Assignment: Decision Tree

1.Difference between Entropy and Gini Impurity

Decision Tree

DATE 06/11/2022

* Assignment : 01

Entropy	Gini Impurity
1. Mathematical formula: $E(S) = \sum_{i=1}^{n} -P_i \log_2 P_i$	1. Mathematical formula. $GI = 1 - (P_Y^2 + P_N)^2$
eg. If we have two class $E(S) = -P_Y \log_2(P_Y) - P_N \log_2(P_N)$	where, GI = gini impurity
where,	P_Y → Probability of 'yes'
E(S) → Entropy	P_N → Probability of 'No'.
P_Y → Probability of 'yes'	
P_N → Probability of 'no'	
2. for binary classification problem E(S) min = 0 & E(S) max = 1 For multiclass classification E(S) min = 0 & E(S) max => 1	2. For binary classification GI min = 0 & GI max = 0.5
3. Entropy is computationaly slower due logarithmic calculation.	3. Gini is computationaly faster as compared to entropy due to squaring.
4. Entropy gives slightly better results and balanced trees.	4. Gini is less accurate as compared to entropy
5. Entropy is used for EDA	5. Gini is used to minimize missclassification.

classmate

PAGE

2.Take a tennis dataset and build a decision tree from scratch with mathematical representation

Dataset:

Day	outlook	temperature	humidity	wind	Decision
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
7	overcast	cool	normal	strong	Yes
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	Yes
11	sunny	mild	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes
14	rainfall	mild	high	strong	No

Step 1: Selecting root node.
Calculate gini index of each feature and select one which has less gini index.

a) Outlook decision

		T.V.	P(Y)	P(N)
sunny	Y=2, N=3	5	2/5	3/5
overcast	Y=4, N=0	4	4/4=1	0
rainfall	Y=3, N=2	5	3/5	2/5
		14		

$G.I_{sunny} = 1 - P_Y^2 - P_N^2 = 1 - \frac{4}{25} - \frac{9}{25} = 0.48$

$GI_{overcast} = 1 - 1 - 0 = 0$

$GI_{rainfall} = 1 - \frac{9}{25} - \frac{4}{25} = 0.48$

$Gini_{column} = \sum \frac{\text{no. of instance for class}}{\text{total no. of instance}} \times Gini_{class}$

$Gini_{outlook} = \frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48$

$Gini_{outlook} = 0.34$

DATE

temperature	decision	T.V.	P(Y)	P(N)
hot	Y=2, N=2	4	2/4	2/4
mild	Y=4, N=2	6	4/6	2/6
cool	Y=3, N=1	4	3/4	1/4
		14		

$$Gini_{hot} = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$$

$$Gini_{mild} = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.444$$

$$Gini_{cool} = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$$

$$Gini_{temp} = \frac{4}{14} \times 0.5 + \frac{6}{14} \times 0.44 + \frac{4}{14} \times 0.375 = 0.43$$

humidity	decision	T.V.	P(Y)	P(N)
high	Y=3, N=4	7	3/7	4/7
normal	Y=6, N=1	7	6/7	1/7
		14		

$$Gini_{high} = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.49$$

$$Gini_{normal} = 1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = 0.24$$

$$Gini_{humidity} = \frac{7}{14} \times 0.49 + \frac{7}{14} \times 0.24 = 0.36$$

DATE

d) Wind

	decision	T.V.	P(Y)	P(N)
weak	Y=6, N=2	8	6/8	2/8
strong	Y=3, N=3	6	3/6	3/6
		14		

$$Gini_{weak} = 1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2 = 0.37$$

$$Gini_{strong} = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$$

$$Gini_{wind} = \frac{8}{14} \times 0.37 + \frac{6}{14} \times 0.5 = 0.42$$

$Gini_{outlook}$	$Gini_{temp}$	$Gini_{humidity}$	$Gini_{wind}$
0.34	0.43	0.36	0.42

Since, $Gini_{outlook}$ has less value compared to other three features selected as root node.

Step 2: Further splitting of root node to get lowest label. (Sunny, overcast, rainfall)

* Splitting Sunny.

a) $Gini_{temp}$ ⇒

temp	decision	TV.	P(Y)	P(N)
hot	Y=0, N=2	2	0	2/2=1
cool	Y=1, N=0	1	1/1=1	0
mild	Y=1, N=1	2	1/2	1/2
		05		

classmate

PAGE

D	outlook	temperatu	humidity	wind	Decisio
1	sunny	hot	high	weak	No
2	sunny	hot	high	strong	No
8	sunny	mild	high	weak	No
9	sunny	cool	normal	weak	Yes
11	sunny	mild	normal	strong	Yes

$Gini_{hot} = 1 - (0)^2 - (1)^2 = 0$

$Gini_{cool} = 1 - (1)^2 - (0)^2 = 0$

$Gini_{mild} = 1 - \left(\frac{1}{4}\right) - \left(\frac{1}{4}\right) = 0.5$

$Gini_{temp} = \frac{2}{5} \times 0 + \frac{1}{5} \times 0 + \frac{2}{5} \times 0.5 = 0.2$

b) $Gini_{humidity} =$

humidity	decision	TV	P(Y)	P(N)
high	Y=0, N=3	3	0/3=0	3/3=1
normal	Y=2, N=0	2/5	2/2=1	0/2=0

$Gini_{high} = 1 - (0)^2 - (1)^2 = 0$

$Gini_{normal} = 1 - (1)^2 - (0)^2 = 0$

$Gini_{humidity} = \frac{3}{5} \times 0 + \frac{2}{5} = 0$

DATE

c) $Gini_{wind}$

wind	decision	TV.	P(Y)	P(N)
strong	Y=1, N=1	2	1/2	1/2
weak	Y=1, N=2	3	1/3	2/3
		05		

$Gini_{strong} = 1 - (\frac{1}{4}) - (\frac{1}{4}) = 0.5$

$Gini_{weak} = 1 - (\frac{1}{9}) - (\frac{4}{9}) = 0.44$

$Gini_{wind} = \frac{2}{5} \times 0.5 + \frac{3}{5} \times 0.44 = 0.46$

Since, $Gini_{temp} = 0.2$, $Gini_{humidity} = 0$, $Gini_{wind} = 0.46$

$Gini_{humidity}$ is selected as leaf node with no further splitting

* Splitting Overcast :
Since overcast has gini value 0, no further splitting required.

D	outlook	temperatu	humidity	wind	Decisio
3	overcast	hot	high	weak	Yes
7	overcast	cool	normal	strong	Yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	Yes

* Splitting of rainfall:

a) Gini temp:

temp.	decision	TV	P(Y)	P(N)
mild	Y=2, N=1	3	2/3	1/3
cool	Y=1, N=1	2	1/2	1/2
		05		

D	outlook	temperatu	humidit	wind	Decisi
4	rainfall	mild	high	weak	Yes
5	rainfall	cool	normal	weak	Yes
6	rainfall	cool	normal	strong	No
10	rainfall	mild	normal	weak	Yes
14	rainfall	mild	high	strong	No

$Gini_{mild} = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.444$

$Gini_{cool} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$

$Gini_{temp} = \frac{3}{5} \times 0.444 + \frac{2}{5} \times 0.5 = 0.466$

b) Gini humidity

humidity	decision	TV	P(Y)	P(N)
high	Y=1, N=1	2	1/2	1/2
normal	Y=2, N=1	3	2/3	1/2
		05		

$Gini_{high} = \frac{2}{5} \quad 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$

$Gini_{normal} = 1 - \left(\frac{4}{9}\right) - \left(\frac{1}{4}\right) = 0.30$

$Gini_{humidity} = \frac{2}{5} \times 0.5 + \frac{3}{5} \times 0.30 = 0.38$

c) Gini wind

wind	decision	TV	P(Y)	P(N)
weak	Y=3, N=0	3	3/3=1	0/3=0
strong	Y=0, N=2	2	0/2=0	2/2=1
		05		

$Gini_{weak} = 1 - (1)^2 - (0)^2 = 0$

$Gini_{strong} = 1 - (0)^2 - (1)^2 = 0$

$Gini_{wind} = \frac{3}{5} \times 0 + \frac{2}{5} \times 0 = 0$

DATE

$Gini_{temp} = 0.46$, $Gini_{humidity} = 0.38$, $Gini_{wind} = 0$

Hence, Gini wind is selected as child node with no further splitting