Decision Tree

24th November 2021

1 Biến đổi lại gini score, gini index và cách xây dựng decision tree

1.1 Gini score

$$giniscore = 1 - \sum_{i=1}^{C} p_i^2 = 1 - \sum_{i=1}^{C} (\frac{n_i}{N})^2$$

n_i là số lượng phần tử của lớp i
N là tổng số phần tử ở node
$N = \sum_{i=1}^{N} n_i$ -> $\sum_{i=1}^{N} p_i = 1$

Set: $a_i = \frac{n_i}{N}$
$S = \sum_{i=1}^{N} a_i^2$
điều kiện: $0 \leq a_i \leq 1$ và $\sum a_i = 1$

1.1.1 S min

We have Bu-nhi-a:

$$(a_1^2 + a_2^2 + ... + a_N^2).(1^2 + 1^2 + ... + 1^2) \geq (a_1 + a_2 + ... + a_N)^2$$

$$<=> S.N \geq 1$$

$$<=> S \geq \frac{1}{N}$$

$S_{min} = \frac{1}{N}$ khi $a_1 = a_2 = ... = a_N = \frac{1}{N}$

1.1.2 S max

$$S = \sum_{i=1}^{N} a_i^2 \leq (\sum_{i=1}^{N} a_i)^2 = \sum_{i=1}^{N} a_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} a_i.a_j$$

dấu "= " xảy ra khi $a_j = 0$ và $a_{ij} = 0$

1.2 Gini index

$$giniindex = giniparent - \sum_{i=1}^{N} \frac{n_i}{N}.g_{c_i}$$

1.3 Xây dựng decision tree

Maximize gini-index