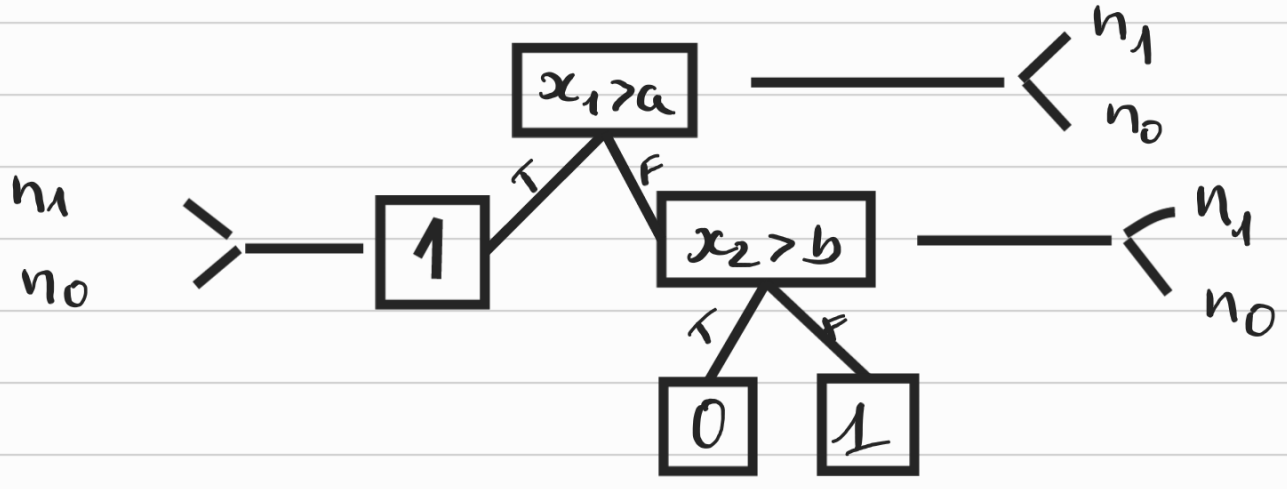
Gini Score , Gini index

- Để lựa chọn được điều kiện phù hợp nhất cho việc phân tách dữ liệu thành các lớp, máy tính dựa vào các chỉ số để đánh giá. Trong đó có chỉ số gini index.

① Xây dựng cây quyết định

VD ta xây dựng 1 cây như sau:



$$\text{Gini Score} = 1 - \sum_{i=1}^{C} (p_i)^2$$

+ C: số lớp cần phân loại

+ $p_i = \frac{n_i}{N}$ $\begin{cases} n_i: \text{số lượng phần tử lớp thứ } i \\ N: \Sigma \text{ số lượng phần tử ở node} \end{cases}$

+ $\sum_{i=1}^{N} p_i = 1$, $N = \sum_{i=1}^{N} n_i$

① Vì $\sum p_i = 1$ và $0 \le p_i \le 1$ nên:

$$\sum_{i=1}^{C} (p_i)^2 \le \left(\sum_{i=1}^{C} p_i\right)^2 = 1$$

$$\Rightarrow \text{gini score} = 1 - \sum_{i=1}^{C} (p_i)^2 \ge 0$$

dấu = xảy ra khi $\exists j \mid p_j = 1$ và $p_k = 0$ $(k \ne j)$

② vì $C > 1$

$$\sum_{i=1}^{C} (p_i)^2 \ge \frac{\left(\sum_{i=1}^{C} p_i\right)^2}{C} = \frac{1}{C} \quad \left(0 < \frac{1}{C} \le 1\right)$$

$$\Rightarrow \text{gini score} = 1 - \sum_{i=1}^{C} (p_i)^2$$

$$\Leftrightarrow \text{gini score} \le 1 - \frac{1}{C}$$

Dấu = xảy ra $\Leftrightarrow$ $p_j = \frac{1}{C}$ $\forall j$

$$\Rightarrow \begin{cases} \max \text{ gini score} = 1 - \frac{1}{C} \\ \min \text{ gini score} = 0 \end{cases}$$

Vậy, gini index = $gini(p) - \sum_{i=1}^{k} \frac{m_k}{M} gini(C_k)$

+ gini(p) = gini score ở node cha

+ k số node con được tách ra

+ gini(C_k) chỉ số gini ở node con thứ k

+ M là số phần tử ở node p

+ m_i là số phần tử ở node con thứ i

+ $\sum_{i=1}^{k} m_i = M$

⇒ Để tìm đk tách, ta sẽ thử tất cả thuộc tính, ứ thuộc tính, thử 1 giá trị của xem gini index của giá trị của nào max thì chọn

[gini score càng nhỏ càng tốt
gini index càng lớn càng tốt