Homework II-KEY

1. Consider the training dataset given below. A, B, and C are the attributes and Y is the class variable. **[10 Points]**

A	B	C	Y
0	1	0	Yes
1	0	1	Yes
0	0	0	No
1	0	1	No
0	1	1	No
1	1	0	Yes

   a. Can you draw a decision tree having 100% accuracy on this training set? If you answer is yes, draw the decision tree in the space provided below. If your answer is no, explain why?
   No. Because Examples #2 and #4 have the same feature vector but different classes.

   b. Which attribute among A, B and C has the highest information gain? Explain your answer.
   Entropy of the entire data set = H(3,3) =1

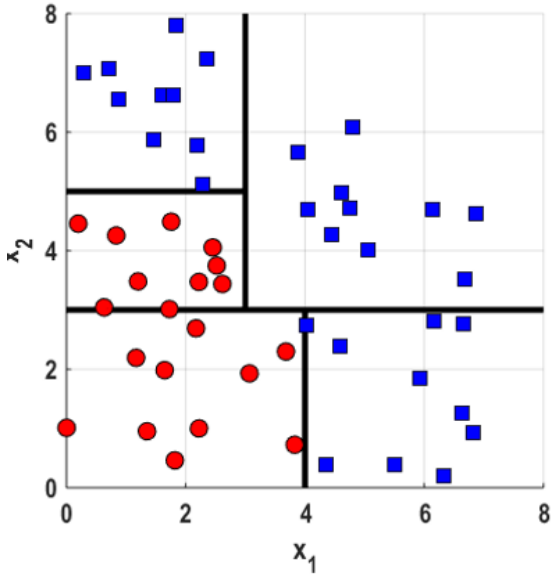
$$IG(A) = H(3,3) - \tfrac{1}{2}H(2,1) - \tfrac{1}{2}H(2,1)$$

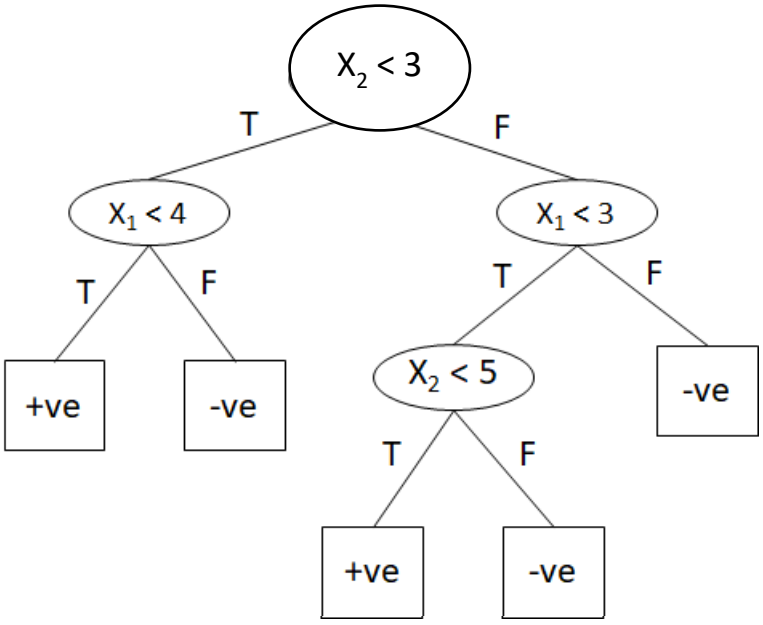
$$IG(B) = H(3,3) - \frac{1}{2}H(2,1) - \frac{1}{2}H(2,1)$$

$$IG(C) = H(3,3) - \frac{1}{2}H(2,1) - \frac{1}{2}H(2,1)$$

   Therefore, all three attributes have the same information gain.

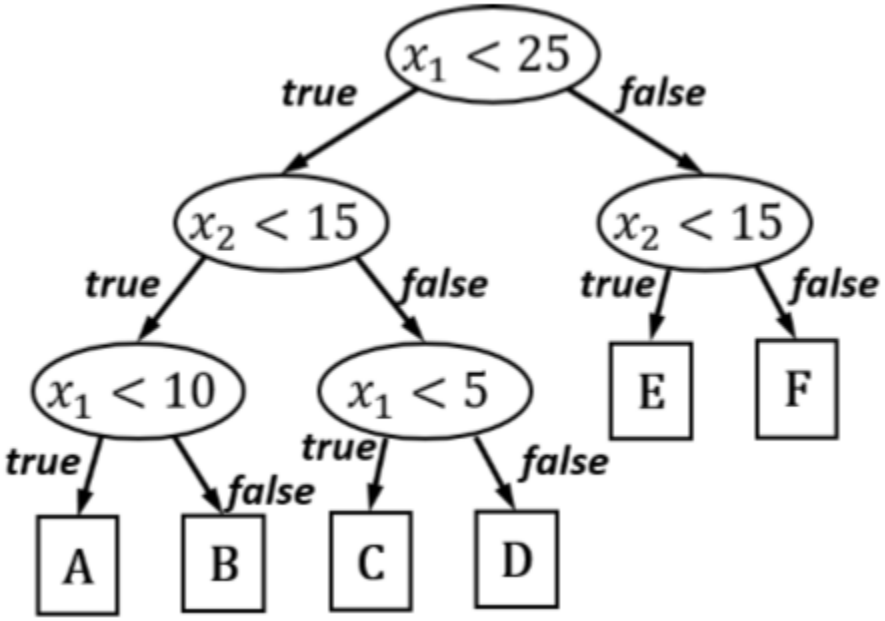
2. Interpreting a decision tree: Consider the decision boundary in Fig. and draw the equivalent decision tree. Red circle are Class +1 and blue squares are class -1.**[5 Points]**

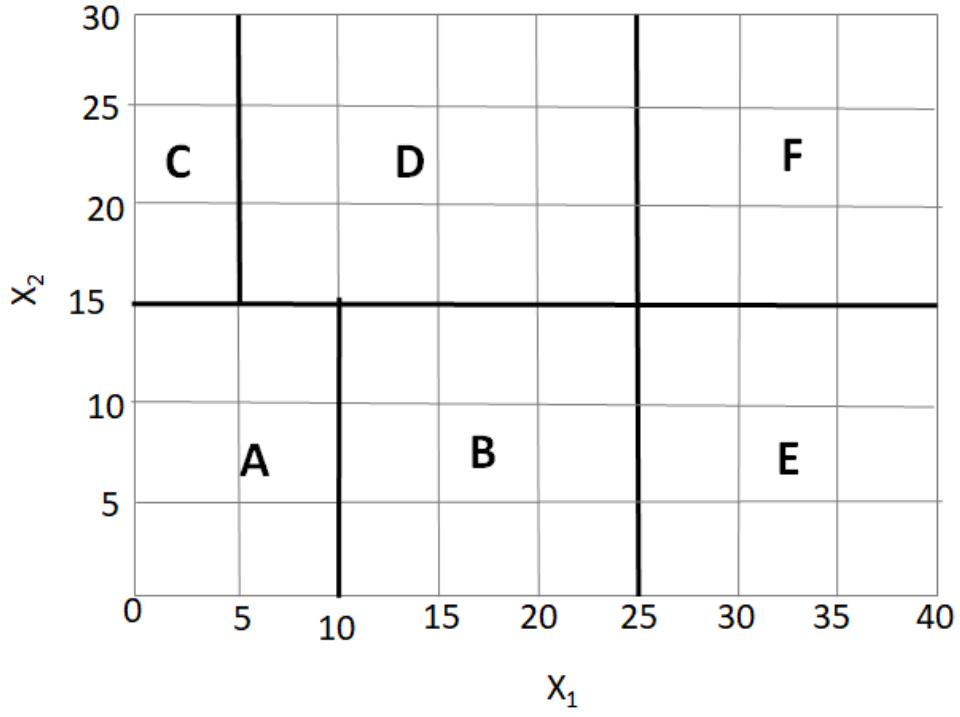




3. Visualizing a decision tree: Consider the decision in Fig and draw the equivalent decision boundary. Make sure to label each decision region with the corresponding leaf node from the decision tree.

[5 Points]





4. Bayes rule for medical diagnosis.
   After your yearly checkup, the doctor has some bad news and some good news. The bad news is that you tested positive for a serious disease, and the test is 99% accurate(i.e. that probability of testing positive given that you have the disease is .99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the changes that you actually have the disease? (Show the calculation as well as giving the final result)

[5 Points]

Let X = 1 represent a positive test outcome, X = 0 represent a negative test outcome, Y = 1 mean you have the disease, and Y= 0 mean you don't have the disease. We are told
P(X = 1 | Y = 1) = 0:99
P(X = 0 | Y = 0) = 0:99
P(Y = 1) = 0:0001
We are asked to compute P(Y = 1|T = 1), which we can do using Bayes' rule:
P(Y = 1| X = 1) = P(X = 1|Y = 1)P(Y = 1) / P(X = 1| Y = 1)P(Y = 1) + P(X = 1 | Y = 0)P(Y = 0)

=
0:99 * 0:0001 /(0:99 * 0:0001 + 0:01 * 0:9999)

= 0:009804
So although you are much more likely to have the disease (given that you have tested positive) than a random member of the population, you are still unlikely to have it.

5. Express the mutual information in terms of the entropies. Show that
   I[X,Y] = H[X] – H[X|Y] = H[Y] – H[Y|X] **[10 Points]**

$$I(X,Y) = \sum_{x,y} p(x,y)\, log \frac{p(x,y)}{p(x)p(y)}$$

$$= \sum_{x,y} p(x,y)\, log \frac{p(x|y)}{p(x)}$$

$$= -\sum_{x,y} p(x,y) log p(x) \; + \; \sum_{x,y} p(x,y) \log p(x|y)$$

$$= -\sum_{x} p(x) log p(x) \; - \; \left(-\sum_{x,y} p(x,y) \log p(x|y) \right)$$

$$= -\sum_{x} p(x) log p(x) \; - \; \left(-\sum_{y} p(y) \sum p(x|y) \log p(x|y) \right)$$

$$= H(X) - H(X|Y)$$