COMP 472: Artificial Intelligence

Decision Trees

Solutions

1 Question

(a) Given the training instances below, use information gain to determine whether 'Outlook' or 'Windy' is the best feature to decide when to play a game of golf.

Outlook	**Temperature**	**Humidity**	**Windy**	**Golf**
sunny	hot	high	false	don't play
sunny	hot	high	true	don't play
overcast	hot	high	false	play
rain	mild	high	false	play
rain	cold	normal	false	play
rain	cold	normal	true	don't play
overcast	cold	normal	true	play
sunny	mild	high	false	don't play
sunny	cold	normal	false	play
rain	mild	normal	false	play
sunny	mild	normal	true	play
overcast	mild	high	true	play
overcast	hot	normal	false	play
rain	mild	high	true	don't play

Solution

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

$H(\text{Golf}|\text{Outlook=sunny}) = H\left(\frac{3}{5}, \frac{2}{5}\right) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.97$
$H(\text{Golf}|\text{Outlook=overcast}) = H\left(0, 1\right) = -\left(0\log_2 0 + 1\log_2 1\right) = 0$
$H(\text{Golf}|\text{Outlook=rain}) = H\left(\frac{2}{5}, \frac{3}{5}\right) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.97$

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

$\text{Gain(Golf, Outlook)} = H(\text{Golf}) - H(\text{Golf}|\text{Outlook}) = 0.94 - 0.69 = 0.25$

$H(\text{Golf}|\text{Windy=true}) = H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$
$H(\text{Golf}|\text{Windy=false}) = H\left(\frac{1}{4}, \frac{3}{4}\right) = 0.81$

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

$\text{Gain(Golf, Windy)} = H(\text{Golf}) - H(\text{Golf}|\text{Windy}) = 0.94 - 0.89 = 0.05$

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

(b) Assume that we build a decision tree with the feature 'Outlook' as root. What would be the best feature to use as root of the sub-tree for the branch 'Outlook=sunny'. Again, use information gain.

Solution

To simplify the notation, let $A = (\text{Golf}|\text{Outlook=sunny})$

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

$H(A|\text{Temperature=hot}) = H\left(\frac{2}{2}, \frac{0}{2}\right) = -\left(1\log_2 1 + 0\log_2 0\right) = 0$
$H(A|\text{Temperature=mild}) = H\left(\frac{1}{2}, \frac{1}{2}\right) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$
$H(A|\text{Temperature=cold}) = H\left(\frac{0}{1}, \frac{1}{1}\right) = -\left(0\log_2 0 + 1\log_2 1\right) = 0$
$H(A|\text{Temperature}) = \frac{2}{5}0 + \frac{2}{5}1 + \frac{1}{5}0 = 0.4$
$\text{Gain}(A, \text{Temperature}) = H(A) - H(A|\text{Temperature}) = 0.97 - 0.4 = 0.57$

$H(A|\text{Humidity=high}) = H\left(\frac{3}{3}, \frac{0}{3}\right) = 0$
$H(A|\text{Humidity=normal}) = H\left(\frac{0}{2}, \frac{2}{2}\right) = 0$
$H(A|\text{Humidity}) = \frac{3}{5}0 + \frac{2}{5}0 = 0$
$\text{Gain}(A, \text{Humidity}) = H(A) - H(A|\text{Humidity}) = 0.97 - 0 = 0.97$

$H(A|\text{Windy=true}) = H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$
$H(A|\text{Windy=false}) = H\left(\frac{2}{3}, \frac{1}{3}\right) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0.27$
$H(A|\text{Windy}) = \frac{2}{5}1 + \frac{3}{5}0.27 = 0.56$
$\text{Gain}(A, \text{Windy}) = H(A) - H(A|\text{Windy}) = 0.97 - 0.56 = 0.41$

'Humidity' is a better feature for this branch since it has a higher information gain.