COMP 472: Artificial Intelligence

Decision Trees

Solutions

Question 1 Given the training instances below, use information theory to find whether 'Outlook' or 'Windy' is the best feature to decide when to play a game of golf.

Outlook	Temperature	Humidity	Windy	Play / Don't Play
sunny	85	85	false	Don't Play
sunny	80	90	true	Don't Play
overcast	83	78	false	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
overcast	64	65	true	Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
rain	75	80	false	Play
sunny	75	70	true	Play
overcast	72	90	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play

$$H(Output) = H\left(\frac{5}{14}, \frac{9}{14}\right) = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right) = 0.94$$

$$H(Output|sunny) = H\left(\frac{3}{5}, \frac{2}{5}\right) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.97$$

$$H(Output|overcast) = H(0, 1) = -(0\log_2 0 + 1\log_2 1) = 0$$

$$H(Output|rain) = H\left(\frac{2}{5}, \frac{3}{5}\right) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.97$$

$$H(Output|Outlook) = \frac{5}{14}H(Output|sunny) + \frac{4}{14}H(Output|overcast) + \frac{5}{14}H(Output|rain)$$

$$H(Output|Outlook) = \frac{5}{14}0.97 + \frac{4}{14}0 + \frac{5}{14}0.97 = 0.69$$

$$gain(Outlook) = H(Output) - H(Output|Outlook) = 0.94 - 0.69 = 0.25$$

$$H(Output|Windy = true) = H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$$

$$H(Output|Windy = false) = H\left(\frac{1}{4}, \frac{3}{4}\right) = 0.81$$

$$H(Output|Windy) = \frac{6}{14}1 + \frac{8}{14}0.81 = 0.89$$

$$gain(Windy) = H(Output) - H(Output|Windy) = 0.94 - 0.89 = 0.05$$

'Outlook' is a better feature because it has a bigger information gain.