**Decision Tree**

**Cây quyết định**

**Cây quyết định là gì ?**

age=?

*middle\_aged*

*youth*

*senior*

Credit\_rating=?

student=?

*excellent*

*fair*

*yes*

*no*

* Một cây quyết định có thể được biểu diễn (diễn giải) bằng một tập các luật IF-THEN

|  |
| --- |
| *Ví dụ :*  IF age = middle\_aged THEN  buy\_computer = yes  ELSE IF age = youth THEN  IF student = yes THEN  buy\_computer = yes  ELSE IF student = no THEN  buy\_computer = no  ELSE IF age = senior THEN  IF credit\_rating = fair THEN  buy\_computer = yes  ELSE IF credit\_rating = excellent THEN  buy\_computer = no |

1. **Thuật toán ID3**

* Entropy của tập S đối với việc phân lớp có c lớp c

*trong đó là tỷ lệ các mẫu trong tập S thuộc vào lớp i, và quy ước*

* Entropy của tập S đối với việc phân lớp có 2 lớp :

|  |
| --- |
| *Ví dụ :*  Tập S gồm 14 mẫu, trong đó 9 mẫu thuộc về lớp c1 và 5 mẫu thuộc về lớp c2.  Entropy của tập S đối với phân lớp có 2 lớp này là:  Info(S) = - (9/14).log2 (9/14) - (5/14).log2 (5/14) = 0.94 |

* Information Gain của biến A đối với tập S:

*trong đó là từng giá trị trong tập các giá trị Value(A)*

* Chọn thuộc tính có Gain(A)lớn nhất làm nút gốc. Và chia thành n node tương ứng với n giá trị của thuộc tính đó. Với tùng nút tính thực hiện với

**Bài tập ví dụ: Tìm cây quyết định với thuật toán ID3 của tập dữ liệu sau:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *ID* | *age* | *income* | *student* | *credit\_rating* | *buys\_computer* |
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle\_aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle\_aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle\_aged | medium | no | excellent | yes |
| 13 | middle\_aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

*Bảng 1*

* **Bước 1:**

Ta thấy nhãn *buys\_computer* có 5 : *no*, 9 : *yes* 🡪 Gain của tập dữ liệu D

=0.940286

* **Xét thuộc tính *age***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *youth* | 2 | 3 | 5 |
| *middle\_aged* | 4 | 0 | 4 |
| *senior* | 3 | 2 | 5 |

= 0.693536

0.693536=0.24675

* **Xét thuộc tính *income***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *high* | 2 | 2 | 4 |
| *medium* | 4 | 2 | 6 |
| *low* | 3 | 1 | 4 |

= 0.911063

0.911063=0.029223

* **Xét thuộc tính *student***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *yes* | 3 | 4 | 7 |
| *no* | 6 | 1 | 7 |

= 0.788450

0.788450=0.151836

* **Xét thuộc tính *credit\_rating***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *fair* | 6 | 2 | 8 |
| *excellent* | 3 | 3 | 6 |

=0.892159

0.892159=0.048127

Ta thấy =0.24675 lớn nhất 🡪 *age* được chọn là nút gốc

age=?

*middle\_aged*

*senior*

*youth*

* **Bước 2**

1. **Xét với tâp Dage=youth**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *ID* | *age* | *income* | *student* | *credit\_rating* | *buys\_computer* |
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |

Ta thấy nhãn *buys\_computer* có 3 : no, 2 : yes 🡪 Gain của tập dữ liệu Dage=youth

= 0.970951

* **Xét thuộc tính *income***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *high* | 0 | 2 | 2 |
| *medium* | 1 | 1 | 2 |
| *low* | 1 | 0 | 1 |

= 0.4

0.4=0.570951

* **Xét thuộc tính *student***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *yes* | 2 | 0 | 2 |
| *no* | 0 | 3 | 3 |

= 0

0=

* **Xét thuộc tính *credit\_rating***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *fair* | 1 | 2 | 3 |
| *excellent* | 1 | 1 | 2 |

=0.950978

0.950978=

Ta thấy = lớn nhất 🡪 *student* được chọn là nút gốc của nhánh *age=youth*

age=?

*Middle\_aged*

*youth*

*senior*

student=?

*yes*

*no*

1. **Xét với tâp Dage=senior**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *ID* | *age* | *income* | *student* | *credit\_rating* | *buys\_computer* |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 10 | senior | medium | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

Ta thấy nhãn *buys\_computer* có 3 : yes, 2 : no 🡪 Gain của tập dữ liệu Dage=senior

= 0.970951

* **Xét thuộc tính *income***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *medium* | 2 | 1 | 3 |
| *low* | 1 | 1 | 2 |

0.950978

0.950978=

* **Xét thuộc tính *student***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *yes* | 2 | 1 | 3 |
| *no* | 1 | 1 | 2 |

0.950978

0.950978=

* **Xét thuộc tính *credit\_rating***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *fair* | 3 | 0 | 3 |
| *excellent* | 0 | 2 | 2 |

=0

0=

Ta thấy = lớn nhất 🡪 *credit\_rating* được chọn là nút gốc của nhánh *age=senior*

age=?

*middle\_aged*

*youth*

*senior*

Credit\_rating=?

student=?

*excellent*

*fair*

*yes*

*no*

Ta thấy không còn nút để xét tiếp( chỉ còn các nút lá) 🡪 thuật toán kết thúc và hình trên chính là cây hoàn chỉnh.

1. **Thuật toán C4.5**

* Là thuật toán cải tiến của ID3
* Giá trị phân chia thông tin của thuộc tính A trên tập S

*trong đó là từng giá trị trong tập các giá trị Value(A)*

* Tỷ số độ lớn thông tin của thuộc tính A
* Để tìm node cha chọn thuộc tính có GainRatio lớn nhất

**Bài tập ví dụ: Tìm cây quyết định với thuật toán C4.5 của tập dữ liệu ở bảng 1:**

* **Bước 1:**

Ta thấy nhãn *buys\_computer* có 5 : no, 9 : yes 🡪 Gain của tập dữ liệu D

=0.940286

* **Xét thuộc tính *age***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *youth* | 2 | 3 | 5 |
| *middle\_aged* | 4 | 0 | 4 |
| *senior* | 3 | 2 | 5 |

= 0.693536

0.693536=0.24675

* **Xét thuộc tính *income***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *high* | 2 | 2 | 4 |
| *medium* | 4 | 2 | 6 |
| *low* | 3 | 1 | 4 |

= 0.911063

0.911063=0.029223

* **Xét thuộc tính *student***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *yes* | 3 | 4 | 7 |
| *no* | 6 | 1 | 7 |

= 0.788450

0.788450=0.151836

* **Xét thuộc tính *credit\_rating***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *fair* | 6 | 2 | 8 |
| *excellent* | 3 | 3 | 6 |

=0.892159

0.892159=0.048127

Ta thấy = lớn nhất 🡪 *age* được chọn là nút gốc.

age=?

*Middle\_aged*

*senior*

*youth*

* **Bước 2**

1. **Xét với tâp Dage=youth**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *ID* | *age* | *income* | *student* | *credit\_rating* | *buys\_computer* |
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |

Ta thấy nhãn *buys\_computer* có 3 : no, 2 : yes 🡪 Gain của tập dữ liệu Dage=youth

= 0.970951

* **Xét thuộc tính *income***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *high* | 0 | 2 | 2 |
| *medium* | 1 | 1 | 2 |
| *low* | 1 | 0 | 1 |

= 0.4

0.4=0.570951

* **Xét thuộc tính *student***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *yes* | 2 | 0 | 2 |
| *no* | 0 | 3 | 3 |

= 0

0=

* **Xét thuộc tính *credit\_rating***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *fair* | 1 | 2 | 3 |
| *excellent* | 1 | 1 | 2 |

=0.950978

0.950978=

Ta thấy = lớn nhất 🡪 *student* được chọn là nút gốc của nhánh *age=youth*

age=?

*middle\_aged*

*youth*

*senior*

student=?

*yes*

*no*

1. **Xét với tâp Dage=senior**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *ID* | *age* | *income* | *student* | *credit\_rating* | *buys\_computer* |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 10 | senior | medium | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

Ta thấy nhãn *buys\_computer* có 3 : yes, 2 : no 🡪 Gain của tập dữ liệu Dage=senior

= 0.970951

* **Xét thuộc tính *income***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *medium* | 2 | 1 | 3 |
| *low* | 1 | 1 | 2 |

0.950978

0.950978=

* **Xét thuộc tính *student***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *yes* | 2 | 1 | 3 |
| *no* | 1 | 1 | 2 |

0.950978

0.950978=

* **Xét thuộc tính *credit\_rating***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *fair* | 3 | 0 | 3 |
| *excellent* | 0 | 2 | 2 |

=0

0=

Ta thấy = lớn nhất 🡪 *credit\_rating* được chọn là nút gốc của nhánh *age=senior*

age=?

*middle\_aged*

*youth*

*senior*

Credit\_rating=?

student=?

*excellent*

*fair*

*yes*

*no*

Ta thấy không còn nút để xét tiếp( chỉ còn các nút lá) 🡪 thuật toán kết thúc và hình trên chính là cây hoàn chỉnh.

1. **Thuật toán CART (Classification and Regression Trees)**

* Thuật toán CART chỉ xây dựng cây nhị phân.
* Thuất toán CART cũng là 1 thuật toán hồi quy.
* Chỉ số Gini

*)+ + …*

* Để chọn thuộc tính có giá trị Gini nhỏ nhất làm node gốc.

**Bài tập ví dụ: Tìm cây quyết định với thuật toán CART của tập dữ liệu sau:**

* **Bước 1:**

Ta thấy nhãn *buys\_computer* có 5 : *no*, 9 : *yes* 🡪 Gini của tập dữ liệu D

0.459184

* **Xét thuộc tính *age***
* **Trường hợp 1**: chia tập dữ liệu thành 2 giá trị *{ youth, midd\_leaged }* và *{ senior }*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *{ youth, midd\_leaged }* | 6 | 3 | 9 |
| *{ senior }* | 3 | 2 | 5 |

* **Trường hợp 2**: chia tập dữ liệu thành 2 giá trị *{youth}* và *{ midd\_leaged , senior}*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *{ youth }* | 2 | 3 | 5 |
| *{ midd\_leaged , senior }* | 7 | 2 | 9 |

* **Trường hợp 3**: chia tập dữ liệu thành 2 giá trị *{ midd\_leaged }* và *{ senior, youth }*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *{ youth , senior }* | 5 | 5 | 10 |
| *{ midd\_leaged }* | 4 | 0 | 4 |

Ta thấy với **Trường hợp 3** cho giá trị nhỏ nhất 🡪 Chọn là giá trị Gini cho thuộc tính *age.*

* **Xét thuộc tính *income***
* **Trường hợp 1**: chia tập dữ liệu thành 2 giá *trị {high, medium}* và *{ low }*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *{high, medium}* | 6 | 4 | 10 |
| *{low}* | 3 | 1 | 4 |

* **Trường hợp 2**: chia tập dữ liệu thành 2 giá trị *{ high }* và *{ medium , low }*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *{ high }* | 2 | 2 | 4 |
| *{ medium , low }* | 7 | 3 | 10 |

* **Trường hợp 3**: chia tập dữ liệu thành 2 giá trị *{ high, low }* và *{ medium }*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *{ high, low }* | 5 | 3 | 8 |
| *{ medium }* | 4 | 2 | 6 |

Ta thấy với **Trường hợp 2** cho giá trị nhỏ nhất 🡪 Chọn là giá trị Gini cho thuộc tính *income.*

* **Xét thuộc tính *student***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *yes* | 3 | 4 | 7 |
| *no* | 6 | 1 | 7 |

* **Xét thuộc tính *credit\_rating***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *fair* | 6 | 2 | 8 |
| *excellent* | 3 | 3 | 6 |

Ta thấy nhỏ nhất 🡪 *age* được chọn là nút gốc.

age=?

*{youth,senior}*

*{ middle\_aged}*

* **Bước 2**

**Xét với tâp Dage={youth, senior)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *ID* | *age* | *income* | *student* | *credit\_rating* | *buys\_computer* |
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 10 | senior | medium | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

Ta thấy nhãn *buys\_computer* có 5 : no, 5 : yes 🡪 Gain của tập dữ liệu Dage={youth,senior}

= 0.5

* **Xét thuộc tính *income***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *High* | 0 | 2 | 2 |
| *Medium* | 3 | 2 | 5 |
| *low* | 2 | 1 | 3 |

***( Vì thuật toán CART chỉ xây dựng cây nhị phân trong khi thuộc tính income có 3 giá tính, nên chia thành các trường hợp để nhóm thành 2 giá trị: {{high, medium}* , *{ low }} , { { high }* , *{ medium , low }} và { { high, low } , { medium } }. Trường hợp nào cho chỉ số Gini nhỏ nhất thì lấy chỉ số Gini của trường hợp đấy làm chỉ số Gini cho thuộc tính income)***

* **Trường hợp 1**: chia tập dữ liệu thành 2 giá trị *{high, medium}* và *{ low }*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *{high, medium}* | 3 | 4 | 7 |
| *{low}* | 2 | 1 | 3 |

* **Trường hợp 2**: chia tập dữ liệu thành 2 giá trị *{ high }* và *{ medium , low }*

|  |  |  |  |
| --- | --- | --- | --- |
|  | yes | no | Tổng bản ghi |
| *{ high }* | 0 | 2 | 2 |
| *{ medium , low }* | 5 | 3 | 8 |

* **Trường hợp 3**: chia tập dữ liệu thành 2 giá trị { high, low } và { medium }

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *{ high, low }* | 2 | 3 | 5 |
| *{ medium }* | 3 | 2 | 5 |

Ta thấy với **Trường hợp 2** cho giá trị nhỏ nhất 🡪 Chọn là giá trị Gini cho thuộc tính *income.*

* **Xét thuộc tính *student***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *yes* | 4 | 1 | 5 |
| *no* | 1 | 4 | 5 |

* **Xét thuộc tính *credit\_rating***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *Fair* | 4 | 2 | 6 |
| *Excellent* | 1 | 3 | 4 |

Ta thấy nhỏ nhất 🡪 *student* được chọn là nút gốc của nhánh *age={ youth,senior}*

age=?

*{ youth, senior }*

*{middle\_aged}*

student=?

*yes*

*no*

* **Bước 3**

1. **Xét với tâp** D**age={youth, senior), student =no**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *ID* | *age* | *income* | *student* | *credit\_rating* | *buys\_computer* |
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 8 | youth | medium | no | fair | no |
| 4 | senior | medium | no | fair | yes |
| 14 | senior | medium | no | excellent | no |

Ta thấy nhãn *buys\_computer* có 4 : no, 1 : yes 🡪 Gini của tập dữ liệu Dage={youth,senior}, student=no

= 0.32

* **Xét thuộc tính *income***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *High* | 2 | 0 | 2 |
| *Medium* | 1 | 2 | 3 |

* **Xét thuộc tính *credit\_rating***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *Fair* | 1 | 2 | 3 |
| *Excellent* | 0 | 2 | 2 |

Ta thấy nhỏ nhất 🡪 *credit\_rating* được chọn là nút gốc của nhánh *age={ youth,senior}, student=no*

age=?

*{ youth, senior }*

*{midd\_leaged}*

student=?

*yes*

*no*

credit\_rating=?

*fair*

*excellent*

1. **Xét với tâp** D**age={youth, senior), student =yes**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *ID* | *age* | *income* | *student* | *credit\_rating* | *buys\_computer* |
| 9 | youth | low | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 10 | senior | medium | yes | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |

Ta thấy nhãn *buys\_computer* có 4 : yes, 1 : no 🡪 Gini của tập dữ liệu Dage={youth,senior}, student=no

= 0.32

* **Xét thuộc tính *income***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *low* | 2 | 1 | 3 |
| *Medium* | 2 | 0 | 2 |

* **Xét thuộc tính *credit\_rating***

|  |  |  |  |
| --- | --- | --- | --- |
|  | *yes* | *no* | *Tổng bản ghi* |
| *Fair* | 3 | 0 | 3 |
| *Excellent* | 1 | 1 | 2 |

Ta thấy nhỏ nhất 🡪 *credit\_rating* được chọn là nút gốc của nhánh *age={ youth,senior}, student=yes*

age=?

*{ youth, senior }*

*{midd\_leaged}*

student=?

*yes*

*no*

credit\_rating=?

credit\_rating=?

*fair*

*excellent*

Ta thấy tập dữ liệu đã hết 🡪 thuật toán kết thúc