Decision Tree Based clustering

ID3 (Iterative Dichotomiser)

It is a decision tree learning algorithm invented by Ross Quinlan used to generate a decision tree from a dataset.

ID3 is the precursor to C4.5 algorithm and is typically used in Machine learning and natural language preprocessing domains.

Example:- Play Tennis Database:-

Day	Outlook	Temp	Humidity	Wind	Play
1	Sunny	80	High	Week	No
2	Sunny	81	High	Strong	No
3	Overcast	82	High	Week	Yes
4	Rain	65	High	Week	Yes
5	Rain	40	Normal	Week	yes
6	Rain	41	Normal	Strong	No
7	Overcast	42	Normal	Strong	Yes
8	Sunny	66	High	Weak	No
9	Sunny	43	Normal	Weak	Yes
10	Rain	67	Normal	Weak	Yes
11	Sunny	68	Normal	Strong	Yes
12	Overcast	69	High	Strong	Yes
13	Overcast	83	Normal	Weak	Yes
14	Rain	70	High	Strong	No

Page No.
Date:

Steps-

(1) Start with a training data set, which we call as 'S'. It should have attributes & classification

(2) Determine the best attribute in the dataset S.

(3) Split S into subsets that correspond to the possible values of the best attribute

(4) Make a decision tree node that contains the best attribute.

(5) Recursively make new decision tree nodes, with the subsets of data created in step 3. Attributes can't be reused.
If there is no more attributes to split on, choose the most popular classification.

⇒ In the above example database attributes are outlook, Temperature, humidity and wind

And The classification is → Yes
→ NO

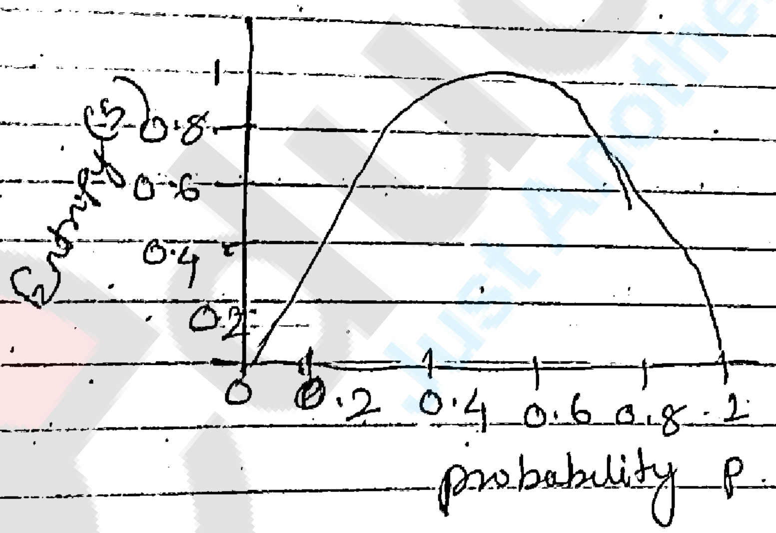
To find out the best attribute we have to calculate (1) Entropy
(2) Information gain

Entropy is an information theory and machine learning sense, measures the homogeneity of data set S's classification.

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

where $c \rightarrow$ corresponds to the number of different classifications.

$p_i \rightarrow$ corresponds to the proportion of the data with the classification i.



Entropy Calculation

In the above example two classes Yes & NO

$$Entropy(S) = \sum_{i=1}^{2} -p_i \log_2 p_i$$

$= -P_{yes} \log_2 P_{yes} - P_{NO} \log_2 P_{NO}$

No of Yes = 9 out of 14

$P_{yes} = \frac{9}{14}$

No of NO = 5

$P_{NO} = \frac{5}{14}$

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

$$\boxed{\text{Entropy}(S) = 0.940}$$

Information Gain measures the reduction in entropy that results from partitioning the data on an attribut A, that is it represents how effective an attribute is at classifying the data.

Given a set of training data S and attribute A the formula for information gain is

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

Calculation of Information Gain for all Attribute

(1) Outlook

Values of Outlook = Sunny, Overcast, Rain

Gain(S, outlook) = Entropy(S) − [|S_sunny|/|S| Entropy(S_sunny) + |S_overcast|/|S| Entropy(S_overcast) + |S_Rain|/|S| Entropy(S_Rain)]

S = 14 (9 yes 5 NO)

S_sunny = 5 (2 yes 3 NO)

⇒ P_yes = 2/5 P_NO = 3/5

S_overcast = 4 (4 yes 0 NO) P_yes = 4/4 = 1 P_NO = 0/4 = 0

S_Rain = 5 (3 yes 2 NO)

P_yes = 3/5 P_NO = 2/5

Entropy(S_sunny) = −P_yes log₂ P_yes − P_NO log₂ P_NO

= −2/5 log₂ 2/5 − 3/5 log₂ 3/5

= .971

Entropy(S_overcast) = −P_yes log₂ P_yes − P_NO log₂ P_NO

= −1 log₂ 1 − 0 log₂ 0

= 0

Entropy(S_Rain) = −P_yes log₂ P_yes − P_NO log₂ P_NO

= −3/5 log₂ 3/5 − 2/5 log₂ 2/5

= .971

Gain(S, outlook) = .940 − [5/14 × 0.971 + 4/14 × 0 + 5/14 × 0.971]

Gain(S, outlook) = 0.247

Page No.
Date: 1 1

Temperature (no calculation for numerical data values)

Gain (S, Humidity) =

Entropy (S) $-\left(\frac{7}{14}\right)$ Entropy (S_{High})

$-\frac{7}{14}$ Entropy (S_{Normal})

[∵ values of = Humidity is High or Normal
High = 7 Normal = 7
(6 Yes 1 No)]

Entropy S_{High} = 7 (3 Yes 4 No)

Entropy (S_{High}) $= -P_{Yes} \log_2 P_{Yes} - P_{No} \log_2 P_{No}$

$= -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7}$

$= 0.985$

Entropy (S_{Normal}) $= -P_{Yes} \log_2 P_{No} - P_{No} \log_2 P_{No}$

$= -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7}$

$= 0.592$

Gain (S, Humidity) =

$0.940 - \frac{7}{14} \times 0.985 - \frac{7}{14} \times 0.592$

$= 0.152$

Wind

Page No. 5
Date: 11

values of wind : weak, strong

$S_{weak} = 8$ (6 yes 2 No)

$S_{strong} = 6$ (3 yes 3 No)

$Entropy(S_{weak}) = -P_{yes}\log_2 P_{yes} - P_{No}\log_2 P_{No}$

$= -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$

$= 0.811$

$Entropy(S_{strong}) = -P_{yes}\log_2 P_{yes} - P_{No}\log_2 P_{No}$

$= -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$

$= 1$

$Gain(S, wind) = Entropy(S) - \frac{8}{14} \times 0.811$

$- \frac{6}{14} \times 1$

$= 0.940 - \left(\frac{8}{14}\right) \cdot 811 - \frac{6}{14}$

$= 0.48$

Among all the information Gain values the Gain(S, Outlook) is high.

→ So Outlook Attribute will be choosen as root.

→ The decision tree is expanded to cover outlook's possible values.