A).P(B) khi và chỉ khi A và B là các biến cố độc lập. Nếu A và B không độc lập thì P(A,B) ≠ P(A).P(B).**

Do đó, Nếu P(A|B) = P(A) thì P(A,B) = P(A).P(B) ⇒ sai.

##### **[Phim] ⇒**

Đúng, vì P(A|B) = P(A) thì A và B độc lập => P(A,B) = P(A).P(B)

slide 23??

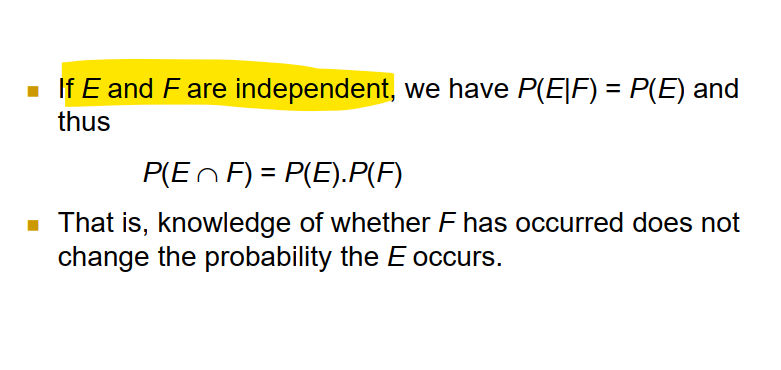
##### **[HX Long]**

Đúng.

Nếu P(A|B) = P(A) thì điều này chỉ ra rằng các biến cố A và B là độc lập, tức là sự xảy ra của biến cố A không ảnh hưởng đến xảy ra của biến cố B, và ngược lại.

Khi hai sự kiện là độc lập, ta có công thức P(A,B) = P(A)P(B|A) = P(A)P(B), vì P(A|B) = P(A) dẫn đến P(B|A) = P(B).

Do đó, nếu P(A|B) = P(A) thì P(A,B) = P(A)P(B).



[Thành]

Đúng.

If P(A|B) = P(A), then P(A,B) = P(A) \* P(B) **only if** A and B are independent events.

#### 4.3 Question

State the difference between k-nearest neighbor algorithm and Naïve Bayes in classification.

*(Nêu sự khác biệt giữa thuật toán k-lân cận gần nhất và Naïve Bayes trong phân loại)*

#### 4.3 Answer

##### **[Tuấn]**

Thuật toán k-lân cận gần nhất (K-NN) và Naïve Bayes đều là các thuật toán học máy thường được sử dụng cho các tác vụ phân loại, nhưng chúng khác nhau trong cách tiếp cận.

Thuật toán k-lân cận gần nhất hoạt động bằng cách tìm ra k ví dụ huấn luyện trong bộ dữ liệu gần nhất với điểm dữ liệu đầu vào dựa trên một phép đo khoảng cách được chọn. Sau đó, thuật toán dự đoán lớp của điểm dữ liệu đầu vào dựa trên lớp đa số của k hàng xóm gần nhất. Giá trị của k được chọn bởi người dùng và việc lựa chọn phép đo khoảng cách có thể ảnh hưởng đáng kể đến hiệu suất của thuật toán.

Trong khi đó, Naïve Bayes là một thuật toán xác suất dự đoán dựa trên xác suất của điểm dữ liệu đầu vào thuộc mỗi lớp. Nó sử dụng định lý Bayes để tính toán xác suất hậu nghiệm của mỗi lớp dựa trên điểm dữ liệu đầu vào và sau đó chọn lớp có xác suất cao nhất làm dự đoán. Naïve Bayes giả định rằng các đặc trưng của điểm dữ liệu đầu vào là độc lập có điều kiện với nhãn lớp, đó là nguyên nhân của việc sử dụng từ "naïve" trong tên gọi của nó.

Một khác biệt chính giữa hai thuật toán này là K-NN là một thuật toán phi tham số, có nghĩa là nó không đưa ra bất kỳ giả định nào về phân phối dữ liệu bên dưới. Ngược lại, Naïve Bayes là một thuật toán tham số giả định một phân phối cụ thể (thường là Gaussian hoặc multinomial) cho các đặc trưng dữ liệu đầu vào.

Một khác biệt khác là K-NN yêu cầu một phép đo khoảng cách để xác định các ví dụ trong bộ dữ liệu gần nhất với điểm dữ liệu đầu vào, trong khi Naïve Bayes thì không cần. Naïve Bayes đơn giản là tính toán xác suất của điểm dữ liệu đầu vào thuộc mỗi lớp dựa trên tần suất quan sát được của các đặc trưng trong dữ liệu huấn luyện.

##### **[Phim] (bổ sung thêm của a Tuấn)**

##### K-NN classifier is a supervised lazy classifier which has local heuristics.

Naive Bayes is an eager learning classifier and it is much faster than K-NN

#### 4.4 Question

State the assumption on the characteristic of the dataset which allows us to apply Naïve Bayes classifier.

*(Nêu giả định về đặc điểm của tập dữ liệu cho phép chúng tôi áp dụng bộ phân loại Naïve Bayes.)*

#### 4.4 Answer

##### **[Tuấn]**

Giả định về đặc điểm của tập dữ liệu cho phép chúng ta áp dụng bộ phân loại Naïve Bayes là các đặc điểm của tập dữ liệu có tính độc lập có điều kiện giữa các đặc trưng (hay còn gọi là các đặc tính), biết trước nhãn lớp. Nói cách khác, bộ phân loại giả định rằng sự xuất hiện hay không xuất hiện của một đặc trưng cụ thể không ảnh hưởng đến sự xuất hiện hay không xuất hiện của bất kỳ đặc trưng nào khác trong tập dữ liệu, biết trước nhãn lớp. Giả định này đơn giản hóa tính toán xác suất hậu nghiệm của một lớp dựa trên các đặc trưng đầu vào, làm cho bộ phân loại Naïve Bayes tính toán hiệu quả và hiệu quả với các tập dữ liệu có số chiều cao. Tuy nhiên, trong thực tế, các đặc trưng của tập dữ liệu có thể tương quan và giả định này không đúng với tất cả các trường hợp, điều này có thể ảnh hưởng đến độ chính xác của bộ phân loại.

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

#### 4.5 Question

Consider the following data set: *(Hãy xem xét tập dữ liệu sau)*

| Feature 1 | Feature 2 | Feature 3 | Class |
| --- | --- | --- | --- |
| 0  1  1  1  0  0 | 0  0  0  1  1  1 | 0  1  0  1  1  1 | 0  1  0  1  1  0 |

If we have a test pattern *P* with feature 1 as 0 and feature 2 as 0 and feature 3 as 1, classify this pattern using Naïve Bayes classifier.

*(Nếu chúng ta có mẫu thử nghiệm P với đặc điểm 1 là 0 và đặc điểm 2 là 0 và đặc điểm 3 là 1, hãy phân loại mẫu này bằng bộ phân loại Naïve Bayes.)*

#### 4.5 Answer

##### **[Tuấn]**

First, we need to calculate the probabilities of each feature given each class:

P(feature 1 = 0 | class = 0) = 2/3

P(feature 1 = 1 | class = 0) = 1/3

P(feature 1 = 0 | class = 1) = 1/3

P(feature 1 = 1 | class = 1) = 2/3

P(feature 2 = 0 | class = 0) = 2/3

P(feature 2 = 1 | class = 0) = 1/3

P(feature 2 = 0 | class = 1) = 1/3

P(feature 2 = 1 | class = 1) = 2/3

P(feature 3 = 0 | class = 0) = 1/3

P(feature 3 = 1 | class = 0) = 2/3

P(feature 3 = 0 | class = 1) = 1/3

**P(feature 3 = 1 | class = 1) = 2/3**

Next, we can calculate the prior probabilities of each class:

P(class = 0) = 3/6 = 1/2

P(class = 1) = 3/6 = 1/2

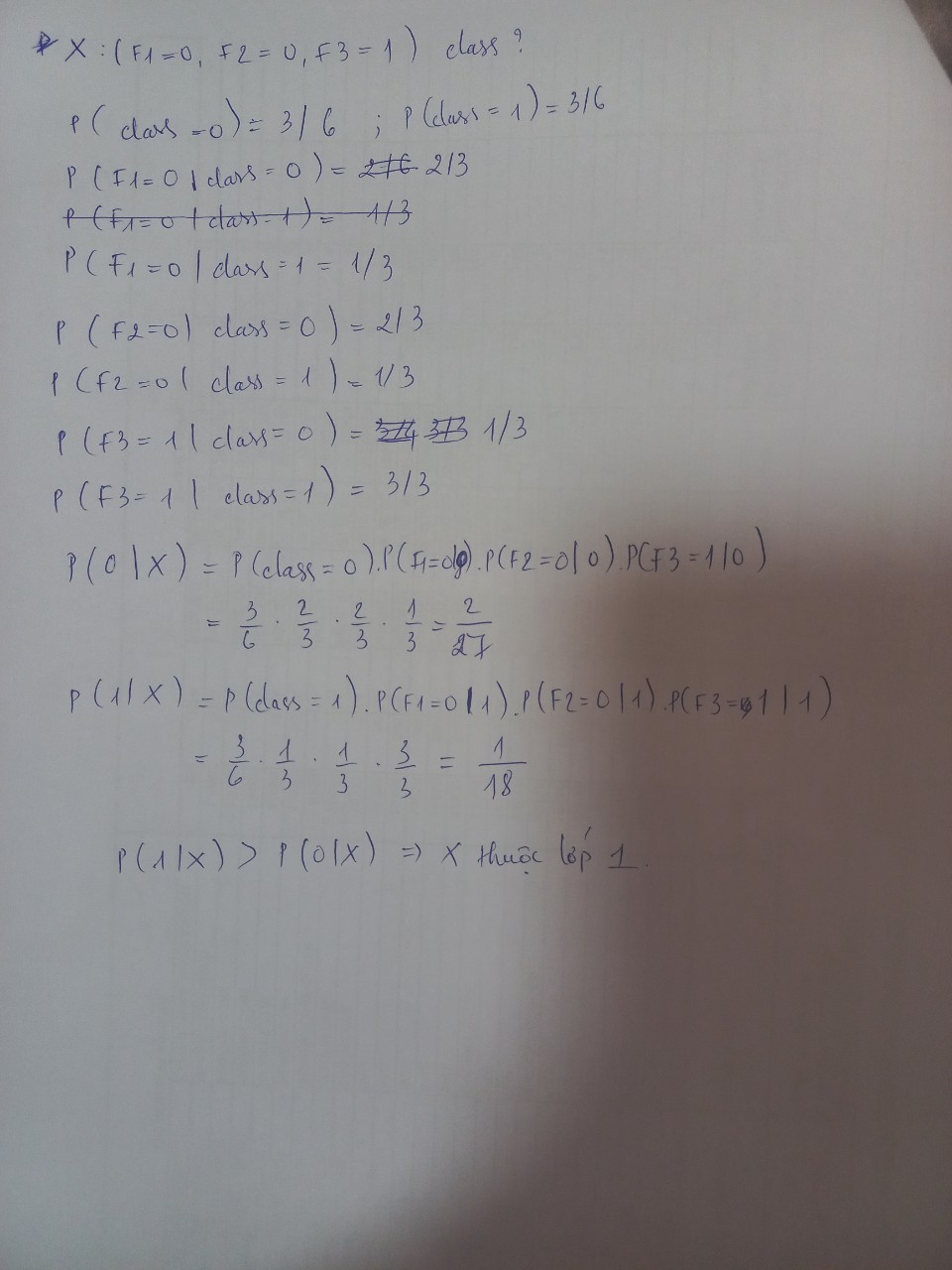
Now we can calculate the likelihood of the features given each class:

P(features = (0, 0, 1) | class = 0) = P(feature 1 = 0 | class = 0) \* P(feature 2 = 0 | class = 0) \* P(feature 3 = 1 | class = 0) = 2/27

P(features = (0, 0, 1) | class = 1) = P(feature 1 = 0 | class = 1) \* P(feature 2 = 0 | class = 1) \* P(feature 3 = 1 | class = 1) = 1/27

→ P(features = (0, 0, 1) | class = 0) > P(features = (0, 0, 1) | class = 1) → pattern P belongs to class 0

##### **[Phim] ⇒ Bổ sung bài a Tuấn: Chỗ tính xác xuất nên cần tính 2 TH thôi, với lại xem lại bảng có sai sót gì ko**

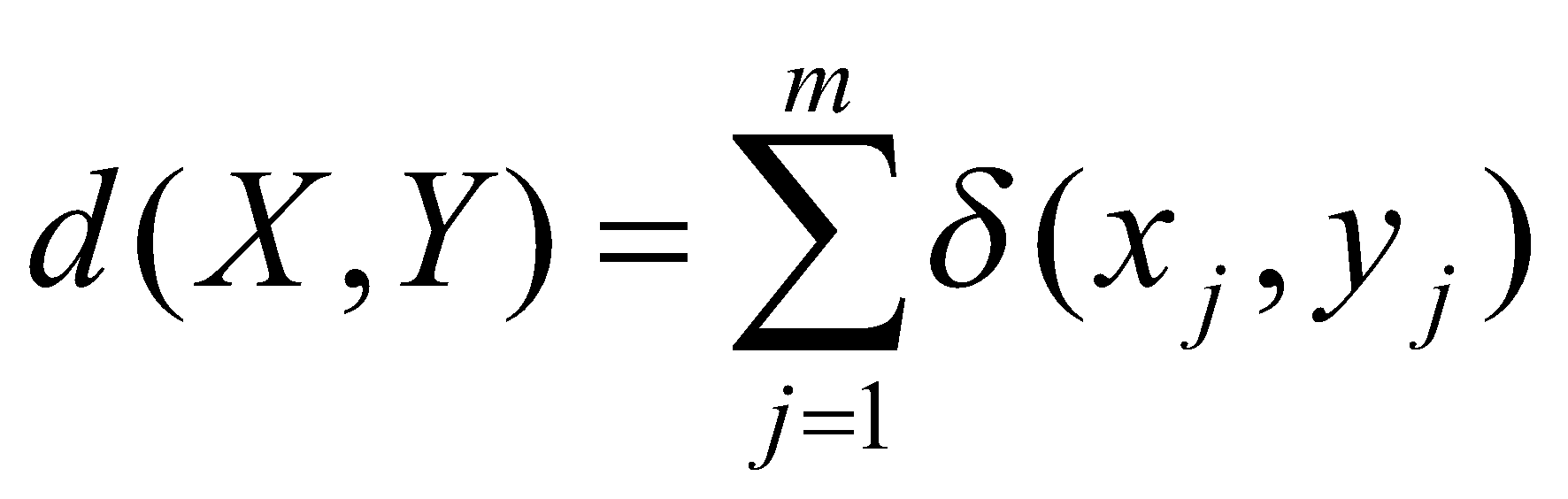


[HX Long]: **X thuộc lớp 0** do P(C=0|X) = 0.074074 > 0.055556 = P(C=1|X)

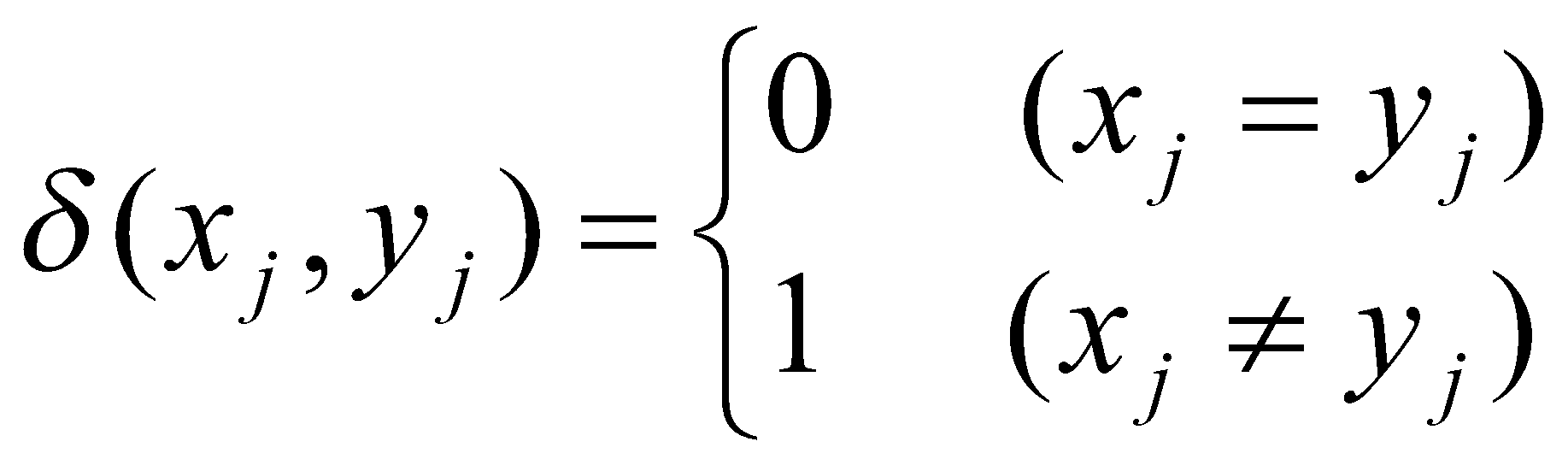
#### 4.6. Question

Given a dataset as in Exercise 4.4. Since the attributes are not continuous, we apply the following method to calculate the distance between two patterns with categorical attributes. Given two patterns, each consists of *m* categorical attributes. The distance between *X* and *Y* is total number of differences between the corresponding attribute values of the two patterns. The total number of differences is smaller, the two patterns more similar. That means:

*Cho một tập dữ liệu như trong Bài tập 4.4. Vì các thuộc tính không liên tục, nên chúng tôi áp dụng phương pháp sau để tính khoảng cách giữa hai mẫu có thuộc tính phân loại. Cho hai mẫu, mỗi mẫu bao gồm m thuộc tính phân loại. Khoảng cách giữa X và Y là tổng số khác biệt giữa các giá trị thuộc tính tương ứng của hai mẫu. Tổng số điểm khác biệt càng nhỏ thì hai mẫu càng giống nhau. Điều đó có nghĩa là:*



where



By using this distance measure, apply 1-nearest neighbor algorithm to classify the test pattern *P* = (0, 0, 1), based on the dataset given in Exercise 4.4.

Compare the results of the two classification method: 1-nearest neighbor algorithm and Naïve Bayes (Exercise 4.4).

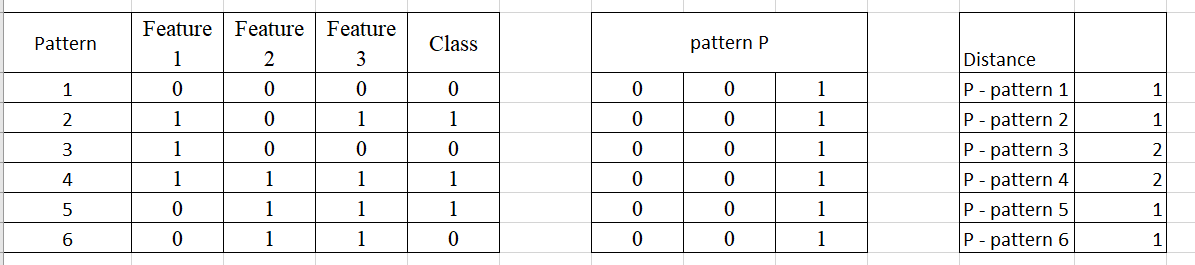
*Bằng cách sử dụng thước đo khoảng cách này, hãy áp dụng thuật toán 1-láng giềng gần nhất để phân loại mẫu kiểm tra P = (0, 0, 1), dựa trên bộ dữ liệu được đưa ra trong Bài tập 4.4.*

*So sánh kết quả của hai phương pháp phân loại: thuật toán 1-láng giềng gần nhất và Naïve Bayes (Bài tập 4.4).*

#### 4.6 Answer

##### **[Tuấn]**

Ta có bảng số liệu:



Áp dụng thuật toán k-NN với k = 1, có 4 pattern có distance = 1 là 1,2,5,6. Ta khó kết luận được class của P trong trường hợp này.

Ta nhận thấy thuật toán Naive Bayes hiệu quả hơn k-NN trong trường hợp này.

##### **[Phim]giống kết quá**

