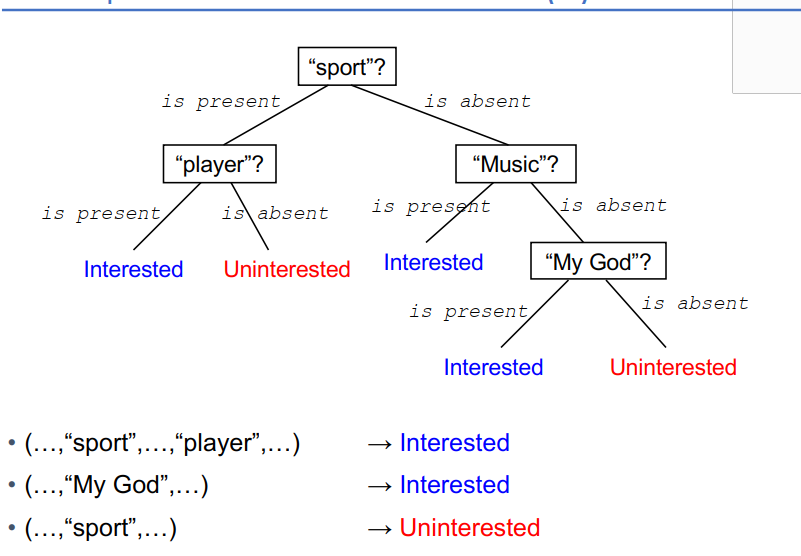
Decision tree: Dùng để biểu diễn 1 func bằng cách sử dụng Tree

+Mỗi cây quyết định có thể hiểu là 1 tập của các IF-Then

* Cây quyết định được sử dụng nhiều trong các ứng dụng thực tế

***Ex:*** ******

Diagram

Description automatically generated

Mỗi nút bên trong đại diện cho một thuộc tính để kiểm tra dữ liệu đến.

Mỗi nhánh/cây con của một nút tương ứng với một giá trị của thuộc tính của nút đó.

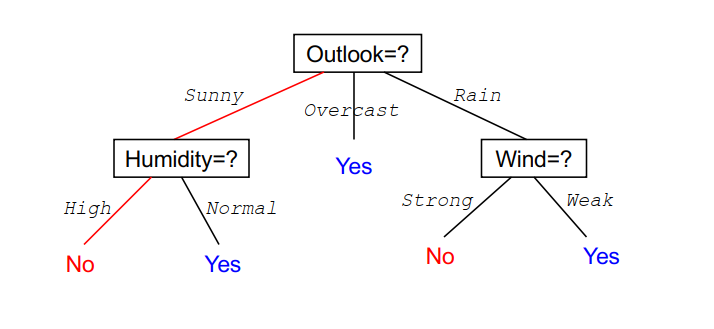
Mỗi nút lá đại diện cho một nhãn lớp.

Khi một cây đã được học, chúng ta có thể dự đoán nhãn cho một phiên bản mới bằng cách sử dụng các thuộc tính của nó để di chuyển từ gốc xuống một chiếc lá.

¨ Nhãn của lá sẽ được sử dụng để gán cho thể hiện mới.

Mỗi đường dẫn từ gốc đến lá là sự kết hợp/VÀ của các bài kiểm tra thuộc tính.

Bản thân cây quyết định là sự phân tách/HOẶC của những liên từ.



Diagram

Description automatically generated

**Học = ID3**

ID3 (Iterative Dichotomiser 3) là một thuật toán tham lam mà được đề xuất bởi Ross Quinlan vào năm 1986.

Nó sử dụng sơ đồ từ trên xuống.

Tại mỗi nút N, chọn thuộc tính kiểm tra A có thể giúp chúng tôi tốt nhất là phân loại cho dữ liệu trong N.

¨ Tạo một nhánh cho mỗi giá trị của A, sau đó tách các dữ liệu vào các nhánh của nó cho phù hợp.

Phát triển cây cho đến khi:

¨ Nó phân loại chính xác tất cả dữ liệu huấn luyện; hoặc là

¨ Tất cả các thuộc tính được sử dụng.

*Lưu ý:* mỗi thuộc tính chỉ được xuất hiện nhiều nhất một lần trong bất kỳ

đường đi của cây.

**ID3\_alg**(*Training\_Set, Class\_Labels, Attributes*)  
Generate the Root of the tree  
If all of *Training\_Set* belong to class c, then Return Root as leaf with label c  
If *Attributes* is empty, then  
Return Root as leaf with label c = **Majority\_Class\_Label**(*Training\_Set*)  
A ← a set of *Attributes* that are best discriminative for *Training\_Set*Let A be the test attributes of Root  
For each value v of A  
Generate a branch of Root which corresponds with v.  
Determine Training\_Setv = { x in *Training\_Set* | xA = v}  
If (Training\_Setv is empty) Then  
Generate a leaf with class label c = **Majority\_Class\_Label**(*Training\_Set*)  
Else  
Generate a subtree by **ID3\_alg**(Training\_Setv, Class\_Labels, Attributes \{A})  
Return Root

Tại mỗi nút, làm thế nào chúng ta có thể chọn một tập hợp các thuộc tính kiểm tra?

¨ Các thuộc tính này phải có tính phân biệt, nghĩa là có thể giúp chúng ta phân loại tốt dữ liệu bên trong nút đó.

Làm thế nào để biết một thuộc tính là phân biệt đối xử?

Ví dụ: giả sử có 2 lớp trong dữ liệu, lớp nào là A1 và A2 nên được chọn làm thuộc tính kiểm tra?

Diagram

Description automatically generated

Lớp A2 sẽ tốt hơn , tại vì dữ liệu A1 chênh lệch giữa c1 và c2 quá ít.và A2 ít nhánh hơn và tối ưu sẽ ok hơn.

***Information gain*** can help.

Information gain: entropy

Entropy đo lường tạp chất/tính không đồng nhất của một tập hợp.

Text, letter

Description automatically generated

Đối với 2 lớp: entropy(S) = - p1log2p1 - p2log2p2

Ý nghĩa của entropy trong Lý thuyết thông tin:

¨ Entropy cho biết số bit trung bình để mã hóa một lớp S.

¨ Entropy của một tin nhắn đo lượng thông tin trung bình

có trong thông điệp đó.

¨ Entropy của một biến ngẫu nhiên x đo độ khó dự đoán của x.

S gồm 14 ví dụ trong đó 9 thuộc lớp c1 và 5 thuộc lớp c2.

¡ Vậy entropy của S là:

Entropy(S)

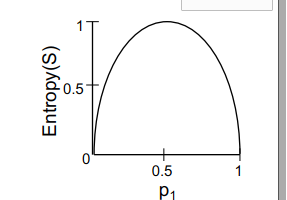
= -(14/9).log2(14/9) -(14/5).log2(14/5)

≈ 0,94

¡ Entropy = 0 nếu tất cả các ví dụ trong S có cùng nhãn.

¡ Entropy = 1 nếu hai lớp trong S có kích thước bằng nhau.

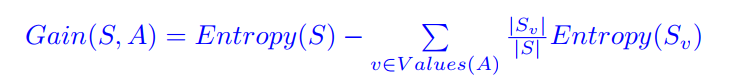
¡ Mặt khác, entropy sẽ luôn thuộc về (0, 1)



Độ lợi thông tin của một thuộc tính trong S:

¨ Đo mức giảm entropy nếu chúng ta chia S thành các tập con theo thuộc tính đó.

Độ lợi thông tin của thuộc tính A trong S được định nghĩa là:



Where Values(A) is the set of all values of A, and  
Sv= {**x** | **x** in S, and xA = v}

Thuật ngữ thứ hai trong Gain(S,A) đo lường thông tin còn lại khi S được chia thành các tập con theo các giá trị của A

Ý nghĩa của Gain(S,A): lượng thông tin trung bình s mất đi khi chia S theo A .

Một tập hợp S các quan sát về một người chơi quần vợt.

Table

Description automatically generated

Text, letter

Description automatically generated

Diagram

Description automatically generated

Diagram

Description automatically generated

Cây quyết định là cây thấp hơn ,bởi vì nó được tối ưu thuật toán tốt và cho ra kết quả đúng ngay cả khi bỏ qua 1 attribute nào đó,…

