ANSWERS

Exercise 1 - Information Gain

We start by finding the entropy of the whole unsplit set and for all the subsets resulting from possible splits. These entropies can be read from the table.

Entropy	**Proportion of whole set**	**Split**
$entropy(whole\_set) = 0.94$	14/14	5, 9
$entropy(Outlook : Sunny) = 0.97$	5/14	2, 3
$entropy(Outlook : Overcast) = 0$	4/14	4, 0
$entropy(Outlook : Rainy) = 0.97$	5/14	3, 2
$entropy(Temp : Hot) = 1$	4/16	2, 2
$entropy(Temp : Mild) = 0.92$	6/14	4, 2
$entropy(Temp : Cool) = 0.81$	4/14	3, 1
$entropy(Humidity : High) = 0.99$	7/14	3, 4
$entropy(Humidity : Normal) = 0.59$	7/14	6, 1
$entropy(Windy : True) = 1$	6/14	3, 3
$entropy(Windy : False) = 0.81$	8/14	6, 2

Now we calculate the information gain that would be achieved by each split.

$IG(Outlook) = entropy(whole\_set) – 5/14 * 0.97 – 4/14 * 0 – 5/14 * 0.97 = 0.25$
$IG(Temp) = entropy(whole\_set) – 4/14 * 1 - 6/14 * 0.92 – 4/14 * 0.81 = 0.03$
$IG(Humidity) = entropy(whole\_set) – 7/14 * 0.99 – 7/14 * 0.59 = 0.15$
$IG(Windy) = entropy(whole\_set) – 6/14 * 1 – 8/14 * 0.81 = 0.05$

The variable with the strongest relationship to the target 'Play' is 'Outlook' as it produces the highest information gain (0.25) with respect to 'Play'.

Exercise 2 - Evidence Lift

The evidence lift is the ratio of the *probability that evidence is present provided an outcome has happened* to the *probability of the same evidence in general.* This can be expressed with the following formula:

$$lift_o(e) = \frac{p(e|o)}{p(e)}$$

1. In the first case, 'square' is evidence and 'thick border' is the outcome. The probability of a square in the entire population is:

   $$p(S) = \frac{17}{26}$$

   The probability of a square among shapes with thick borders is:

   $$p(S|Thk) = \frac{7}{14}$$

   The evidence lift is:

   $$lift_{Thk}(S) = \frac{p(S|Thk)}{p(S)} = \frac{\frac{7}{14}}{\frac{17}{26}} = \frac{13}{17}$$

   The lift is less than 1 which means that it isn't a lift but rather a 'lowering', in that a shape is less likely t be a square if it has thick borders than in general.

2. Here the 'thick border' is the evidence and 'square' is the outcome. The probability of a thick border in the entire population is:

   $$p(Thk) = \frac{14}{26}$$

   The probability of a thick border among squares is:

   $$p(Thk|S) = \frac{7}{17}$$

   The evidence lift is:

   $$lift_S(Thk) = \frac{p(Thk|S)}{p(Thk)} = \frac{\frac{7}{17}}{\frac{14}{26}} = \frac{13}{17}$$

3. The results obtained in the previous two calculations are the same. This is not a random coincidence, as the lift is a symmetrical value with respect to the two events involved and

they can interchangeably play the roles of 'evidence' and 'outcome'. The general formula for lift is:

$$lift(A, B) = \frac{p(A, B)}{p(A)p(B)}$$

It tells us how much more probable the co-occurrence of A and B is ($p(A, B)$) than it would be if A and B were occurring completely independently ($p(A)p(B)$).

The nature of probability allows us to express co-occurrence probability as:

$$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$

The lift is then:

$$lift(A, B) = \frac{p(A|B)p(B)}{p(A)p(B)} = \frac{p(A|B)}{p(A)}$$

or:

$$lift(A, B) = \frac{p(B|A)p(A)}{p(A)p(B)} = \frac{p(B|A)}{p(B)}$$

The two results we have obtained are, respectively, the lift provided by A for B and the lift provided by B for A.

4. We perform the lift calculation as in previous cases:

$$p(Thk) = \frac{14}{26}$$

$$p(Thk|C) = \frac{7}{9}$$

$$lift_C(Thk) = \frac{p(Thk|C)}{p(Thk)} = \frac{\frac{7}{9}}{\frac{14}{26}} = \frac{13}{9} = 1.44$$

This lift value says that given a shape is a circle it is almost 1.5 times more likely to be thick bordered than the general population of shapes.