DECISION TREE

Representation

1. Suppose you have a learning problem in which each instance has 4 Boolean attributes and a Boolean class label. What would be the number of nodes in the DT and how many of these would be leaf nodes?

2. In case of question 1 above, how many possible labeling (i.e. class assignment) of the instances are possible.
* Another way of asking the same question is "In how many ways can the leaves be labeled?" *

Entropy

1. What is the entropy of a fair coin?
Hint: Two classes, p(x) = 0.5 for each

2. What is the entropy of a fair dice?

3. A random variable is defined as:

$$X = \begin{cases} a & \text{with probability } 1/2, \\ b & \text{with probability } 1/4, \\ c & \text{with probability } 1/8, \\ d & \text{with probability } 1/8. \end{cases}$$

What is its entropy? Would the value be greater or smaller if all the values had an equal probability?

4. For a two class mixture, when is the value of entropy maximized? What about an "n-class" mixture?

Conditional Probability

1. Doctors find that people with high cholesterol (C) almost invariably ate hamburgers (H), thus p(Hamburger Eater|C) = 0.9. The probability of an individual having C is currently rather low, about one in 100,000. Assuming lots of people eat hamburgers i.e. p(H) = 0.5, what is the probability that someone who eats hamburgers (H) will get the disease C.
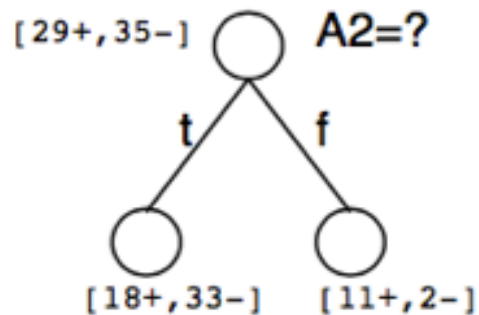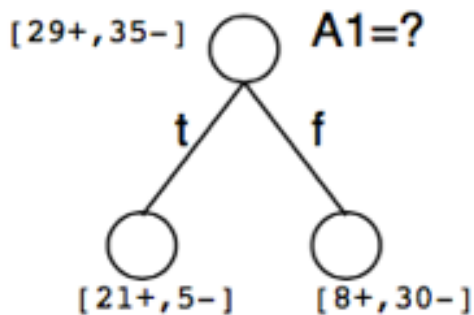
That is, find the following:

P(C | H) = ?

2. From a deck of 52 cards, 2 cards are missing. You randomly draw a card from the deck, what is the probability that it will be a spade?

Hint: You have to consider three separate cases – 0, 1, or 2 spades may be missing. Then add up their conditional probabilities.
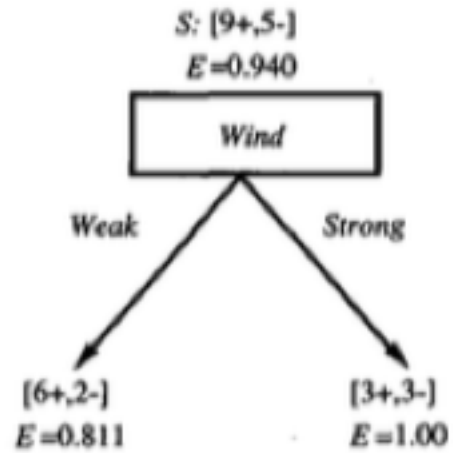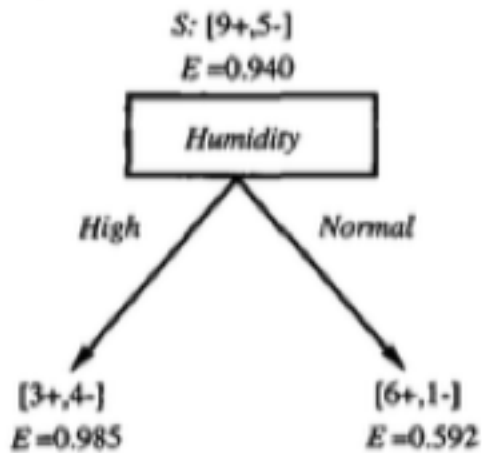
Information Gain

1. Compute the IG for the scenario below:

That is, there are two attributes A1 and A2. You have to decide which one gives you higher information gain?

First compute original entropy. Then compute IG for both cases.

2. Consider the scenario below:

S: [9+,5-]
E=0.940

Humidity

High   Normal

[3+,4-]   [6+,1-]
E=0.985   E=0.592

S: [9+,5-]
E=0.940

Wind

Weak   Strong

[6+,2-]   [3+,3-]
E=0.811   E=1.00

The entropy values are provided. Which attribute would provide more IG -> Humidity or Wind?

3. Consider the dataset shown below:

Instance	Classification	a_1	a_2
1	+	T	T
2	+	T	T
3	−	T	F
4	+	F	F
5	−	F	T
6	−	F	T

a. What is the entropy of the dataset with respect to the target classification attribute.

b. How much IG do you obtain by sorting on a1 and sorting on a2? Which one should you prefer?

Applying ID3 Algorithm

See the attached PowerPoint.