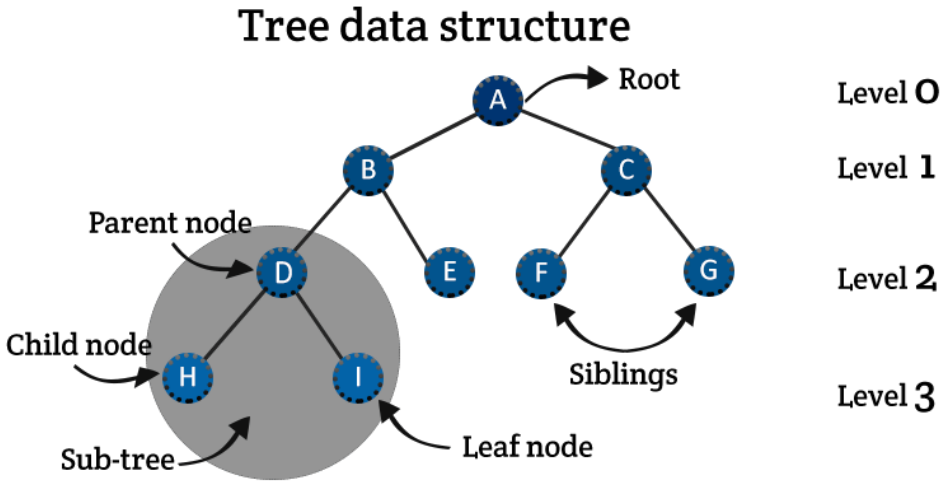
All about Decision Tree

A decision tree is a very specific type of probability tree that enables you to make a decision about some kind of process. It is used to break down complex problems or branches. Each branch of the decision tree could be a possible outcome.



All about Decision Tree

A decision tree is a very specific type of probability tree that enables you to make a decision about some kind of process. It is used to break down complex problems or branches. Each branch of the decision tree could be a possible outcome.

- Supervised
- Classification
- Entropy
- Information Gain (IG)
- Gini Index

All about Decision Tree

Problem Data Set

Class

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

All about Decision Tree in Machine Learning

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$IG(Y, X) = E(Y) - E(Y|X)$$

$$Gini\ index = 1 - \sum_{i=1}^{n} p_i^2$$

AIQuest

www.aiquest.org

① Integer Number :

$\log_2 4 = \log_2 2^2$

$= 2 \log_2 2$

$= 2$

Because, $\log_2 2 = 1$

Or, You can follow :

$\log_2 4 = \frac{\log 4}{\log 2}$

$= 2$

Base change rule :

$\log_b a = \frac{\log a}{\log b}$ → 10 Base

② Fraction Number :

$\log_2 (\frac{1}{4}) = \frac{\log(\frac{1}{4})}{\log 2}$ ← Number, Base

$= \frac{\log 1 - \log 4}{\log 2}$

$= \frac{0 - \log 2^2}{\log 2}$

$= \frac{-2 \log 2^2}{\log 2}$

$= -2$ (Use Calculator)

rule : $\log(\frac{M}{N}) = \log M - \log N$

All about Decision Tree

Wear Jacket?		
1	YES	3 Times
2	NO	4 Times

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

E(Y) = Entropy Before Partition
E(Y|X) = Entropy After Partition
Target, E(Y) >> E(Y|X)

Entropy Before Partition:

Entropy of Wear Jacket:

= Entropy (4, 3)
= Entropy (– (Pi $\log_2$ Pi) + (– Pi $\log_2$ Pi))
= (-4/7 $\log_2$ 4/7) + (-3/7 $\log_2$ 3/7)
= (-.57 $\log_2$ .57) + (-.43 $\log_2$ .43)
= .985 (Entropy Before Partition)

All about Decision Tree

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Outlook
E (Outlook, Sunny) = -(1/4 $\log_2$ ¼ + ¾ $\log_2$ ¾) = .812
E (Outlook, Cloudy) = -(2/3 $\log_2$ 2/3 + 1/3 $\log_2$ 1/3) = .918
Info Gain (S, Outlook) = E(S) – (4/7 * .812) – (3/7 * .918) = .985 – (4/7 * .812) – (3/7 * .918) = .127

Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

All about Decision Tree

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Temperature
E (Temperature, Cold) = -(1/4 $\log_2$ ¼ + ¾ $\log_2$ ¾) = .812
E (Temperature, Warm) = -(0/3 $\log_2$ 0/3 + 3/3 $\log_2$ 3/3) = 0.00
Info Gain (S, Temperature) = E(S) – (4/7 * .812) – (3/7 * 0) = .985 – (4/7 * .812) – (3/7*0) =.520

Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

All about Decision Tree

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Routine
E (Routine, Indoor) = -(1/4 $\log_2$ ¼ + ¾ $\log_2$ ¾) = .812
E (Routine, Outdoor) = -(2/3 $\log_2$ 2/3 + 1/3 $\log_2$ 1/3) = .918
Info Gain (S, Routine) = E(S) – (4/7*.812) – (3/7 * .918) = .985 – (4/7*.812) – (3/7 * .918) =.127

Problem Data Set

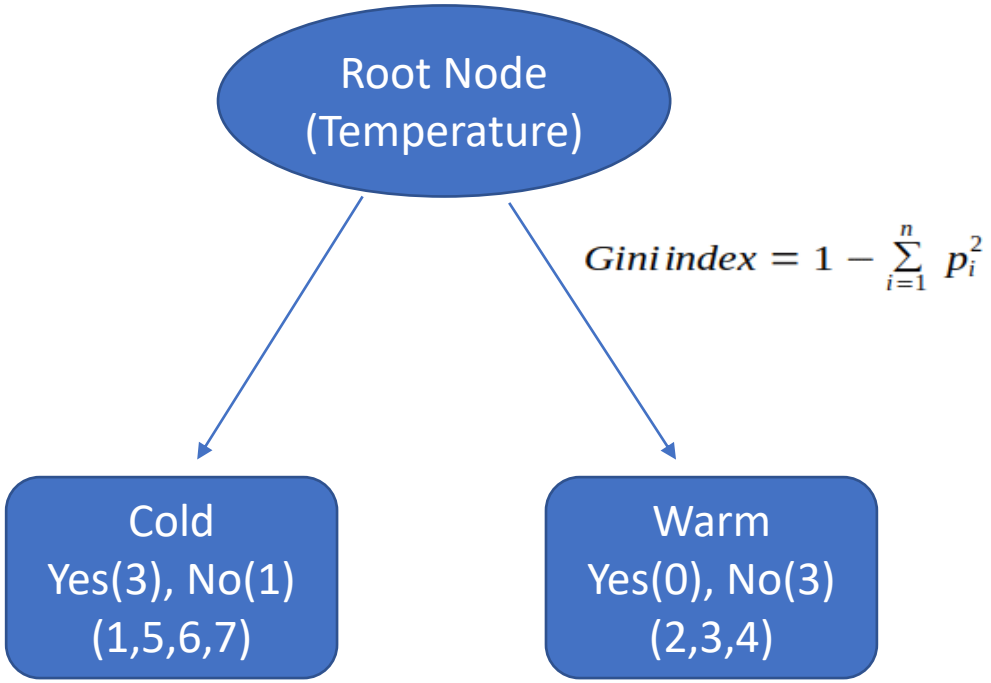
Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

All about Decision Tree

Root Node Selection Table

Outlook	Temperature	Routine
E (Outlook, Sunny) = -(1/4 $\log_2$ ¼ + ¾ $\log_2$ ¾) = .812	E (Temperature, Cold) = -(1/4 $\log_2$ ¼ + ¾ $\log_2$ ¾) = .812	E (Routine, Indoor) = -(1/4 $\log_2$ ¼ + ¾ $\log_2$ ¾) = .812
E (Outlook, Cloudy) = -(2/3 $\log_2$ 2/3 + 1/3 $\log_2$ 1/3) = .918	E (Temperature, Warm) = -(0/3 $\log_2$ 0/3 + 3/3 $\log_2$ 3/3) = 0.00	E (Routine, Outdoor) = -(2/3 $\log_2$ 2/3 + 1/3 $\log_2$ 1/3) = .918
Info Gain (S, Outlook) = E(S) – (4/7 * .812) – (3/7 * .918) = .985 – (4/7 * .812) – (3/7 * .918) = .127	Info Gain (S, Temperature) = E(S) – (4/7 * .812) – (3/7 * 0) = .985 – (4/7 * .812) – (3/7*0) =.520	Info Gain (S, Routine) = E(S) – (4/7*.812) – (3/7 * .918) = .985 – (4/7*.812) – (3/7 * .918) =.127

All about Decision Tree

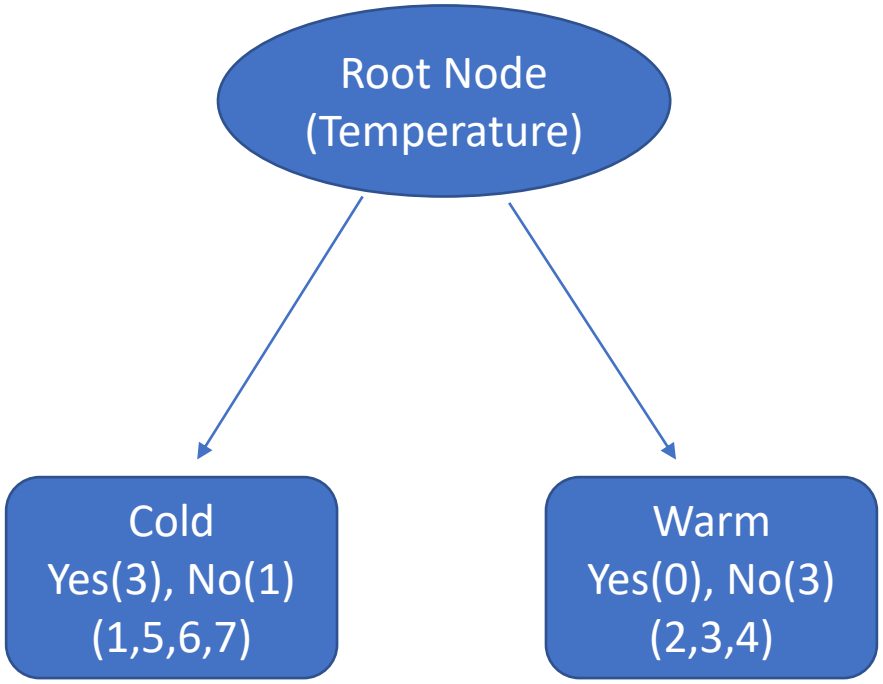


$$Gini\,index = 1 - \sum_{i=1}^{n} p_i^2$$

Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

All about Decision Tree



Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

All about Decision Tree

Entropy of New Subset:

S2 = Entropy (1,3)
= Entropy (– (Pi $\log_2$ Pi) + (– Pi $\log_2$ Pi))
= (-1/4 $\log_2$ 1/4) + (-3/4 $\log_2$ 3/4)
= (-.25 $\log_2$ .25) + (-.75 $\log_2$ .75)
= .812 (Entropy for New Subset)

Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

All about Decision Tree

E (Routine, Indoor) = $-(1/2 \log_2 ½ + 1/2 \log_2 1/2)$ = 1
E (Routine, Outdoor) = $-(2/2 \log_2 2/2 + 0/2 \log_2 0/2)$ = 0
Info Gain (S2, Routine) = E(S2) – 2/4 * 1 – 2/4 *0 = .812 – 2/4 * 1 – 2/4 *0 = .312

Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

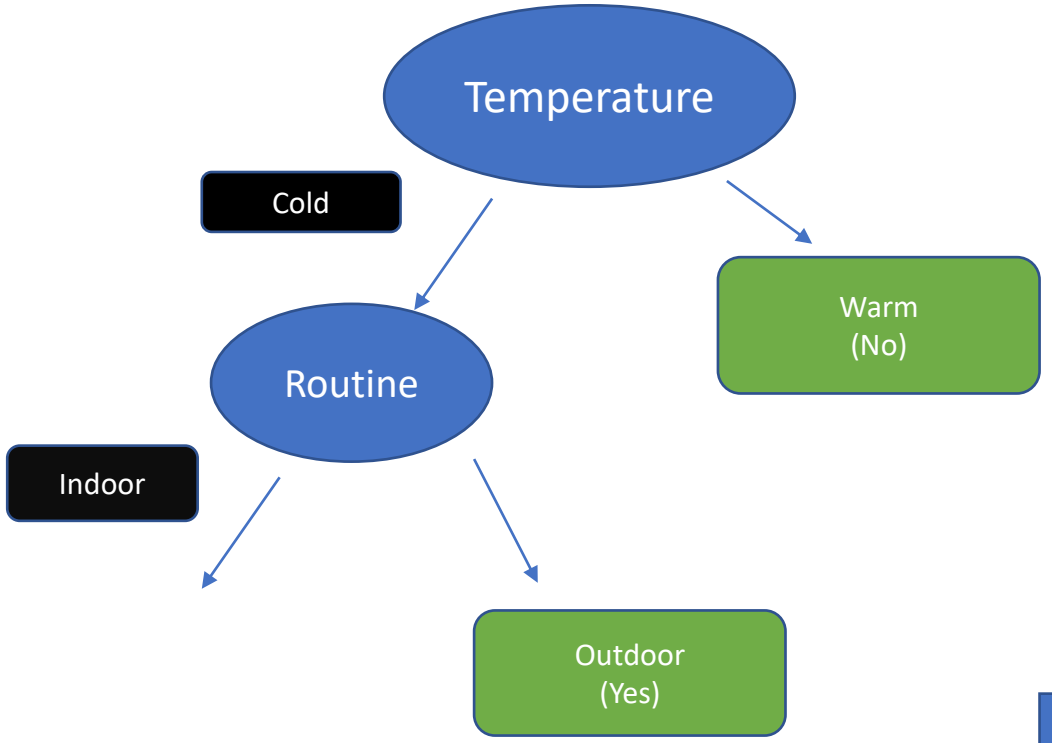
All about Decision Tree

E (Outlook, Sunny) = $-(1/2 \log_2 ½ + 1/2 \log_2 1/2)$ = 1
E (Outlook, Cloudy) = $-(2/2 \log_2 2/2 + 0/2 \log_2 0/2)$ = 0
Info Gain (S2, Outlook) = E(S2) – 2/4 * 1 – 2/4 *0 = .812 – 2/4 * 1 – 2/4 *0 = .312

Problem Data Set

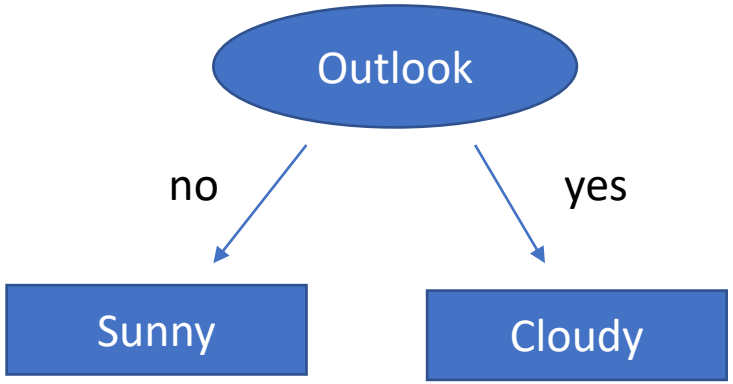
Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

All about Decision Tree



Sunny, Cold , Indoor= ??

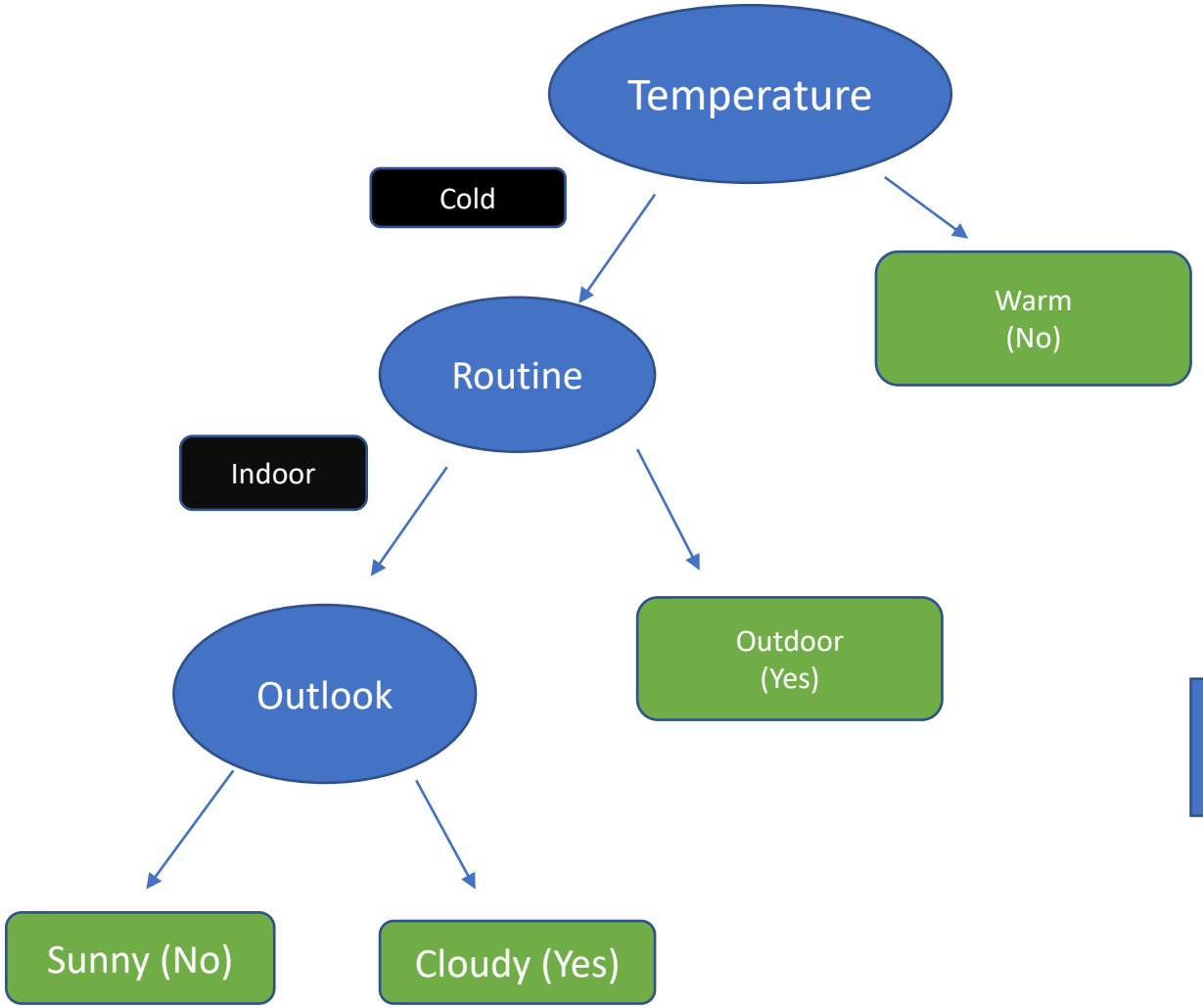
All about Decision Tree



Problem Data Set

Days	Outlook	Temperature	Routine	Wear Jacket?
1	Sunny	Cold	Indoor	No
2	Sunny	Warm	Outdoor	No
3	Cloudy	Warm	Indoor	No
4	Sunny	Warm	Indoor	No
5	Cloudy	Cold	Indoor	Yes
6	Cloudy	Cold	Outdoor	Yes
7	Sunny	Cold	Outdoor	Yes

All about Decision Tree



Sunny, Cold , Indoor= ??