Final Exam

CSCI 561 Fall 2022: Foundation of Artificial Intelligence

Instructions:

1. Maximum credits/points for this midterm: 100 points.
2. No books (or any other material) are allowed.
3. All the questions in this exam **are going to be auto-graded.** This means that you should **exactly follow the instructions** in entering your results.
4. You are allowed to use a calculator.
5. Some questions have hints. Be sure to check them before solving the problem.
6. Adhere to the Academic Integrity Code.
7. Please make sure that you write the answers in the format discussed.

Problems	100 Percent Total
1 – General AI Knowledge	18%
2 – Decision Trees	12%
3 – Neural Networks	15%
4 – Bayesian Networks	15%
5 – Probability Theory	15%
6 – HMM, Temporal Model	15%
7 – Naive Bayes	10%

1. True/False [18%]

For each of the statements below, fill in the bubble **T** if the statement is always and unconditionally true, or fill in the bubble **F** if it is always false, sometimes false, or just does not make sense:

1. Both deductive and inductive learning agents learn new rules/facts from a dataset.
2. Learning is useful as a system construction method, because we only need to expose the agents to reality without any manual input.
3. In the ID3 algorithm, we need to choose the attribute that has the largest expected information gain.
4. Both perceptron and decision tree learning can learn majority function (output 1 if and only if more than half of n binary variables are 1) easily. It is representable within a perceptron and only needs a few branches in DTL (Decision Tree Learning).
5. The process of learning of a neural network happens in both the feed-forward (prediction) part and back-propagation part.
6. The basic principles of deep learning are similar to those of basic neural networks, but deep learning has newer methods for larger datasets.
7. Probabilities of propositions may change with new evidence.
8. A complete probability model specifies every entry in the joint distribution for all the variables.
9. When calculating a probability distribution, normalization will be needed in the end to make the distribution sum to 1. However, even if you use the inference rules properly, the normalization may not be preserved.
10. Probabilistically speaking, two coin tosses are conditionally independent.
11. The reward for a probabilistic decision making model can be given from states R(s), stats-action R(s, a), or transition R(s, a, s').
12. The principle of a MEU (Maximum Expected Utility) is that a rational agent should always choose the action that maximizes the utility.
13. The major difference between a POMDP (Partially Observed Markov Decision Process) and a general MDP (Markov Decision Process) is merely a sensor model P(e|s).
14. States transit randomly for Markov Chains and Hidden Markov Models.
15. In HMM (Hidden Markov Models), there are two important independence properties. The first is that the future depends only on the present; the second is that observations are independent of each other.
16. Forward procedure computes all $\alpha_t(s_i)$ on state trellis, while the Viterbi algorithm only computes the best for each step.

17. Discrete valued dynamic Bayes nets are not HMMs (Hidden Markov Models).
18. For Bayesian learning, we are given a set of new data D, background knowledge X, and are supposed to predict a concept C where P(C|DX) is the most probable.

2. Decision Trees [12%]

Lyft wants to analyze if a student at USC gets a lyft depending on if it is raining around the university, if the destination is near or far and whether or not the ride was free. They have provided the training data below and they need your help to train a machine to decide whether a student gets a lyft.

Note:

(for calculations, always take digits up to 3 decimal places and drop the rest without rounding. Eg. 0.9737 becomes 0.973)

For all the following questions, use log base 2
(Use Table 2.1 to answer Q1-3)

(Table 2.1)

#	Rain	Free?	Near?	Takes Lyft
1	Yes	No	Yes	Yes
2	No	No	Yes	No
3	Yes	No	No	Yes
4	Yes	No	No	Yes
5	No	Yes	Yes	Yes
6	Yes	Yes	Yes	Yes
7	No	Yes	Yes	No
8	Yes	Yes	No	Yes
9	No	Yes	Yes	No
10	Yes	Yes	Yes	Yes

Q1. Calculate the information conveyed by the distribution of the Takes Lyft column to 3 decimal places [2%]

1. 0.879
2. 0.933
3. 1
4. 0.325

Q2. Which would be the best attribute to split on? (Assume this attribute to be X for further questions) [4%]

A. Rain
B. Free
C. Near

Q3. What is the value of Remainder(Free) ? (Ans up to 3 decimal places) [2%]

a. 0.423
b. 0.634
c. 0.875
d. 0

Q4. Assume that the Entropy and Remainder values of the given training data is:
(Use Table 2.2 to answer questions Q4 a. and Q4 b.)
Entropy = 0.910
Remainder(X) = 0.230
Remainder(Y) = 0.510
Remainer(Z) = 0.810

(Table 2.2)

Sr No	X	Y	Z	is Correct?
1	TRUE	TRUE	FALSE	Yes
2	FALSE	TRUE	TRUE	No
3	FALSE	TRUE	TRUE	No
4	TRUE	FALSE	TRUE	Yes
5	TRUE	FALSE	TRUE	Yes
6	FALSE	FALSE	TRUE	No
7	TRUE	TRUE	TRUE	Yes
8	FALSE	FALSE	FALSE	No
9	TRUE	TRUE	FALSE	Yes

a. Which would be the worst attribute to split on for the given data? [2%]

1. X
2. Y
3. Z

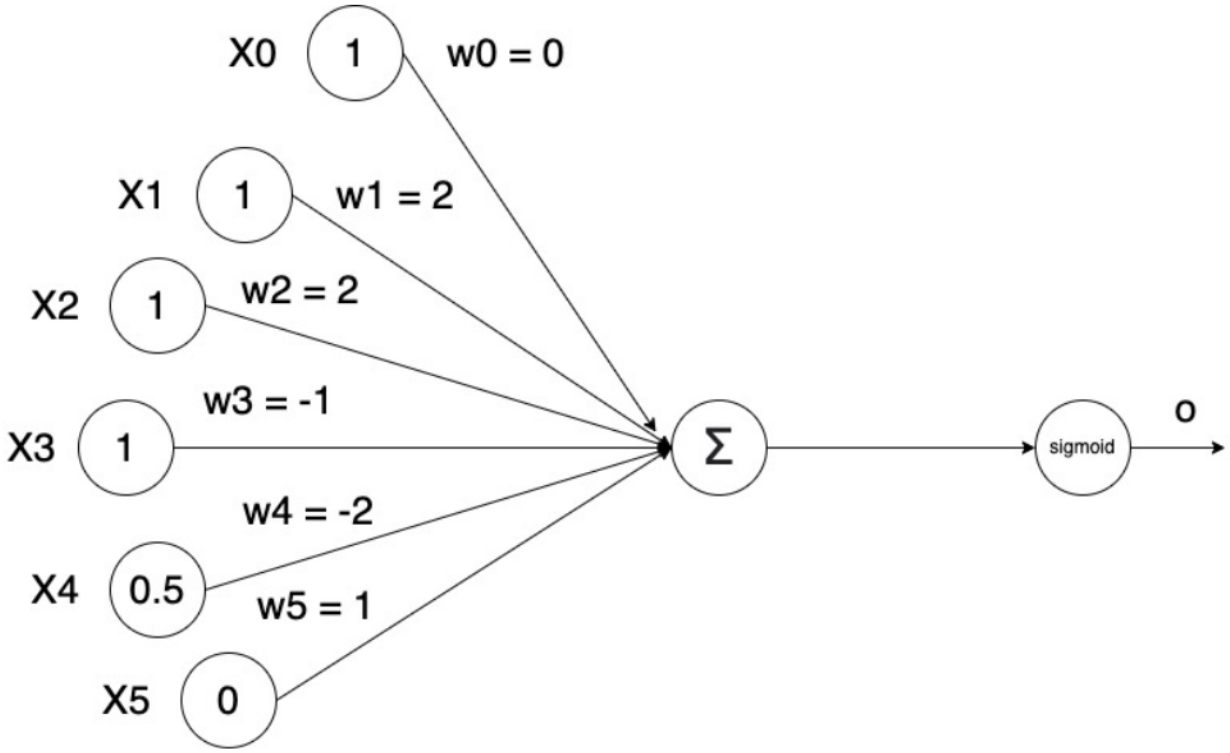
b. What output (IS Correct) would the machine give after being trained (assuming the calculations are correct for the root node) for the test data where X=False, Y= False and Z=True. [Hint: The decision tree learned uses only one attribute]

[2%]

1. Yes
2. No

3. Neural Networks [15%]

For the following perceptron model,



Loss function, $P = -(d-o)^2/2$
Learning rate, $\alpha = 100$
Activation function = sigmoid(x) = $1/(1+e^{-x})$
Expected output d = 1

1. For the given inputs X0 = 1, X1 = 1, X2 = 1, X3 = 1, X4 = 0.5, X5 = 0 if the value of the predicted output "o" after forward propagation is 0.9, calculate the updated weights after running backpropagation:

Note: (for calculations, always take digits up to 3 decimal places and drop the rest without rounding. Eg. 0.9737 becomes 0.973)

[2% each]

Original Solution:

1. Minimizing with 0.9

Original Weight (w_i)	Rate (alpha)	Input (x_i)	alpha * dP/dw = alpha * (d-o)*o*(1-o)*x_i	New Weight

2. Maximizing with 0.9

Original Weight (w_i)	Rate (alpha)	Input (x_i)	alpha * dP/dw = alpha * (d-o)*o*(1-o)*x_i	New Weight

3. Minimizing with 0.88

Original Weight (w_i)	Rate (alpha)	Input (x_i)	alpha * dP/dw = alpha * (d-o)*o*(1-o)*x_i	New Weight

4. Maximizing with 0.88

Original Weight (w_i)	Rate (alpha)	Input (x_i)	alpha * dP/dw = alpha * (d-o)*o*(1-o)*x_i	New Weight

2. [3%] Calculate the predicted output for the updated weights using the same input values:

1. Using (1)

2. Using (2)

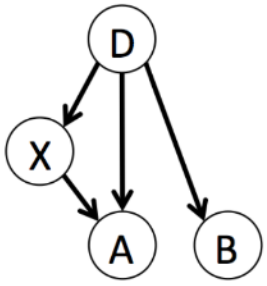
3. Using (3)

4. Using (4)

4. Bayesian Networks [15%]

Round off your final answer to three decimal places. Ex 0.36899 should be chopped as 0.368, 0.1 should be written as as 0.100.

$P(A\|D,X)$			
$+d$	$+x$	$+a$	0.9
$+d$	$+x$	$-a$	0.1
$+d$	$-x$	$+a$	0.8
$+d$	$-x$	$-a$	0.2
$-d$	$+x$	$+a$	0.6
$-d$	$+x$	$-a$	0.4
$-d$	$-x$	$+a$	0.1
$-d$	$-x$	$-a$	0.9



$P(D)$	
$+d$	0.1
$-d$	0.9

$P(X\|D)$		
$+d$	$+x$	0.7
$+d$	$-x$	0.3
$-d$	$+x$	0.8
$-d$	$-x$	0.2

$P(B\|D)$		
$+d$	$+b$	0.7
$+d$	$-b$	0.3
$-d$	$+b$	0.5
$-d$	$-b$	0.5

(a) [4 pts] What is the probability of having disease D and getting a positive result on test A?

P(+d, +a) = $\sum_x$P(+d,x,+a) = $\sum_x$ P(+a|+d,x)P(x|+d)P(+d) = P(+d)$\sum_x$ P(+a| + d,x)P(x| + d) =

(0.1)((0.9)(0.7) + (0.8)(0.3)) = 0.087

(b) [4 pts] What is the probability of not having disease D and getting a positive result on test A?

P(−d,+a) = $\sum_x$ P(−d,x,+a) = $\sum_x$P(+a| − d,x)P(x| − d)P(−d) = P(−d)$\sum_x$ P(+a| − d,x)P(x| − d) =

(0.9)((0.6)(0.8) + (0.1)(0.2)) = 0.450

(c) [3 pts] What is the probability of having disease D given a positive result on test A?

P(+d| + a) = P(+a,+d) / P(+a) = P(+a,+d)/ $\sum_d$P(+a,d) = 0.087/(0.087+0.45) ≈ 0.162

(d) [4 pts] What is the probability of having disease D given a positive result on test B?

P (+d| + b) = P (+b|+d)P (+d)/P(+b) = P (+b|+d)P (+d)/ $\sum_{d}$P(+b|d)P(d)=

(0.7)(0.1)/((0.7)(0.1)+(0.5)(0.9))=0.1346 =0.134

5. Probability theory[15%]

Ash enters a reality game where 30 different types of Pokémon are present, he sets his journey to chase down powerful ones. Being a Pokémon capturing expert he can capture every Pokémon possible. Each one that he finds and battles against has an equally likely chance of being one of the 30 types.

1. **(5%)** If Ash captures 5 Pokémon, what is the probability that he has captured at least 2 of the same type.

Solution: 0.296

$$Probability\ of\ all\ 5\ different\ types = \frac{30}{30} \times \frac{29}{30} \times \frac{28}{30} \times \frac{27}{30} \times \frac{26}{30} = 0.7037$$

$$Probability\ of\ at\ least\ 2\ of\ the\ same\ type = 1 - (0.7037) = 0.296$$

2. **(5%)** Assuming there is no limit on the number of Pokémon that Ash can capture, how many would he have to take down to make the probability of at least 2 of the same type more likely than not. (Answer as a whole number without decimal point precision)

Solution: 7

Since we know that

$$P(at\ least\ 2\ of\ the\ same\ type) = 1 - P(all\ k\ different\ types)$$

$$P(all\ k\ different\ types) = \frac{30-0}{30} \times \frac{30-1}{30} \times \frac{30-2}{30} \times \frac{30-i}{30} \ldots\ where\ 0 \leq i < k$$

We need to find the smallest k where $P(all\ k\ different\ types) < 0.5$ so that the first expression becomes > 0.5. We can find this iteratively,

$$Probability\ of\ all\ 5\ different\ types = \frac{30}{30} \times \frac{29}{30} \times \frac{28}{30} \times \frac{27}{30} \times \frac{26}{30} = 0.70$$

$$Probability\ of\ all\ 6\ different\ types = \frac{30}{30} \times \frac{29}{30} \times \frac{28}{30} \times \frac{27}{30} \times \frac{26}{30} \times \frac{25}{30} = 0.59$$

$$Probability\ of\ all\ 7\ different\ types = \frac{30}{30} \times \frac{29}{30} \times \frac{28}{30} \times \frac{27}{30} \times \frac{26}{30} \times \frac{25}{30} \times \frac{24}{30} = 0.47$$

Smallest k that makes $P(all\ k\ different\ types) < 0.5$ is 7.

3. **(5%)** Ash loses a battle after 2 powerful Pokémon types Darkrai and Scizor are added to the game. Probability that the Pokémon that he lost to being Darkrai is 70% and Scizor is 30%. Darkrai has a special attack which it uses with a probability of 0.90. The same special attack is occasionally used by Scizor with a probability of 0.08. If Ash loses because of the special attack, what is the probability that the Pokémon that he encountered is Darkrai.

Solution: 0.963

Say,

$Let\ D\ stand\ for\ an\ event\ of\ a\ Pokemon\ being\ Darkrai\ P(D) = 0.7$

$Let\ S\ stand\ for\ an\ event\ of\ a\ Pokemon\ being\ Scizor\ P(S) = 0.3$

$Let\ A\ stand\ for\ an\ event\ where\ a\ special\ attack\ is\ seen$

$P(A \mid D) = 0.90$

$P(A \mid S) = 0.08$

$$P(D \mid A) = \frac{P(A \mid D) \times P(D)}{P(A \mid D) \times P(D) + P(A \mid S) \times P(S)}$$

$$P(D \mid A) = \frac{0.90 \times 0.7}{0.90 \times 0.7 + 0.08 \times 0.3}$$

$$P(D \mid A) = 0.9633$$

6. HMM - Temporal Modeling[15%]

An environment is defined as follows:
States = {S1, S2, S3}
Observations = {a, b, c}

Transition probabilities: T(<initial state>, <final state>)

	S1	S2	S3
S1	0.0	0.5	0.5
S2	1.0	0.0	0.0
S3	0.0	1.0	0.0

Emission/Observation Probabilities: E(<state>, <observation>)

	a	b	c
S1	0.5	0.5	0.0
S2	0.3	0.3	0.4
S3	0.25	0.0	0.75

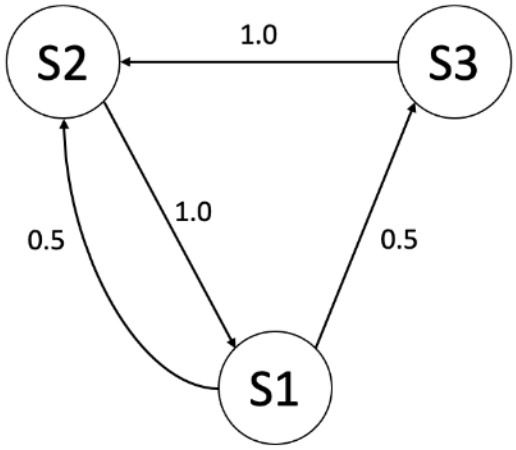
Initial State Probabilities: π(<state>)

	π
S1	0.25
S2	0.75
S3	0.0

Answer the below questions based on the information above:

1. [2%] How many non-zero edges will be there in the state diagram of this environment?
   [ANS] 4 or 11(if including emmision edges)

Explanation:



2. [4%] Select all the possible state paths for the Observation Sequence O = a, c, a in the list below.
   a. S1, S2, S1
   b. S1, S3, S2
   c. S3, S2, S1
   d. S1, S2, S1
   e. S2, S1, S2
   f. S2, S1, S3
   g. S1, S2, S3

3. [5%] What is the probability of the observation sequence O = a, c, a?

   [ANS] **0.027**

   Explanation:

   P(O) = $\sum_{s} P(O,\ s)$ = P(O, s=S1, S2, S1) + P(O, s=S1, S3, S2)

   P(O) = $\pi(S1).E(S1,\ a).\ T(S1,\ S2).E(S2,\ c).\ T(S2,\ S1).E(S1,\ a)$
   $+\ \ \pi(S1).E(S1,\ a).T(S1,\ S3).E(S3,\ c).\ T(S3,\ S2).E(S2,\ a)$

   P(O) = 0.25*0.5*0.5*0.4*1.0*0.5 + 0.25*0.5*0.5*0.75*1.0*0.3

   P(O) = 0.01250 + 0.01406 = **0.02656**

4. [2%] What is the most probable path for getting the observation sequence O = a, c, a?

a. S1, S2, S1

b. S1, S3, S2 explanation: P(O, s=S1, S3, S2) = 0.01406 is the maximum

c. S3, S2, S1

d. S1, S2, S1

e. S2, S1, S2

f. S2, S1, S3

g. S1, S2, S3

5. [2%] The classroom slides defined some general problems for temporal models. Which general problem does the above question (i.e, determining the most probable state path for getting the observation sequence O = a, c, a), come in:

   a. State Explanation ($P(X_{1:t}|E_{1:t})$)

   b. State Estimation ($P(X_t|E_{1:t})$)

   c. State Prediction ($P(X_{t+k}|E_{1:t}),\ k > 0$)

   d. State Smoothing ($P(X_k|E_{1:t}),\ k < t$)

   e. Model Learning ($P(M_t|E_{1:t})$)

7. Naive Bayes[10%]

As a star student in AI class, you have been recruited by the local weather station to help predict the weather. They give you access to their secret data store, displayed below. Only consider data from this table when making your assessments. Simplify all fractions if possible. (Answers in red)

Temp above 75	Humidity above 65	Pressure below 40	USC won football	**Storm next day?**
0	1	1	1	0
1	0	0	0	0
1	0	1	1	0
0	1	1	1	1
1	0	1	0	1
1	0	0	1	0
0	0	1	1	0

Use the following abbreviations for the columns:

T = temp above 75
H = humidity above 65
P = pressure below 40
W = USC won football
S = Storm next day?

1. (45) What is the maximum likelihood estimate for a storm the next day? [2%]
   P(S = 1) = _ / _ A: 2/7 [1%]
   P(S = 0) = _ / _ A: 5/7 [1%]

2. (46) What is the conditional probability of USC winning football given there will be a storm the next day? [2%]
   P(W = 1 | S = 1) = _ / _ A: 1/2 [1%]
   P(W = 1 | S = 0) = _ / _ A: 4/5 [1%]

3. (47) Although the news team seems insistent, you recommend **not** using the USC won football column to predict storms. Select all the **correct** reasons you can give them as to why you ignore this data. [3%]
    - The USC won football column does not contain any information and won't produce better results.
    - Even if this column contains information, it is likely irrelevant to the predicted variable and may cause noisy predictions.
    - Using more data typically causes overfitting, leading to incorrect future predictions.
    - Using more data typically causes underfitting, decreasing future performance.

    OR

    - The USC won football column does not contain any information and won't produce better results.
    - Even if this column contains information, it is likely irrelevant to the predicted variable and may cause noisy predictions.
    - Using more data typically causes overfitting, leading to incorrect future predictions.
    - Using more data typically causes underfitting, decreasing future performance.

4. (48) Using the 3 approved columns (T,H,P), use the Naive Bayes method to determine the joint probabilities P(S=1, X) and P(S=0, X), where X is (T = 1, H = 1, P = 0), then select the predicted class c* (S=1 or S=0). Report your answer to 3 significant digits (ex: 0.0123). [3%]

    P(S = 1, X) = ____ 0
    P(S = 0, X) = ____ 0.0342 (0.034 <= A <= 0.035)
    c* = ____ 0

P(X | S = 1) => P(T = 1 | S = 1) * P(H = 1 | S = 1) * P(P = 0 | S = 1) = 1/2 * 1/2 * 0 = 0
P(X | S = 0) => P(T = 1 | S = 0) * P(H = 1 | S = 0) * P(P = 0 | S = 0) = 3/5 * 1/5 * 2/5 = 6/125

Use: c* = argmax P(c) * P(d|c)
=> P(S = 1 , X) = P(S = 1) * P(X | S = 1) = 2/7 * 1/4 = 0
=> P(S = 0 , X) = P(S = 0) * P(X | S = 0) = 5/7 * 6/125 = 12/875 = 0.0342
=> c* => S = 0

(Full marks given, original answer was incorrect)