[17%] True/False Questions

ML and Decision Tree:

1- Generalization of a data set is to overfill the data by a function. (False)

2- Both deductive and inductive learning are guaranteed to be correct. (False)

3- The labels of examples are determined by a learning procedure. (False)

NN and Deep Learning:
4- Deep Learning cannot automatically change the topological structure of a NN. (True)

5- NN are often successful for both supervised and unsupervised learning. （T）

Probability:

6- Probability theory can be used for both incremental(on-line) and batch(off-line) learning. (T)

7- If A and B are independent, then they are mutually exclusive. （F）

8- Mutual exclusiveness is not the same as probability independence. （T）

Bayesian Networks:

9- Given a fully defined Bayesian Network, we can do all the probabilistic reasonings of the random variables in the Network. (t)

10- A Bayesian Network is not semantically equivalent to a fully joint probability distribution table. （F）

Maximal Expected Utility (MEU) and Rationality

11- Expected Utilities have nothing to do with the probabilistic outcomes of the actions. (f)

12- Rational agents choose actions without knowing the expected utilities of the states. (F)

Bayesian Learning/Classifier

13- A Naive Bayesian Classifier assumes, as much as it can, about the independence of the given variables. (T)

14- The Bayesian Rule cannot be derived from the Product Rule. (F)

Enumeration algorithm

15- The enumeration algorithm uses only the marginalization rule (F)

Approximation algorithms (approach to the real probability)

16- An approximation reasoning algorithm for BBnets does not need to be provable to approach the true probability when the sample size goes to infinity. (f)

Hidden Markov Models

17- The results of the Viterbi algorithm do not include the distribution of all possible sequences of state transitions for a given observation sequence. (T)

[14%] Decision Tree Learning (Monday Section)

You are given the task of classifying if a person buys a computer based on 4 attributes **[Person's Age, Income, Is he a student, Credit Rating]** as shown in the following table :

As you can observe - "age" can be 21-30 or 31-40, "Income" can be high or low, "is he a student" can be yes or no and the "credit rating" can be excellent or poor. The data provided is fictitious and for the purpose of testing your knowledge about decision trees only.

Index	Age	Income	Is he a student	Credit rating	Buys a computer
1	31-40	high	yes	excellent	***yes***
2	21-30	high	yes	poor	***yes***
3	21-30	low	no	excellent	***no***
4	31-40	high	no	poor	***no***
5	31-40	low	yes	excellent	***yes***
6	21-30	low	yes	poor	***no***
7	31-40	low	no	excellent	***yes***
8	21-30	low	no	poor	***no***

Answer the following questions based on the information provided above :

The following values are provided for your reference -

0 * log (0) = 0
1/4 * log (1/4) = - 0.5
2/4 * log (2/4) = - 0.5
3 /5 * log (3 /5) = -0.44
2 /5 * log(2 /5) = -0.53
3/4 * log (3/4) = -0.3
4/4 * log (4/4) = 0
2/3 * log(2/3) = -0.39
1/3 * log(1/3) = -0.53

Please note we would only be grading based on log base 2 values (using which the above values were calculated). Please do not use any other log base.

[1%][A] Calculate the entropy of the decision [Buys a computer]

Entropy (Buys a computer) = [-0.5 * log(0.5) - 0.5 log(0.5)] = 1

For B,C,D partial marking - 1% for correct formula application/correct raw values, 1% for correct final answer

[2%][B] Calculate the entropy remaining (remainder) if “income” is chosen as the splitting attribute.

Entropy (income) = 3/8 * [- 2/3 * log (2/3) - 1/3 * log(1/3)] + 5/8 * [- 2/5 * log(2/5) - 3/5 * log(3/5)] = 3/8 * [0.39 + 0.53] + 5/8 * [0.53 + 0.44] = 0.95

[2%][C] Calculate the entropy remaining (remainder) if “is he a student” is chosen as the splitting attribute.

Entropy (is he a student) = ½ * [- 3/4 * log(3/4) - 1/4 * log(1/4)] + ½ * [- 1/4 * log (1/4) - 3/4 * log(3/4)] = 0.5 * (0.3+0.5) + 0.5 * (0.3+0.5) = 0.8

[2%][D] Calculate the entropy remaining (remainder) if “credit rating” is chosen as the splitting attribute.

Entropy (credit rating) = ½ * [- 3/4 * log(3/4) - 1/4 * log(1/4)] + ½ * [- 1/4 * log (1/4) - 3/4 * log(3/4)] = 0.5 * (0.3+0.5) + 0.5 * (0.3+0.5) = 0.8

[3%][E] Calculate the information gain for each attribute

InfoGain (income) =
InfoGain (is he a student) =
InfoGain (credit rating) =

InfoGain (income) = **1 - 0.95 = 0.05 [1%]**
InfoGain (is he a student) = **1 - 0.8 = 0.2 [1%]**
InfoGain (credit rating) = **1 - 0.8 = 0.2 [1%]**

[2%][F] Which attribute should NOT be chosen to split the dataset and why?
In case of a tie, please mention all such attributes.

Income[1%]. Reason : It has lowest information gain.[1%]

[2%][G] How many distinct decision trees are possible with n Boolean attributes?

1. n
2. 2 ^ n
3. 2 ^ (2 ^ n)
4. n!

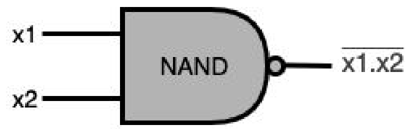
Option 3

[14%] Neural Networks

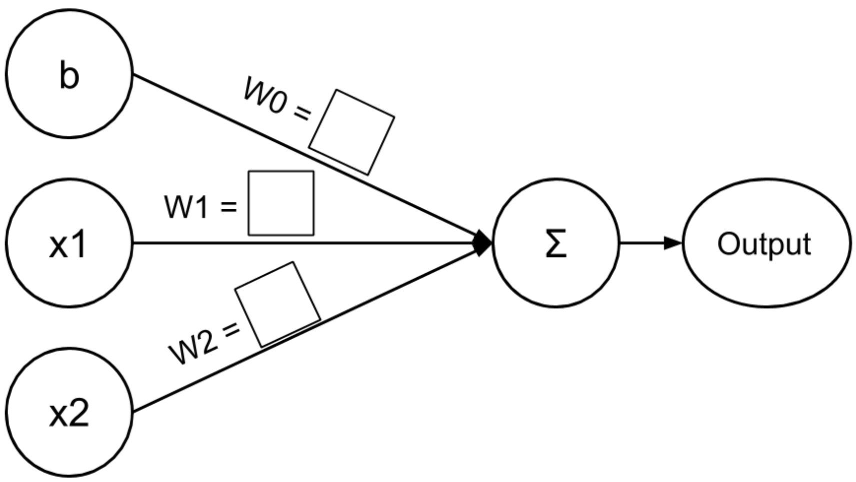
Given the truth tables and neural network structures below, write the weights and bias in the empty squares on the edge that will be equivalent to the required logic gate. Assume the following activation function:

$$f(x) = \begin{cases} 1 & for\ x \geq 0 \\ 0 & for\ x < 0 \end{cases}$$

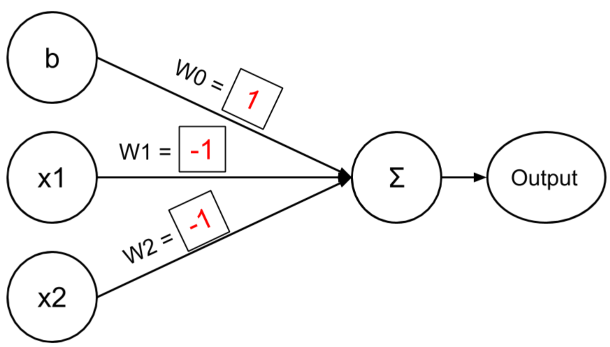
1a. [3%] NAND Gate



x1	x2	Out
0	0	1
0	1	1
1	0	1
1	1	0



[No partial credits. There are more than one correct solutions, solutions are correct as long as it satisfies the truth table]



1b. [4%] Assuming the initial weights are all 1 and that the learning rate is 0.5, what would be the weights after training on the correct value for x1=1 and x2=1? Please show your work below only for one iteration. (Write the updated w1, w2, w0 and output of the perceptron for the updated weights)

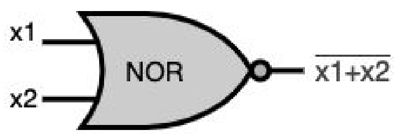
Solution:

x1	x2	y	hw(x)
1	1	0	1

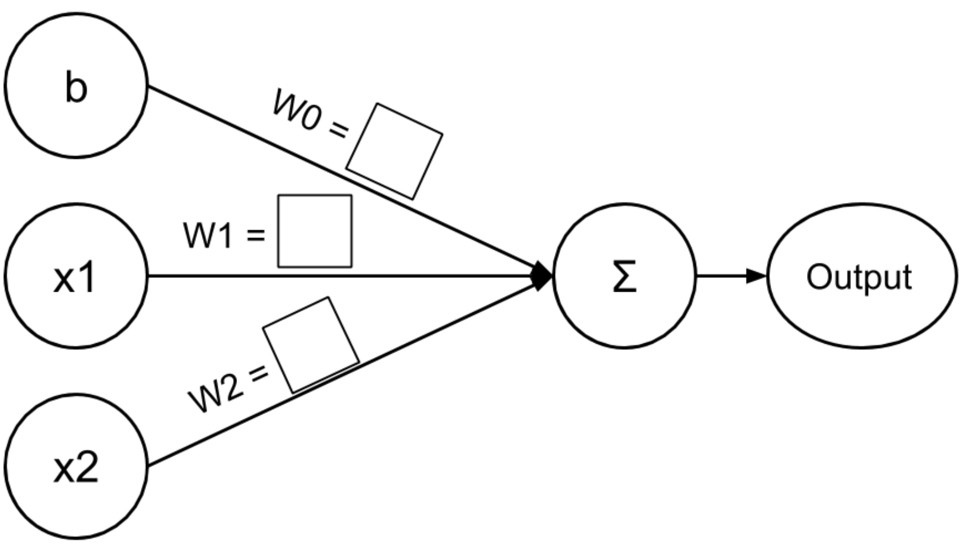
wi ← wi + α(y − hw(x))∗xi
1st iteration:
W0 = 1 + 0.5(0-1)*1 = 0.5
W1 = 1 + 0.5(0-1)*1 = 0.5
W2 = 1 + 0.5(0-1)*1 = 0.5
Output = 1

[1% for each correct value]

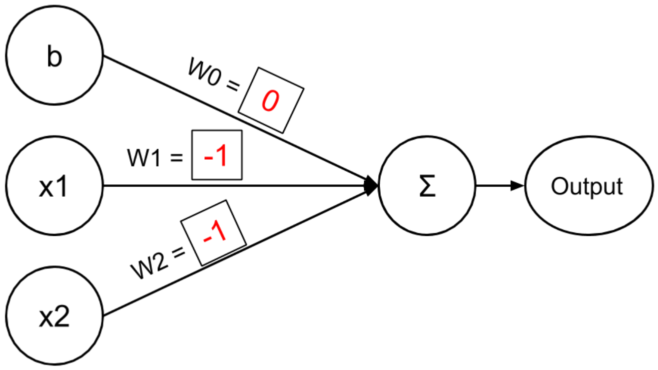
2a. [3%] NOR Gate



x1	x2	Out
0	0	1
0	1	0
1	0	0
1	1	0



- [No partial credits. There are more than one correct solutions, solutions are correct as long as it satisfies the truth table]



2b. [4%] Assuming the initial weights are all -1 and that the learning rate is 0.5, what would be the weights after training on the correct value for x1=0 and x2=0? Please show your work below only for one iteration.(Write the updated w1, w2, w0 and output of the perceptron for the updated weights)

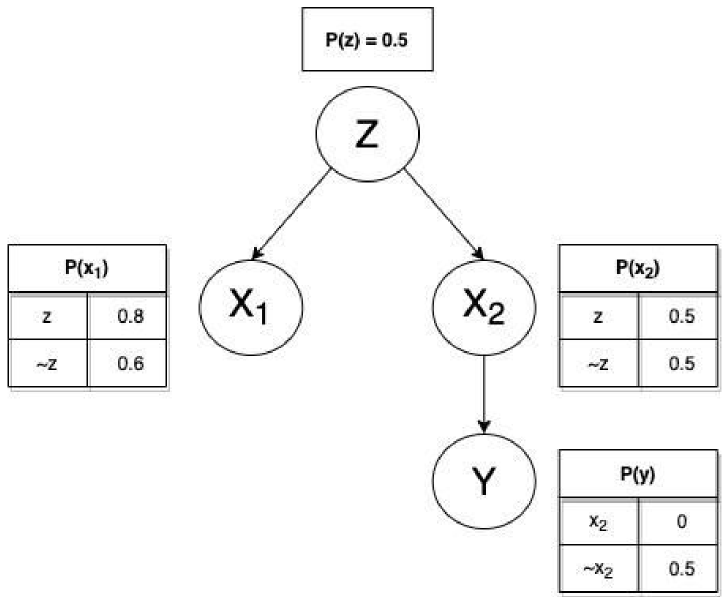
Solution:

x1	x2	y	hw(x)
0	0	1	0

wi ← wi + α(y − hw(x))∗xi
1st iteration:
W0 = -1 + 0.5(1-0)*1 = -0.5
W1 = -1 + 0.5(1-0)*0 = -1
W2 = -1 + 0.5(1-0)*0 = -1
Output = 0
[1% for each correct value]

[15%] Bayesian Networks

Consider the following Bayesian Network with binary random variables
$Z \in \{z, \tilde{}z\}$, $X_1 \in \{x_1, \tilde{}x_1\}$, $X_2 \in \{x_2, \tilde{}x_2\}$, $Y \in \{y, \tilde{}y\}$:



2.a. [3%] Let's look at solving the query $P(X_1|Y = y)$ **using variable elimination**. The hidden variables in this query are X_2 and Z. Using variable elimination, we can either eliminate them in the order X_2, Z , or in the order Z, X_2. Using which order is more computationally efficient? Justify your answer.

1% for correct answer. 2% for proper justification. Justifications based on a comparison of 1) the number of steps and 2) the size of the largest factor are both acceptable. A response like " should be expanded later because it has more children than " is NOT a valid justification.

a) $P(X_1|y)$.

$$= \alpha \sum_{z} P(X_1|z).P(z). \sum_{x_2} P(x_2|z).P(y|x_2).$$ [Eliminate x_2 first, then z]

No of steps = 2^2 (Join) + 2 (Eliminate) = 6.

No of steps = 2^2 (Join) + 2 (Eliminate) = 6. ∴ Total = 6+6 = 12.

$$= \alpha \sum_{x_2} P(y|x_2). \sum_{z} P(X_1|z). P(z). P(x_2|z)$$ [Eliminate z first, then x_2]

No of steps = 2^3 (Join) + 4 (Eliminate) = 12

No of steps = 2^2 (Join) + 2 (Eliminate) = 6. ∴ Total = 12 + 6 = 18

∴ Eliminating x_2 first and then z is more efficient.

2.b. [12%] Using the order that you said was more computationally efficient in 2.a., find the solution to the query $P(X_1|Y = y)$ **using variable elimination**. For full credit, clearly show all your work.

2% for correct expression. 4% for each Join + Eliminate step. 2% for correct normalization.

b) $P(X_1 | y)$.

$$= \alpha \sum_{z} P(X_1|z) \cdot P(z) \cdot \sum_{x_2} P(x_2|z) \cdot P(y|x_2)$$

$$f(z,y) = \sum_{x_2} P(x_2|z) \cdot P(y|x_2)$$

X_2	Z	$P(X_2\|Z)$
x_2	z	0.5
x_2	$\neg z$	0.5
$\neg x_2$	z	0.5
$\neg x_2$	$\neg z$	0.5

X_2	$P(y\|X_2)$
x_2	0
$\neg x_2$	0.5

$\xrightarrow{\text{Join}}$

X_2	Z	
x_2	z	0
x_2	$\neg z$	0
$\neg x_2$	z	0.25
$\neg x_2$	$\neg z$	0.25

$\xrightarrow{\text{Eliminate}}$

Z	$f(Z,y)$
z	0.25
$\neg z$	0.25

$$P(X_1, y) = \sum_{z} P(X_1|z) \cdot P(z) \cdot f(z,y)$$

X_1	Z	$P(X_1\|Z)$
x_1	z	0.8
x_1	$\neg z$	0.6
$\neg x_1$	z	0.2
$\neg x_1$	$\neg z$	0.4

Z	$P(Z)$
z	0.5
$\neg z$	0.5

Z	$f(Z,y)$
z	0.25
$\neg z$	0.25

$\xrightarrow{\text{Join}}$

X_1	Z	
x_1	z	0.1
x_1	$\neg z$	0.075
$\neg x_1$	z	0.025
$\neg x_1$	$\neg z$	0.05

$\xrightarrow{\text{Eliminate}}$

X_1	$P(X_1,y)$
x_1	0.175
$\neg x_1$	0.075

$$P(X_1|y) = \alpha \begin{bmatrix} 0.175 \\ 0.075 \end{bmatrix}, \text{ where } \alpha = \frac{1}{(0.175 + 0.075)} = \frac{1}{0.250}$$

$$= \begin{bmatrix} 175/250 \\ 75/250 \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix}$$

[18%] Probability Theory

Q1[9%] The Tübatulabal tribe is made up of only three mutually exclusive groups, called bands i.e ['Bankalachi', 'Pahkanapil', 'Palagewan'] (there are no other bands within the Tübatulabal tribe). The weather in the Tübatulabal tribe is either 'Sunny' or 'Rainy' (there are no other weather conditions). The mood of the people in the Tübatulabal tribe is either 'Happy' or 'Sad' (there are no other moods). Let's Consider 'band', 'mood' and 'weather' as the random variables.

Also, you are given following probabilities :-

P(weather = Sunny)
P(mood = Happy)
P(mood = Happy, 'band' = Bankalachi | weather = Sunny)
P(mood = Happy, 'band' = Pahkanapil | weather = Sunny)
P(mood = Happy, 'band' = Palagewan | weather = Sunny)

In reference to "Tübatulabal tribe" above, please answer the following (Please consider the probabilities given only in context of the above details for within Tübatulabal tribe):-

A. P(mood = Happy | weather = Sunny)
B. P(weather = Sunny | mood = Happy)
C. P('band' = Bankalachi) + P('band' = Pahkanapil) + P('band' = Palagewan)

1) P(mood = Happy | weather = Sunny) = P(mood = Happy, 'band' = Bankalachi | weather = Sunny) + P(mood = Happy, 'band' = Pahkanapil | weather = Sunny) + P(mood = Happy, 'band' = Palagewan | weather = Sunny)
2) P(weather = Sunny | mood = Happy) = P(mood = Happy | weather = Sunny) P(weather = Sunny) / P(mood = Happy)
3) P('band' = Bankalachi) + P('band' = Pahkanapil) + P('band' = Palagewan) = 1

3% Each, (no partial marking)

Q2[9%] Given the following joint distribution, answer the following questions:-

	fever		~fever	
	Runny nose	~Runny nose	Runny nose	~Runny nose
flu	0.108	0.012	0.072	0.008
~flu	0.016	0.064	0.144	0.576

Calculate :-

1) **P(~flu | fever)**
2) **P(fever or flu)**
3) **P(fever)**

1) P(~flu | fever) = P(~flu and fever) / P(fever) = 0.4
2) P(fever or flu) = 0.28
3) P(fever) = 0.2

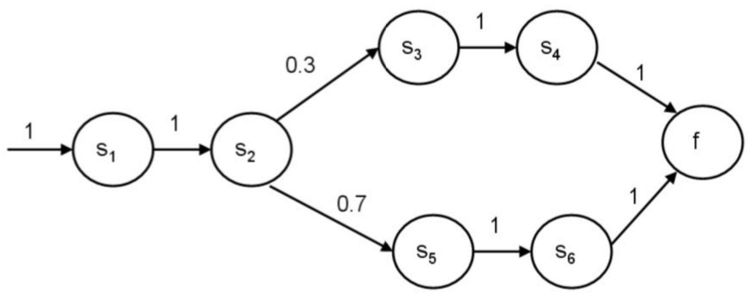
3% Each

[12%] HMM

Consider the HMM defined by the transition and emission probabilities in the table below. This HMM has six states (plus a start and end states) and an alphabet with four symbols (A,C, G and T). Thus, the probability of transitioning from state S_1 to state S_2 is 1, and the probability of emitting A while in state S_1 is 0.5.

	0	S_1	S_2	S_3	S_4	S_5	S_6	f	A	C	G	T
0	0	1	0	0	0	0	0	0				
S_1	0	0	1	0	0	0	0	0	0.5	0.3	0	0.2
S_2	0	0	0	0.3	0	0.7	0	0	0.1	0.1	0.2	0.6
S_3	0	0	0	0	1	0	0	0	0.2	0	0.1	0.7
S_4	0	0	0	0	0	0	0	1	0.1	0.3	0.4	0.2
S_5	0	0	0	0	0	0	1	0	0.1	0.3	0.3	0.3
S_6	0	0	0	0	0	0	0	1	0.2	0.3	0	0.5

Here is the state diagram:



For each of the pairs below, place "<" , ">" or "=" between the right and left components of each pair. Note: Only one of the 3 symbols are required for your answer. Hint: Think before directly calculating the values.

(a) P (O_1 =A, O_2 =C, O_3 =T, O_4 =A, q_1 =S_1, q_2 =S_2) ( ) P(O_1 =A, O_2 =C ,O_3 =T, O_4 =A | q_1 =S_1, q_2 =S_2)
(b) P (O_1 =A, O_2 =C, O_3 =T, O_4 =A, q_3 =S_3, q_4 =S_4 ( ) P(O_1 =A, O_2 =C , O_3 =T, O_4 =A | q_3 =S_3, q_4 =S_4)
(c) P (O_1 =A, O_2 =C, O_3 =T, O_4 =A, q_3 =S_3, q_4 =S_4) ( ) P(O_1 =A, O_2 =C , O_3 =T, O_4 =A , q_3 =S_5, q_4 =S_6)
(d) P (O_1 =A, O_2 =C, O_3 =T, O_4 =A) ( ) P(O_1 =A, O_2 =C, O_3 =T, O_4 =A , q_3 =S_3, q_4 =S_4)
(e) P(O_1 =A, O_2 =C, O_3 =T, O_4 =A) ( ) P(O_1 =A, O_2 =C, O_3 =T, O_4 =A| q_3 =S_3, q_4 =S_4)
(f) P(O_1 =A,O_2 =C,O_3 =T,O_4 =A) ( ) P(O_1 =A,O_2 =T,O_3 =T,O_4 =G)

Solutions:

2% per correct answer

a. =
P(A,C,T,A,S_1,S_2) = P(A,C,T,A|S_1,S_2)P(S_1,S_2) = P(A,C,T,A|S_1,S_2), since P(S_1,S_2) = 1

b. <
As in (b), P(A,C,T,A,S_3,S_4) = P(A,C,T,A|S_3,S_4)P(S_3,S_4) however, since P(S_3,S_4) = 0.3, then the right hand side is bigger.

c. $<$
   The first two emissions (A and C) do not matter since they are the same. Thus, the right hand side translates to$P(O_3 = T, O_4 = A, q_3 = S_3, q_4 = S_4) = P(O_3 = T, O_4 = A | q_3 = S_3, q_4 = S_4) P(S_3, S_4) = 0.7*0.1*0.3 = 0.021$ while the right hand side is $0.3 * 0.2 * 0.7 = 0.042$.

d. $>$

   Here the left hand side is: $P(A, C, T, A, S_3, S_4) + P(A, C, T, A, S_5, S_6)$. The right side of the summation is the right hand side above. Since the left side of the summation is greater than 0, the left hand side is greater.

e. $<$
   The left hand side is: $P(A,C,T,A,S_3,S_4)+P(A,C,T,A,S_5,S_6)=P(A,C,T,A|S_3,S_4)P(S_3,S_4)+P(A,C,T,A|S_5,S_6)P(S_5,S_6)$. Since $P(A,C,T,A|S_3,S_4) > P(A,C,T,A|S_5,S_6)$ the left hand side is lower from the right hand side.

f. $<$
   Since the first and third letters are the same, we only need to worry about the second and fourth. The left hand side is: $0.1*(0.3*0.1+0.7*0.2) = 0.017$ while the right hand side is: $0.6*(0.7*0+0.3*0.4) = 0.072$.

[10%] Naive Bayes

Suppose we are given the following data, where A, B, C ∈ { 0, 1 } are independent random variables, and y is a binary output whose value we want to predict.

How would a naive Bayes classifier predict y if the input is { A = 0, B = 0, C = 1 } ? Assume that in case of a tie the classifier always prefers to predict 0 for y.

A	B	C	Y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

Rubric:

$P(Y = 0 \mid A = 0, B = 0, C = 1) = P(A = 0, B = 0, C = 1 \mid Y = 0)\, P(Y = 0) / P(A = 0, B = 0, C = 1)$

$P(Y = 1 \mid A = 0, B = 0, C = 1) = P(A = 0, B = 0, C = 1\ Y = 1)\, P(Y = 1) / P(A = 0, B = 0, C = 1)$

$P(A = 0 \mid Y = 0)P(B = 0 \mid Y = 0)P(C = 1 \mid Y = 0)P(Y = 0) = 2/3 * 1/3 * 1/3 * 3/7 = 2/63$. (3 points)

For example: ⅔ : 0.5 point, ⅓: 0.5 point, ⅓: 0.5 point, 3/7: 0.5 point

finally 2/63: 1 point

$P(A = 0 \mid Y = 1)P(B = 0 \mid Y = 1)P(C = 1 \mid Y = 1)P(Y = 1) = 1/4 * 2/4 * 2/4 * 4/7 = 1/28$ (3 points)

Same as above:

1/4 : 0.5 point, 2/4: 0.5 point, 2/4: 0.5 point, 4/7: 0.5 point

finally 1/28: 1 point

$P(A = 0 \mid Y = 0)P(B = 0 \mid Y = 0)P(C = 1 \mid Y = 0)P(Y = 1) > P(A = 0 \mid Y = 1)P(B = 0 \mid Y = 1)P(C = 1 \mid Y = 1)P(Y = 1)$ (optional)

$\therefore Y = 1$. (4 points)