1. [10%] General AI Knowledge

1(T) In Backpropagation learning, weights on the arcs are modified based on the error values propagated from output nodes to input nodes.

2(T) In Decision Tree learning, the best splitting attribute is the one that gives the greatest information gain.

3(F) In an SVM defined by decision boundary $\mathbf{w}^T\mathbf{x} + b = 0$, the *support vectors* are those examples in the training set that lie *exactly* on this decision boundary.

4(F) A and B are independent if and only if P(A|B)=P(AB)/P(B)

5(T) Bayesian Networks are a compact way to represent a fully joint probability distribution using conditional independence that holds among the variables.

6(F) Value Iteration for MDP always finds the optimal policy, when run to convergence.

7(T) Policy iteration makes use of the current utility values of the states.

8(F) The number of parameters in a Bayesian network is exponential in the total number of arcs in the graph.

9(T) It is better to use Neural Networks than Decision Trees to learn the function majority(x1,…,xi,...,xn) where all xi are binary, and the output of the majority function is 1 if the majority of inputs are 1 and 0 otherwise.

10(T) The EM algorithm is a probability based learning framework that can find the best model M from data D that achieves the maximum P(M | D).

2. [14%] Decision Tree Learning

You are given the task of classifying whether you should drive fast or slow based on 3 attributes **[Road Type, Obstruction on the Road, Speed Limit]** as shown in the following table :

Index	Road Type	Obstruction	Speed Limit	Drive Speed
1	steep	yes	yes	***slow***
2	steep	no	yes	***slow***
3	flat	yes	no	***fast***
4	steep	no	no	***fast***
5	flat	yes	yes	***slow***
6	flat	no	no	***fast***

As you can observe - Road Type can be steep or flat, Obstruction can be yes or no (indicating whether a road has an obstruction or not) and Speed Limit can be yes or no (indicating if speed limit has been imposed on a road or not).

Answer the following questions based on the information provided above :

The following values are provided for your reference -

0 * log (0) = 0
1/6 * log (1/6) = - 0.13
2/6 * log (2/6) = - 0.16
3/6 * log (3/6) = -0.15

4/6 * log (4/6) = - 0.12
5/6 * log (5/6) = - 0.07
6/6 * log(6/6) = 0

[Note : If you have consistently used the same log base throughout the solution, no points have been deducted for using the bases of 10,2 or e. If you have used log values with different bases i.e inconsistently, some points have been deducted while considering the max out of all cases.]

[2%][A] Calculate the initial entropy of the decision [Drive Speed]

Entropy (Drive Speed) = [-0.5 * log(0.5) - 0.5 log(0.5)] =
[log base 10 = 0.30, log base 2 = 1, log base 3 = 0.693]
Rubric : 1% for correct answer and 1% approach

[2%][B] Calculate the entropy remaining (remainder) if Road Type is chosen as the splitting attribute.

Entropy (Road Type) = ½ * [- ⅔ * log(⅔) - ⅓ * log(⅓)] + ½ * [- ⅓ * log (⅓) - ⅔ * log(⅔)] =

1. **log base 10 = 0.5 * (0.12 + 0.16) + 0.5 * (0.16 + 0.12) = 0.28**
2. **log base 2 = 0.5 * (0.53 + 0.39) + 0.5 * (0.39 + 0.53) = 0.92**
3. **log base e = 0.5 * (0.37 + 0.27) + 0.5 * (0.37 + 0.27) = 0.64**

Rubric : 1% for correct answer and 1% for correct approach

[2%][C] Calculate the entropy remaining (remainder) if Obstruction is chosen as the splitting attribute.

Entropy (Obstruction) = ½ * [- ⅓ * log (⅓) - ⅔ * log(⅔)] + ½ * [- ⅔ * log(⅔) - ⅓ * log(⅓)] =

1. **log base 10 = 0.5 * (0.12 + 0.16) + 0.5 * (0.16 + 0.12) = 0.28**
2. **log base 2 = 0.5 * (0.53 + 0.39) + 0.5 * (0.39 + 0.53) = 0.92**
3. **log base e = 0.5 * (0.37 + 0.27) + 0.5 * (0.37 + 0.27) = 0.64**

Rubric : 1% for correct answer and 1% for correct approach

[2%][D] Calculate the entropy remaining (remainder) if Speed Limit is chosen as the splitting attribute.

Entropy (Speed Limit) =
½ * [- 0 * log (0/3) - 3/3 * log(3/3)] + ½ * [- 3/3 * log(3/3) - 0/3 * log(0/3)] =
[for all log bases = 0]
Rubric : 1% for correct answer and 1% for correct approach

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

[0.5% for using correct approach i.e Entropy - Remainder(Attribute) and 0.5% for using correct values and correct answer] [Each]

InfoGain (Road Type) =

log base 10 : 0.3 - 0.28 = 0.02 [1%]

log base 2 : 1 - 0.92 = 0.08 [1%]

log base e : 0.693 - 0.64 = 0.053 [1%]

InfoGain (Obstruction) =

log base 10 : 0.3 - 0.28 = 0.02 [1%]

log base 2 : 1 - 0.92 = 0.08 [1%]

log base e : 0.693 - 0.64 = 0.053 [1%]

InfoGain (Speed Limit) =

log base 10 : 0.3 - 0 = 0.3 [1%]

log base 2 : 1 - 0 = 1 [1%]

log base e : 0.693 - 0 = 0.693 [1%]

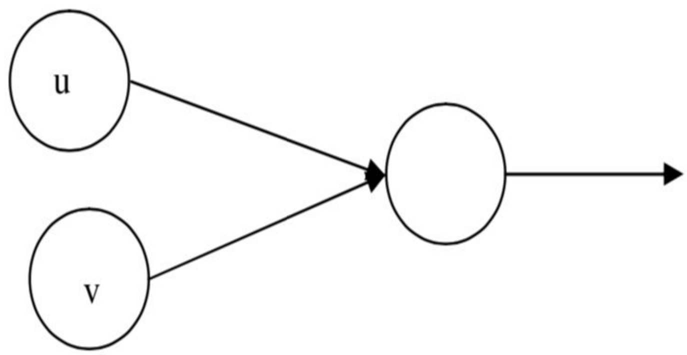
[3%][F] Which attribute should be chosen to split the dataset and why?

[Note : In case of a tie, resolve by taking attribute names in alphabetical order]

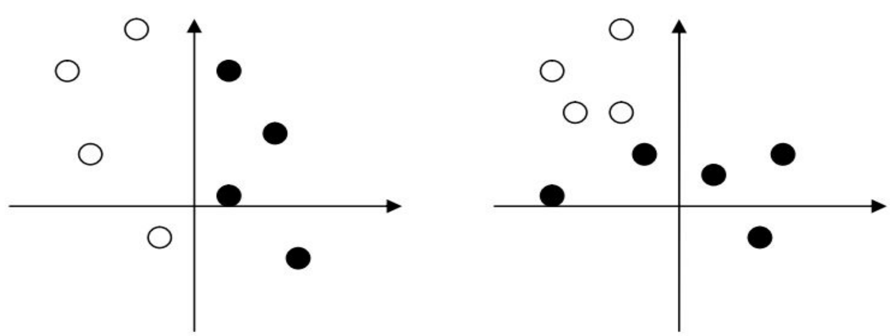
Speed Limit.[1.5%] Reason : It has maximum information gain [1.5%]

3. [16%] Neural Networks

Consider a problem in which each data point has two coordinates u and v. We wish to learn a classifier for this problem by using a linear perceptron network with inputs u and v and weights Wu and Wv on the two connections. We use a threshold of 0 so that the output of the network is +1 (class 1) if the output unit is greater than or equal to 0 and -1 (class 2) otherwise.



[4%][A] Can the network distinguish the two classes in the cases illustrated below? Why/Why not in each case?



Case I

Case II

RUBRIC:

Case1: Yes [-1% if wrong]

Case2: No [-1% if wrong]

Reason: Without a constant term, the decision boundary must go through the origin. [-2% if wrong]

[4%][B] Assume that we augment the network with an additional input unit with constant value 1 (we'll call the weight on the additional connection W). How does the answer to (a) change and why?

RUBRIC:

Case1: Yes [-1% if wrong]

Case2: Yes [-1% if wrong]

Reason: With a constant term the decision boundary can go anywhere though it must still be linear [-2% if wrong]

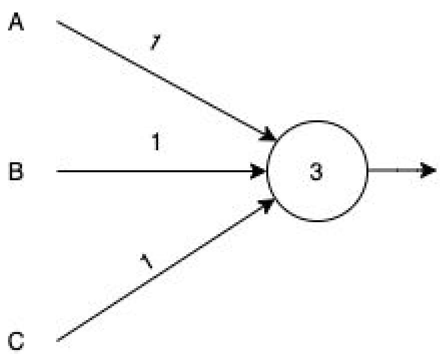
You are building a set of perceptrons to mimic logic circuits. For each perceptron, assume its firing is based on the threshold rule i.e

$$f(x) = \begin{cases} 1 & \text{if } x \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

Assume all inputs have value either 0 or 1. Show how perceptrons can be used to mimic the following logic sentences.

You are free to choose the threshold.

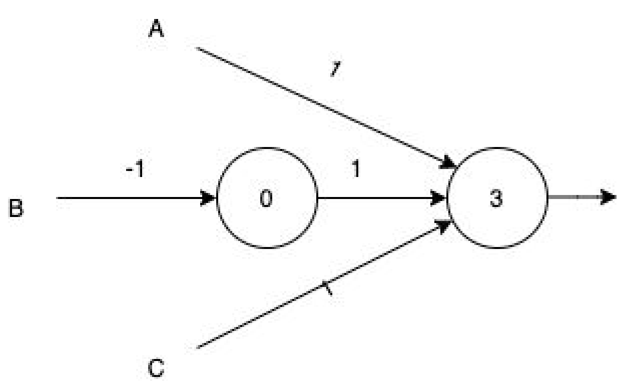
[4%] [C] A ∧ B ∧ C



RUBRIC: Any other valid network which follows the truth table also gets full credits.

[Full or zero credits]

[4%] [D] A ∧ ¬B ∧ C



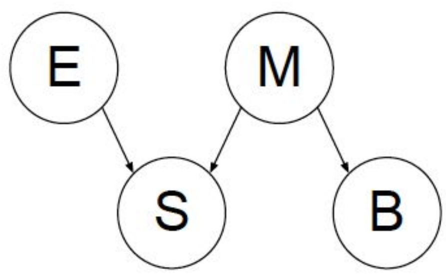
RUBRIC: Any other valid network which follows the truth table also gets full credits.
[Full or zero credits]

4. [10%] Bayes Nets

A dusty sky (S) can be caused either by elephants migrating (E), or because of a wind storm that always happens in the month of May (M). The May wind storm also causes buses to take twice as long on their routes (B). The Bayesian network and corresponding conditional probability tables for this situation are shown below. For each part, you should give the arithmetic expression (in terms of numbers from the tables below, or from the numerical answer of other questions) and a numerical answer (e.g. 0.81). Be sure to show your work.

$P(E)$	
$+e$	0.4
$-e$	0.6

$P(S \mid E, M)$			
$+e$	$+m$	$+s$	1.0
$+e$	$+m$	$-s$	0.0
$+e$	$-m$	$+s$	0.8
$+e$	$-m$	$-s$	0.2
$-e$	$+m$	$+s$	0.3
$-e$	$+m$	$-s$	0.7
$-e$	$-m$	$+s$	0.1
$-e$	$-m$	$-s$	0.9



$P(M)$	
$+m$	0.1
$-m$	0.9

$P(B \mid M)$		
$+m$	$+b$	1.0
$+m$	$-b$	0.0
$-m$	$+b$	0.1
$-m$	$-b$	0.9

RUBRIC: full credit if correct arithmetic expression, -1% for each error (NOTE: no response is required to give explanation as seen in answer)

[2%] [A] Compute the following entry from the joint distribution:

$$P(-e, -s, -m, -b) = P(-e)P(-m)P(-s \mid -e, -m)P(-b \mid -m) = (0.6)(0.9)(0.9)(0.9) = 0.4374$$

by expanding the joint according to the chain rule of conditional probability.

RUBRIC: full credit if correct arithmetic expression, -1% for each error

[2%][B] What is the probability that the buses take twice as long on their routes?

$$P(+b) = P(+b \mid +m)P(+m) + P(+b \mid -m)P(-m) = (1.0)(0.1) + (0.1)(0.9) = 0.19$$

by marginalizing out m according to the law of total probability.

RUBRIC: full credit if correct arithmetic expression, -1% for each error

[2%] [C] What is the probability that the May wind storm is occurring, given that the buses are taking twice as long on their routes?

$P(+m|+b) = \frac{P(+b|+m)P(+m)}{P(+b)} = \frac{(1.0)(0.1)}{0.19} \approx .5263$
by the definition of conditional probability.

RUBRIC: full credit if correct arithmetic expression, -1% for each error

[2%] [D] What is the probability that the May wind storm is occurring, given that there is a dusty sky, buses are taking twice as long on their routes, and there are elephants migrating?

$$P(+m|+s,+b,+e) =$$

$$\frac{P(+m,+s,+b,+e)}{\sum_m P(m,+s,+b,+e)} = \frac{P(+e)P(+m)P(+s|+e,+m)P(+b|+m))}{\sum_m P(+e)P(m)P(+s|+e,m)P(+b|m)}$$
$$= \frac{(0.4)(0.1)(1.0)(1.0)}{(0.4)(0.1)(1.0)(1.0)+(0.4)(0.9)(0.8)(0.1)}$$
$$= \frac{0.04}{0.04+0.0288} \approx .5814$$

RUBRIC: full credit if correct arithmetic expression, -1% for each error

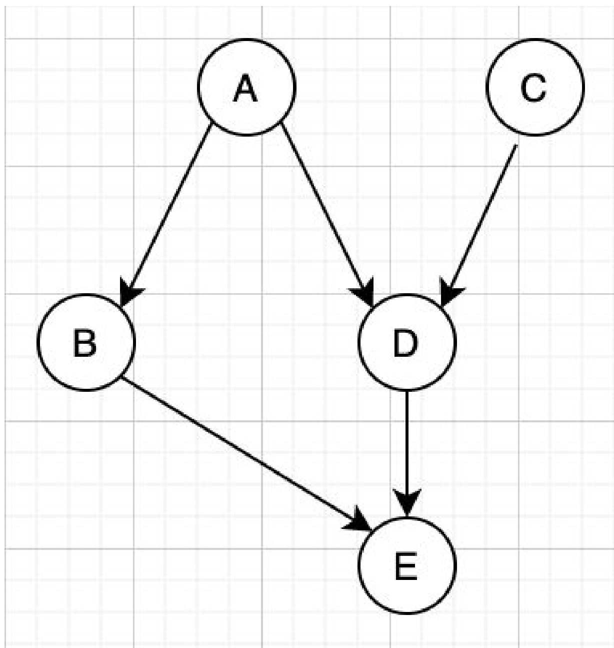
[2%] [E] What is the probability that elephants migrating are present, given that the May dust storm is occurring?

$P(+e|+m) = P(+e) = 0.4$
The first equality holds true as we have $E \perp\!\!\!\perp M$ (E is independent of M), which can be inferred from the graph of the Bayes' net.

RUBRIC: full credit if correct arithmetic expression, -1% for each error

5. [10%] Enumeration

Consider the following Bayesian Network containing 5 Boolean random variables:



[4%][A] Enumerate P(C | ¬E, B, A) using all the random variables given in the above Bayesian Network [You will have to represent this in terms of an expression of joint probabilities with each probability containing all the present variables].

RUBRIC: -1% for each error

{ P(C, ¬E, B, A, D) + P(C, ¬E, B, A, ¬D) } ÷ { P(C, ¬E, B, A, D) + P(C, ¬E, B, A, ¬D) + P(¬C, ¬E, B, A, D) + P(¬C, ¬E, B, A, ¬D) }

[4%][B] Represent each of the joint probabilities you got in the above expression as expressions of conditional probabilities, using only the information represented by the above Bayesian Network [to obtain full points, each expression for every joint probability above must be expressed using only the terms in the CPT associated with the above network. You can assume that the CPT, representing conditional dependencies, for the above network is provided].

RUBRIC: -1% for each error

P(C, ¬E, B, A, D) = P(A) . P(C) . P(D|A,C) . P(B|A) . P(¬E|D)

$P(C, \neg E, B, A, \neg D) = P(A) . P(C) . P(\neg D|A,C) . P(B|A) . P(\neg E|\neg D)$

$P(\neg C, \neg E, B, A, D) = P(A) . P(\neg C) . P(D|A,\neg C) . P(B|A) . P(\neg E|D)$

$P(\neg C, \neg E, B, A, \neg D) = P(A) . P(\neg C) . P(\neg D|A,\neg C) . P(B|A) . P(\neg E|\neg D)$

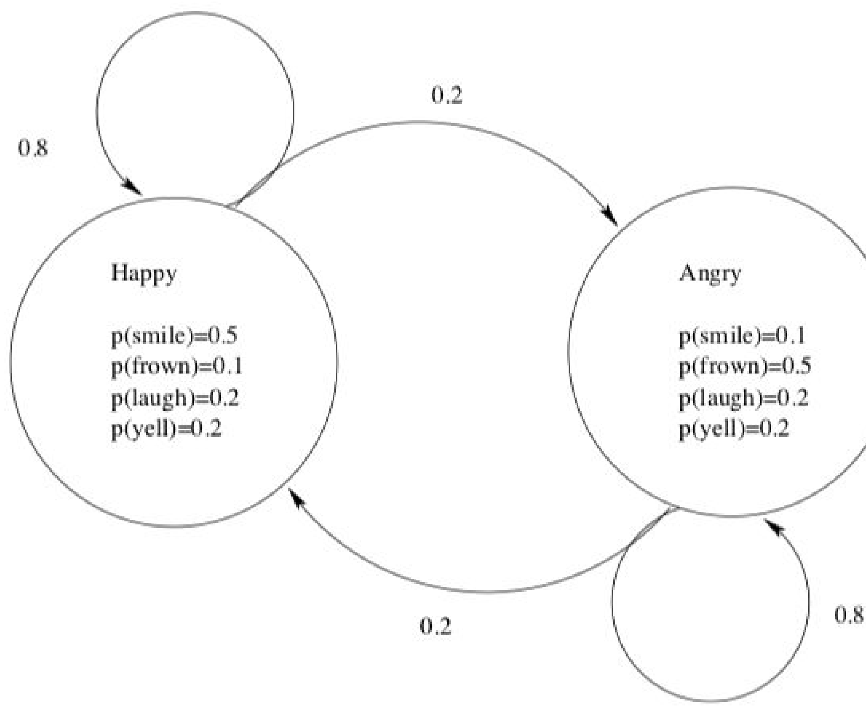
[2%][C] Using only the probability P(A,C|B) and P(~A,C|B), infer probability P(C | B).

RUBRIC: -1% for each error

$P(C | B) = P(A, C | B) + P(\neg A, C | B)$

6. [10%] HMM

Andrew Viterbi lives a simple life. Some days he's Angry and some days he's Happy. But he hides his emotional state, and so all you can observe is whether he smiles, frowns, laughs, or yells. We start on day 1 in the Happy state, and there's one transition per day.



Definitions:

q_t = state on day t.

O_t = observation on day t.

[1%] [A] What is $P(q_2 = \text{Happy})$?
RUBRIC: no credit if incorrect

Answer: 0.8

[2%][B] What is $P(O_2=\text{frown})$?
RUBRIC: full credit if correct arithmetic expression, -1% for each error

Answer: $(8/10 * 1/10) + (2/10 * 1/2)$ = 18/100

[2%][C] What is $P(q_2 = \text{Happy}|O_2 = \text{frown})$?

RUBRIC: full credit if correct arithmetic expression, -1% for each error

Answer: $P(O_2 = \text{frown} \mid q_2 = \text{Happy}) * P(q_2 = \text{Happy}) / P(O_2 = \text{frown})$

$= (8/10 * 1/10) / (18/100) = 4/9$

[2%][D] What is $P(O_{100} = \text{yell})$?

RUBRIC: full credit if correct arithmetic expression, -1% for each error

Answer: $= P(O_{100} = \text{yell} \mid q_{100} = \text{Happy})*P(q_{100} = \text{Happy}) + P(O_{100} = \text{yell} \mid q_{100} = \text{Angry})*P(q_{100} = \text{Angry})$

$= 2/10 * (P(q_{100} = \text{Happy}) + P(q_{100} = \text{Angry})) = 2/10*1 = 2/10$

[3%][E] Assume that O_1 = frown, O_2 = frown, O_3 = frown, O_4 = frown, and O_5 = frown. What is the most likely sequence of states?

Answer: HAAAA

RUBRIC: -1% for each error

7. [10%] MDP

The gridworld MDP shown below operates like the one we saw in class.

The states are grid squares, identified by their row and column number (row first). The agent always starts in state (1,1), marked with the letter S. There are two terminal goal states, (2,4) with reward +10 and (2,3) with reward -6. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.)

The transition function is such that the intended agent movement (North, South, West, or East) happens with probability **.8**. With probability **.1** each, the agent ends up in one of the states perpendicular to the intended direction. If a collision with a wall happens, the agent stays in the same state.

↓	←	**-6**	**+10**
S →	→	→	↑

[4%][A] Draw the optimal policy for this grid, using the grid world above.
RUBRIC: -1% for each error

[6%][B] Suppose the agent knows the transition probabilities. Give the first 3 rounds of value iteration updates for each state, with a discount of 0.8. (Assume V0 is 0 everywhere and compute Vi for times i = 1, 2, 3). **RUBRIC: -1% for each error**

State	(1,1)	(1,2)	(1,3)	(1,4)	(2,1)	(2,2)	(2,3)	(2,4)
V_0	0	0	0	0	0	0	0	0
V_1	0	0	0	**8**	0	0	0	0
V_2	0	0	**4.52**	**8**	0	0	0	0
V_3	0	**3.616**	**4.52**	**8**	0	0	0	0

8. [10%] Naïve Bayes

When there are clouds in the sky, there is a 15% chance that it will rain. If there are no clouds in the sky, there is only a 3% chance that it will rain. There are clouds in the sky on 2 out of 5 days. If it rained today, what is the probability that there were clouds in the sky?

[2%][A] Write the formula used in Bayes' theorem for this problem

RUBRIC: -1% for each error

$$P(clouds \mid rain) = \frac{P(rain\ AND\ clouds)}{P(rain)} = \frac{P(rain \mid clouds)*P(clouds)}{P(rain \mid clouds)*P(clouds)+P(rain \mid \neg clouds)*P(\neg clouds)}$$

[3%][B] Use Bayes' theorem to calculate the probability that there were clouds in the sky

RUBRIC: Full credit for correct arithmetic expression, -1% for each error

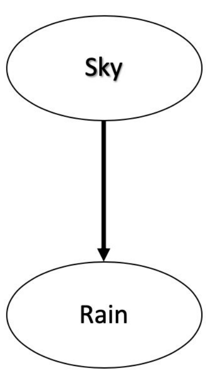
$$P(clouds \mid rain) = \frac{0.15*0.4}{(0.15*0.4)+(0.03*0.6)} = \frac{6}{6+1.8} = \frac{6}{7.8} = \frac{30}{39}$$

or

$$P(clouds \mid rain) = \frac{0.15*0.4}{(0.15*0.4)+(0.03*0.6)}$$

[5%][C] Model the situation as a Bayesian network with 2 nodes, and give the conditional probability tables for both nodes.

RUBRIC: -1% for each error



	clouds	no clouds
	0.4	0.6

Sky	rain	no rain
clouds	0.15	0.85
No clouds	0.03	0.97

9. [10%] Multiple Choice related to Discussions

Each question has zero or more correct choices. Circle the letters (a., b., c., etc) of all correct choices. Partial credit: beware that you lose 1% for each wrong answer, up to losing 2% for each question (so: all correct = 2%; 1 mistake = 1%, 2+ mistakes = 0%).

Note: Both not selecting correct answers as well as selecting incorrect answers are considered as 'mistakes'.

1. In the discussion, we discussed probability and reasoning with uncertainty, A and B are independent if P(AB)=P(A)P(B), and they are mutually exclusive if P(AB)=0. Please circle all that are true:

+a. If A and B are mutually exclusive, then they are not independent

b. If A and B are independent, then they are mutually exclusive

c. Mutually exclusiveness is equivalent to probability independence

+d. Mutually exclusiveness is not the same as probability independence

e. None of the above

2. In the discussion, we discussed the Bayesian Network representation and reasoning. Please circle all that are true:

a. A Bayesian Network is not semantically equivalent to a fully joint probability distribution table

+b. A Bayesian Network makes use of probability independence information

+c. The size of a Bayesian Network is usually smaller than its corresponding fully joint probability distribution table

d. Given a fully joint probability distribution table, we cannot do all the probabilistic reasonings without any Bayesian Networks

+e. Given a fully defined Bayesian Network, we can do all the probabilistic reasonings of the random variables in the Network

3. In the discussion, we discussed the artificial Neural Networks (NN). Please circle all that are true:

+a. The learning occurred in a NN are the adjustments of the weights inside NN and nothing else

+b. Deep Learning cannot automatically change the topological structure of a NN

+c. Neural Networks are mostly successful for both supervised and unsupervised learning

d. After the trainings are completed, a deployed NN can continue learn by itself

e. Deep Learning cannot be used to automatically learn useful features from the raw data for classification

4. In the discussion, we discussed Maximal Expected Utility (MEU) and Rationality. Please circle all that are true:

+a. Expected Utilities can be used by an agent to select actions intelligently

+b. If an agent follows the principle of MEU, then it behaves rationally

c. Expected Utilities have nothing to do with the probabilistic outcomes of the actions

+d. In today's AI research, Rationality is fully defined by MEU

e. Rationality can be defined without knowing the utilities of the states

5. In the discussion, we discussed the Bayesian Classifier. Please circle all that are true:

a. Bayesian Classifier can only be used when random variables are mutually independent

+b. Naive Bayesian Classifier often assumes independence as much as it can

+c. Bayesian Classifier makes use of probability theory

+d. Bayesian Classifier can deal with uncertainties and noises in the training data

e. The Bayesian Rule cannot be derived from the basic Product Rule