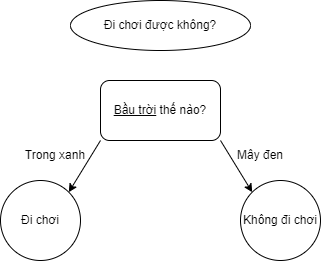
Decision Tree: ID3

Thuật toán ID3 (Iterative Dichotomiser 3) là một kỹ thuật xây dựng cây quyết định dựa trên cơ chế chọn lọc các đặc trưng để tối ưu hóa thông tin nhận vào (Information Gain). Thuật toán này thường được dùng trong phân lớp (classification).

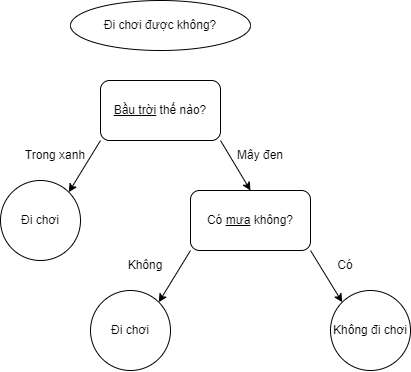
Ý tưởng của thuật toán là mình cần tìm ra đặc trưng tốt nhất làm nút gốc dựa trên một tiêu chuẩn nhất định. Với một đặc trưng được chọn, mình sẽ chia dữ liệu vào các nút con tương ứng với các giá trị của đặc trưng ấy, mà rồi tiếp tục áp dụng phương pháp này cho mỗi nút con. Việc chọn đặc trưng tốt nhất ở mỗi bước được gọi là một cách chọn tham lam. Sau tất cả thì mình muốn tìm đặc trưng có lượng thông tin nhận vào tốt nhất làm nút gốc.

Trong cây quyết định, mỗi đặc trưng trong tập dữ liệu là một *câu hỏi*, và các dữ liệu được phân chia vào các nút con tượng trưng cho *các câu trả lời* cho *câu hỏi* ấy.



*Đặc trưng “Bầu trời” và 2 lớp của nó (“Trong xanh” và “Mây đen”) dưới dạng cây quyết định đưa ra khả năng đi chơi.*

Trong cây quyết định, một phép phân chia tốt nhất là khi dữ liệu trong mỗi nút con hoàn toàn phụ thuộc vào một lớp. Cùng lúc đó, nút con này có thể được gọi là một nút lá, tức là một nút không còn khả năng phân chia dữ liệu.



*Đặc trưng “Bầu trời”, và 1 đặc trưng mới, “Mưa”, quyết định sự kiện đi chơi.*

Theo sơ đồ trên dựa trên tập dữ liệu, trong đó:

* Đặc trưng “Bầu trời”: nút gốc/nút cha (root node/parent node).
* Đặc trưng “Mưa”: nút con (child node) của nút gốc.
* Các thuộc tính “Đi chơi” và “Không đi chơi” của biến mục tiêu “Đi chơi?”: các nút lá (leaf node).
* Các mũi tên: Chỉ ra sự kiện sẽ xảy ra dựa trên các lớp của các đặc trưng.

Nếu vậy dữ liệu trong các nút con sẽ bị trộn lẫn với nhau theo một tỉ lệ lớn, mình coi rằng phép phân chia ấy thực sự chưa tốt.

Vì thế, mình cần có một hàm đo lường độ tinh khiết (purity), hoặc ngược lại, độ vẩn đục (impurity) của phép phân chia. Phép đo lường này cho giá trị thấp nhất nếu dữ liệu trong mỗi nút con nằm trong cùng một lớp (tinh khiết nhất), và sẽ cho giá trị cao nếu mỗi nút con chứa dữ liệu của nhiều lớp khác nhau.

Vì dụ cơ bản về độ tinh khiết trong dữ liệu nằm ở hình thứ 2 trong tài liệu này:

* Sự kiện “Bầu trời trong xanh” hoàn toàn sẽ dẫn đến sự kiện “đi chơi”.
* Sự kiện “Bầu trời mây đen” phải nhờ đến sự kiện “Có mưa” hay “Không mưa” thì mới có thể quyết định được khả năng đi chơi của sự kiện.

Nhưng để mà thực sự quyết định được đặc trưng nào trong tập dữ liệu làm nút gốc, nút con, mình sẽ sữ dụng một hàm số đặc biệt thường được dùng nhiều trong lý thuyết thông tin, hàm **Entropy**. hàm này đo lường độ vẩn đục trong một tập dữ liệu..Mình có công thức tính hàm Entropy như sau:

Trong đó:

* H(S): Entropy của tập dữ liệu S.
* pi : Tỉ lệ của lớp i trong 1 đặc trưng của tập dữ liệu. (với 0 ≤ pi ≤ 1)

Trong thuật toán ID3, hàm Entropy

Xét 1 tập dữ liệu như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **id** | **outlook** | **temperature** | **humidity** | **wind** | **play** |
| 1 | sunny | hot | high | weak | no |
| 2 | sunny | hot | high | strong | no |
| 3 | overcast | hot | high | weak | yes |
| 4 | rainy | mild | high | weak | yes |
| 5 | rainy | cool | normal | weak | yes |
| 6 | rainy | cool | normal | strong | no |
| 7 | overcast | cool | normal | strong | yes |
| 8 | sunny | mild | high | weak | no |
| 9 | sunny | cool | normal | weak | yes |
| 10 | rainy | mild | normal | weak | yes |
| 11 | sunny | mild | normal | strong | yes |
| 12 | overcast | mild | high | strong | yes |
| 13 | overcast | hot | normal | weak | yes |
| 14 | rainy | mild | high | strong | no |