Sudha Amarnath

SAN JOSÉ STATE
UNIVERSITY

Graduate and Extended Studies

FA19: CMPE-297 Sec 01 - Special Topics

Prof. Chandrasekar Vuppalapati

Assignment #2

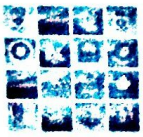
Sudha Amarnath
ID: 013709956

ICC Worldcup 2015 PlayCricket - Decisiontree

Please Develop Decision tree for ICC world cup 2015 PlayCricket - Decision tree

Deliverable:

- Decision tree calculations




Training example for the target match "PlayCricket"

Match Day	Outlook	Temperature	Humidity	Wind	Play Cricket
Day 1	Sunny	Hot	High	Weak	No
Day 2	Sunny	Hot	High	Strong	No
Day 3	Overcast	Hot	High	Weak	Yes
Day 4	Rain	Mild	High	Weak	Yes
Day 5	Rain	Cool	Normal	Weak	Yes
Day 6	Rain	Cool	Normal	Strong	No
Day 7	Overcast	Cool	Normal	Strong	Yes
Day 8	Sunny	Mild	High	Weak	No
Day 9	Sunny	Cool	Normal	Weak	Yes
Day 10	Rain	Mild	Normal	Weak	Yes
Day 11	Sunny	Mild	Normal	Strong	Yes
Day 12	Overcast	Mild	High	Strong	Yes
Day 13	Overcast	Hot	Normal	Weak	Yes
Day 14	Rain	Mild	High	Strong	No

1. classification using ID3 algorithm to build a decision tree for the given table. This uses Entropy and information gain as metric.

2.

Independent variables	Dependent variable
- outlook - Temperature - Humidity - Wind	- play Cricket

3. To calculate the root for the decision tree, calculate Entropy of the target (Yes = 9, No = 5)

- The expected information needed to classify a tuple in the data partition D, is given by

$$\text{Entropy}(D) = -\sum_{i=1}^{m} P_i \log(P_i)$$

Entropy (Play Cricket) = Entropy (5, 9)

= Entropy (0.36, 0.64)

$= -(0.36 \log_2 0.36) - (0.64 \log_2 0.64)$

= 0.94

4. The dataset is split on the different attributes The Entropy is calculated for each branch and total entropy is calculated.

- From the table derive the data for the attribute outlook.

Outlook	play		Total
	Yes	NO	
Sunny	2	3	5
Overcast	4	0	4
Rainy	3	2	5
			14

- The average weighted entropy is calculated using:

$$\text{Entropy}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Entropy}(D_j)$$

Entropy(D, outlook) = $5/14 * Entropy(2,3) + (4/14) *$ Entropy(4,0) + $(5/14) *$ Entropy(3,2)

$= (5/14)(-(2/5)\log(2/5) - (3/5)\log(3/5)) + (4/14)(0) + (5/14)(-(3/5)\log(3/5) - (2/5)\log(2/5))$

$= 0.693$

- Information gain for the above data is calculated using

$$\text{Gain}(A) = \text{Entropy}(D) - \text{Entropy}_A(D)$$

Gain (D, outlook) = 0.940 - 0.693 = 0.247

- calculate information gain for each of the attributes

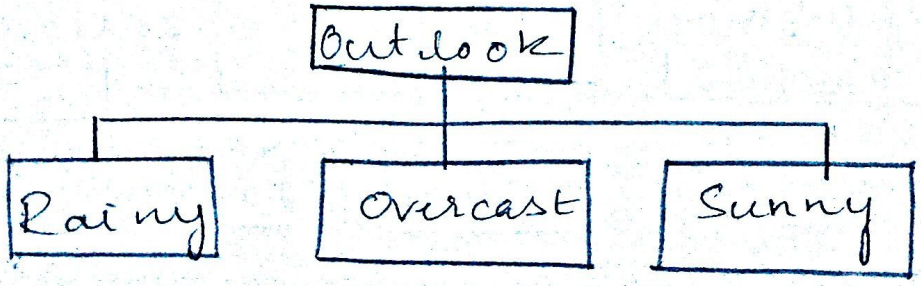
Outlook	Playcricket	
	Yes	No
Sunny	2	3
overcast	4	0
Rainy	3	2
Gain = 0.247		

Temperature	Playcricket	
	Yes	No
Hot	2	2
Mild	4	2
cool	3	1
Gain = 0.029		

Humidity	Playcricket	
	Yes	No
High	3	4
Normal	6	1
Gain =	0.152	

Wind	Playcricket	
	Yes	No
Weak	6	2
Strong	3	3
Gain = 0.048		

5. choose the attribute with largest Information Gain as the decision node (root) Repeat the same step on each branch by dividing the dataset



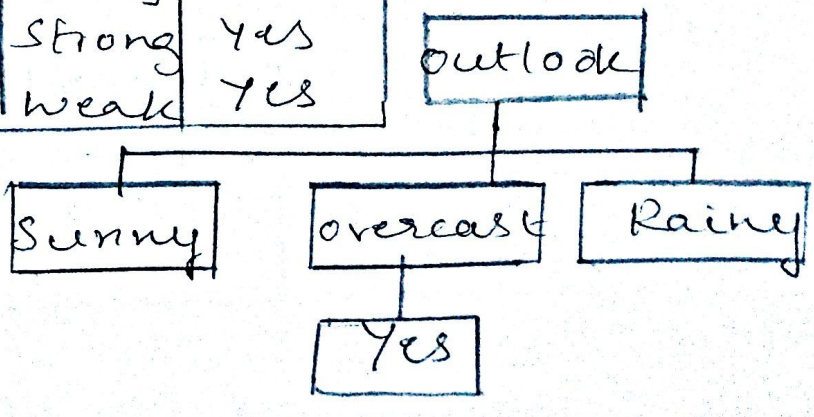
Outlook	Temperature	Humidity	Wind	PlayCricket
Sunny	Hot	High	weak	NO
Sunny	Hot	High	Strong	NO
Sunny	Mild	High	weak	NO
Sunny	cool	Normal	weak	Yes
Sunny	Mild	Normal	strong	Yes

overcast	Hot	High	weak	Yes
overcast	cool	Normal	Strong	Yes
overcast	Mild	High	Strong	Yes
overcast	Hot	Normal	weak	Yes

Rain	Mild	High	weak	Yes
Rain	cool	Normal	weak	yes
Rain	cool	Normal	Strong	No
Rain	Mild	Normal	weak	Yes
Rain	Mild	High	Strong	NO

- The branch node with Entropy 0 is the leaf node

Temperature	Humidity	Wind	PlayCricket
Hot	High	weak	Yes
cool	Normal	Strong	Yes
Mild	High	Strong	Yes
Hot	Normal	weak	Yes



6. Entropy more than 0 needs further splitting. Consider the table for attribute with outlook with tuple sunny.

Temperature	Humidity	Wind	Playcricket
Hot	High	weak	NO
Hot	High	Strong	NO
Mild	High	weak	NO
cool	Normal	weak	Yes
Mild	Normal	Strong	Yes

Temp	Play		Total
	Yes	NO	
hot	0	2	2
cool	1	1	2
mild	1	0	1
			5

$E(Sunny) = (-(2/5)\log(2/5) - (3/5)\log(3/5)) = 0.971$

$E(Sunny, temperature) = (2/5) * E(0,2) + (2/5) * E(1,1) + (1/5) * E(1,0) = 2/5 = 0.4$

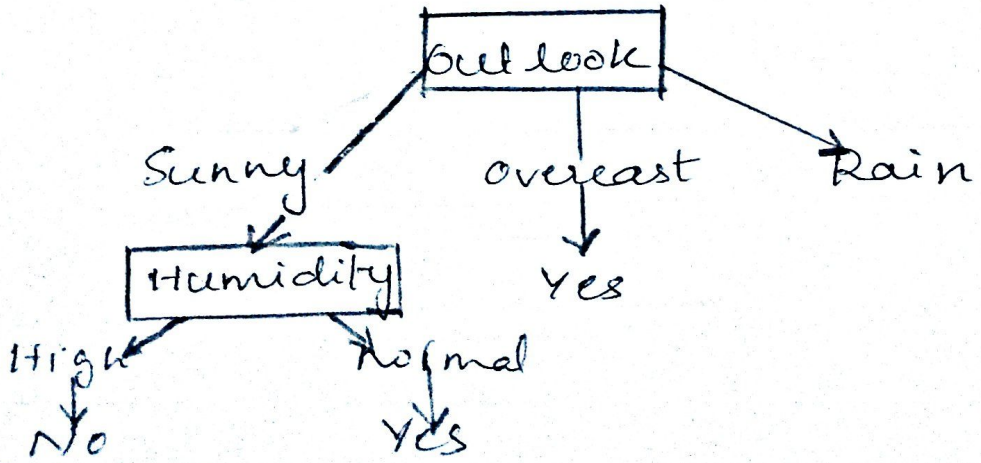
$Gain(Sunny, temperature) = 0.971 - 0.4 = 0.571$

Similarly for other attributes we get

$Gain(Sunny, Humidity) = 0.971$

$Gain(Sunny, Wind) = 0.020$

- Gain (Sunny, Humidity) has the largest value, so it is the node that comes under Sunny



7- The ID3 algorithm is run recursively on non-leaf branches until the data is classified

- Finally the tree looks as below