National University of Singapore
School of Computing
CS3243: Foundations of Artificial Intelligence
Tutorial 10

Readings: AIMA Chapter 18

1. (Question 18.4 from AIMA) We never test the same attribute twice along one path in a decision tree. Why not?

2. Given the following sets of class values. For each set, calculate the Information Content (Entropy) given by: $I(\frac{p}{p+n}, \frac{n}{p+n}) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$. Explain your findings.

   (a) {p, p, p, p}

   (b) {p, p, p, n}

   (c) {p, p, n, n}

   (d) {p, n, n, n}

   (e) {n, n, n, n}

3. Given the following training data set about exotic dishes, we want to predict whether or not a dish is appealing given attributes 'Temperature', 'Taste', and 'Size'.

ID	Temperature	Taste	Size	Appealing
1	Hot	Salty	Small	No
2	Cold	Sweet	Large	No
3	Cold	Sweet	Large	No
4	Cold	Sour	Small	Yes
5	Hot	Sour	Small	Yes
6	Hot	Salty	Large	No
7	Hot	Sour	Large	Yes
8	Cold	Sweet	Small	Yes
9	Cold	Sweet	Small	Yes
10	Hot	Salty	Large	No

   (a) What is the information gain $IG(A)$ associated with choosing the attribute 'Taste' as the root of the decision tree?

   (b) Draw a decision tree with 'Taste' as the root.

   (c) Use the decision tree to predict the class value for the record given by

ID	Temperature	Taste	Size
11	Hot	Salty	Small
12	Cold	Sweet	Large

4. The loans department of a bank has the following past loan processing records each containing an applicant's income, credit history, debt, and the final approval decision. These records can serve as training examples to build a decision tree for a loan advisory system.

Income	Credit History	Debt	Decision
$0-5K$	Bad	Low	Reject
$0-5K$	Good	Low	Approve
$0-5K$	Unknown	High	Reject
$0-5K$	Unknown	Low	Approve
$0-5K$	Unknown	Low	Approve
$0-5K$	Unknown	Low	Reject
$5-10K$	Bad	High	Reject
$5-10K$	Good	High	Approve
$5-10K$	Unknown	High	Approve
$5-10K$	Unknown	Low	Approve
Over $10K$	Bad	Low	Reject
Over $10K$	Good	Low	Approve

   (a) Construct a decision tree based on the above training examples. (Note: $\log_2 \frac{x}{y} = \log_2 x - \log_2 y$, $\log_2 1 = 0$, $\log_2 2 = 1$, $\log_2 3 = 1.585$, $\log_2 4 = 2$, $\log_2 5 = 2.322$, $\log_2 6 = 2.585$, $\log_2 7 = 2.807$, $\log_2 8 = 3$, $\log_2 9 = 3.170$, $\log_2 10 = 3.322$, $\log_2 11 = 3.459$, and $\log_2 12 = 3.585$)

   (b) What is decision tree classifier's decision for a person who has 4K yearly income, a good credit history and a high amount of debt?

   (c) Use Naïve Bayes to calculate the decision. Does it differ from the decision tree classifier?

5. The Fit And Trim club is recruiting people, and we believe that weight and height may be two of their primary recruitment criteria. We managed to obtain some data on their past recruitment history on the following six people.

Person	Weight (kg)	Height (cm)	Accepted?
A	79	189	Yes
B	76	170	Yes
C	77	155	Yes
D	72	163	No
E	73	195	No
F	70	182	No

   We want to use the data to predict whether Ah Bui, whose weight is 78 kg and whose height is 179 cm, will be recruited into the Fit And Trim club.

   (a) We can consider each person's weight and height as a point in the two-dimensional Euclidean space. By using the Nearest Neighbour learning algorithm on the raw data, predict whether Ah Bui will be recruited into the Fit And Trim club. Is there a problem with this prediction? (The Euclidean distance between two points (w_1, h_1) and (w_2, h_2), representing the weights and heights of two individuals, can be calculated as: $\sqrt{(w_1 - w_2)^2 + (h_1 - h_2)^2}$).

(b) Now normalize the data such that the weights and heights are in the range of 0 to 1, by scaling them in the following way:

If the weight is w, then the normalized weight is $(w - 70)/10$.
If the height is h, then the normalized height is $(h - 150)/50$.

Reapply the Nearest Neighbour algorithm on the normalized data to predict whether Ah Bui will be recruited into the Fit And Trim club. Did normalizing the data improve the prediction?