Machine Learning

Lecturer: Doctor Bui Thanh Hung
Data Science Laboratory
Faculty of Information Technology
Industrial University of Ho Chi Minh city
Email: hung.buithanhcs@gmail.com (buithanhhung@iuh.edu.vn)
Website: https://sites.google.com/site/hungthanhbui1980/

Tự cài đặt giải thuật Decision Tree (ID3) theo frame code sau:

Top-Down Tree Construction

```
BuildTree(Node t, Training database D, Split Selection Method S)
    (1) Apply S to D to find splitting criterion
    (2) if (t is not a leaf node)
    (3)     Create children nodes of t
    (4)     Partition D into children partitions
    (5)     Recurse on each partition
    (6) endif
```

Sử dụng Entropy và Information Gain để Split, trong đó:

Entropy: Given a set S of positive and negative examples of some target concept (a 2-class problem), the entropy of set S relative to this binary classification is

$$E(S) = - p(P)\log2\ p(P) – p(N)\log2\ p(N)$$

Information Gain:

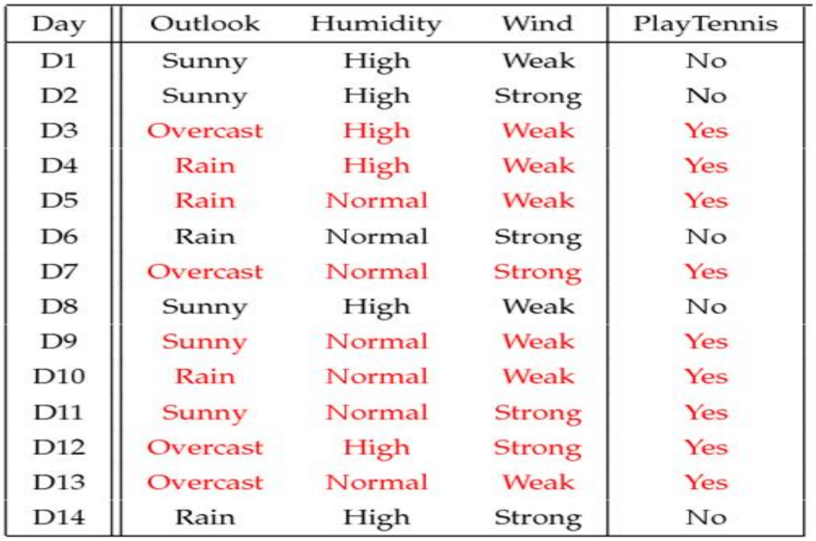
$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Information gain measures the expected reduction in entropy, or uncertainty.

- ✓ Values(A) is the set of all possible values for attribute A, and Sv the subset of S for which attribute A has value v Sv = {s in S | A(s) = v}.
- ✓ the first term in the equation for *Gain* is just the entropy of the original collection *S*
- ✓ the second term is the expected value of the entropy after S is partitioned using attribute A

Viết các hàm (không sử dụng thư viện):

1- Tính Entropy
2- Tính Information gain
3- Xây dựng cây Decision Tree
4- Áp dụng giải thuật này cho bộ dữ liệu sau:



Day	Outlook	Humidity	Wind	PlayTennis
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No