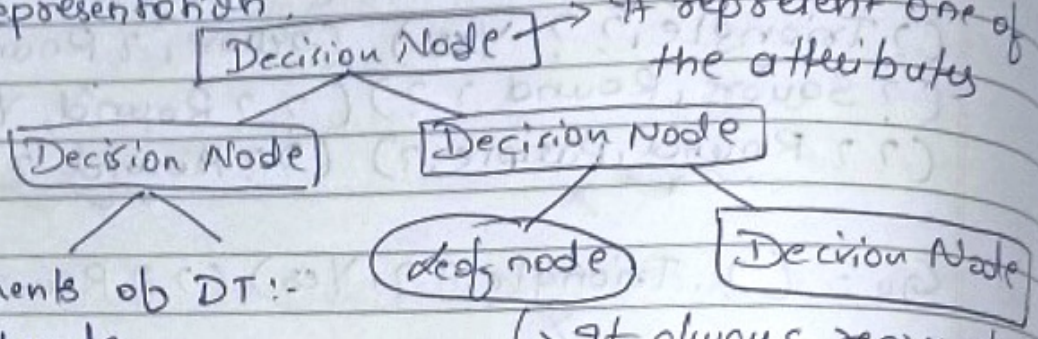
28/12/2023

Decision Tree learning.

-> popular supervised classification Algorithm.

-> inductive inference. (final conclusion)

-> Representation.



Components of DT:-

- Root node
- Internal node
- Leaf node
- Branches

A DT is a tree in which each branch represents a choice b/w a number of alternatives and each leaf node represents a decision.

A DT is a tree in which each node represents a feature or attribute, each branch represents a decision or rule each leaf represents an outcome.

Components of a DT:-

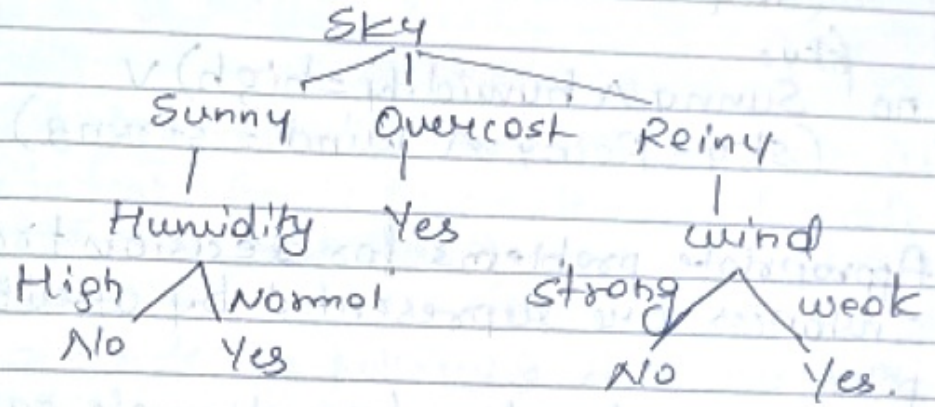
→ Root node refers to the start of the decision tree

→ Node is a condition with multiple outcomes in the tree

→ Leaf is the final decision or the end point of a node

→ In General, A DT represents a disjunction of conjunction of constraints on the attribute values of the instances.

→ Each path from the root to a leaf corresponds to a conjunction of attribute value & the tree itself corresponds to a disjunction of these conjunction.



(Rule / Expression: (Sky = Sunny ∧ Humidity = Normal) ∨ (Sky = Overcast) ∨ (Sky = Raining ∧ Wind = Weak)

Appropriate problems

1) Instances are represented by attribute value pairs
2) The target function has discrete output values.
3) Disjunctive descriptions may be required
4) Training data may contain errors & may contain missing attribute value.) write once again.

Rule/expression:-

yes: (sky = Sunny ∧ Humidity = Normal) ∨
(sky = overcast) ∨
(sky = Rain ∧ wind = weak)

sky =
no Sunny ∧ humidity = high) ∨
(sky = Rainy ∧ wind = strong)

Appropriate problems for decision tree learning

1) Instances are represented by attribute value pairs
2) The target function has discrete output values
3) Disjunctive description may be required.
4) Training data may contains errors and may contain missing attribute values.

29/12/2023

ID3 (Iterative Dichotomiser 3)
C4.5 (Successor of ID3)
CART (Classification & Regression Tree)

ID3

Based on the statistical valu of each attribute we decide which attribute should select as a root node. To do this we use diff Algorithms

* ID3.

→ learns the DT by constructing them top down begining with the Question "which attribute should be at the root of the tree". To answer this Question each attribute is tested using a statistical test. To determine how well it classifies training examples

* ID3(examples, target attribute, attributes)

- examples are the training examples
- target attributes is the att whose value has to be predicted by the tree
- att is the list of other attributes that may be tested by the tree.

→ Given a dataset how we build a tree using ID3
- The initial step is to create a root node.
- The att that best classifies the training example is taken as the root node. In order to choose the best attribute the following steps are followed by ID3.

Step 1:- Compute the entropy for the dataset i.e ENTROPY(S) S → Dataset

step 2:- for every att or feature
→ Calculate the entropy for all other values i.e entropy of A.
→ Calculate the average information entropy for the current attribute
→ Calculate Gain for the current attribute
step 3:- Pick the att with the highest gain value
step 4:- Repeat until we get the desired output.

Entropy :- Given a collec" S containing +ve & -ve examp the entropy of (S) is given as:
Entropy(S) = $-P_+ \log_2 P_+ - P_- \log_2 P_-$ when
S is a sample of training examples
P_+ is the proportion of +ve "
P_- " " " -ve "

Entropy is defined as the ~~entropy~~ amount of uncertainty in the dataset.

* It can also be represented as

$$\text{Entropy} = \sum_{i=1} -p_i \times \log_2(p_i)$$

$$\text{Entropy} = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

Average Info Entropy is calculated as

$$I(\text{Attribute}) = \sum \frac{p_i + n_i}{p+n} \text{Entropy}(A)$$

Info Gain / Gain

It is the difference in the entropy before & after splitting the dataset on attribute A.

$$\text{Gain} = \text{Entropy}(S) - I(\text{Attribute})$$

D

Sl	Outlook	Temp	humidity	wind	Playtennis
1	Sunny	hot	high	weak	No -
2	Sunny	hot	high	strong	No -
3	Overcast	hot	high	weak	Yes -
4	Rainy	mild	high	weak	Yes -
5	Rainy	cool	normal	weak	Yes -
6	Rainy	cool	normal	strong	No -
7	Overcast	cool	normal	strong	Yes -
8	Sunny	mild	high	weak	No -
9	Sunny	cool	normal	weak	Yes -
10	Rainy	mild	normal	weak	Yes -
11	Sunny	mild	normal	strong	Yes -
12	Overcast	mild	high	strong	Yes -
13	"	hot	normal	weak	Yes -
14	Rainy	mild	high	strong	No -

p = 9
n = 5
p+n = 14

$$\log_2\frac{9}{14} = \frac{\ln(9/14)}{\ln(2)} = \frac{-0.1518}{0.3610}$$

Step 1:- Calculated entropy for entire dataset

$$\text{Entropy}(S) = -\frac{9}{9+5}\log_2\left(\frac{9}{9+5}\right) - \frac{5}{9+5}\log_2\left(\frac{5}{9+5}\right)$$

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

$$= (-0.643)(-0.637) - (0.357)(-1.486)$$

$$= 0.4096 + 0.5305$$

$$= 0.940.$$

of S

→ Entropy is 0 when all the members belongs to class.

→ Entropy is 1 " the collec" contains equal no. of +ve & -ve examples

→ If the collec" contains unequal no of +ve & -ve examples then the entropy will be b/w 0 & 1.

22/01/2024

Step 2: calculate entropy for each attribute

outlook	+ve	-ve
Sunny	2	3
Rainy	3	2
Overcast	4	0

calculate entropy(outlook = sunny)

$$= \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

$$= -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

$$= 0.4(-1.3219) - (0.6)(-0.2218)/\log_2 2 = (0.737$$

$$0.5287 + 0.3388\,(0.4422)$$

$$= 0.971$$

entropy(outlook = rainy)

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

$$= -(0.6)(-0.7379) - (0.4)(-1.3219)$$

$$= 0.4422 + 0.5287 = 0.971$$

entropy(outlook = overcast)

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

$-1(0) - 0$

$= 0$

Calculate average Information Entropy.

$I(\text{Outlook}) = \frac{P_{sunny} + n_{sunny}}{p+n}\text{Entropy(outlook=sunny)} + \frac{P_{sunny} + n_{sunny}}{p+n}\text{(Entropy(outlook=sunny))} + \frac{P_{overcast} + n_{overcast}}{p+n}$

(Entropy(outlook = overcast))

$= \frac{2+3}{14}(0.971) + \frac{3+2}{14}(0.971) + \frac{4+0}{14}(0)$

$= 0.357(0.971) + 0.357(0.971) + 0$

$= 0.3466 + 0.3466$

$= 0.693$

Calculate Gain for the outlook

Gain = Entropy(S) − I(Outlook)

$= 0.940 - 0.693$

Gain = 0.247

Calculate Entropy for Temp

Temp	p	n
hot	2	2
mild	4	2
cool	3	1

$En(\text{Temp} = \text{hot}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)$

$= -0.5(-1) - 0.5(-1)$

$= 0.5 + 0.5$

$= 1$

$En(\text{Temp} = \text{mild}) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)$

$= \frac{2}{3}(0.66)(0.353) - 0.33(0)$

$= 0.232$

$= 0.923$

$En(\text{Temp} = \text{cold}) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$

$= -(0.75)(\cdot$

$= 0.811$

$I(\text{Temp}) = \frac{2+2}{14}(1) + \frac{4+2}{14}(0.92) + \frac{3+1}{14}(0.811)$

$0.285 + 0.394 + 0.231$

$= 0.912$

Gain $= 0.940 - 0.912$

$= 0.028$

humidity	+ve	-ve
high	3	4
normal	6	1

$-\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)$

$-(0.428)(-1.222) - (0.571)(-0.807)$

$= 0.985$ $\quad 0.523 + 0.460$

$-\frac{6}{7}\log\left(\frac{6}{7}\right) - \frac{1}{7}\log\left(\frac{1}{7}\right)$

$-(0.857)(-2.564) - 0.142(-2.807)$

$= -2.197 + 0.398$ $\quad = 0.591$

$I(\text{humidity}) = \frac{3+4}{14}(0.985) + \frac{6+1}{14}(0.591)$

$= 0.788$

$0.940 - 0.788$

Gain = 0.152

Wind	+ve	-ve
weak	6	2
Strong	23	3

$-\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$

$= -(0.75)(-0.415) - (0.25)(-2)$

$= 0.3112 + 0.5$

$= 0.811$

$-\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$

~~$(0.4)(-1.321) - (0.6)(-0.769)$~~

$-0.5(-1) - 0.5(-1)$

$0.5 + 0.5 = 1$

$I = \frac{6+2}{14}(0.811) + \frac{3+3}{14}(1)$

$= 0.571(0.811) + 0.428$

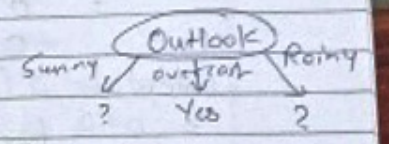
$0.463 + 0.428$

$= 0.891$

Gain = $0.940 - 0.891$

$= 0.049$

Gain	Attribute	
0.247	Outlook	✓ root node (high gain value)
0.028	temp	
0.152	humidity	
0.049	wind	



02/01/2024

Write table when outlook = sunny.

Outlook	Temp	Humidity	wind	Play tennis
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

p = 2 n = 3

Entropy(Sunny)

$= -\frac{2}{2+3}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$-0.4(-1.3219) - 0.6(-0.7369)$

$+0.5288 + 0.442$

$= 0.971$

Temp	+ve	-ve
hot	0	2
mild	1	1
cool	1	0

temp hot $= -\frac{0}{0+2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\left(\log_2\left(\frac{2}{2}\right)\right)$

$+0(0) - 1(0)$

$= 0$

temp = mild $\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$

$= 1$

temp= cool $\frac{1}{0}\log_2\left(\frac{1}{0}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)$

$= 0 \times 0 -$

$= 0.$

Calculate I(temp)

$= \frac{0+2}{2+3}\log(0) + \frac{1+1}{2+3}(1) + \frac{1+0}{5}(0)$

$= 0 + 0.4 + 0$

I(temp) = 0.4

Calculate Gain

Gain = Entropy(Sunny) - I (temp)

$= 0.971 - 0.4$

Gain = 0.571

Humidity	+ve	-ve
high	0	3
normal	2	0

Hum = high = $\frac{0}{0+3}\log\left(\frac{0}{3}\right) - \frac{3}{0+3}\log\left(\frac{3}{3}\right)$

$= 0 \times (\ ) - 3(0)$

$= 0$

Hum = normal = $\frac{2}{0+2}\left(\log\frac{2}{0}\right) - \frac{0}{0+2}\log_2\left(\frac{0}{2}\right)$

$= 2(0) - 0 \times 1$

$= 0$

Calculate.

$I(\text{Humidity}) = \frac{0+3}{5}(0) + \frac{2+0}{5}(0) = 0$

Gain = 0.971 - 0

Gain = 0.971

wind	+ve	-ve
weak	1	2
strong	1	1

weak $= -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)$

$0.33(-1.584) - 0.67(-0.584)$

$0.5228 + 0.3911 = 0.912$

$= 0.918$

strong = 1

$I(\text{wind}) = \frac{1+2}{5}(0.918) + \frac{1+1}{5}(1)$

$= \frac{3}{5}(0.918) + \frac{2}{5}(1)$

$0.6\ (0.918)$

$= 0.550 + 0.4$

$= 0.950.$

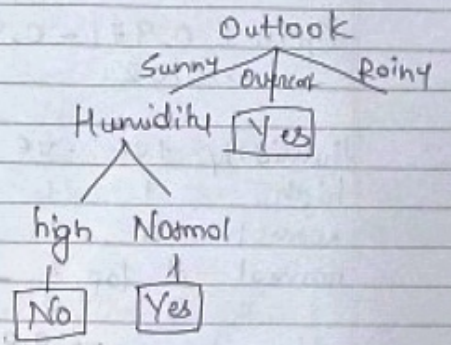
Gain = 0.971 - 0.950

Gain. = 0.02

Temp 0.571

Humidity 0.971

Wind 0.02



Outlook	temp	Humidity	wind	Play tennis
Rainy	mild	high	weak	Yes
Rainy	cool	normal	weak	Yes
Rainy	cool	normal	strong	No
Rainy	mild	normal	weak	Yes
Rainy	mild	high	strong	No.

temp	+ve	-ve
mild	2	1
cool	1	1

mild $-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}$

$0.66(-0.5849) - 0.33\log(-1.584)$

$0.386 + 0.522$

$= 0.918$

cool = 1

$I(temp) = \frac{3}{5}(0.971) + \frac{2}{5}(1)$

$0.6(0.978) + 0.4$

$0.586 + 0.4$

$= 0.95$

Gain $= 0.971 - 0.95$

$= 0.02.$

Humidity	+ve	-ve
high	1	1
normal	2	1

normal $\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$

$= 0.918$

high = 1

$I(humidity) = \frac{1+1}{5}(1) + \frac{2+1}{5}(0.918)$

$= 0.95$

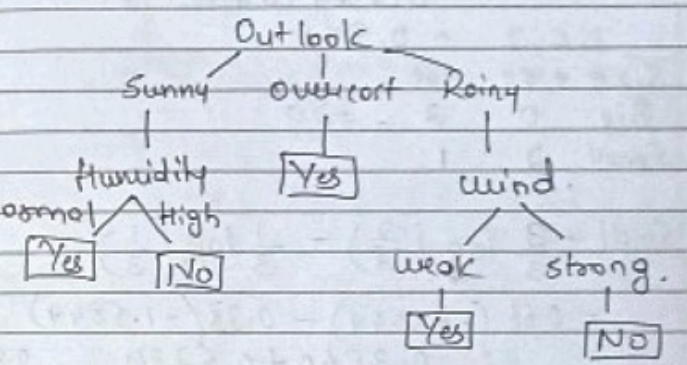
Gain $= 0.02$

wind	+ve	-ve	
weak	3	0	= 0
strong	0	2	= 0

$I = 0$

$G = 0.971.$

	Gain
Temp	0.02
Humidity	0.02
wind	0.971



—— I^st Internals.

05/01/2023

Issues in Decision Tree Learning.

* Avoiding overfitting the data.
* Incorporating Continous valued attributes
* Alternate measures for selecting the attribute
* Handling training examples with missing attribute value
* " Attributes with differ in cost.

Ex

Size	Color	Shape	Class / Label
Big	Red	Circle	No
Small	Red	Triangle	No
Small	Red	Circle	Yes
Big	Blue	Circle	No
Small	Blue	Circle	Yes

+ve -ve
2 3

Entropy (S)

$-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$

$-0.4(-1.3219) - 0.6(-0.7369)$

$0.52896 + 0.4421$

$= 0.971$

Size	+ve	-ve	
Big	0	2	=> 0
Small	2	1	

Small $-\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$

$-0.66(-0.5849) - 0.33(-1.5849)$

$0.3860 + 0.5230$ 0.909)

$= 0.98$

$I(\text{Size}) = \frac{0+2}{5}(0) + \frac{2+1}{5}(0.918)$

$0 + 0.6(0.918)$

$= 0.5508$

Gain $= 0.971 - 0.5508$

$= 0.42$

Color	+ve	-ve	
Red	2	1	=> 0.918
Blue	1	1	=> 1

$\frac{2+1}{5}(0.918) + \frac{1+1}{5}(1)$

$= 0.5508 + 0.4$

$I(\text{Color}) = 0.9508 = 0.951$

Gain $= 0.971 - 0.9508$

$= 0.02$

Shape	+ve	-ve	
Circle	1	3	=> 0.811
Triangle	0	1	=> 0.

$-\frac{1}{4}\log_2\frac{1}{4}$

$-0.2(-2.3219) +$ 0.811

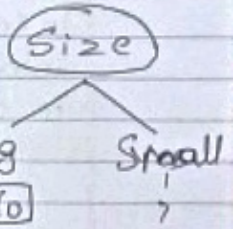
$I(\text{Shape}) = 0.8$

Gain $= 0.171$ $(0.971 - 0.8)$

Size 0.42
Color 0.02
Shape 0.171

Size	Color	Shape	Label
Small	Red	Triangle	No
Small	Red	Circle	Yes
Small	Blue	Circle	Yes

+ve -ve
2 1



Color	+ve	-ve	
Red	1	1	=> 1
Blue	1	0	=> 0

$I(\text{Color}) = \frac{1+1}{3}(1) + \frac{1+0}{3}(0)$

$= 0.66$

Gain $0.918 - 0.66$

Gain $= 0.258$

Entropy $(S)_{\text{small}}$

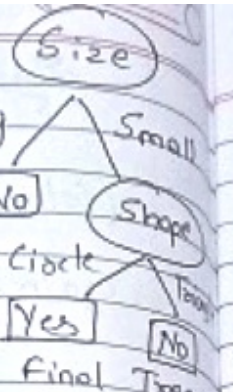
$-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$

$= 0.918.$

Shape	+ve	-ve	
Triangle	0	1	=> 0
Circle	2	0	=> 0

$I(Shape) = 0$

Gain = 0.918



Final Tree

Continued.

Issues in Decesion ~~Training~~ Tree learning

* Incorporating Continued valued attributes.
  => If we have " " s we can make it as Discrete value by using Discrete valued intervals.

  ex:- Age - 4, 10, 14, 24, 24, 34, 64, 84, 95

  [child, Young, Adult, Teenage Senior Citizen]
  0-5 6-12 20-35 13-19 60-79 80-100

In some attributes we can find Discrete valued intervals are we other method where we find to calculate threshold value 'c'

step1: sort

2:. Look for where value classification is changing

3:. And calculate Gain as the previous sum.

Temp	40	48	60	72	80	90
Play Tennis	No	No	Yes	Yes	Yes	No

$\frac{48+60}{2} = 54$

Threshold value 'c'

54

48 48 | 60 72 80 90

$\frac{80+90}{2} = 85$

85

K-NN Algorithm:-

-> K-Nearest Neighbour Algorithm
-> Multi Classification
-> K stands for no. of neighbours
-> The minimum value of K is always 3

Distance Measures.

-> Euclidean Distance

$$E = \sqrt{(x_2 - x_1)^2 + (Y_2 - Y_1)^2}$$

-> Manhattan Distance

$$|x_2 - x_1| + |Y_2 - Y_1|$$

Value of K will be changed based on the Accuracy

Confusion Matrix

	0	1
0	00	01
1	10	11

0 -> 0 -> True +ve
0 -> 1 -> False Neg
1 -> 0 -> " +ve
1 -> 1 True -ve

No of Correct predictions

$$\text{Accuracy} = \frac{00 + 11}{00 + 01 + 10 + 11}$$

Issues continued.

3) Alternate measures for selecting the attribute.
   Information Gain

It favours for the attributes which has more value.

Disad:- Broader tree at one level.

Alternate measure is Gain Ratio.

↓

Split information

$$\text{Split info} = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log \frac{|S_i|}{|S|}$$

where S_1 to S_c are 'c' subsets of examples resulting from partitioning 'S' by the 'c' valued attribute A.

$$\text{Gain Ratio}(S,A) = \frac{\text{Gain}(S,A)}{\text{Split info}(S,A)}$$

28/01/2024

1) Handle the training examples with missing attribute

Gain(S, A) A(x) = Temp / Humidity / Wind / Outlook (Hot, Mild, Cool; High, Normal; Strong, Weak; Sunny, Rain, Overcast)

~~Training examples~~ <x, c(x)>, A(x) => Null

2 strategies

1) whichever value is max in that attribute then fill it with that max value irrespective of c(x)

2) c(x) of that instance if it +ve or -ve check the target concept whichever is +ve take or -ve then fill it correspondingly. acc to c(x) value

Handling attributes with diff costs.

Some times " are associated with cost/weight.

ep:- to predict medical disease

Temp | Pulse | BloodTest | Biopsy | Target

$$\frac{\text{Gain}^2(S,A)}{\text{Cost}(A)}$$

null

Based on ~~Get~~ value.

imp Avoiding Overfitting in the Data

less more

~~more~~ error in train data & ~~less~~ errors in test data.

when we apply hypothesis on training sample it show less error

" " " some " " testing sample it shows more error because it has learnt more than cut is required.

Given an hypothesis h, it said to overfit the training examples if the hypothesis h is performing better over the training set than over the testing set. So Training error(h) < Testing error(h).

error less accuracy more.

why is this overfitting causing in ID3

-> Noise.

-> If Size of data set is small they are not in a position to test.

eg:- 1 (Noise)

(outlook = sunny, Temp = hot, Humidity = Normal, wind = storm, Play Tennis = No)

(for the Tree)