PH-227
AI and Data Science

- Aftab Alam
- Email : aftab@iitb.ac.in
- Ext. : 5564 or 8564

TAs:

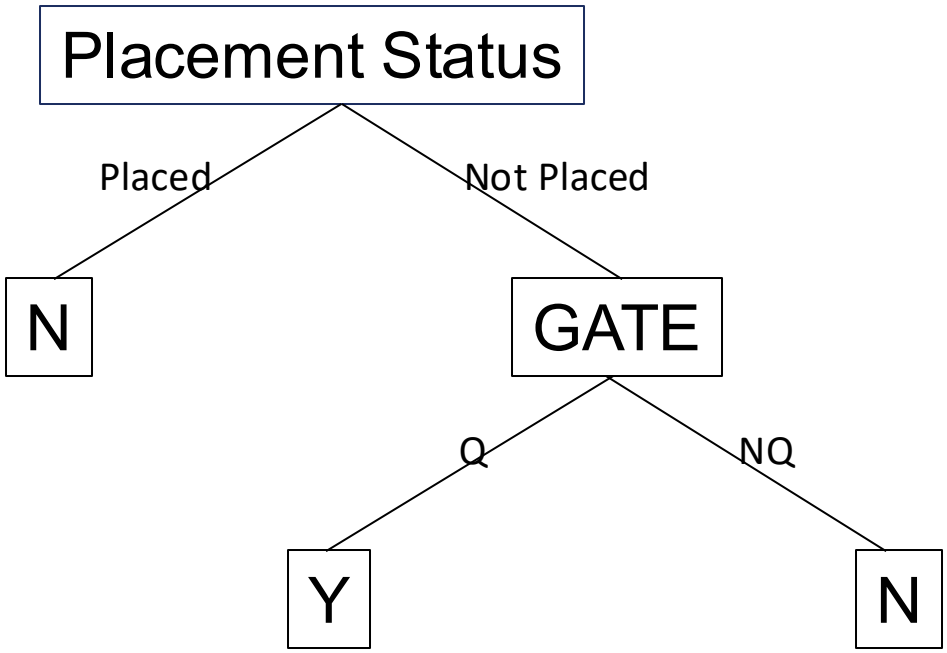
Yashowardhan, Divyansh, Matam, Peela, Piyush

A Quick Recap

Decision Tree

Decision Tree Algorithm is used for both classification and regression task.

Example:



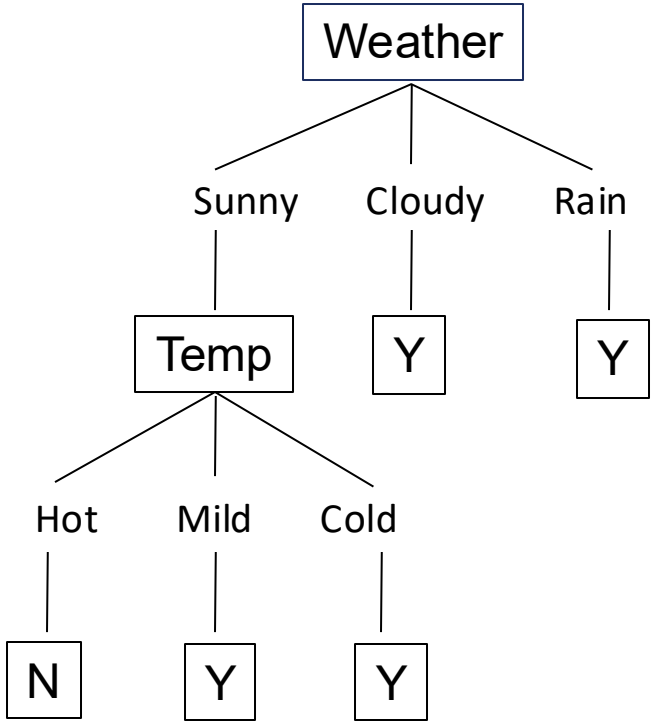
Decision Tree (ALGORITHM)

Decision Tree Algorithm is used for both classification and regression task.

Few Key Attributes:

- Tree Structure
- Decision Nodes
- Leaf Nodes
- Splitting
- Entropy and Information Gain
- Pruning

Example:



Iterative Dichotomiser 3 (ID3) Algorithm

Example:

Day	Weather	Temp	Humidity	Wind	Play Football?
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Cloudy	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Cloudy	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Cloudy	Mild	High	Strong	Yes
Day13	Cloudy	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Entropy Gain and Information Gain of Weather

Entropy of entire Dataset: $S\{\overset{y}{+9},\overset{n}{-5}\} = -(9/14)*\log_2(9/14)-(5/14)*\log_2(14)$
$=0.94$

Entropy of other attributes:
Entropy of Sunny, $S\{+2,-3\} = -(2/5)*\log_2(2/5)-(3/5)*\log_2(3/5)=0.97$ ➔ Ent(S)
Entropy of Cloudy $S\{+4,0\} = 0$ ➔ Ent(C)
Entropy of Rain $S\{+3,-2\} = 0.97$ ➔ Ent(R)

Information Gain =
Entropy (entire data) – (5/14) * Ent(S) – (4/14) * Ent(C)- (5/14)*Ent(R)
= 0.247

Day	Weather	Temp	Humidity	Wind	Play Football?
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Cloudy	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Cloudy	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Cloudy	Mild	High	Strong	Yes
Day13	Cloudy	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Entropy Gain and Information Gain of Temperature

Entropy of entire Dataset: $S\{\overset{y}{+9},\overset{n}{-5}\} = -(9/14)*\log_2(9/14)-(5/14)*\log_2(14)$
$=0.94$

Entropy of other attributes:
Entropy of Hot, $S\{+2,-2\} = -(2/4)*\log_2(2/4)-(2/4)*\log_2(2/4)=1.0$ ➔ Ent(H)
Entropy of Cool, $S\{+3,-1\} = -(3/4)*\log_2(3/4)-(1/4)*\log_2(1/4)=1.0$ ➔ Ent(C)
Entropy of Mild, $S\{+4,-2\} = -(4/6)*\log_2(4/6)-(2/6)*\log_2(2/6)=0.92$ ➔ Ent(M)

Information Gain =
Entropy (entire data) – (4/14) * Ent(H) – (4/14) * Ent(C)- (6/14)*Ent(M)
= 0.029

Day	Weather	Temp	Humidity	Wind	Play Football?
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Cloudy	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Cloudy	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Cloudy	Mild	High	Strong	Yes
Day13	Cloudy	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Entropy Gain and Information Gain of Humidity

Entropy of entire Dataset: $S\{\overset{y}{+9},\overset{n}{-5}\} = -(9/14)*\log_2(9/14)-(5/14)*\log_2(14)$
$=0.94$

Entropy of other attributes:

Entropy of High, $S\{+3,-4\} = -(3/7)*\log_2(3/7)-(4/7)*\log_2(4/7)=0.98$ ➔ Ent(H)
Entropy of Normal $S\{+6,-1\} = -(6/7)*\log_2(6/7)-(1/7)*\log_2(1/7)=0.59$ ➔ Ent(N)

Information Gain =
Entropy (entire data) – (7/14) * Ent(H) – (7/14) * Ent(N)
= 0.15

Day	Weather	Temp	Humidity	Wind	Play Football?
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Cloudy	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Cloudy	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Cloudy	Mild	High	Strong	Yes
Day13	Cloudy	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Entropy Gain and Information Gain of Wind

Entropy of entire Dataset: $S\{+9,-5\} = -(9/14)*\log_2(9/14)-(5/14)*\log_2(14)$ (y n)
$=0.94$

Entropy of other attributes:

Entropy of Strong, $S\{+3,-3\} = -(3/6)*\log_2(3/6)-(3/6)*\log_2(3/6)=1.0$ ➔ Ent(S)
Entropy of Weak, $S\{+6,-2\} = -(6/8)*\log_2(6/8)-(2/8)*\log_2(2/8)=0.81$ ➔ Ent(W)

Information Gain =
Entropy (entire data) – (6/14) * Ent(S) – (8/14) * Ent(W)
= 0.048

Day	Weather	Temp	Humidity	Wind	Play Football?
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Cloudy	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Cloudy	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Cloudy	Mild	High	Strong	Yes
Day13	Cloudy	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

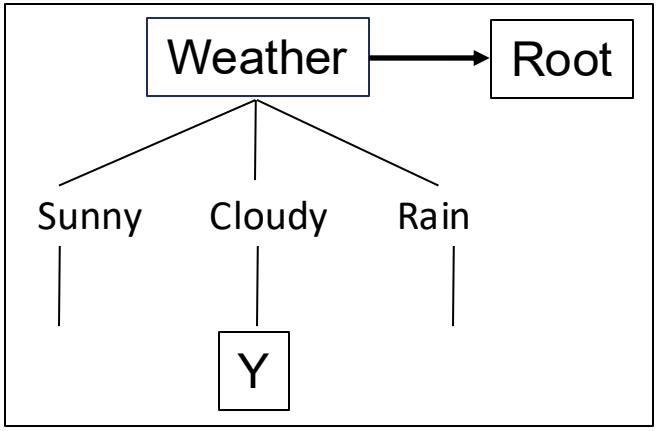
Comparison of Information Gain for different Attributes

IGain(Weather)=**0.246**

IGain(Temperature)=0.029

IGain(Humidity)=0.15

IGain(Wind)=0.048



Day	Weather	Temp	Humidity	Wind	Play Football?
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Cloudy	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Cloudy	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Cloudy	Mild	High	Strong	Yes
Day13	Cloudy	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Day	Weather	Temp	Humidity	Wind	Play Football?
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes

Day	Weather	Temp	Humidity	Wind	Play Football?
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day10	Rain	Mild	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Information Gain of Temperature w.r.to Sunny

y n
Entropy of entire Dataset: $S\{+2,-3\} = -(2/5)*\log_2(2/5)-(3/5)*\log_2(3/5) = 0.97$

Entropy of other attributes:
Entropy of Hot, $S\{0,-2\} = 0.0$ ➔ Ent(H)
Entropy of Mild, $S\{+1,-1\} = -(1/2)*\log_2(1/2)-(1/2)*\log_2(1/2)=1.0$ ➔ Ent(M)
Entropy of Cool, $S\{+1,0\} = 0.0$ ➔ Ent(C)

Information Gain =
Entropy (entire data) – (2/5) * Ent(H) – (2/5) * Ent(M)- (1/5)*Ent(C)
= 0.57

Day	Weather	Temp	Humidity	Wind	Play Football?
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes

Information Gain of Humidity w.r.to Sunny

Entropy of entire Dataset: $S\{+2,-3\} = -(2/5)*\log_2(2/5)-(3/5)*\log_2(3/5)$ (y n)
$=0.97$

Entropy of other attributes:
Entropy of High, S{0,-3} = 0.0 ➔ Ent(H)
Entropy of Normal, S{+2,0} = 0.0 ➔ Ent(N)

Information Gain =
Entropy (entire data) – (3/5) * Ent(H) – (2/5) * Ent(N)
= 0.97

Day	Weather	Temp	Humidity	Wind	Play Football?
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes

Information Gain of Wind w.r.to Sunny

Entropy of entire Dataset: $S\{+2,-3\} = -(2/5)*\log_2(2/5)-(3/5)*\log_2(3/5)$ (y n)
$=0.97$

Entropy of other attributes:
Entropy of Strong, $S\{+1,-1\} = -(1/2)*\log_2(1/2)-(1/2)*\log_2(1/2) = 1.0$ ➔ Ent(S)
Entropy of Weak, $S\{+1,-2\} = -(1/3)*\log_2(1/3)-(2/3)*\log_2(2/3) = 0.92$ ➔ Ent(W)

Information Gain =
Entropy (entire data) – (2/5) * Ent(H) – (3/5) * Ent(N)
= 0.019

Day	Weather	Temp	Humidity	Wind	Play Football?
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes

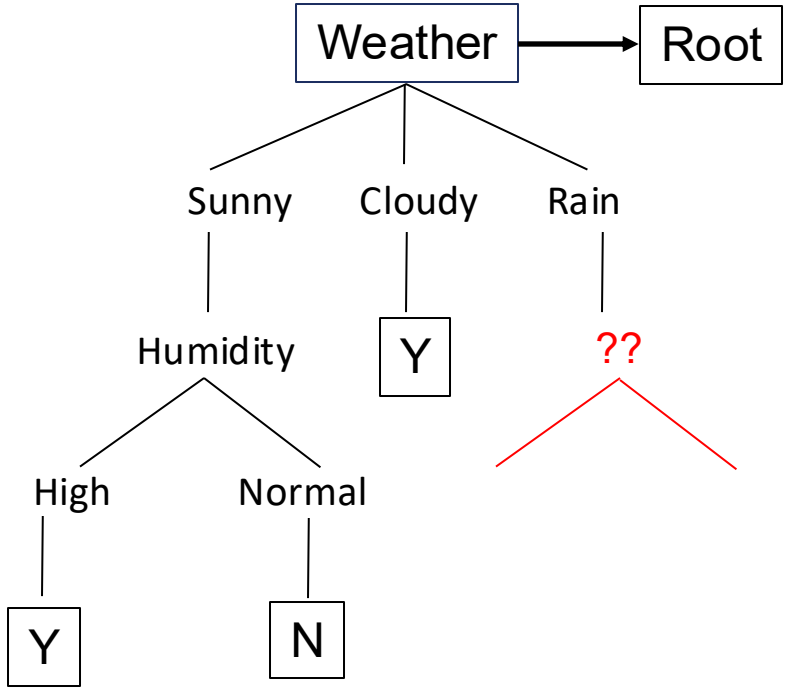
Comparison of Information Gain w.r.to Sunny

$IGain(S_{Sunny}, Temp) = \mathbf{0.57}$

$IGain(S_{Sunny}, Humidity) = \mathbf{0.97}$

$IGain(S_{Sunny}, Wind) = 0.15$

Day	Weather	Temp	Humidity	Wind	Play Football?
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes



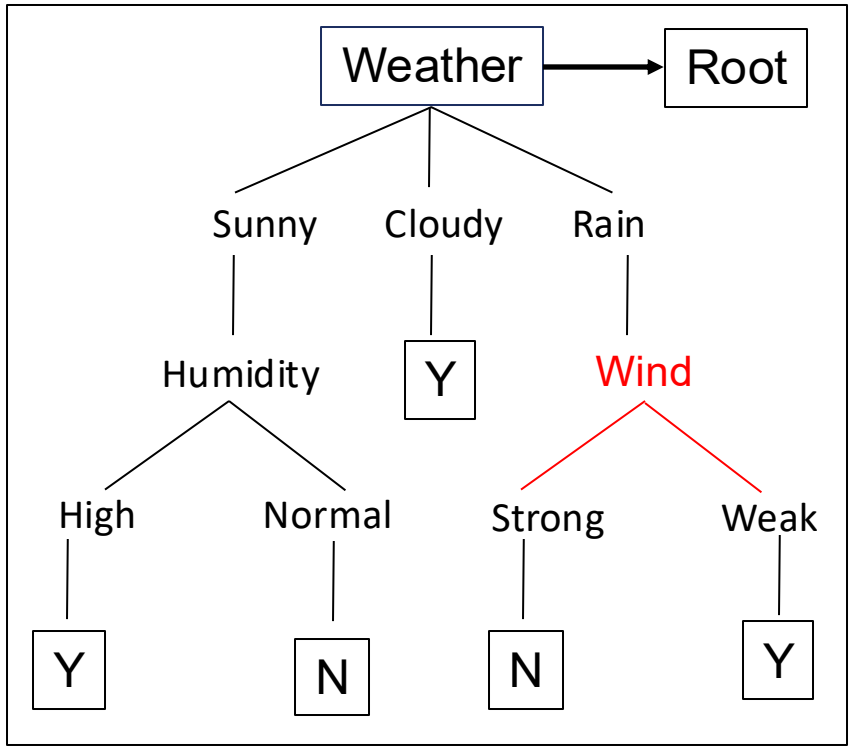
Comparison of Information Gain w.r.to Rain

$IGain(S_{Rain}, Temp) = \mathbf{0.019}$

$IGain(S_{Rain}, Humidity) = \mathbf{0.019}$

$IGain(S_{Rain}, Wind) = \mathbf{0.97}$

Day	Weather	Temp	Humidity	Wind	Play Football?
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day10	Rain	Mild	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No



Decision Tree