SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601 – Data Science and Analytics

III Year / VI Semester

Unit 3 – PREDICTIVE MODELING AND MACHINE LEARNING

Topic 5: Decision Tree

Decision Tree – General Algorithm

- In general, the objective of a decision tree algorithm is to construct a tree T from a training set S.

- If all the records in S belong to some class C, or if S is sufficiently pure (greater than a preset threshold), then that node is considered a leaf node and assigned the label C.

- In contrast, if not all the records in S belong to class C or if S is not sufficiently pure, the algorithm selects the next most informative attribute A and

- partitions S according to A‘s values. The algorithm constructs subtrees , T1,T2… for the subsets of S recursively until one of the following criteria is met:
  - All the leaf nodes in the tree satisfy the minimum purity threshold.
  - The tree cannot be further split with the preset minimum purity threshold.
  - Any other stopping criterion is satisfied (such as the maximum depth of the tree).

Decision Tree – General Algorithm

- The first step in constructing a decision tree is to choose the most informative attribute.
- A common way to identify the most informative attribute is to use entropy-based methods.
- The entropy methods select the most informative attribute based on two basic measures:
  - Entropy, which measures the impurity of an attribute
  - Information gain, which measures the purity of an attribute

Decision Tree – General Algorithm

- Entropy is a measure of the randomness in the information being processed.

- The higher the entropy, the harder it is to draw any conclusions from that information.

Example Flipping a coin

- The entropy for a completely pure variable is 0 and is 1 for a set with equal occurrences for both the classes (head and tail, or yes and no).

- Given a class and its label x=X, let P(x) be the probability of x. Hx the entropy of , is defined as

$$H_X = -\sum_{\forall x \in X} P(x) \log_2 P(x)$$

Decision Tree – General Algorithm

- As an example of a binary random variable, consider tossing a coin with known, not necessarily fair, probabilities of coming up heads or tails.
- Let x=1 represent heads and x=0 represent tails. The entropy of the unknown result of the next toss is maximized when the coin is fair.
- That is, when heads and tails have equal probability $P(x=1)=P(x=0)=0.5$

  Entropy $H_X=-(0.5\times\log_2 0.5+0.5\times\log_2 0.5)=1$.
- On the other hand, if the coin is not fair, the probabilities of heads and tails would not be equal and there would be less uncertainty.
- As an extreme case, when the probability of tossing a head is equal to 0 or 1, the entropy is minimized to 0.

Decision Tree – General Algorithm

Conditional Entropy

The next step is to identify the conditional entropy for each attribute.

Given an attribute X, its value x , its outcome Y, and its value y, conditional entropy is the remaining entropy of given , formally defined as

$$H_{Y|X} = \sum_{x} P(x) H(Y|X = x)$$

$$= -\sum_{\forall x \in X} P(x) \sum_{\forall y \in Y} P(y|x) \log_2 P(y|x)$$

The information gain of an attribute A is defined as the difference between the base entropy and the conditional entropy of the attribute $InfoGain_A = H_S - H_{S|A}$

Information gain compares the degree of purity of the parent node before a split with the degree of purity of the child node after a split.

At each split, an attribute with the greatest information gain is considered the most informative attribute. Information gain indicates the purity of an attribute.

Decision Tree – General Algorithm

Predictors				Target
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Decision Tree – General Algorithm

Step 2: Calculating Entropy for the classes (Play Golf)

In this step, you need to calculate the entropy for the Play Golf column and the calculation step is given below.

$$Entropy(PlayGolf) = \boldsymbol{E}(5,9)$$

$$\boldsymbol{E}(\boldsymbol{PlayGolf}) = E(5,9)$$

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

$$= -(0.357\ \log_2 0.357) - (0.643\ \log_2 0.643)$$

$$= \mathbf{0.94}$$

Decision Tree – General Algorithm

Step 2: Calculating Entropy for the classes (Play Golf)

In this step, you need to calculate the entropy for the Play Golf column and the calculation step is given below.

$$Entropy(PlayGolf) = \boldsymbol{E}(5,9)$$

$$\boldsymbol{E}(\boldsymbol{PlayGolf}) = E(5,9)$$

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

$$= -(0.357\ \log_2 0.357) - (0.643\ \log_2 0.643)$$

$$= \mathbf{0.94}$$

Decision Tree – General Algorithm

Step 3: Calculate Entropy for Other Attributes After Split

For the other four attributes, we need to calculate the entropy after each of the split.

- E(PlayGolf, Outloook)
- E(PlayGolf, Temperature)
- E(PlayGolf, Humidity)
- E(PlayGolf,Windy)

The entropy for two variables is calculated using the formula.

(don't worry if about this formula, its really easy doing the calculation 😇

$$Entropy(S,T) = \sum_{c \in T} P(c)E(c)$$

There to calculate **E**(PlayGolf, Outlook), we would use the formula below:

$$\boldsymbol{E}(PlayGolf, Outlook) = \boldsymbol{P}(Sunny)\boldsymbol{E}(Sunny) + \boldsymbol{P}(Overcast)\boldsymbol{E}(Overcast) + \boldsymbol{P}(Rainy)\boldsymbol{E}(Rainy)$$

Decision Tree – General Algorithm

Which is
the same as:

$$E(PlayGolf, Outlook) = \mathbf{P}(Sunny)\,\mathbf{E}(3,2) + P(Overcast)\,\mathbf{E}(4,0) + \mathbf{P}(rainy)\,E(2,30$$

This formula may look unfriendly, but it is quite clear. The easiest way to approach this calculation is to create a frequency table for the two variables, that is PlayGolf and Outlook.

This frequency table is given below:

		PlayGolf(14)		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5

Table 3: Frequency Table for Outlook

Using this table, we can then calculate E(PlayGolf, Outlook), which would then be given by the formula below

$$E(PlayGolf, Outlook) = \frac{5}{14}E(3,2) + \frac{4}{14}E(4,0) + \frac{5}{14}E(2,3)$$

Decision Tree – General Algorithm

Let's go ahead to calculate E(3,2)

We would not need to calculate the second and the third terms! This is because

E(4, 0) = 0

E(2,3) = E(3,2)

Isn't this interesting!!!

$$E(Sunny) = E(3,2)$$

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

$$= -(0.60\ \log_2 0.60) - (0.40\log_2 0.40)$$

$$= -(0.60 * 0.737) - (0.40 * 0.529)$$

$$= \mathbf{0.971}$$

Just for clarification, let's show the the calculation steps

The calculation steps for E(4,0):

$$E(Overcast) = E(4,0)$$

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

$$= -(0) - (0)$$

$$= \mathbf{0}$$

Decision Tree – General Algorithm

The calculation step for E(2,3) is given below

$$E(Rainy) = E(2,3)$$

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

$$= -(0.40\ \log_2 0.40) - (0.6\log_2 0.60)$$

$$= 0.971$$

Time to put it all together.

We go ahead to calculate the **E**(PlayGolf, Outlook) by substituting the values we calculated from **E**(Sunny), **E**(Overcast) and **E**(Rainy) in the equation:

$$E(PlayGolf, Outlook) = P(Sunny)\ E(3,2) + P(Overcast)\ E(4,0) + P(rainy)\ E(2,3)$$

$$E(PlayGolf, Outlook) = \frac{5}{14}E(3,2) + \frac{4}{14}E(4,0) + \frac{5}{14}E(2,3)$$

$$= \frac{5}{14}0.971 + \frac{4}{14}0.0 + \frac{5}{14}0.971$$

$$= 0.357 * 0.971 + 0.0 + 0.357 * 0.971$$

$$= 0.693$$

Decision Tree – General Algorithm

E(PlayGolf, Temperature) Calculation

Just like in the previous calculation, the calculation of E(PlayGolf, Temperature) is given below. It

It is easier to do if you form the frequency table for the split for Temperature as shown.

		PlayGolf(14)		
		Yes	No	
Temperature	Hot	2	2	4
	Cold	3	1	4
	Mild	4	2	6

Table 4: Frequency Table for Temperature

$$E(PlayGolf, Temperature) = P(Hot)\ E(2,2) + P(Cold)\ E(3,1) + P(Mild)\ E(4,2)$$

Decision Tree – General Algorithm

$$E(\text{PlayGolf, Temperature}) = \mathbf{4/14 * E(Hot)} + \mathbf{4/14 * E(Cold)} + \mathbf{6/14 * E(Mild)}$$

$$E(\text{PlayGolf, Temperature}) = \mathbf{4/14 * E(2, 2)} + \mathbf{4/14 * E(3, 1)} + \mathbf{6/14 * E(4, 2)}$$

$$E(\text{PlayGolf, Temperature}) = 4/14 * -(2/4 \log 2/4) - (2/4 \log 2/4)$$
$$+ 4/14 * -(3/4 \log 3/4) - (1/4 \log 1/4)$$
$$+ 6/14 * -(4/6 \log 4/6) - (2/6 \log 2/6)$$

$$E(\text{PlayGolf, Temperature}) = 5/14 * 1.0$$
$$+ 4/14 * 1.811$$
$$+ 5/14 * 0.918$$
$$= \mathbf{0.911}$$

Decision Tree – General Algorithm

E(PlayGolf, Humidity) Calculation

Just like in the previous calculation, the calculation of E(PlayGolf, Humidity) is given below. It

It is easier to do if you form the frequency table for the split for Humidity as shown.

		PlayGolf(14)		
		Yes	No	
Humidity	High	3	4	7
	Normal	6	1	7

Table 5: Frequency Table for Humidity

Decision Tree – General Algorithm

$$E(\text{PlayGolf, Humidity}) = \mathbf{7/14} * \mathbf{E}(\text{High}) + \mathbf{7/14} * \mathbf{E}(\text{Normal})$$

$$E(\text{PlayGolf, Humidity}) = \mathbf{7/14} * \mathbf{E}(3, 2) + \mathbf{7/14} * \mathbf{E}(4, 0)$$

$$E(\text{PlayGolf, Humidity}) = 7/14 * -(3/7 \log 3/7) - (4/7 \log 4/7)$$
$$+ 7/14 * -(6/7 \log 6/7) - (1/7 \log 1/7)$$

$$E(\text{PlayGolf, Humidity}) = 7/14 * 0.985$$
$$+ 7/14 * 0.592$$
$$= \mathbf{0.788}$$

Decision Tree – General Algorithm

E(PlayGolf, Windy) Calculation

Just like in the previous calculation, the calculation of E(PlayGolf, Windy) is given below. It

It is easier to do if you form the frequency table for the split for Windy as shown.

		PlayGolf(14)		
		Yes	No	
Windy	TRUE	3	3	6
	FALSE	6	2	8

Table 6: Frequency Table for Windy

Decision Tree – General Algorithm

$$E(\text{PlayGolf, Windy}) = \mathbf{6/14} * \mathbf{E}(\text{True}) + \mathbf{8/14} * \mathbf{E}(\text{False})$$

$$E(\text{PlayGolf, Windy}) = \mathbf{6/14} * \mathbf{E}(3, 3) + \mathbf{8/14} * \mathbf{E}(6, 2)$$

$$E(\text{PlayGolf, Windy}) = 6/14 * -(3/6 \log 3/6) - (3/6 \log 3/6)$$
$$+ 8/14 * -(6/8 \log 6/8) - (2/8 \log 2/8)$$

$$E(\text{PlayGolf, Windy}) = 6/14 * 1.0$$
$$+ 8/14 * 0.811$$
$$= \mathbf{0.892}$$

Decision Tree – General Algorithm

So now that we have all the entropies for all the four attributes, let's go ahead to summarize them as shown in below:

1. E(PlayGolf, Outloook) = **0.693**
2. E(PlayGolf, Temperature) = **0.911**
3. E(PlayGolf, Humidity) = **0.788**
4. E(PlayGolf,Windy) = **0.892**

Decision Tree – General Algorithm

Step 4: Calculating Information Gain for Each Split

The next step is to calculate the information gain for each of the attributes. The information gain is calculated from the split using each of the attributes. Then the attribute with the largest information gain is used for the split.

The information gain is calculated using the formula:

$$Gain(S,T) = Entropy(S) - Entropy(S,T)$$

For example, the information gain after spliting using the Outlook attibute is given by:

$$Gain(PlayGolf, Outlook) = Entropy(PlayGolf) - Entropy(PlayGolf, Outlook)$$

Decision Tree – General Algorithm

So let's go ahead to do the calculation

Gain(PlayGolf, Outlook) = Entropy(PlayGolf) − Entropy(PlayGolf, Outlook)

= 0.94 − 0.693 **= 0.247**

Gain(PlayGolf, Temperature) = Entropy(PlayGolf) − Entropy(PlayGolf, Temparature)

= 0.94 − 0.911 **= 0.029**

Gain(PlayGolf, Humidity) = Entropy(PlayGolf) − Entropy(PlayGolf, Humidity)

= 0.94 − 0.788 **= 0.152**

Gain(PlayGolf, Windy) = Entropy(PlayGolf) − Entropy(PlayGolf, Windy)

= 0.94 − 0.892 **= 0.048**

Having calculated all the information gain, we now choose the attribute that gives the highest information gain after the split.

Decision Tree – General Algorithm

Step 5: Perform the First Split

Draw the First Split of the Decision Tree

Now that we have all the information gain, we then split the tree based on the attribute with the highest information gain.

From our calculation, the highest information gain comes from Outlook. Therefore the split will look like this:

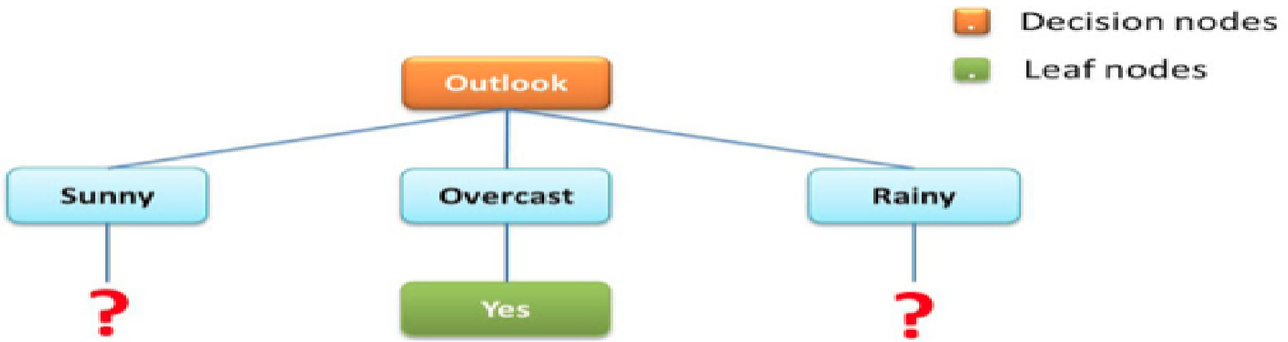

Figure 2: Decision Tree after first split

Decision Tree – General Algorithm

Step 6: Perform Further Splits

The Sunny and the Rainy attributes needs to be split

The Rainy outlook can be split using either Temperature, Humidity or Windy.

Quiz 1: What attribute would best be used for this split? Why?

Answer: **Humidity**. Because it produces homogenous groups.

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No

Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Table 8: Split using Humidity

The Rainy attribute could be split using High and Normal attributes and that would give us the tree below.

Decision Tree – General Algorithm

Outlook	Temperature	Humidity	Windy	Play Golf
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Table 7: Initial Split using Outlook

Decision Tree – General Algorithm

Step 6: Perform Further Splits

The Sunny and the Rainy attributes needs to be split

The Rainy outlook can be split using either Temperature, Humidity or Windy.

Quiz 1: What attribute would best be used for this split? Why?

Answer: **Humidity**. Because it produces homogenous groups.

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Table 8: Split using Humidity

The Rainy attribute could be split using High and Normal attributes and that would give us the tree below.

Decision Tree – General Algorithm

Outlook	Temperature	Humidity	Windy	Play Golf
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	High	TRUE	No

Table 9: Split using Windy Attribute

If we do the split using the Windy attribute, we would have the final tree that would require no further splitting! This is shown in Figure 4

Decision Tree – General Algorithm

Step 7: Complete the Decision Tree

The complete table is shown in Figure 4

Note that the same calculation that was used initially could also be used for the further splits. But that would not be necessary since you could just look at the sub table and be able to determine which attribute to use for the split.

Quiz: What does each of he color represent in the tree?

Leave your answer in the comment box below

THANK YOU