प्र What is decision tree i~~nduction~~

Ans:

A decision tree is a structure that includes a root node, branches & leaf nodes. It's a top down recursive divide & conquer algorithm.

Decision Tree रुप

→ Each internal node denotes a test on an attribute.

→ Each branch denotes the outcome of a test.

→ Each leaf node holds a class label.

→ The topmost node in the tree is the root node.

प्र Define classification & it's step

Ans: Classification is a data mining function. It describes & distinguishes data classes.

It's a two step Process.

Step-1: Learning step → classification model, classifier is constructed. Using training dataset.

Step-2: Classification step → Classification model is used to Predict class lables for given data.

Decision Tree Algorithm এর ব্যবহৃত Common Symbol

Symbol	Term
D	Training Data
x_i	A data instance
X	A subset of instances
A_j	A feature
a_j^i	A feature's value
C_i	A class label
DT	A decision tree

- $D = \{x_1, x_2, \dots, x_n\}$, D হলো traing data.
- x_i কে vector দিয়ে represent করা হয়, $x_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$
- $D \Rightarrow \{A_1, A_2, A_3, \dots A_n\}$; dataset ; D তে — A_n সংখ্যক attribute থাকতে পারে,
- প্রতিটি attribute এর different ; attribute value থাকে, $\{A_{i1}, A_{i2}, A_{ik}\}$
- Dataset এ class এর একটি Column থাকবে, $C = \{C_1, C_2, C_m\}$
- প্রতিটি x_i instance একটি Predefined class label c_i এর সাথে থাকে,

- Decision Tree এর প্রতিটি leaf decision হয়,
- Decision Tree এর little Prior knowledge দরকার হয়,
- Decision Tree মূলত Class Predict করে

~~প্র Decision~~ প্র Write the advantages of decision tree.

Ans:

~~The decisions of~~

The advantages are

a. simple to understand

b. Easy to implement

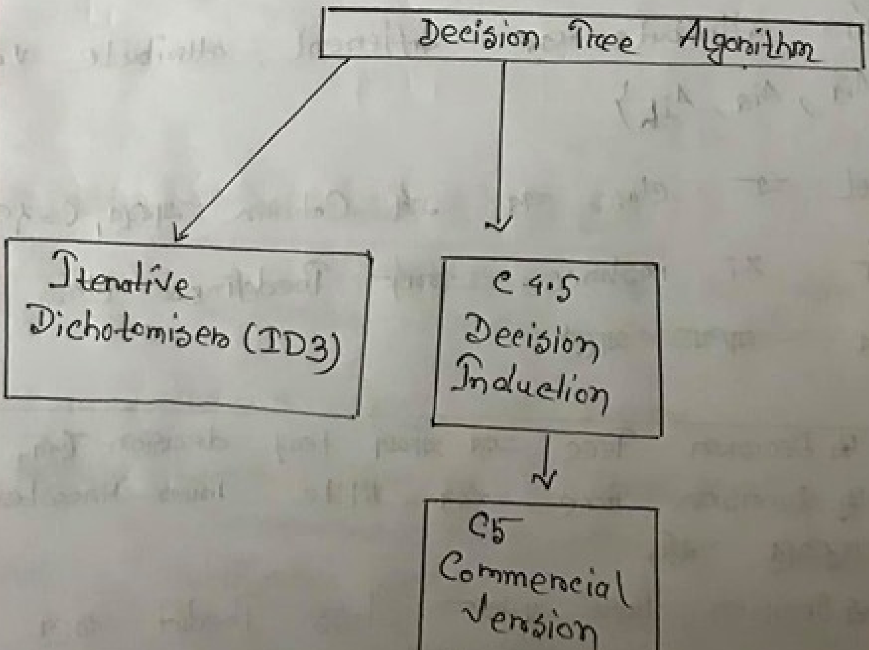
c. Requiring a little Prior knowledge.

d. Can handle numerical & categorical data.

e. ~~R~~ Robust.

f. Dealing with large & noisy datasets.

প্র Decision Tree এর কয়েকটি version আছে।



প্র Write the Working Process of Iterative Dichotomiser3 (ID3)

Ans:

ID3 iteratively Partition the data into smaller subsets until all the subsets belong to a single class.

ID3 এর সমীকরণ

a. $\text{Info}(D) = -\sum_{i=1}^{n} P_i \cdot \log_2 P_i$ // Prior Probability Calculate করছে.

→ i class value / attribute value represent করে

→ এবং D হচ্ছে Total data / Training data

b. $\text{Info}_A(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} \times \text{Info}(D)$

c. $\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$

⇒ Info(D) এর অন্য আরেকটা version হল Entropy. যেটার formula ; $\text{Entropy} = -\sum P(X) \log P(X)$.

প্র Define Entropy

Ans.

It is the measure of impurity, disorder, uncertainty in a bunch of examples

প্র What is information gain.

Ans:

It measures how much information a feature gives us about a class.

Why Information Gain matters.

Ans:

It is the main key that is used by Decision Tree algorithms to construct a decision tree. It always tries to maximize information gain. An attribute with highest information gain splited first.

- Info gain dataset কে more appropriately ভাগ করে,
- ID3 চলতে থাকবে যতক্ষণ পর্যন্ত না সবগুলো instance একটা Particular class এ belong করবে,

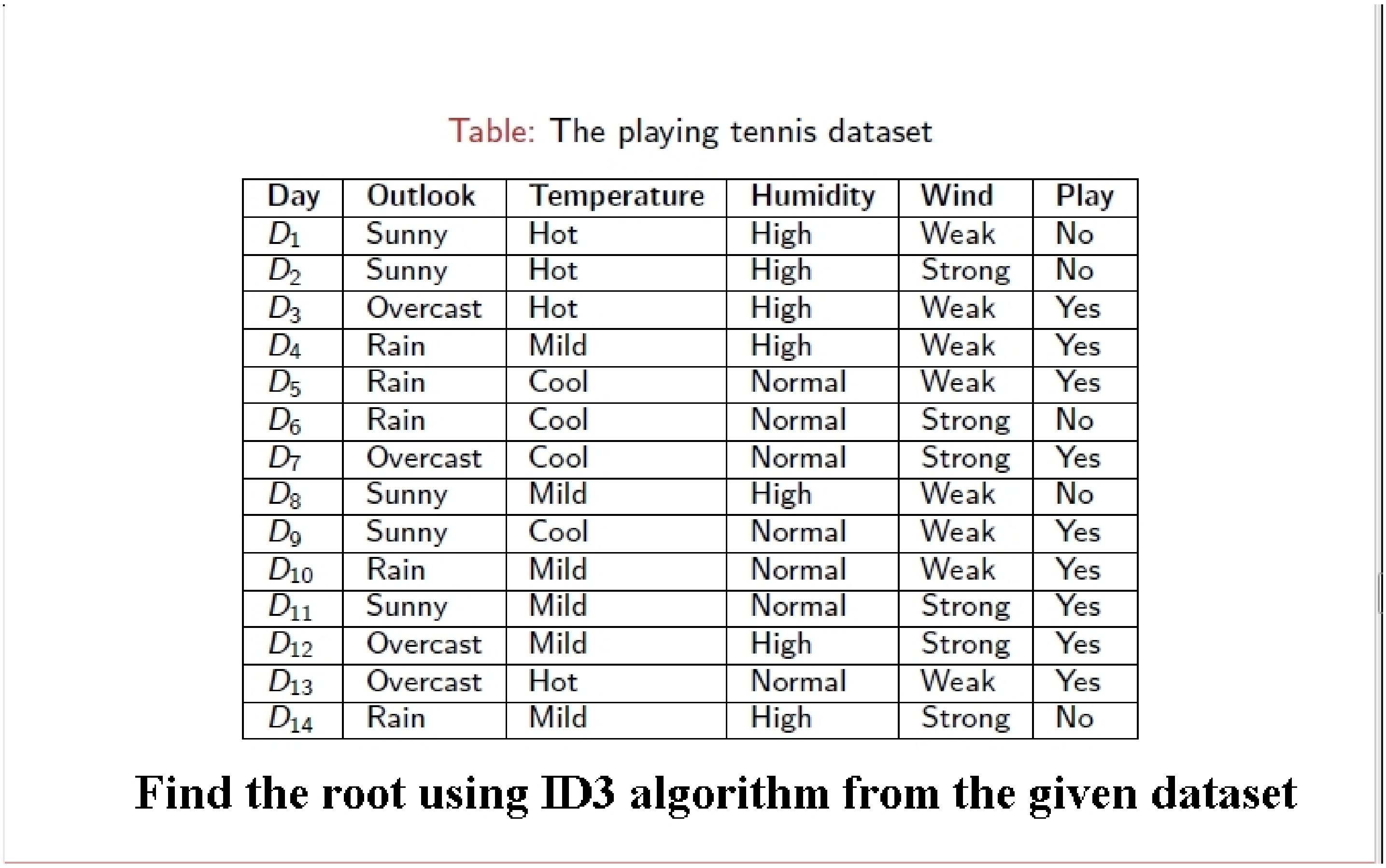


Table: The playing tennis dataset

Day	Outlook	Temperature	Humidity	Wind	Play
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	Weak	Yes
D_4	Rain	Mild	High	Weak	Yes
D_5	Rain	Cool	Normal	Weak	Yes
D_6	Rain	Cool	Normal	Strong	No
D_7	Overcast	Cool	Normal	Strong	Yes
D_8	Sunny	Mild	High	Weak	No
D_9	Sunny	Cool	Normal	Weak	Yes
D_{10}	Rain	Mild	Normal	Weak	Yes
D_{11}	Sunny	Mild	Normal	Strong	Yes
D_{12}	Overcast	Mild	High	Strong	Yes
D_{13}	Overcast	Hot	Normal	Weak	Yes
D_{14}	Rain	Mild	High	Strong	No

Find the root using ID3 algorithm from the given dataset

Page→52

~~$Info(D) = -\frac{9}{14}\log_2(9/14) - 5/14 \log_2(5/14)$~~

~~$= -0.64 \times (-0.191) - 0.35 \times (-0.45)$~~

~~$= 0.122 + 0.15$~~

$$Info(D) = -\frac{9}{14}\log_2(9/14) - \frac{5}{14}\log_2(5/14)$$

$$= -0.64 \times (-0.64) - 0.35 \times (-1.51)$$

$$= 0.40 + 0.52$$

$$= 0.92$$

Attribute ⇒ Outlook

$$Info_{Outlook}(D) = \underbrace{\frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)}_{\text{Sunny}} +$$

$$\underbrace{\frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right)}_{\text{Overcast}} +$$

$$\underbrace{\frac{5}{14} \times \left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}\right)}_{\text{Rain}}$$

$$= 0.35 \times \{-0.4 \times (-1.32) - 0.6 \times (-0.73)\} +$$

$$0.28 \times \{-1 \times (0) - 0 \times \log_2(0/4)\} +$$

$$0.35 \times \{-0.6 \times (-0.73) - 0.4 \times (-1.32)\}$$

$$= 0.35 \times 0.966 + 0.28 \times 0 + 0.35 \times 0.966$$

$$= 0.3381 + 0.3381 = 0.6762$$

$$\text{Gain(Outlook)} = \text{Info}(D) - \text{Info}_{\text{Outlook}}(D)$$

$$= 0.92 - 0.6762$$

$$= 0.2438$$

Attribute → Temparature

$$\text{Info}_{\text{Temparature}}(D) = \underbrace{\frac{4}{14} \times \left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right)}_{\text{Hot}} +$$

$$\underbrace{\frac{6}{14} \times \left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right)}_{\text{Mild}} +$$

$$\underbrace{\frac{4}{14} \times \left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right)}_{\text{Cool}} +$$

$$= 0.28 \times \{-0.5 \times (-1) - 0.5(-1)\} + 0.42 \times \{-0.66 \times (-0.599)$$
$$- 0.33 \times (-1.59)\} + 0.28 \times \{-0.75 \times (-0.41) - 0.25 \times (-2$$

~~= 0.28 (0.5~~

$$= 0.28 \times 1 + 0.42 \times 0.92004 + 0.28 \times 0.8075$$

$$= 0.28 + 0.386568 + 0.2261$$

$$= 0.892$$

$$\text{Gain (Temparature)} = \text{Info}(D) - \text{Info}_{\text{Temparature}}(D) = 0.92 - 0.892$$
$$= 0.028$$

Attribute → Humidity

$$\text{Info}_{\text{Humidity}}(D) = \underbrace{\frac{7}{14} \times \left(-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}\right)}_{\text{High}} +$$

$$\underbrace{\frac{7}{14} \times \left(-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right)}_{\text{Normal}}$$

$$= 0.5 \times \{-0.42 \times (-1.25) - 0.57 \times (-0.81)\} +$$

$$0.5 \times \{-0.85 \times (-0.23) - 0.14 \times (-2.83)\}$$

$$= 0.5 \times 0.98 + 0.5 \times 0.5917$$

$$= 0.49 + 0.29585$$

$$= 0.78$$

$$\text{Gain}(\text{Humidity}) = \text{Info}(D) - \text{Info}_{\text{Humidity}}(D)$$

$$= 0.92 - 0.785$$

$$= 0.135$$

Attribute ⇒ Wind

$$\text{Info}_{\text{Wind}}^{(D)} = \underbrace{\frac{8}{14}\left(-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right)}_{\text{Weak}} +$$

$$\underbrace{\frac{6}{14}\left(-\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\text{Log}_2\frac{3}{6}\right)}_{\text{Strong}}$$

$$= 0.57 \times \{-0.75 \times (-0.41) - 0.25 \times (-2)\} +$$

$$0.42 \times \{-0.5 \times (-1) - 0.5 \times (-1)\}$$

$$= 0.57 \times 0.8075 + 0.42 \times 1$$

$$= 0.460 + 0.42$$

$$= 0.88$$

$$\text{Gain}(\text{Wind}) = \text{Info}(D) - \text{Info}_{\text{wind}}(D)$$

$$= 0.92 - 0.88$$

$$= 0.04$$

So, Gain of

Outlook	= 0.2438
Temparature	= 0.028
Humidity	= 0.135
Wind	= 0.04

Page->56

->Outlook এর gain বেশি, তাই এটা হবে decision tree এর node.

Ay—