1 Entropy

1: How would you informally describe entropy

In class we defined Entropy as being a measure of hidden information or uncertainty.

2: Give the formula for entropy of a distribution

X is a discrete random variable with pmf $p_X(x)$, the formula for its entropy is as follows:

$$H(X) = -\sum_x p(x)log(p(x))$$

3: Describe the distribution with would have the least entropy. Describe the distribution having the maximum entropy (hint - question 5)

The minimum value a distribution can have is 0. This happens when their is only one output for an event i.e. its probability is 100%
The maximum value a distribution can have is 1. This means that the data set would have an equal probability for each output in the feature?

4: A distribution has x% "true" and (100-x)% false, and has entropy h. Describe another distribution with different the percentage of true being something other than x that also has the same entropy.

If x% is true and (100-x)% is false with entropy h, then (100-x)% being true and x% being false would yield the same entropy h.

5: A distribution d1 has 50%"a" and 50%"b". What is its entropy? A distribution d2 has 25%"a", 25%"b", 25%"c" and 25%"d". What is its entropy?

The entropy for both distributions is 0 because their is an equal probability for every output possible in both cases.

2 Mutual Information

1: How will you informally describe mutual information?

In regard to two random variables, mutual information measures the amount of information that one variable contains about the other.

2: Consider distributions X and Y (both binary). Give the formula for mutual information between X and Y. Give the formula for mutual distribution between Y

and X. Will they be same or different? If they will be same, prove this; if they will be different, give one counter examples.

Mutual Information between X and Y:
$MI(Y,X) = MI(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$, where H is entropy.

3: What will the maximum and minimum values of mutual information be?

Minimum value: 0
Maximum value: $\min[H(var1), H(var2)]$ or the smallest entropy between the two features.

4: Let X and Y be as follows.

x	y
A	T
A	T
B	T
B	T
C	F
C	F
D	F
D	F

Give the conditional entropy and information gain in both directions, I.e. how much information you gain about y by knowing about x; as well as how much information you gain about x by knowing about y.

$$
\begin{aligned}
H(Y|X) &= -[P(A)[P(T|A)log(P(T|A)) + P(F|A)log(P(F|A))] \\
&\quad + P(B)[P(T|B)log(P(T|B)) + P(F|B)log(P(F|B))] \\
&\quad + P(C)[P(T|C)log(P(T|C)) + P(F|C)log(P(F|C))] \\
&\quad + P(D)[P(T|D)log(P(T|D)) + P(F|D)log(P(F|D))]] \\
H(Y|X) &= -4[\frac{1}{4}((1) * log(1) + 0 * log(0))] \\
H(Y|X) &= 0 \\
MI(X,Y) &= H(Y) - H(Y|X) \\
MI(X,Y) &= 1 - 0 \\
MI(X,Y) &= 1
\end{aligned}
$$

$$
\begin{aligned}
H(X|Y) &= -(P(T)[P(A|T)log(P(A|T)) + P(B|T)log(P(B|T)) + P(C|T)log(P(C|T)) + P(D|T)log(P(D|T))] \\
&\quad + P(F)[P(A|F)log(P(A|F)) + P(B|F)log(P(B|F)) + P(C|F)log(P(C|F)) + P(D|F)log(P(D|F))]) \\
&= -(\frac{1}{2}[\frac{1}{2}log(\frac{1}{2}) + \frac{1}{2}log(\frac{1}{2}) + 0 \times log(0) + 0 \times log(0)] + \frac{1}{2}[0 \times log(0) + 0 \times log(0) + \frac{1}{2}log(\frac{1}{2}) + \frac{1}{2}log(\frac{1}{2})]) \\
&= 0.301029995664 \\
H(X) &= 1 \\
M(Y,X) &= H(X) - H(X|Y) \\
&= 1 - 0.301029995664 \\
&= 0.698970004336
\end{aligned}
$$

Given resources say $MI(X,Y)$ is equal to $MI(Y,X)$. We also know that X gives us all the information we need to know about Y but Y only narrows the probability of getting X correct is 50%. We're not sure why we are getting different mutual information numbers, we tried asking for help but got no response. Feedback would be helpful here.

5: Consider the following distributions over Boolean. X and not(X). That is

x	Notx
T	F
F	T
F	T
F	T
T	F

What is the information gain either way?

Information gain is maximum both ways. We can determine x if we know $!x$, and we can also determine $!x$ if we know x with 100% accuracy.

6: If the information gain of a feature about the output is 0, does this mean that the feature is irrelevant and should be thrown away?

No the feature is not irrelevant, this means that the two features are independent of one another

3 Cross Entropy

1: How would you intuitively describe cross entropy?

Cross entropy is the difference between two probabilities.

2: Give the formula of cross entropy between 2 distributions

$$H(p,q) = -\sum_{x} p(x)log(q(x))$$

3: When is the cross entropy maximum and when is it minimum?

Minimum value: 0 (when p(x) is 0)
Maximum value: ∞ (when q(x) is 0)

4: Consider distributions X and Y (both binary). Give the formula for mutual information between X and Y. Give the formula for cross entropy distribution between Y and X. Will they be same or different? If they will be same, prove this; if they will be different, give one counter examples.

Mutual Information: $H(X) - H(X|Y) = H(Y) - H(Y|X)$
Cross Entropy: $-\sum_x p(x)log(q(x))$

Mutual Information and Cross Entropy are not the same since Mutual Information is the same going either from X to Y or Y to X by definition. However, Cross Entropy is not bidirectional.

4 Decision Trees

1: Describe in words the decision tree algorithm?

Choose the "best" decision attribute, lets call it X, to be be your root value. Then, for each value of X create a new descendant node. Then sort the training examples to leaf nodes, if they are classified perfectly, stop, else, recurse algorithm over new leaf nodes.

2: When does the algorithm stop?

When either the data is unambiguous, or there are no remaining features

3: What is the maximum depth and maximum number of nodes that may be generated?

The maximum depth of a decision tree is N - 1 where N is the number of training samples. The maximum number of nodes in a decision tree is 2 to the power of the depth of the tree.

4: If the table you created in naive baysian is all that you know of the data, how much of the decision tree can you construct?

With a Naive Baysian table, you could determine which feature has the lowest entropy output and determine the first node of the tree, but since we do not know how the data is split after the first node, we cannot determine anything else about the decision tree.

5: What is the bias of the decision tree algorithm? How is this reflected in the algorithm? Why is this a greedy algorithm?

The decision tree algorithm makes the most optimal decision at each node of the tree, which means it is making locally optimal decisions recursively in the assumption that this will lead to the best solution. The algorithm is recursive so only works on one subset of the tree at a time, and only considers the best result for subset. In other words, it uses a divide and conquer approach, making it a greedy algorithm.

6: Construct the decision tree by hand showing all your computation.

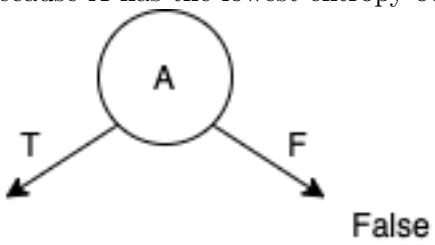
A	B	C	Op
T	T	T	T
T	T	F	T
T	F	T	T
T	F	F	F
F	T	T	F
F	T	F	F
F	F	T	F
F	F	F	F

$A = T$:
op = { T, T, T, F }
Entropy = $-(\frac{3}{4}log(\frac{3}{4} + \frac{1}{4}log(\frac{1}{4})$
= 0.24219
$A = F$
op = { F, F, F, F }
Entropy = 0
H(op) = $\frac{1}{2}0.24219 + \frac{1}{2}0 = 0.12211$

$B = T$
op = { T, T, F, F }
Entropy = 1
$B = F$:
op = { T, F, F, F }
Entropy = $-(\frac{3}{4}log(\frac{3}{4} + \frac{1}{4}log(\frac{1}{4})$
= 0.24219
H(op) = $\frac{1}{2}0.24219 + \frac{1}{2}1 = 0.19897$

$C = T$
op = { T, T, F, F }
Entropy = 1
$C = F$:
op = { T, F, F, F }
Entropy = $-(\frac{3}{4}log(\frac{3}{4} + \frac{1}{4}log(\frac{1}{4})$
= 0.24219
H(op) = $\frac{1}{2}0.24219 + \frac{1}{2}1 = 0.19897$

Because A has the lowest entropy output, it becomes the root node of our decision tree like so:

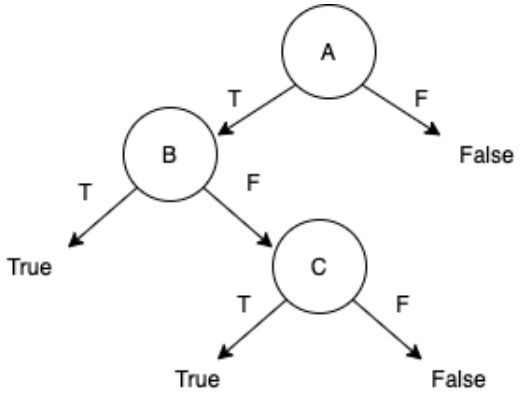


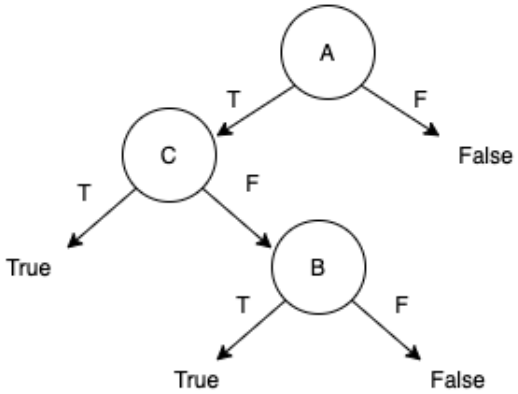
Now using the op set from A=T:
$B = T$
op = { T, T }
Entropy = 0
$B = F$
op = { T, F }
Entropy = 1
H(op) = $\frac{1}{2}0 + \frac{1}{2}1 = 0.5$

$C = T$
op = { T, T }
Entropy = 0
$C = F$
op = { T, F }
Entropy = 1
H(op) = $\frac{1}{2}0 + \frac{1}{2}1 = 0.5$

Both of these values are equal so choose one to be the next node. Let B be the next node, using the op set for A=T and B=T
$C = T$
op = { T }
Entropy = 0
$C = F$
op = { F }
Entropy = 0
Data is unambiguous so the tree stops.

Now let C be the next node instead of B, using the op set for A=T and C =T
$B = T$
op = { T }
Entropy = 0
$B = F$
op = { F }
Entropy = 0
Data is unambiguous so the tree stops.
So both of the following trees are valid:


A
T
F
False
B
T
F
True
C
T
F
True
False


A
T
F
False
C
T
F
True
B
T
F
True
False