Decision Tree

Friday, January 15, 2021 8:17 AM

ID3

(1) Compute the ENTROPY for data-set ENTROPY(S)

(2) For every attribute/feature

(A) Calculate entropy for all other values ENTROPY(A)

(B) Take AVERAGE INFORMATION ENTROPY for the current attribute

(C) Calculate GAIN for the current attribute

(3) Pick the HIGHEST GAIN ATTRIBUTE

(4) REPEAT until we get the tree we desired.

$$\text{Entropy}(S) = \sum -p(I).\log_2 p(I)$$

$S = p+n$ (Total sample space); Yes $\downarrow p$ / No $\downarrow n$

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

($\frac{p}{p+n}$, $\frac{n}{p+n}$ ← S; $\frac{p}{p+n}$, $\frac{n}{p+n}$ → S)

$$\text{Entropy}(A) = \sum_{i=1}^{x} \frac{p_i+n_i}{p+n} I(p_i,n_i) \quad ---- (II)$$

(A: Attribute)

$$Gain(S,A) = Entropy(S) - \sum \left[P(S/A) \cdot Entropy(S/A) \right]$$

$$\left\{ \log_2 x = \frac{\log_{10} x}{\log_{10} 2} \right\}$$

$$Gain = Entropy(S) - I(Attribute)$$

$$I(p,n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$

$$Entropy(Decision) = -P(yes) \times \log_2 p(yes) - p(no) * \log_2 P(no)$$

yes = 9
No = 5
S = yes+No = 9+5 = 14

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

$$= 0.94$$

Attributes = {Outlook, Temperature, Humidity, Wind}

For each attribute (let say wind)

- Calculate Entropy for each wind i.e strong & weak

Wind	Play Tennis
Weak	No
Weak	Yes
Weak	Yes
Weak	Yes
Weak	No
Weak	Yes
Weak	Yes
Weak	Yes

Wind	Play Tennis
Strong	No
Strong	No
Strong	Yes
Strong	Yes
Strong	Yes
Strong	No

wind	P	n	Entropy
strong	3	3	1
weak	6	2	0.811

$$\text{Entropy}(S, w=\text{strong}) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right)\times 2 = ?$$

$$\text{Entropy}(S, w=\text{weak}) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) = ?$$

Average Information Entropy,

$$I(\text{Wind}) = \frac{P_{\text{strong}} + n_{\text{strong}}}{\cdot}\cdot \text{Entropy}(\text{Wind}=\text{strong}) +$$

$$\frac{p+n}{}$$

$$\frac{p_{weak}+n_{weak}}{p+n} \cdot \text{Entropy (wind: weak)}$$

$$= \frac{3+3}{9+5} \times 1 + \frac{6+2}{9+5} \cdot 0.811$$

$$= 0.892$$

$\text{Gain}^{(S,A)} = \text{Entropy}(S) - \text{IC Attribute}) = 0.94 - 0.892$
$= 0.048$

Humidity	Yes	No	Entropy
High	3	4	0.985
Normal	6	1	0.591

Entropy (humidity)
$= 0.940$

IC(Humidity) $= 0.788$, Gain $= 0.152$

Temperature	Yes	NO	Entropy
Cool	3	1	0.811
Hot	2	2	

Mild	4	2	0.918

Entropy (temp) = 0.940

I(temperature) = 0.911, Gain = 0.029

Outlook

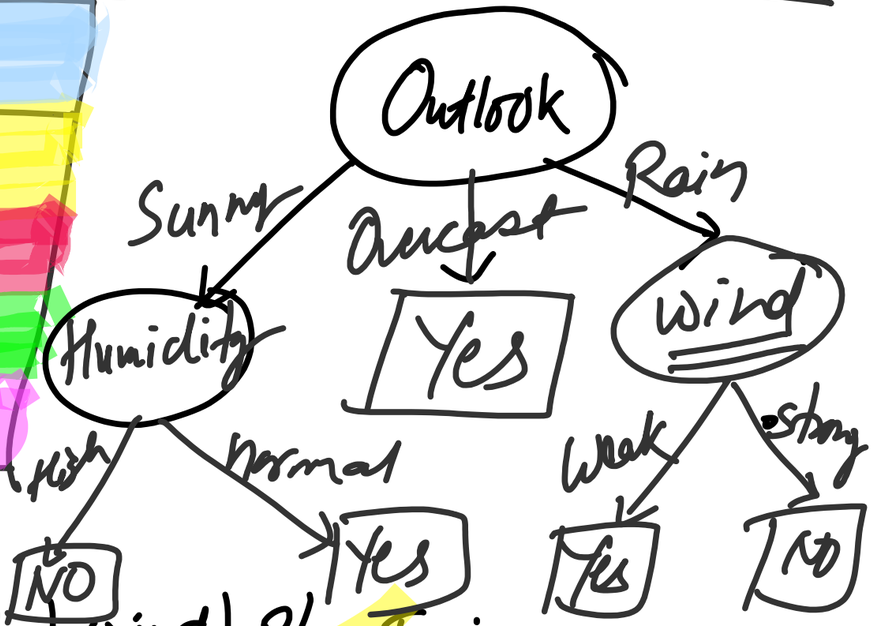
Entropy = 0.941

I(outlook) = 0.693

Gain(outlook) = 0247

Outlook	Yes	No	Entropy
Sunny	2	3	0.971
Overcast	4	0	0
Rain	3	2	0.971

Attribute	Gain
Outlook	0.247
Temperature	0.029
Humidity	0.152
Wind	0.048



Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No

Sunny	Cold	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

Humidity	Yes	No	Entropy
High	0	3	?
Normal	2	0	?

Entropy = 0.971

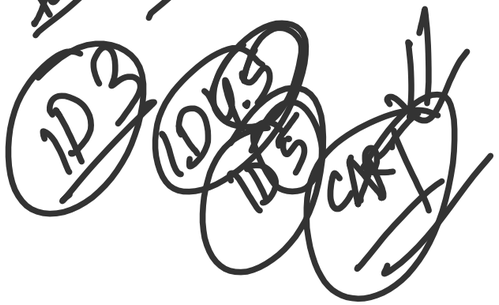
I(Humidity) = 0

Gain(Humidity) = 0.971

Outlook = Sunny

Attributes	Gain
Temperature	0.571
Humidity	0.971
Wind	0.02

Decision Tree



For the following Medical Data, create decision tree using ID3 algorithm.

Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
Yes	Yes	Yes	Yes	Yes	Strep Throat
No	No	No	Yes	Yes	Allergy
Yes	Yes	No	Yes	No	Cold
Yes	No	Yes	No	No	Strep Throat
No	Yes	No	Yes	No	Cold

No	No	No	Yes	No	Allergy
No	No	Yes	No	No	Strep Throat
Yes	No	No	Yes	Yes	Allergy
No	Yes	No	Yes	Yes	Cold
Yes	Yes	No	Yes	Yes	Cold

Total sample space (S) = Strep Throat + Allergy + Cold = 10

(0) (0) (4)

$$\text{Info Gain}(S) = -\left[\frac{3}{10}\log_2\left(\frac{3}{10}\right) + \frac{3}{10}\log_2\frac{3}{10} + \frac{4}{10}\log_2\frac{4}{10}\right] = 1.57$$

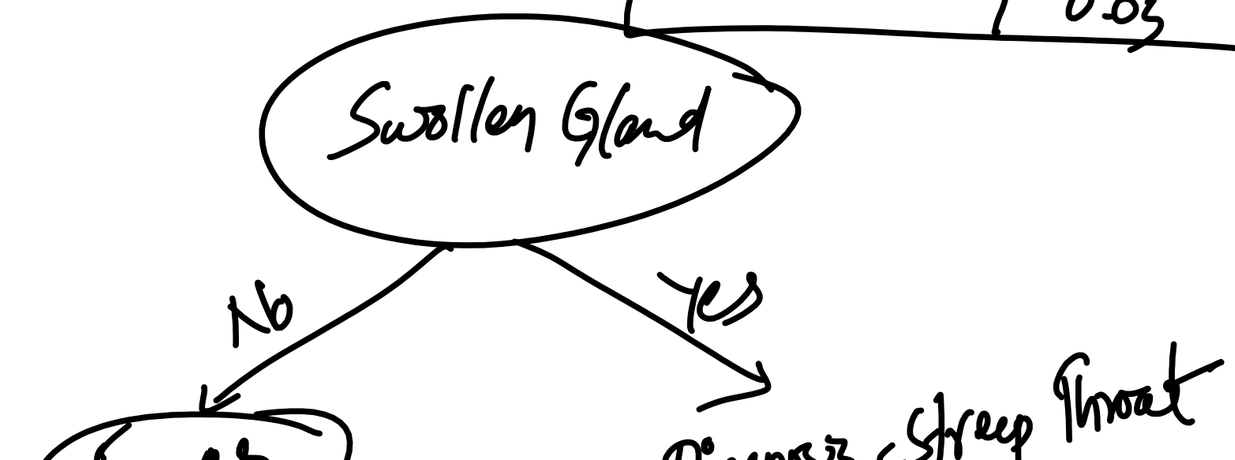
Finding splitting attribute: (Select attribute with highest gain)

(1) Sore Throat:-

Decision	S.T.	A	C
Yes	2	1	2
NO	1	2	2

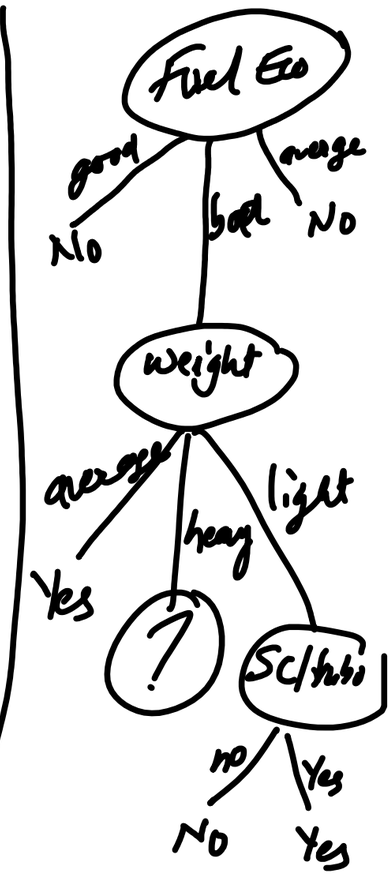
Gain of each attribute

Attribute	Gain
Sore Throat	0.05
Fever	0.72
Swollen Glands	0.88
Congestion	0.45
Headache	0.05



ID3 Question from the following data, find which car is fast using decision tree with ID3 algorithm.

Model	Engine	SC/Turbo	Weight	Fuel Eco.	Fast
Prius	small	no	average	good	no
Civic	small	no	light	average	no
WRX STI	small	yes	average	bad	yes
M3	medium	no	heavy	bad	yes
RS4	large	no	average	bad	yes
GTI	medium	no	light	bad	no
XJR	large	yes	heavy	bad	no
S500	large	no	heavy	bad	no
911	medium	yes	light	bad	yes
Corvette	large	no	average	bad	yes
Insight	small	no	light	good	no
RSX	small	no	average	average	no
IS350	medium	no	heavy	bad	no
MR2	small	yes	average	average	no
E320	medium	no	heavy	bad	no



CART

- CART (for classification and regression tree) is often used as generic acronomy for decision tree, although it is a specific implementation.
- Similar to C4.5, CART can handle continuous attributes

- CART uses the Gini diversity index which is defined in equation as

$$Gini_X = 1 - \sum_{x \in X} P(x)^2$$

- CART constructs a sequence of subtrees, uses cross-validation to estimate the misclassification cost of each subtree and chooses the one with the lowest cost.

CART 2 Formulas

Gini index is a metric for classification task in CART

① Gini index (attribute = value) $= 1 - \sum_{i=1}^{N} (P_i)^2$

② Gini index (attribute) $= \sum P_v \times GI(v)$

v = value

Outlook	Yes	No	No. of metrices
Sunny	2	3	5
Overcast	4	0	4
Rain	3	2	5

$$Gini(outlook = sunny) = 1 - \sum_{i=1}^{N} (P_i)^2$$

$$= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

$$Gini(outlook = overcast) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2$$

$$= 0$$

$$Gini(outlook = rain) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2$$

$$= 0.48$$

Now calculate weighted sum of Gini Index for outlook

Calculate weighted sum of Gini index for outlook,

$$\text{Gini(outlook)} = \sum_{v \in \text{value}} P_v \times GI(v) = \underbrace{\frac{5}{14} \times 0.48}_{v=\text{sunny}} + \underbrace{\frac{4}{14} \times 0}_{v=\text{overcast}} + \underbrace{\frac{5}{14} \times 0.48}_{v=\text{rain}} = 0.342$$

Temperature	Yes	No	No. of metrices
Hot	2	2	4
Mild	4	2	6
Cool	3	1	4

$$\text{Gini(temperature=hot)} = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$$

$$\text{Gini(temperature=mild)} = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.444$$

$$\text{Gini(temperature=cool)} = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$$

Weighted sum for temperature, $\text{Gini(temperature)} = \frac{4}{14} \times 0.5 + \frac{6}{14} \times 0.444 + \frac{4}{14} \times 0.375$

$$= 0.440$$

Humidity	Yes	No	No. of metrics
High	3	4	7
Normal	6	1	7

$$\text{Gini(humidity=high)} = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.489$$

$$\text{Gini(humidity=normal)} = 1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = 0.244$$

Weighted sum of Gini index for humidity,

Gini(Humidity) $= \frac{7}{14} \times 0.489 + \frac{7}{14} \times 0.244 = 0.367$

Wind	Yes	No	No. of instances
Weak	6	2	8
strong	3	3	6

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

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

weighted sum of Gini Indices for wind,

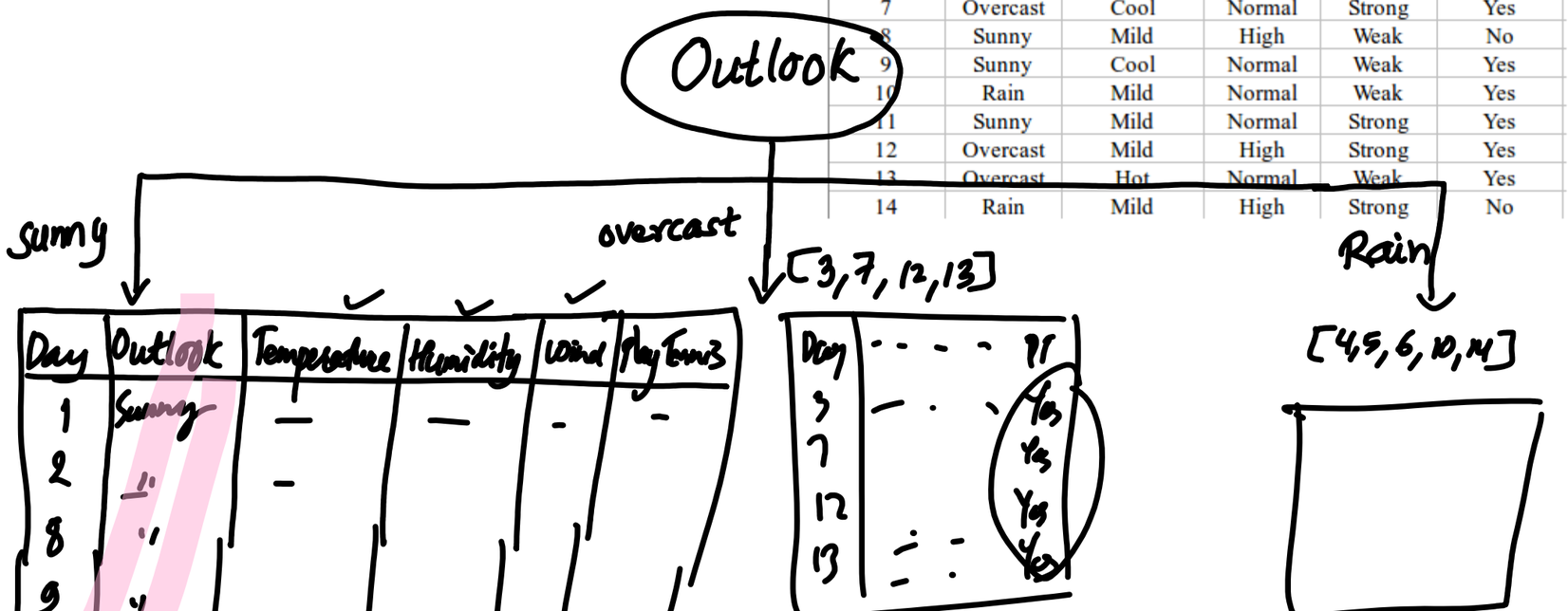
Gini(wind) $= \frac{8}{14} \times 0.375 + \frac{6}{14} \times 0.5 = 0.428$

Decision for node value

Features	Gini Index
Outlook	0.342
Temperature	0.440
Humidity	0.367
Wind	0.428

⇒ lowest value is the decision factor in CART

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



11	1		

Sunny

Temperature	Yes	No	No. of instances	
Hot	0	2	2	0
Mild	1	0	1	0
Cool	1	1	2	0.5

Gini (Outlook = sunny and temperature = hot)

$$= 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0$$

Weighted sum

Gini (Outlook = sunny) $= \frac{2}{5} \times 0 + \frac{1}{5} \times 0 + \frac{2}{5} \times 0.5$

$= 0.2$

Outlook = Sunny

Features	Gini Index
Temperature	
Humidity	
Wind