06 June 2023 20:58

Weather	Humidity	Wind	play
sunny	high	weak	no
sunny	high	strong	no
cloudy	high	weak	yes
rainy	high	weak	yes
rainy	normal	weak	yes
rainy	normal	strong	no
cloudy	normal	strong	yes
sunny	high	weak	no
sunny	normal	weak	yes
rainy	normal	weak	yes
sunny	normal	strong	yes
cloudy	high	strong	yes
cloudy	normal	weak	yes
rainy	high	strong	no

Play

9 Yes 5 No

$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$

$H(S) = -P(Yes)\log_2 P(Yes) - P(No)\log_2 P(No)$

$H(S) = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right)$

$P(Yes) = \frac{9}{14}$

$P(No) = \frac{5}{14}$

$H(S) = 0.97$

weather

Sunny | Cloudy | Rainy

2 Yes / 3 No | 4 Yes / 0 No | 3 Yes / 2 No.

Pure

$H(Sunny) = -\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) = 0.97$

$H(cloudy) = -\left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right) - 0 = 0$

$H(Rainy) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) = 0.97$

$Gain = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$.

$= 0.94 - \left(\frac{S_v}{5}\right)H(Sunny) - \left(\frac{S_v}{5}\right)H(cloudy) - \left(\frac{S_v}{5}\right)H(Rainy)$

$= 0.94 - \left(\frac{5}{14}\right)(0.97) - \left(\frac{4}{14}\right)(0) - \left(\frac{5}{14}\right)(0.97)$

Gain = 0.247 → weather

Gain

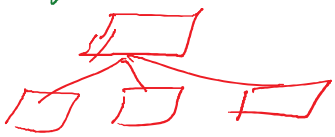
Gain(weather) = 0.247

Gain(wind) = 0.12

Gain(Humidity) = 0.07

out of all these, weather feature is giving maximum information gain.

weather is the root node



Limitation with Entropy

Computation time taken will be very high because of log function.

Gini Index / Gini Impurity

Gini Index / Gini Impurity

Since computation time is high in Entropy we don't use entropy instead we use Gini.

Gini

* It measures impurity of the split.
* It is used to decide the root node

Steps

→ Find Gini of each category in Column
→ Find weighted Gini of that Column.
→ Repeat above steps for all the Columns
→ Select the Column which has less weighted Gini.

$$Gini = 1 - \sum P_i^2 = 1 - P_+^2 - P_-^2$$

P_i → probability = $\frac{}{}$ classes

$$\text{weighted Gini} = \sum \frac{S_v}{S} Gini(S)$$

Note → Select the variable which has less weighted Gini as a root node.

weighted Gini (weather) = 0.43
weighted Gini (windy) = 0.8
weighted Gini (Humidity) = 0.5

root node

wind

Weather	Humidity	Wind	play
sunny	high	weak	no
sunny	high	strong	no
cloudy	high	weak	yes
rainy	high	weak	yes
rainy	normal	weak	yes
rainy	normal	strong	no
cloudy	normal	strong	yes
sunny	high	weak	no
sunny	normal	weak	yes
rainy	normal	weak	yes
sunny	normal	strong	yes
cloudy	high	strong	yes
cloudy	normal	weak	yes
rainy	high	strong	no

weak: 6 Yes, 2 No
strong: 3 Yes, 3 No

$$Gini = 1 - [P(Yes)]^2 - P[No]^2$$

$$Gini(weak) = 1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2 = 0.375$$

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

$$\text{weighted Gini} = \sum \frac{|S_v|}{|S|} Gini(S_v)$$

S → no Sample before split

$$\text{weighted Gini} = \left(\frac{S_v}{S}\right) Gini(weak) + \left(\frac{S_v}{S}\right) Gini(strong)$$

S → Sample before split

S_v → no of samples after split

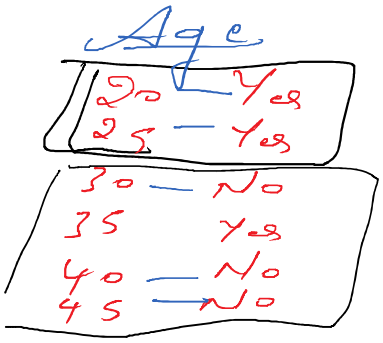
weighted Gini $= \left(\frac{S_v}{S}\right) Gini(weak) + \left(\frac{S_v}{S}\right) Gini(strong)$

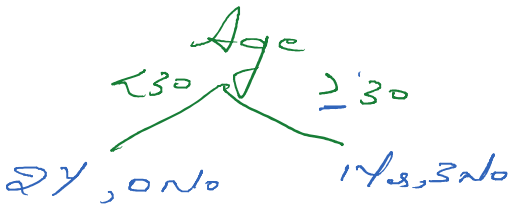
weighted Gini $= \left(\frac{8}{14}\right)(0.375) + \frac{6}{14}(0.5) = 0.43$

S = 14

Wind [14]

weak	strong
8	6
$S_v = 8$	$S_v = 6$





Imp

* Scaling and handling outliers are not required for Decision Tree algorithms.

Overfitting

→ Training score is high and testing score is less.

→ Model works extremely well on train data and score will be high, but during testing it will make wrong prediction eventually testing score will be less

→ Low bias and high variance → overfitting

Bias → Training error

Variance → Testing error

Underfitting

* Training score will be less and testing score is less.

* high bias, high variance

Score Problem:

* high bias, high variance

How do you say that model is Generalized

→ Training score and testing score should be high.

→ Low bias, Low variance

→ high bias, Low variance → Training score can be less and testing score should be high.

How to overcome overfitting?

→ Hyperparameters