**Chapter 5 (Decision Trees)**

[**5.1. Question 1**](#_v9m825fu0k20)

[**5.1 Answer 1**](#_bh3p3y2pc4p6)

[[Nghia Pham] 1](#_5x16k1qtzsm7)

[[Tú] 1](#_a6fx7s1stmmy)

[(Nhân) 2](#_u3ozpz907qx7)

[**5.2. Question 2**](#_lkfxr6fkn0or)

[**5.2 Answer 3**](#_a48h6yeu0748)

[[Tú] 3](#_4d1dhvbumzo)

[[Author 2] 3](#_ubxpgsy65nl2)

[**5.3. Question 4**](#_t853ut7winb5)

[**5.3 Answer 5**](#_kpbv1dh9gfuz)

[[Tú] 5](#_tb0lrh4eu1tx)

[**5.4 Question 6**](#_i4bvh7uqv72y)

[**5.4 Answer 6**](#_bixjhgfbw4w)

[[Nghia Pham] 6](#_goua7zx6bt60)

[[Tú] 6](#_jrmw3y7x5sx8)

[[Nhân] 6](#_nh0rg3xuejw)

#### 5.1. Question

Determine the entropy impurity for the following distributions.

a) The dataset has 1/2 of patterns belonging to the first class, 1/4 of patterns belonging to the second class, 1/8 of patterns belonging to the third class, 1/16 of patterns belonging to the fourth class, and 1/16 of patterns belonging to the fifth class.

b) The dataset consists of five classes and each class has 1/5 of patterns.

*Xác định tạp chất entropy cho các phân phối sau.*

*a) Tập dữ liệu có 1/2 số mẫu thuộc loại thứ nhất, 1/4 số mẫu thuộc loại thứ hai, 1/8 số mẫu thuộc loại thứ ba, 1/16 số mẫu thuộc loại thứ tư và 1/16 số mẫu thuộc lớp thứ năm.*

*b) Tập dữ liệu gồm 5 lớp và mỗi lớp có 1/5 số mẫu.*

#### 5.1 Answer

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

Công thức:

Entropy = -∑(p\_i \* log\_2(p\_i))

where p\_i is the proportion of the instances of class i in a set

1. Entropy impurity = - (1/2)log2(1/2) - (1/4)log2(1/4) - (1/8)log2(1/8) - (1/16) log2(1/16) - (1/16)log2(1/16) = 1.9375
2. Entropy impurity = 5 \* (- (1/5)log2(1/5)) = 2.3219

##### **[Tú]**

a) Entropy impurity theo công thức:

i(N1) = -0.5 × log2(0.5) - 0.25 × log2(0.25) - 0.125 × log2(0.125) - 0.0625 × log2(0.0625) -0.0625 × log2(0.0625) = 1.875

b) Tương tự:

i(N2) = -0.2 × log2(0.2) × 5 = 2.3219

##### **(Nhân)**

a) The entropy impurity at a node *N* is *i*(*N*) and is given by

*P*(*w1*) = 1/2 = 0,5; *P*(*w2*) = 1/4 = 0,25; *P*(*w3*) = 1/8 = 0,125; *P*(*w4*) = 1/16 = 0,0625; *P*(*w3*) = 1/16 = 0,0625.

And entropy *i*(*N*) = -0,5log20.5 - 0,25log20,25 – 0,125log20,125 – 0,0625log20,0625 – 0,0625log20,0625 = **1.9375.**

b) The entropy impurity at a node *N* is *i*(*N*) and is given by

*P*(*w1*) = *P*(*w2*) = *P*(*w3*) = *P*(*w4*) = *P*(*w5*) = 1/5 = 0,2

And entropy *i*(*N*) = -5\*0,2log20.2 = **2.3219**

#### 

#### 5.2. Question

Consider the following data set for a binary classification problem. Each pattern has two binary attributes and one class label (+ or -).

*Hãy xem xét tập dữ liệu sau đây cho một vấn đề phân loại nhị phân. Mỗi mẫu có hai thuộc tính nhị phân và một nhãn lớp (+ hoặc -).*

| A | B | Class label |
| --- | --- | --- |
| T | F | + |
| T | T | + |
| T | T | + |
| T | F | - |
| T | T | + |
| F | F | - |
| F | F | - |
| F | F | - |
| T | T | - |
| T | F | - |

Let use the information gain when determining the splitting attribute. Which of the features is selected as splitting attribute at the root node in the decision tree for the data set.

*(Hãy sử dụng thông tin thu được khi xác định thuộc tính phân tách. Tính năng nào được chọn làm thuộc tính phân tách tại nút gốc trong cây quyết định cho tập dữ liệu.)*

#### 5.2 Answer

##### **[Tú]**

Có 4 nhãn thuộc lớp + và 6 nhãn thuộc lớp – trong 10 mẫu nhãn lớp

i(N) = -(4/10) × log2(4/10) – (6/10) × log2(6/10) = **0.971**

**Với thuộc tính A**,

+ T có 7 mẫu trong đó có 4 thuộc lớp + và 3 thuộc lớp –

i(NT) = -(4/7) × log2(4/7) – (3/7) × log2(3/7) = 0.9852

+ F có 3 mẫu trong đó có 3 thuộc lớp – nên entropy i(NF) bằng 0

Vậy nên drop in entropy hay Gain được tính:

Gain(A) (hay ∆i(N)) = 0.971 – 0.9852 × (7/10) – 0 × (3/10) = **0.2814**

**Với thuộc tính B**,

+ T có 4 mẫu trong đó có 3 thuộc lớp + và 1 thuộc lớp –

i(NT) = -(3/4) × log2(3/4) – (1/4) × log2(1/4) = 0.8113

+ F có 6 mẫu trong đó có 1 thuộc lớp + và 5 thuộc lớp –

i(NT) = -(1/6) × log2(1/6) – (5/6) × log2(5/6) = 0.6500

Vậy Gain được tính:

Gain(B) = 0.971 – 0.8113 × (4/10) – 0.65 × (6/10) = **0.2565**

Vậy thuộc tính B được chọn để làm thuộc tính phân lớp ở nút gốc.

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

Split point  là thuộc tính có info gain lớn nhất vì denta(i) max => tìm min ∑(p\_j \* i(Nj))

**Với thuộc tính A**,

+ T có 7 mẫu trong đó có 4 thuộc lớp + và 3 thuộc lớp –

i(NT) = -(4/7) × log2(4/7) – (3/7) × log2(3/7) = 0.9852

+ F có 3 mẫu trong đó có 3 thuộc lớp – nên entropy i(NF) bằng 0

**Với thuộc tính B**,

+ T có 4 mẫu trong đó có 3 thuộc lớp + và 1 thuộc lớp –

i(NT) = -(3/4) × log2(3/4) – (1/4) × log2(1/4) = 0.8113

+ F có 6 mẫu trong đó có 1 thuộc lớp + và 5 thuộc lớp –

i(NT) = -(1/6) × log2(1/6) – (5/6) × log2(5/6) = 0.6500

Vậy ∑(p\_j \* i(Nj)) được tính:

Vậy thuộc tính B được chọn để làm thuộc tính phân lớp ở nút gốc.

#### 5.3. Question

Consider the following Weather dataset for a binary classification problem. Each pattern has four discrete attributes and one class label (Yes or No).

*Hãy xem xét tập dữ liệu Thời tiết sau đây cho một vấn đề phân loại nhị phân. Mỗi mẫu có bốn thuộc tính rời rạc và một nhãn lớp (Có hoặc Không).*

| Outlook | Temperature | Humidity | Windy | Play Tenis |
| --- | --- | --- | --- | --- |
| Sunny | Hot | High | False | No |
| Sunny | Hot | High | False | No |
| Overcast | Hot | High | False | Yes |
| Rainy | Mild | High | False | Yes |
| Rainy | Cool | Normal | False | Yes |
| Rainy | Cool | Normal | True | No |
| Overcast | Cool | Normal | True | Yes |
| Sunny | Warm | High | False | No |
| Sunny | Cool | Normal | False | Yes |
| Rainy | Warm | Normal | False | Yes |
| Sunny | Warm | Normal | True | Yes |
| Overcast | Warm | High | True | Yes |
| Overcast | Hot | Normal | False | Yes |
| Rainy | Warm | High | True | No |

Let use the information gain when determining the splitting attribute. Which of the features is selected as splitting attribute at the root node in the decision tree for the data set.

*(Hãy sử dụng thông tin thu được khi xác định thuộc tính phân tách. Tính năng nào được chọn làm thuộc tính phân tách tại nút gốc trong cây quyết định cho tập dữ liệu.)*

#### 5.3 Answer

##### 

##### **[Tú]**

Ta có 14 nhãn lớp trong đó có 5 thuộc lớp No và 9 thuộc lớp Yes

i(N) = -(5/14) × log2(5/14) – (9/14) × log2(9/14) = **0.9403**

Tương tự như câu trên ta có 4 thuộc tính cần xét:

**Outlook:**

+ 5 Sunny: 3 No, 2 Yes

i(Nsunny) = -(3/5) × log2(3/5) – (2/5) × log2(2/5) = 0.971

+ 5 Rainy: 2 No, 3 Yes

i(Nrainy) = -(2/5) × log2(2/5) – (3/5) × log2(3/5) = 0.971

+ 4 Overcast: 4 Yes nên i(Novercast) = 0

Gain (Outlook) = 0.9403 – 5/14 × 0.971 – 5/14 × 0.971 – 4/14 × 0 = **0.2467**

**Temperature:**

+ 4 Hot: 2 No, 2 Yes

i(Nhot) = -(2/4) × log2(2/4) – (2/4) × log2(2/4) = 1

+ 4 Cool: 1 No, 3 Yes

i(Ncool) = -(1/4) × log2(1/4) – (3/4) × log2(3/4) = 0.8113

+ 1 Mild: 1 Yes nên i(Nmild) = 0

+ 5 Warm: 2 No, 3 Yes

i(Nwarm) = -(2/5) × log2(2/5) – (3/5) × log2(3/5) = 0.971

Gain (Temperature) = 0.9403 – 4/14 × 1 – 4/14 × 0.8113 – 1/14 × 0 – 5/14 ×0.971 = **0.076**

**Humidity:**

+ 7 High: 4 No, 3 Yes

i(Nhigh) = -(4/7) × log2(4/7) – (3/7) × log2(3/7) = 0.9852

+ 7 Normal: 1 No, 6 Yes

i(Nnormal) = -(1/7) × log2(1/7) – (6/7) × log2(6/7) = 0.5917

Gain (Humidity) = 0.9403 – 7/14 × 0.9852 – 7/14 × 0.5917 = **0.15185**

**Windy:**

+ 9 False: 3 No, 6 Yes

i(Nhigh) = -(3/9) × log2(3/9) – (6/9) × log2(6/9) = 0.9183

+ 5 True: 2 No, 3 Yes

i(Nnormal) = -(2/5) × log2(2/5) – (3/5) × log2(3/5) = 0.971

Gain (Windy) = 0.9403 – 9/14 × 0.9183 – 5/14 × 0.971 = **0.0032**

Vậy thuộc tính Outlook được chọn làm thuộc tính phân lớp ở nút gốc

#### 5.4 Question

(True/false) The depth of a learned decision tree can be larger than the number of training examples used to create the tree.

*(Đúng/sai) Độ sâu của cây quyết định đã học có thể lớn hơn số ví dụ huấn luyện được sử dụng để tạo cây.*

#### 5.4 Answer

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

False vì một tranining example không xuất hiện quá 1 lần trong cây -> chiều sâu của cây không thể vượt quá số traning examples.

##### **[Tú]**

Sai vì mỗi nhánh của cây tương được với một mẫu huấn luyện, do đó, nếu có n mẫu huấn luyện, thì cây có chiều cao tối đa là n.

#### 

##### **[Nhân]**

Sai vì số nhánh ví dụ huấn luyện trong cây quyết định luôn lớn hơn độ sâu của cây quyết định.