Q1-1: How many distinct (binary classification) decision trees are possible with *4* Boolean attributes? Here distinct means representing different functions.

1. 2^4
2. 2^8
3. 2^{16}
4. 2^{32}

Q1-1: How many distinct (binary classification) decision trees are possible with *4* Boolean attributes? Here distinct means representing different functions.

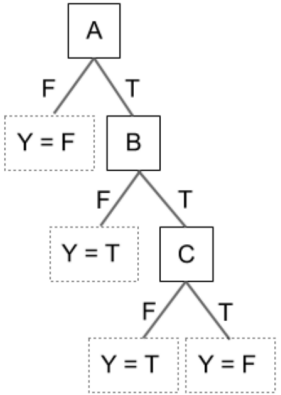
1. 2^4
2. 2^8
3. 2^{16} 
4. 2^{32}

#distinct decision trees
= #distinct Boolean functions
= #functions of 2^4 = 16 inputs, binary label for each input
= $\mathbf{2^{16}}$

Q1-2: Following decision tree (DT) classifies Y as T/F using the 3 binary variables - A, B, C with zero training error. Is this DT unique? If no, which variable can be removed from the DT and still get a zero-training error DT?

1. Yes, it's unique.
2. No, A can be removed.
3. No, B can be removed.
4. No, C can be removed.

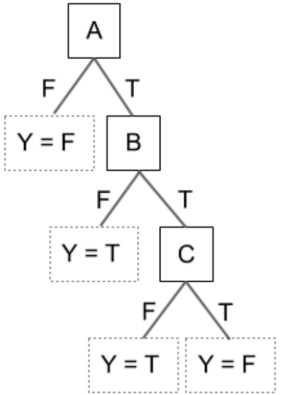
A	B	C	Y
F	F	F	F
T	F	T	T
T	T	F	T
T	T	T	F

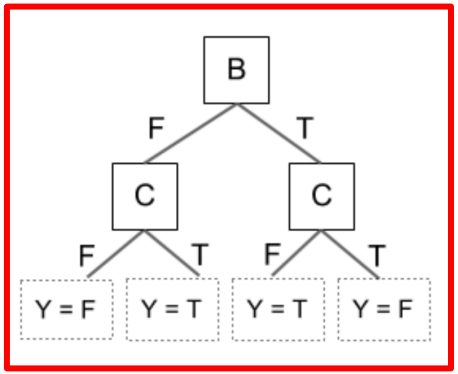


Q1-2: Following decision tree (DT) classifies Y as T/F using the 3 binary variables - A, B, C with zero training error. Is this DT unique? If no, which variable can be removed from the DT and still get a zero-training error DT?

1. Yes, it's unique.
2. No, A can be removed. ⬅
3. No, B can be removed.
4. No, C can be removed.

A	B	C	Y
F	F	F	F
T	F	T	T
T	T	F	T
T	T	T	F





Notice, Y is just a function of B, C: Y = B xor C. Hence, A can be removed and we can build a DT of depth 2 as shown.

Q2-1: Which of the following statement is true?

1. Higher the entropy, lower the uncertainty.
2. Entropy of a variable with an impossible event is not defined.
3. Big DT are preferred over small DT, because it tells us how each variable is affecting the decision.
4. If there are M possible values for a random variable, then the maximum possible entropy is $\log_2(M)$.

Q2-1: Which of the following statement is true?

1. Higher the entropy, lower the uncertainty.
2. Entropy of a variable with an impossible event is not defined.
3. Big DT are preferred over small DT, because it tells us how each variable is affecting the decision.
4. If there are M possible values for a random variable, then the maximum possible entropy is $\log_2(M)$.

1. Higher the entropy, higher the uncertainty (Entropy is highest when all the events are equiprobable).
2. Entropy of an impossible event (i.e. p = 0), is 0.
3. Occam's razor principle.
4. Same as 1. Hence, max possible entropy = $\sum_M -1/M \ \log_2(1/M) = \log_2(M)$.

Q2-2: Following is the table which predicts if people pass CS760 (Yes or No) based on their previous GPA (High, Medium or Low) and if they studied or not. Calculate the entropy *H(Passed | Studied). Assume equal probability for each event in the table. Take* $log_2(1/3) = -1.5$ & $log_2(2/3) = -0.5$

GPA	Studied	Passed
L	N	N
L	Y	Y
M	N	N
M	Y	Y
H	N	Y
H	Y	Y

1. ~0.66
2. ~0.92
3. ~0.42
4. ~0.14

Q2-2: Following is the table which predicts if people pass CS760 (Yes or No) based on their previous GPA (High, Medium or Low) and if they studied or not. Calculate the entropy *H(Passed | Studied). Assume equal probability for each event in the table. Take* $\log_2(1/3) = -1.5$ *&* $\log_2(2/3) = -0.5$

GPA	Studied	Passed
L	N	N
L	Y	Y
M	N	N
M	Y	Y
H	N	Y
H	Y	Y

1. ~0.66
2. ~0.92
3. ~0.42 
4. ~0.14

H(Passed | Studied)
= p(Studied = Y).H(Passed | Studied = Y) + p(Studied = N).H(Passed | Studied = N)

H(Passed | Studied = N)
= p(Passed = Y | Studied = N).$\log_2$(p(Passed = Y | Studied = N)) +
p(Passed = N | Studied = N).$\log_2$(p(Passed = N | Studied = N))
= - [⅓ $\log_2$ ⅓ + ⅔ $\log_2$ ⅔]

Similarly, H(Passed | Studied = Y) = 0.
Hence, H(Passed | Studied) = -½ . [⅓ $\log_2$ ⅓ + ⅔ $\log_2$ ⅔] = 0.5 x [0.5 + 0.33] = 0.42 (approx)

Q3-1: Which of the following statement is true?

1. Splitting occurs on the attribute which has the lowest InfoGain.
2. ID3 always gives a minimal depth Decision tree.
3. The range of InfoGain is [0, 1].
4. Gain ratio is less biased towards tests with many outcomes than InfoGain.

Q3-1: Which of the following statement is true?

1. Splitting occurs on the attribute which has the lowest InfoGain.
2. ID3 always gives a minimal depth Decision tree.
3. The range of InfoGain is [0, 1].
4. Gain ratio is less biased towards tests with many outcomes than InfoGain. 

1. Splitting occurs on the attribute which has the highest InfoGain.
2. ID3 is a greedy algorithm. It doesn't necessarily gives minimal Depth DT.
3. It can be as large as H(Y).
4. This is true. Gain ratio rectifies the issue of bias by normalization.

Q3-2: Following is the table which predicts if people pass CS760 (Yes or No) based on their previous GPA (High, Medium or Low) and if they studied or not.
Which attribute should be chosen for splitting first? *Assume equal probability for each event in the table. Note that we know H(Passed | Studied) =* ***0.42****, H(Passed) =* ***0.92***

1. GPA
2. Studied
3. Any of them

GPA	Studied	Passed
L	N	N
L	Y	Y
M	N	N
M	Y	Y
H	N	Y
H	Y	Y

Q3-2: Following is the table which predicts if people pass CS760 (Yes or No) based on their previous GPA (High, Medium or Low) and if they studied or not.
Which attribute should be chosen for splitting first?
Given H(Passed | Studied) = ***0.42****, H(Passed) =* ***0.92***

GPA	Studied	Passed
L	N	N
L	Y	Y
M	N	N
M	Y	Y
H	N	Y
H	Y	Y

1. GPA
2. Studied 
3. Any of them

H(Passed | GPA = L)
= p(Passed = Y | GPA=L).$\log_2$(p(Passed = Y | GPA=L)) +
p(Passed = N | GPA=L).$\log_2$(p(Passed = N | GPA=L))
= - [0.5 $\log_2$ 0.5 + 0.5 $\log_2$ 0.5] = 1

Similarly, H(Passed | GPA = M) = 1, H(Passed | GPA = H) = 0.
Hence, H(Passed | GPA) = 1/3*1+1/3*1+1/3*0 = 0.66 (approx)

InfoGain(GPA) = H(Passed) - H(Passed | GPA) = 0.92 - 0.66 = 0.26
InfoGain(Studied) = H(Passed) - H(Passed | Studied) = 0.92 - 0.42 = 0.5
InfoGain(Studied) > InfoGain(GPA), therefore, splitting should happen on attribute 'Studied'.