Decision-Trees-2

Monday, 30 August 2021 9:30 PM

Agenda :

1. Building DT formation.
2. Gini Impurity
3. Code - DT
4. Underfitting vs Overfitting
5. Hyper parameters
6. Practical aspects of DT
   - Features
   - Categorical data
   - Categorical features
   - Time Space Complexity
   - Pruning - bottom shift.
   - Feature importance.

Dataset (n = 100)

Gender	Age	Target
F	>30	1
M	>30	1
F	<30	0
M	<30	1
F	<30	0
M	<30	1
F	<30	0
M	<30	1
.	.	.
.	.	.
.	.	.

After splitting :

Gender (Parent Impurity)

	Gender		Total
	Female	Male	
Target Variable 1	30	40	70
0 (Others)	20	10	30
Total	50	50	100

Age (Parent Impurity)

	Age		Total
	>30	<30	
Target Variable 1	40	30	70
0 (Others)	10	20	30
Total	50	50	100

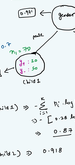
Parent →

$$H(\text{Parent}) \Rightarrow -\sum_{i=1}^{2} p_i \log p_i$$

$$\Rightarrow -[0.5 \log 0.5 + 0.5 \log 0.5]$$

$$\Rightarrow \underline{0.91}$$

Let's use gender to split data



$$H(CL1|L2) \Rightarrow -\sum_{i=1}^{2} p_i \log p_i$$

$$\Rightarrow -[0.36 \log(0.36) + 0.42 \log(0.42)]$$

$$\Rightarrow 0.89$$

$$H(CL1|L2) \Rightarrow 0.918$$

Combined Entropy : $\left[\frac{n_1 H(CL1|L2) + n_2 H(CL2|L2)}{n}\right]$ (Add values)

Combined Entropy (weighted entropy) $\Rightarrow \frac{n_1}{n} H(CL1|L2) + \frac{n_2}{n} H(CL2|L2)$

$$\Rightarrow 0.7 \times 0.89 + 0.3 \times 0.918$$

$$\Rightarrow 0.88$$

Information Gain $\Rightarrow$ Reduction in Entropy

$$I_G \Rightarrow \underbrace{H(\text{Parent})}_{\text{Entropy before split}} - \underbrace{\text{weighted entropy}}_{\text{Entropy after split}}$$

$$\Rightarrow 0.91 - 0.88$$

$$\Rightarrow \boxed{0.03}$$

$$I_G(\text{gender}) \Rightarrow 0.03$$

[Maximizing gain = Reduction in Entropy]

Let's try with age

$$H(CL1|L2) \Rightarrow -[0.66 \log 0.66 + 0.033 \log 0.033]$$

$$\Rightarrow 0.65$$

$$H(CL1|L2) \Rightarrow 0.811$$

Weighted entropy $\Rightarrow w_1 \cdot H(CL1) + w_2 \cdot H(CL2)$

$$\Rightarrow 0.6 \times 0.65 + 0.4 \times 0.811$$

$$\Rightarrow \boxed{0.714}$$

Info Gain (age < 30) $\Rightarrow H(\text{Parent}) - \text{weight entropy of children}$

$$\Rightarrow 0.91 - 0.714$$

$$\Rightarrow \boxed{0.257}$$

I_G (gender) = 0.091

I_G (age<30) = 0.257 (Pick this feature)

age <30
Yes
No
y1 = 40
y0 = 10
y1 = 10
y0 = 30
y1 = 50
y0 = 50
y1 = 0
y0 = 20
y1 = 2
y0 = 0
y1 = 8
y0 = 10
y1 = 2
y0 = 0
y1 = 6
y0 = 10
y1 = 2
y0 = 0
y1 = 4
y0 = 10
y1 = 2
y0 = 0
y1 = 2
y0 = 10
y1 = 2
y0 = 0
y1 = 2
y0 = 10
y1 = 2
y0 = 0
y1 = 2