Unit-3

Entropy and Information Gain

Q) Why entropy and Information gain required?

In the given dataset we have 4 attributes in this which attribute to be selected as root node which has highest priority. Once we calculate the information gain of all attribute we can draw decision tree.

Day	Outlook	Temp	Humidity	Wind	Play tennis
01	Sunny	Hot	High	Weak	No
02	Sunny	Hot	High	Strong	No
03	Overcast	Hot	High	Weak	Yes
04	Rain	Mild	High	Weak	Yes
05	Rain	Cool	Normal	Weak	Yes
06	Rain	Cool	Normal	Strong	No
07	Overcast	Cool	Normal	Strong	Yes
08	Sunny	Mild	High	Weak	No
09	Sunny	Cool	Normal	Weak	Yes
010	Rain	Mild	Normal	Weak	Yes
011	Sunny	Mild	Normal	Strong	Yes
012	Overcast	Mild	High	Strong	Yes
013	Overcast	Hot	Normal	Weak	Yes
014	Rain	Mild	High	Strong	No

If you want to calculate the information gain, we need to calculate entropy. Given the entropy we can calculate the information gain and given the information gain we can select the important attributes for our decision tree.

What is Entropy?

→ It measures the impurity of a collection of examples.

S → a collection of training examples

P_+ → Proportion of positive example in S.

P_- → Proportion of negative example in S.

$$\text{Entropy}(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

Important characteristics of entropy examples

only +ve ⇒ $\text{Entropy}([14+, 0-]) = -14/14 \log_2 (14/14) - 0 \log_2 0 = 0$

only -ve ⇒ $\text{Entropy}([0+, 14-]) = -0 \log_2 (0) - 14/14 \log_2 (14/14) = 0$

Combination ⇒ $\text{Entropy}([9+, 5-]) = -9/14 \log_2 (9/14) - 5/14 \log_2 (5/14)$

$= 0.94$

Equal ⇒ $\text{Entropy}([7+, 7-]) = -7/14 \log_2 (7/14) - 7/14 \log_2 (7/14)$

$= \frac{1}{2} + \frac{1}{2} = 1$

What is Information gain?

→ It measures the expected reduction in entropy caused by partioning the example according to an attribute.

More precisely the information gain, Gain (S, A) of an attribute A, relative to collection of example S is defined as

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

Now let's take one example to calculate information gain of attribute Wind.

values [wind] = weak, strong

S = [9+, 5-]

S_{weak} ← [6+, 2-]

S_{strong} ← [3+, 3-]

$$Gain(S, wind) = Entropy(S) - \sum_{v \in (weak, strong)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= Entropy(S) - (8/14)\, Entropy(S_{weak}) - (6/14)\, Entropy(S_{strong})$$

$$= 0.940 - \left(\frac{8}{14}\right) \times 0.811 - \left(\frac{6}{14}\right) \times 1.00$$

$$= 0.048$$

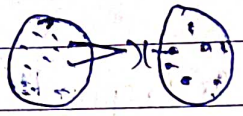
This is how we can calculate the information gain of each attribute and after calculations we can select which attribute has maximum portion and we can build the decision tree.

Q) Nearest Neighbour Algorithm

→ query ⇒ X = [Maths = 6, CS = 8), (K = 3)

	Maths	CS	Result
1)	4	3	Fail
2)	6	7	Pass
3)	7	8	Pass
4)	5	5	Fail
5)	8	8	Pass

For the given dataset and the query given X we need to check whether this student is passed or failed. So here K is 3, it means whatever the result we get should be nearer to it's neighbours.



so to compute this we use euclidian distance

$$d = \sqrt{(|x_{0_1} - x_{A1}|)^2 + (|x_{0_2} - x_{A2}|)^2}$$

(I) $\sqrt{(6-4)^2 + (8-3)^2} = \sqrt{29} = 5.38$

(II) $\sqrt{(6-6)^2 + (8-7)^2} = \sqrt{1} = 1$

(III) $\sqrt{(6-7)^2 + (8-8)^2} = 1$

(IV) $\sqrt{(6-5)^2 + (8-5)^2} = \sqrt{10} = 3.16$

(V) $\sqrt{(6-8)^2 + (8-8)^2} = 2$

so here, the nearest value II is pass, III is pass & V is pass & there is no fail here. So 3>0. So we declare X student is passed.

Probabilistic Reasoning

Baye's Theorem

Describes the probability of an event; based on prior knowledge of conditions that might be related to the event.

Theorem is stated as follows

$$P(B/A) = \frac{P(A/B) \cdot P(B)}{P(A)}$$

P(B/A) ⇒ Likelihood (Probability of evidence)
P(B) ⇒ Prior Probability of B.
P(A/B) ⇒ Probability of "A" is TRUE given that "B" is TRUE (Posterior Probability of B).

Q:- What is the probability that person has disease dengue with neck pain?

Given:-
→ 80% of time dengue causes neck pain.

$$P(a/b) = 0.8$$

$$P(\text{dengue}) = \frac{1}{30000} \Rightarrow P(b) = \frac{1}{30000}$$

$$P(\text{neckpain}) = 0.02 \Rightarrow P(a) = 0.02$$

a → person has neck pain

b → person has dengue

$$P(b/a) = \frac{P(a/b) \cdot P(b)}{P(a)} = \frac{0.8 \times (1/30000)}{0.02}$$

$$= 0.00133$$

Application of Bayes Theorem:-

1) Robot / Automation m/c } - next step is calculated based on prior step.

2) Fore casting
   ↳ weather forecasting

3) Medical diagnosis

4) Spam filtering

5) Credit card fraud detection -

Q) Linear Regression

House Price Prediction

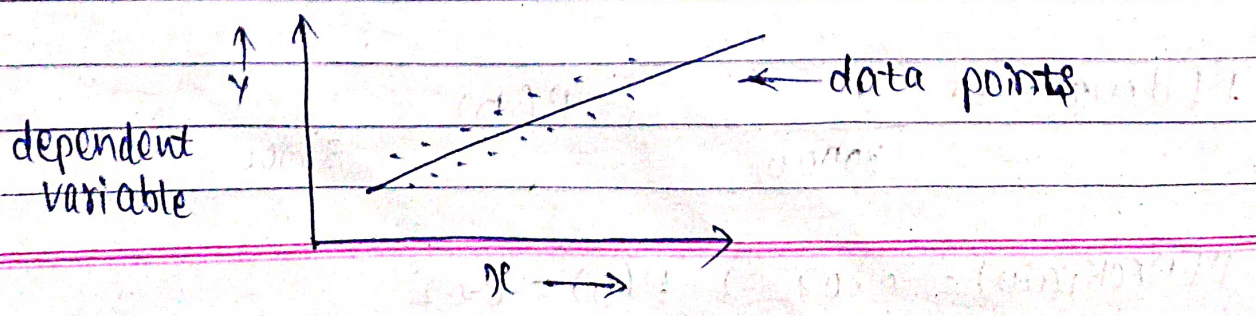
→

To predict the price of house.

we have dependent and independent variable.

Price of house is dependent variable and other factors like near highway, other facilities / factors are independent variable.

· Dependent variable is continuous in nature



when we have only one independent & one dependent variable then it is simple linear equation.

dependent variable → $y = \alpha_0 + \alpha_1 x_1$, α ↔ coefficients

x_1 ← Independent variable

dependent → $y = mx + c$

m ← slope, c ← intercept, x ← Independent

If we have more than one independent and only one dependent variable then it is ~~linear~~ Multiple linear Regression.

~~dependent variable~~ → ~~$y = \alpha_0 + \alpha_1 x_1$~~

~~Independent~~

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_m x_m$$

y → dependent variable
x_i → Independent variable
α_i → Regression coefficient

Regression Coefficient : Your feature or independent variable how much impact it puts on your dependent variable that what α or regression coefficient.

Entropy = Entrophy (old literature)

(1) Entropy and gain information

→ Ex:- Training data tuples from the All Electronics customer database.

PID	age	income	student	credit_rating	class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31...40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31...40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31...40	medium	no	excellent	yes
13	31...40	high	yes	fair	yes
14	>40	medium	no	excellent	no

For age = "<=30":

$S_{11} = 2$, $S_{21} = 3$ $I(S_{11}, S_{21}) = 0.971$

For age = "31...40":

$S_{12} = 4$ $S_{22} = 0$ $I(S_{12}, S_{22}) = 0$

For age = ">40":

$S_{13} = 3$ $S_{23} = 2$ $I(S_{13}, S_{23}) = 0.971$

The expected information needed to classify a given sample if the sample are partitioned according to age is

$$E(age) = \frac{5}{14} I(s_{11}, s_{21}) + \frac{4}{14} I(s_{12}, s_{22}) + \frac{5}{14} I(s_{13}, s_{23})$$

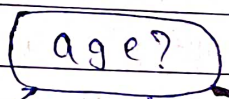
$$= 0.694$$

Hence, the gain in information from such a partitioning would be:

$$Gain(age) = I(s_1, s_2) - E(age) = 0.246$$

Similarly, we can compute Gain (income) = 0.029, Gain (student) = 0.151 & Gain (credit_rating) = 0.048

Since age has the highest information gain among the attributes, it is selected as the test attribute.



<= 30

income	student	credit_rating	class
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

31...40

income	student	credit_rating	class
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

>40

income	student	credit_rating	class
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

theAttribute 'age' has the highest information gain & therefore becomes a test attribute at root node of the decision tree. Branches are grown for each value of age.

In Summary, decision tree induction algorithms have been used for classification in a wide range of application domains such system do not use domain knowledge.
The learning & classification steps of decision tree induction are generally fast.

Q) Naive Bayesian Classification for all Electronics Customer database.

→ Attributes → age, income, student and credit_rating.
The class label attribute, buys_computer has two distinct values (namely, {yes, no}). Let C_1 correspond to the class buys_computer = "yes" and C_2 correspond to buys_computer = "no".

The unknown sample we wish to classify as

X = (age = "<=30", income = "medium", student = "yes", credit_rating = "fair").

We need to maximize $P(X/C_i)\,P(C_i)$, for $i = 1, 2$.

$P(C_i)$, the prior probability of each class.

$P(\text{buys\_computer} = \text{"yes"}) = \frac{9}{14} = 0.643$

$P(\text{buys\_computer} = \text{"no"}) = \frac{5}{14} = 0.357$

Other Conditional Probabilities:

P(age = "<30" / buys_computer = "yes") = 2/9 = 0.222
P(age = "<30" / buys_computer = "no") = 3/5 = 0.600
P(income = "medium" / buys_computer = "yes") = 4/9 = 0.444
P(income = "medium" / buys_computer = "no") = 2/5 = 0.400
P(student = "yes" / buys_computer = "yes") = 6/9 = 0.667
P(student = "yes" / buys_computer = "no") = 1/5 = 0.200

P(credit_rating = "fair" / buys_computer = "yes") = 6/9 = 0.667
P(credit_rating = "fair" / buys_computer = "no") = 2/5 = 0.400

Using the above probabilities we obtain

P(X / buys_computer = "yes") = 0.222 * 0.444 * 0.667 * 0.667 = 0.044
P(X / buys_computer = "no") = 0.600 * 0.400 * 0.200 * 0.400 = 0.019

P(X / buys_computer = "yes") P(buys_computer = "yes") = 0.044 * 0.643 = 0.028

P(X / buys_computer = "no") P(buys_computer = "no") = 0.019 * 0.357 = 0.007

Therefore the naive Bayesian classifies predicts buys_computer = "yes" for sample X.