Thuật toán cây quyết định là một cây trong đó các nút đại diện cho các tính năng (thuộc tính), nhánh đại diện cho quyết định (quy tắc) và các nút lá đại diện cho các kết quả (rời rạc và liên tục).

Có nhiều thuật toán khác nhau được sử dụng để tạo cây quyết định từ dữ liệu, một số thuật toán như sau:

Classification and regression tree CART

ID 3(Information Gain)

CHAID

ID 4.5

Trong phần trình bày này đề cập đến 2 thuật toán:

Thuật toán CART

* GINI Index

Thuật toán ID3

* Hàm Entropy
* Information Gain

# Thuật toán CART

- Được sử dụng để tạo ra cả cây phân loại và cây hồi quy.

- Được sử dụng cho binary classification.

- Nó sử dụng bình phương nhỏ nhất làm số liệu để chọn các tính năng trong trường hợp Cây hồi quy.

*Dùng thuật toán CART để tạo ra classification tree trong bộ dữ liệu thời tiết với mục tiêu là thời tiết có đảm bảo cho việc đi chơi hay không (có hay không).*

Table

Description automatically generated

Theo bộ dữ liệu thời tiết thì có thể dễ dàng thấy được **outlook** sẽ bao gồm các feature: **sunny, overcast, rainfall.**

Table

Description automatically generated

Công thức tính giá trị GINI

**- Tính giá trị Gini cho Outlook**

**- Tính giá trị Gini cho Temperature**

Table

Description automatically generated

**- Tính giá trị Gini cho Humidity**

Chart, box and whisker chart

Description automatically generated

**- Tính giá trị Gini cho Wind**

Table

Description automatically generated

Table

Description automatically generated

Dựa vào bảng trên ta thấy Outlook có giá trị GINI là thấp nhất nên chọn Outlook là root của cây

Diagram

Description automatically generated

Sau khi tìm được root là Outlook thì chúng ta sẽ tính Giá trị GINI cho từng dữ liệu phụ của Outlook bao gồm: **sunny, overcast và rainfall.**

Table

Description automatically generated

**Trường hợp sunny:**

* **Chỉ số Gini cho Temperature khi outlook = sunny**

Table

Description automatically generated

* **Chỉ số Gini cho Humidity khi outlook = sunny**

Table

Description automatically generated

* **Chỉ số Gini cho Wind khi outlook = sunny**

Table

Description automatically generated

Ta thấy **Humidity** có giá trị nhỏ nhất vì thế node tiếp theo là **Humidity**

Diagram

Description automatically generated

**Trường hợp overcast:**

Table

Description automatically generated

Ta thấy trong trường hợp **overcast** thì quyết định luôn là **YES**, do đó giá trị của mỗi feature đều là 0 hay nghĩa là node lá.

Diagram

Description automatically generated

**Trường hợp High Humidity**

Table

Description automatically generated

**Trường hợp Normal Humidity**

Table

Description automatically generated

Từ 2 bảng trên ta luôn thấy **High Humidity** thì quyết định luôn là **NO**, còn **Normal Humidity** thì quyết định luôn là **YES**, vậy nên có 2 node lá

Diagram

Description automatically generated

**Trường hợp outlook = rainfall (**cần tìm giá trị GINI cho temperature, humidity và wind**)**

Table

Description automatically generated

* **Chỉ số Gini cho Temperature khi outlook = rainfall**

Calendar

Description automatically generated with low confidence

* **Chỉ số Gini cho Humidity khi outlook = rainfall**

Calendar

Description automatically generated with medium confidence

* **Chỉ số Gini cho Wind khi outlook = rainfall**

Calendar

Description automatically generated

Table

Description automatically generated

Giá trị GINI của **Wind** là nhỏ nhất nên node tiếp theo là **Wind**

Diagram

Description automatically generated

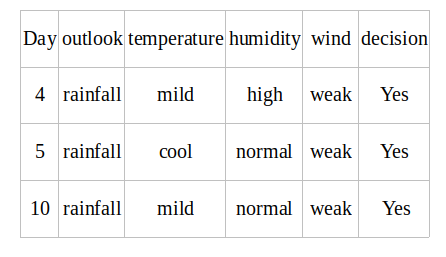
Tiếp theo là phân tích 2 thuộc tính có trong Wind là strong và weak

**Trường hợp strong**

Table

Description automatically generated

**Trường hợp weak**



Dựa vào 2 bảng trên thì ta thấy nếu **wind: strong** thì quyết định là **NO** ngược lại nếu **wind: weak** thì quyết đinh là **YES** => có 2 node lá tiếp theo

Diagram

Description automatically generated

# Thuật toán ID3

Câu hỏi đặt ra: Làm thế nào để chọn ra được node root

Thuộc tính phân loại tốt nhất dữ liệu đào tạo, hãy sử dụng thuộc tính này thuộc tính ở gốc của cây.

Vậy làm thế nào để chọn ra được thuộc tính tốt nhất để làm node root, đó là lý do mà thuật toán ID3 được tạo ra.

Thuật toán tìm ra cây quyết định sẽ theo quy tắc tại các **node root** thì độ hỗn loạn của dữ liệu sẽ là **cao nhất** và không biết nó thuộc về lớp nào, còn tại các **node lá** thì độ hỗn loạn của dữ liệu **thấp nhất** và biết nó thuộc về lớp nào *(có nghĩa là giảm Entropy về 0* ***nhanh nhất có thể, độ sâu của cây quyết định thấp nhất****).*

Thuật toán dừng khi tất cả các thuộc tính thuộc cùng 1 lớp (hay nói các cách là đến node lá)

**Entropy:** Là thước đo về độ hỗn loạn của thông tin. *(là một khái niệm được lấy trong môn vật lý cụ thể là môn nhiệt động lực được đo bằng tổng động năng của các hạt bên trong một khối vật chất, Entropy càng cao thì độ chuyển động của các hạt càng cao hay nói cách khác là các hạt chuyển động càng hỗn loạn dẫn đến việc rất khó để nhận biết được 1 hạt đang ở vị trí nào).*

Hay Entropy còn được hiểu là khả năng mà đoán đúng được thông tin.

VD: Trong một cuộc bầu cử, ban đầu khi chưa phiếu bầu chưa được kiểm định thì độ hỗn loạn của các phiếu bầu cao dẫn đến khả năng đoán chính xác phiếu bầu cho ai rất thấp, khi số lượng phiếu được kiểm định xong tăng lên thì Entropy cũng sẽ tăng và Entropy = 0 khi tất cả các phiếu đã được kiểm định.

Công thức toán học cho Entropy như sau:

Average Information

**Information Gain:**

It is the measure of decrease in entropy after the dataset is split.

Diagram

Description automatically generated

Nhận xét Entropy càng cao thì mức độ rối loạn của dữ liệu càng lớn.

Sử dụng thuật toán ID3 để tạo cây quyết đinh trên tập dữ liệu thời tiết

Table

Description automatically generated

Tính Entropy tổng

Tính Entropy cho mỗi thuộc tính

* **Outlook**

Table

Description automatically generated

Tính Average Information Entropy

Tính Gain:

* **Temperature**

|  |  |  |  |
| --- | --- | --- | --- |
| Temperature | p | n | Entropy |
| Hot | 2 | 2 | 1 |
| Mild | 4 | 2 | 0.918 |
| Cool | 3 | 1 | 0.811 |

Tính Average Information Entropy

Tính Gain:

* **Humidity**

|  |  |  |  |
| --- | --- | --- | --- |
| Humidity | p | n | Entropy |
| High | 3 | 4 | 0.985 |
| Normal | 6 | 1 | 0.591 |

Tính Average Information Entropy

Tính Gain:

* **Windy**

|  |  |  |  |
| --- | --- | --- | --- |
| Windy | p | n | Entropy |
| Strong | 3 | 3 | 1 |
| Weak | 6 | 2 | 0.811 |

Tính Average Information Entropy

Tính Gain:

|  |  |
| --- | --- |
| Attributes | Gain |
| Outlook | 0.247 |
| Temperature | 0.029 |
| Humidity | 0.152 |
| windy | 0.048 |

Node root: Outlook

Diagram

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outlook | Temperature | Humidity | Windy | Decision |
| Overcast | Hot | High | Weak | Yes |
| Overcast | Cool | Normal | Strong | Yes |
| Overcast | Mild | High | Strong | Yes |
| Overcast | Hot | Normal | Weak | Yes |

Khi Outlook = overcast thì quyết đinh luôn là YES nên ta có node lá

Diagram

Description automatically generated

* **Tính Entropy khi Outlook = Sunny**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outlook | Temperature | Humidity | Windy | Decision |
| Sunny | Hot | High | Weak | No |
| Sunny | Hot | High | Strong | No |
| Sunny | Mild | High | Weak | No |
| Sunny | Cool | Normal | Weak | Yes |
| Sunny | Mild | Normal | Strong | Yes |

|  |  |  |  |
| --- | --- | --- | --- |
| outlook | p | n | Entropy |
| sunny | 2 | 3 |  |

* **Tính Entropy cho Humidity khi Outlook = Sunny**

|  |  |  |
| --- | --- | --- |
| Outlook | Humidity | Decision |
| Sunny | High | No |
| Sunny | High | No |
| Sunny | High | No |
| Sunny | Normal | Yes |
| Sunny | Normal | Yes |

|  |  |  |  |
| --- | --- | --- | --- |
| Temperature | p | n | Entropy |
| High | 0 | 3 | 0 |
| Normal | 2 | 0 | 0 |

* **Tính Entropy cho Windy khi Outlook = Sunny**

|  |  |  |
| --- | --- | --- |
| Outlook | Temperature | Decision |
| Sunny | Strong | No |
| Sunny | Strong | Yes |
| Sunny | Weak | No |
| Sunny | Weak | No |
| Sunny | Weak | Yes |

|  |  |  |  |
| --- | --- | --- | --- |
| Temperature | p | n | Entropy |
| Strong | 1 | 1 | 1 |
| Weak | 1 | 2 | 0.918 |

* **Tính Entropy cho Temperature khi Outlook = Sunny**

|  |  |  |
| --- | --- | --- |
| Outlook | Temperature | Decision |
| Sunny | Cool | Yes |
| Sunny | Hot | No |
| Sunny | Hot | No |
| Sunny | Mild | No |
| Sunny | Mild | Yes |

|  |  |  |  |
| --- | --- | --- | --- |
| Temperature | p | n | Entropy |
| Cool | 1 | 0 | 0 |
| Hot | 0 | 2 | 0 |
| Mild | 1 | 1 | 1 |

|  |  |
| --- | --- |
| Attributes | Gain |
| Temperature | 0.571 |
| Humidity | 0.971 |
| Windy | 0.02 |

**Humidity** được chọn làm node tiếp theo.

|  |  |  |
| --- | --- | --- |
| Outlook | Humidity | Decision |
| Sunny | High | No |
| Sunny | High | No |
| Sunny | High | No |
| Sunny | Normal | Yes |
| Sunny | Normal | Yes |

Khi đó ta thấy ứng với Humidity = Hight thì quyết định là No, Humidity = Normal thì quyết định là Yes. Do đó chúng ta có 2 node lá tiếp theo

Diagram

Description automatically generated

* **Tính Entropy khi Outlook = Rainy**

Table

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| outlook | p | n | Entropy |
| rainy | 3 | 2 |  |

* **Tính Entropy cho Humidity khi Outlook = Rainy**

|  |  |  |
| --- | --- | --- |
| outlook | Humidity | Decision |
| Rainy | High | Yes |
| Rainy | High | No |
| Rainy | Normal | Yes |
| Rainy | Normal | No |
| Rainy | Normal | Yes |

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | p | n | Entropy |
| Strong | 1 | 1 | 1 |
| Weak | 2 | 1 | 0.918 |

* **Tính Entropy cho Windy khi Outlook = Rainy**

|  |  |  |
| --- | --- | --- |
| outlook | Windy | Decision |
| Rainy | Strong | No |
| Rainy | Strong | No |
| Rainy | Weak | Yes |
| Rainy | Weak | Yes |
| Rainy | Weak | Yes |

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | p | n | Entropy |
| Strong | 0 | 2 | 0 |
| Weak | 3 | 0 | 0 |

* **Tính Entropy cho Temperature khi Outlook = Rainy**

|  |  |  |
| --- | --- | --- |
| outlook | Temperature | Decision |
| Rainy | Mild | Yes |
| Rainy | Cool | Yes |
| Rainy | Cool | No |
| Rainy | Mild | Yes |
| Rainy | Mild | No |

|  |  |  |  |
| --- | --- | --- | --- |
| Attributes | p | n | Entropy |
| Cool | 1 | 1 | 1 |
| Mild | 2 | 1 | 0.918 |

|  |  |
| --- | --- |
| Attributes | Gain |
| Humidity | 0.02 |
| Windy | 0.971 |
| Temperature | 0.02 |

Windy là node tiếp theo

Diagram

Description automatically generated

<https://blog.tensorflow.org/2021/05/introducing-tensorflow-decision-forests.html>

<https://www.yworks.com/pages/interactive-decision-tree-diagrams>

<https://www.youtube.com/watch?v=0dVZxET_oLM&t=68s>

<https://github.com/thangnch/MIAI_DecisionTree/blob/main/Tree.ipynb>

https://ankitnitjsr13.medium.com/math-behind-decision-tree-algorithm-2aa398561d6d

<https://trituenhantao.io/kien-thuc/decision-tree/>

<https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1>

https://towardsdatascience.com/entropy-and-information-gain-in-decision-trees-c7db67a3a293