Decision Tree

CS7DS1
Bahman Honari

Outline

- Introduction
- Example
- Principles
  - Entropy
  - Joint and Conditional Entropy
  - Information Gain
- Example and Demonstration
- DIY

What is a Decision Tree?

- Decision trees are powerful and popular tools for classification and prediction.
- Decision trees represent *rules*, which can be understood by humans and used in knowledge system such as database.

Requirements

- **Predictor Attribute:** Objects or cases must be presented in terms of different attributes (Temp, Humidity, etc.). Each attribute measures some important feature of an object and will be limited here to taking a (usually small) set of discrete, mutually exclusive values.
- **Predefined Target:** The target variable is a discrete variable (binary or multiclass). Each object in the target variable has one of a set of mutually exclusive classes.
- **Sufficient data:** Enough training cases, whose their target class is known, should be provided to build the model.

Example: J. R. Quinlan - 1986

A small data set with the 'Saturday morning' attributes.
For each observation, the value of each attribute is shown, together with the class of the target variable.

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

Analysis Methodology

- The concepts of **Entropy** and **Information Gain** are employed to construct a decision tree.

Information Function

- To quantify the information conveyed by each attribute, a function is defined as $I(s)$, which represents how much information is gained by knowing the predictor attribute, such that:

(1) $I(.)$ is a decreasing function of the probability $\mathrm{p_i}$, with $I(.) = 0$ if $p_i = 1$;

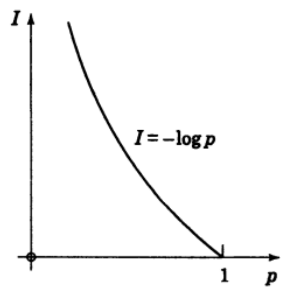
(2) $I(s_i s_j) = I(s_i) + I(s_j)$

Information Function – Cont.

- Condition (1) asserts that the greater the probability of an event, the less information it conveys, and an inevitable event conveys no information
- Condition (2) asserts that since we have independent (predictor) variables, the amount of information gained by knowing about the outcome of two variables is the sum of the two individual amounts of information.
- These properties suggest the information function as below:

$$I(s_i) = log\frac{1}{p_i} = -\,log\,p_i$$

Information Function – Cont.

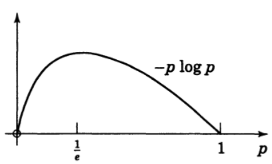


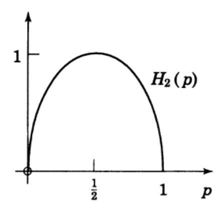
Entropy

$$H(X) = \sum_{x} P(x)\,.\,log\,\frac{1}{P(x)}$$

$$= -\sum_{x} P(x)\,.\,log\,P(x) = -\,E[log\,P(X)]$$

Entropy – Cont.





Entropy – Cont.

- Example 1. Find the entropy for an unbiased die.

How to Build a Decision Tree

- To build a decision tree, we need to calculate two types of entropy:

a) Entropy using the frequency table of one attribute: H(X)

b) Entropy using the frequency table of two attributes: H(X,Y)

Entropy – Cont.

Entropy using the frequency table of one attribute: $H(Y)$

$$H(Y) = \sum_{y} P(Y)\,.\,log\,\frac{1}{P(y)}$$
$$= -\sum_{y} P(y)\,.\,log\,P(y) = -\,E[log\,P(Y)]$$

Entropy – Cont.

Entropy using the frequency table of two attributes: $H(X,Y)$

$$H(X,Y) = -\sum_{y}\sum_{x} P(x,y)\,.\,log\,P(x,y) = -\,E[log\,P(X,Y)]$$

Entropy – Cont.

But, let's first have a look to the concept of the conditional Entropy $H(Y|x)$ and $H(Y|X)$.

$$H(Y|\,x) = \;H(Y|X=x) = -\sum_{y} P(y|x)\,.\,logP(y|x)$$

$$H(Y|X) = \sum_{x} P(x)\,.\,H(Y|X=x) = -\sum_{x} P(x)\,.\sum_{y} P(y|x)\,.\,logP(y|x)$$
$$= -\sum_{y}\sum_{x} P(x)P(y|x)\,.\,logP(y|x) = -\sum_{x}\sum_{y} P(x,y).\,logP(y|x)$$

Entropy – Cont.

Now, let’s g back again to:

$$H(X,Y) = -\sum_{y}\sum_{x} P(x,y)\,.\,log\,P(x,y)$$

$$= -\sum_{y}\sum_{x} P(x,y)\,.\,log\big(P(x).\,P(y|x)\big)$$

$$= -\sum_{y}\sum_{x} P(x,y)\,.\,log\,P(x) - \sum_{y}\sum_{x} P(x,y)\,.\,log\,P(y|x)$$

$$= -\sum_{x} P(x)\,.\,log\,P(x) + H(Y|X) = H(X) + H(Y|X)$$

$$\boldsymbol{H(X,Y) = H(X) + H(Y|X)}$$

Mutual Information of X and Y

The mutual information between random variables X and Y with joint probability mass function $P(X,Y)$ and marginal probability mass functions $P(x)$ and $P(y)$ is defined as:

$$I(X,Y) = \sum_{y}\sum_{x} P(x,y) log \frac{P(x,y)}{P(x).\,P(y)}$$

Mutual Information of X and Y – Cont.

$$I(x,y) = \sum_{x}\sum_{y} P(x,y)\,log \frac{P(x,y)}{P(x).\,P(y)} = \sum_{x}\sum_{y} P(x,y) log \frac{P(y|x)}{P(y)}$$

$$= -\sum_{x}\sum_{y} P(x,y) log P(y) + \sum_{x}\sum_{y} P(x,y) log P(y|x)$$

$$= -\sum_{y} P(y) \log(P(y) + \sum_{x}\sum_{y} P(x,y) \log P(y|x)$$

$$I(X,Y) = \boldsymbol{H(Y) - H(Y|X)} \geq 0$$

Entropy – Cont.

The difference between $H(Y)$ and $H(Y|X)$ explains the reduction in the level of uncertainty by adding the information of X as a predictor of Y.

$$\boldsymbol{Information\ Gain(X) = H(Y) - H(Y|X)}$$

Decision Tree – Cont.

Using the dataset:

Play Golf	
Yes	9
No	5

$$H(Y) = -\left(\frac{9}{14}\log\left(\frac{9}{14}\right) + \frac{5}{14}\log\left(\frac{5}{14}\right)\right) = 0.940$$

Decision Tree – Cont.

Using the dataset:

		Play Golf: Yes	Play Golf: No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
		9	5	14

$$H(Y|X_1) = -\frac{5}{14}\left(\frac{3}{5}\ log\frac{3}{5} + \frac{2}{5}\ log\frac{2}{5}\right) - \frac{4}{14}(0\ log0 + 1\ log1) - \frac{5}{14}\left(\frac{2}{5}\ log\frac{2}{5} + \frac{3}{5}\ log\frac{3}{5}\right)$$

$$= 0.693$$

Decision Tree – Cont.

Also:

		Play Golf: Yes	Play Golf: No	
Temp	Hot	2	2	4
	Mild	4	2	6
	Cool	3	1	4
		9	5	14

$$H(Y|X_2) = -\frac{4}{14}\left(2 \times \frac{2}{4}\ log\frac{2}{4}\right) - \frac{6}{14}\left(\frac{4}{6}\ log\frac{4}{6} + \frac{2}{6} log\frac{2}{6}\right) - \frac{4}{14}\left(\frac{3}{4}\ log\frac{3}{4} + \frac{1}{4}\ log\frac{1}{4}\right)$$

$$= 0.911$$

Decision Tree – Cont.

and:

		Play Golf: Yes	Play Golf: No	
Humidity	Normal	6	1	7
	High	3	4	7
		9	5	14

$$H(Y|X_3) = -\frac{7}{14}\left(\frac{6}{7}\ log\frac{6}{7} + \frac{1}{7}\ log\frac{1}{7}\right) - \frac{7}{14}\left(\frac{3}{4}\ log\frac{3}{4} + \frac{1}{4} log\frac{1}{4}\right) = 0.788$$

Decision Tree – Cont.

finally:

		Play Golf Yes	Play Golf No	
Windy	TRUE	6	2	8
	FALSE	3	3	6
		9	5	14

$$H(Y|X_4) = -\frac{8}{14}\left(\frac{6}{8}\,log\frac{6}{8} + \frac{2}{8}\,log\frac{2}{8}\right) - \frac{6}{14}\left(2 \times \frac{3}{6}\,log\frac{3}{6}\right) = 0.892$$

Decision Tree – Cont.

Summary of results:

Play Golf (Y)	–	Outlook (X1)	Temp (X2)	Humidity (X3)	Windy (X4)
Joint H	0.94	0.693	0.911	0.788	0.892
Gain	0	0.247	0.029	0.152	0.048

Decision Tree – Cont.

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Overcast	Hot	High	False	Yes
Overcast	Cool	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

Decision Tree – Cont.

Predictors				Target
Outlook	Temp	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

Decision Tree – Cont.

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

Decision Tree – Cont.