Midterm 3

CSCI 561 Fall 2023: Foundation of Artificial Intelligence

Instructions:

1. Maximum credits/points for this midterm: 100 points.
2. No books (or any other material) are allowed.
3. Adhere to the Academic Integrity Code.

Problems	100% Percent Total
1- AI Knowledge (T/F)	10
2- Decision Trees	15
3- Neural Networks	15
4- Bayesian Networks	15
5- Probability Theory	15
6- HMM, Temporal Model	15
7- Naive Bayes	10
8- Multiple Choice	5

1. True/False Questions [10%]

1) Hierarchical clustering methods require a predefined number of clusters, much like k-means.
Ans) F

2) We expect a model with a high variance to generalize better than a model with a high bias.
Ans) F

3) For a discrete Bayesian network with n variables, the amount of space required to store the "joint" distribution table is O(n).
Ans) F [d^n]

4) Probability theory can be used for both incremental (on-line) and batch (off-line) learning.
Ans) True

5) A perceptron is guaranteed to learn a given linearly separable function within a finite number of training steps.
Ans) True

6) Linear regression is a useful model for making predictions but is limited by the fact that we cannot interpret the model to make inferences about the relationships between uncorrelated input features and the output.
Ans) False

7) Expected Utilities can be used by an agent to select actions rationally.
Ans) True

8). SVMs can only classify linearly separable functions.
Ans) False [Non-linear SVMs]

9) For reinforcement learning, we need to know the transition probabilities between the states before we start.
Ans) False [Q learning]

10) Occam's Razor prefers the simpler hypothesis over the more complex hypothesis consistent with the data.
Ans). True [Slides 22-Learning pg 16]

2. Decision Trees [15%]

In the state of Minnesota, it is said that there are only two seasons: Winter and Road Construction. The state government is trying to analyze when it is safe to send construction teams to repair roads damaged by winter weather. They have provided the training data below, and they need your help to train a machine to decide whether to send workers out to repair roads.

(Table 2.1)

#	Below Freezing	Heavy Traffic	Wet Roads	Repair Roads
1	Yes	Yes	Yes	No
2	No	Yes	Yes	No
3	Yes	No	Yes	No
4	Yes	Yes	No	No
5	No	Yes	No	Yes
6	Yes	No	No	No
7	No	No	Yes	Yes
8	No	No	No	Yes
9	No	Yes	No	No
10	Yes	Yes	No	Yes

Note:

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

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

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

1. 0.970
2. 0.2
3. 1
4. 0.990

Solution:

Entropy = $-4/10 \log_2(4/10) - 6/10 \log_2(6/10) = 0.970$

Q2. Which would be the best attribute to split on? (Assume this attribute to be X for further questions) [5%]

A. Below Freezing
B. Heavy Traffic
C. Wet Roads

Solution:

Entropy = $-4/10 \log_2(4/10) - 6/10 \log_2(6/10) = 0.970$

Remainder(Below Freezing) = 5/10 [- 1/5 log2(1/5) - ⅘ log2(⅘)] + 5/10 [-3/5 log2(3/5) – ⅖ log2(2/5)] = 0.846
IG(Rain) = 0.124

Remainder(Heavy Traffic) = 6/10 [- 2/6 log2(2/6) – 4/6 log2(4/6)] + 4/10 [-2/4 log2(2/4) – 2/4 log2(2/4)] = 0.950

IG(Heavy Traffic) = 0.020

Remainder(Wet Roads) = 4/10 [-1/4 log2(1/4) - ¾ log2(¾)] + 6/10 [- 3/6 log2(3/6) – 3/6 log2(3/6)] = 0.924

IG(Wet Roads) = 0.046

X =Below Freezing (Since IG Below Freezing is highest)

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

a. 0.020
b. 0.027
c. 0
d. 0.950

Remainder(Heavy Traffic) = 6/10 [- 2/6 log2(2/6) – 4/6 log2(4/6)] + 4/10 [-2/4 log2(2/4) – 2/4 log2(2/4)] = 0.950

Q4. After splitting on attribute X, what attributes are optimal to split on when X = No? [5%]

A. Below Freezing
B. Heavy Traffic
C. Wet Roads

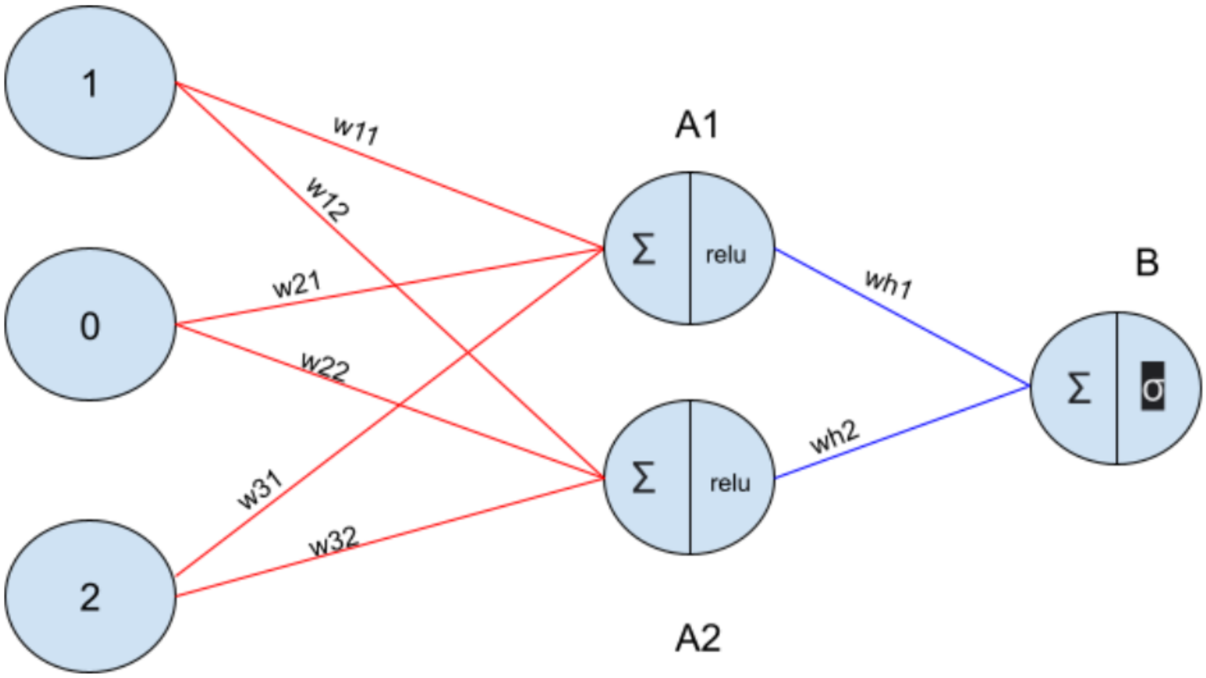
Remainder(Heavy Traffic) = 3/5 [- 1/3 log2(1/3) – 2/3 log2(2/3)] + 2/5 [-2/2 log2(2/2)] = 0.550

IG(Heavy Traffic) = 0.400

Remainder(Wet Roads) = 2/5 [-1/2 log2(1/2)-1/2 log2(1/2)] + 3/5 [- 1/3 log2(1/3) – 2/3 log2(2/3)] = 0.950

IG(Wet Roads) = 0

3. Neural Networks [15%]



Note:

(for the final answer, always take digits up to 3 decimal places and drop the rest without rounding. E.g., 0.9737 becomes 0.973)

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

Given is a Neural network with an input layer, a hidden layer, and an output layer. The weights are given below.

w11 = 1
w12 = -1
w21 = 2
w22 = 1
w31 = -1
w32 = 2
wh1 = 1
wh2 = -1

After the first layer ReLU activation function is applied to the layers. And after the output layer sigmoid activation function is applied.

We use loss function: $L = -(d-o)^2/2$

Learning rate = 1

The true label d = 1

o is the predicted label.

Activation Functions:

ReLU(x) = max{0,x}

[For example ReLU(5)=5 and ReLU(-2) = 0]

Sigmoid(x) = 1/(1+e^(-x))

Derivative of Sigmoid(x) = Sigmoid(x)(1-Sigmoid(x)).

Derivative of ReLu(x) = {1 if x>0, Else 0}

1). Calculate output after first pass for A1 (after relu), A2(after relu), B (after sigmoid) [3 marks]

2). Do backpropagation and calculate the derivative δL/δwh1, δL/δwh2. After that tell their updated weights. [6 marks] (Use the answer (after truncation) from question 1 while solving)

3). Do backpropagation and calculate the derivative δL/δw11, δL/δw32. After that tell their updated weights. [6 marks](Use the answer (after truncation) from question 1 while solving)

Solution

1.
   a. A1 = relu(1*1 + 0*2 + 2*(-1)) = 0
   b. A2 = relu(1*(-1) + 0*1 + 2*2) = 3
   c. B = sigmoid(1*0+(-1)*3) = 0.047425 = 0.047 (rounded off)
2.
   The way to calculate updated w =
   W = w -learning_rate*(δL/δw)
   a. δL/δwh1 = δL/δout * δout/δnet * δnet/δwh1
   = (d-o)*σ(x)*(1-σ(x))*A1
   = (1-0.047)*σ(3)*(1-σ(3))*A1
   δL/δwh1=0
   wh1 = 1 [No updates]
   b. δL/δwh2 = δL/δout * δout/δnet * δnet/δwh2
   = (d-o)*σ(x)*(1-σ(x))*A2
   = (1-0.047)*σ(3)*(1-σ(3))*A2
   = (1-0.047)*0.952*(1-0.952)*3
   =0.953*0.952*0.048*3
   = 0.043*3
   δL/δwh2=0.129
   wh2= -1 - 1*0.129 = -1.129
   wh2= -1.129

3.

a. $\delta L/\delta w11 = \delta L/\delta out * \delta out/\delta net * \delta net/\delta outh1 * \delta outh1/\delta neth1 * \delta neth1/\delta w11$

$= (d-o)*\sigma(x)*(1-\sigma(x))* wh1 * \text{der of relu}(x) * Input1$
$= (1-0.047)*\sigma(3)*(1-\sigma(3))*1*0*1$
$=0$
$\delta L/\delta w11 =0$
w11 = 1 [No updates]

b. $\delta L/\delta w32 = \delta L/\delta out * \delta out/\delta net * \delta net/\delta outh2 * \delta outh2/\delta neth2 * \delta neth2/\delta w32$
$= (d-o)*\sigma(x)*(1-\sigma(x))* wh2 * \text{der of relu}(x) * Input3$
$= (1-0.047)*\sigma(3)*(1-\sigma(3))*(-1) * 1 * 2$
$= 0.043*(-2)$
$\delta L/\delta w32=-0.086$
w32= 2+0.086 = 2.086

4. Bayesian Networks [15%]

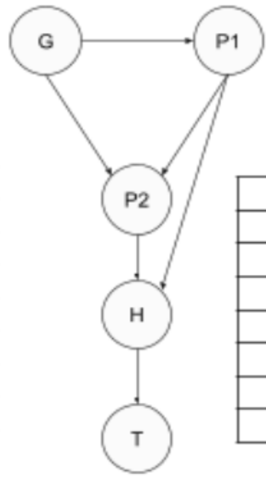
Consider the following Bayesian Network with the boolean variables having the following semantics.
G: Has GPA > 3.9
P1: Has Professor 1 Recommendation
P2: Has Professor 2 Recommendation
H: Selected for Honors Program
T: Offered Teaching Assistantship

P(G)

+g	0.6
-g	0.4

P(P1|G)

+g	+p1	0.8
+g	-p1	0.3
-g	+p1	0.1
-g	-p1	0.9



P(P2|P1,G)

+g	+p1	+p2	0.7
+g	+p1	-p2	0.3
+g	-p1	+p2	0.5
+g	-p1	-p2	0.5
-g	+p1	+p2	0.1
-g	+p1	-p2	0.9
-g	-p1	+p2	0.2
-g	-p1	-p2	0.8

P(H|P1,P2)

+p1	+p2	+h	0.9
+p1	+p2	-h	0.1
+p1	-p2	+h	0.7
+p1	-p2	-h	0.3
-p1	+p2	+h	0.4
-p1	+p2	-h	0.6
-p1	-p2	+h	0.1
-p1	-p2	-h	0.9

P(T|H)

+h	+t	0.6
+h	-t	0.4
-h	+t	0.1
-h	-t	0.9

(Note: For the below questions, truncate only the final answer to 3 decimals. Ex: 0.3956 -> 0.395, 0.29 -> 0.290)

1. What is the probability of a student having a GPA > 3.9, Professor 1 recommendation, and getting selected for the Honors Program? (3 points)

Ans: P(+g,+p1,+h) = P(+g,+p1,+p2,+h) + P(+g,+p1,-p2,+h)

= P(+g) * P(+p1/+g) * P(+p2/+p1,+g)* P(+h/+p1,+p2)
+ P(+g) * P(+p1/+g) * P(-p2/+p1,+g)* P(+h/+p1,-p2)
= 0.6 * 0.8 * 0.7 * 0.9 + 0.6 * 0.8 * 0.3 * 0.7
= 0.3024 + 0.1008
= 0.403 (0.4032 truncated to three decimal places)

2. What is the probability of a student having a GPA > 3.9 and not getting selected for the Honors Program?
(5 points)

Ans: P(+g, -h) = P(+g,+p1,+p2,-h) + P(+g,+p1,-p2,-h) + P(+g,-p1,+p2,-h) + P(+g,-p1,-p2,-h)
= P(+g) * P(+p1/+g) * P(+p2/+p1,+g)* P(-h/+p1,+p2)

+ P(+g) * P(+p1/+g) * P(-p2/+p1,+g)* P(-h/+p1,-p2)
+ P(+g) * P(-p1/+g) * P(+p2/-p1,+g)* P(-h/-p1,+p2)
+ P(+g) * P(-p1/+g) * P(-p2/-p1,+g)* P(-h/-p1,-p2)

= 0.6 * 0.8 * 0.7 * 0.1 + 0.6 * 0.8 * 0.3 * 0.3 + 0.6 * 0.3 * 0.5 * 0.6 + 0.6 * 0.3 * 0.5 * 0.9
= 0.0336 + 0.0432 + 0.054 + 0.081
= 0.211

3. Find P(+g|+p1) (3 points)

Ans: P(+g/+p1) = P(+p1/+g) * P(+g) /P(+p1)

P(+p1) = P(+p1/+g) P(+g) + P(+p1/-g) P(-g)
= 0.8 * 0.6 + 0.1 * 0.4
= 0.52

P(+g/+p1) = 0.8 * 0.6 / 0.52
= 0.923

4. Consider the following statements:
S1: Probability of student getting TAship provided that the student has been selected for Honors Program
S2: Probability of student being selected for Honors Program given that the student has Professor 1 recommendation but not Professor 2.

Is S1 < S2 ? (Answer should be either True or False) (2 points)

Ans: True

5. State whether the following statement is true or false
P(P1,P2|H) = P(P1|H) * P(P2|H) (1 point)
Ans: False
P(T|G,P1,P2,H) = P(T|H) (1 point)
Ans: True

5. Probability Theory [15%]

Suppose you have a deck of 52 playing cards consisting of 4 suits (hearts, diamonds, clubs, spades), each with 13 cards. You shuffle the deck thoroughly and draw 5 cards without replacement.

1. [4%] A Flush in poker is a hand where all 5 cards are of the same suit. What is the probability of getting a Flush? (Truncate the answer to 6 digits after the decimal point)

Ways to get a Flush = ${}^{13}C_5$ * 4 = 1,287 * 4 = 5148 ways

Ways to pick 5 cards from a deck = ${}^{52}C_5$ = 2598960 ways

Prob = (5148/2598960) = **0.0019807 or 0.001981**

Correct Answer: **0.001980 or 0.0019807**

2. [5%] A full house in poker is a hand where three cards share one rank, and two cards share another rank. What is the probability of getting a full house? (Truncate the answer to 6 digits after the decimal point)

   In the context of playing cards, "rank" refers to the numerical or alphabetical value assigned to a card within a specific suit. For example, if you have two cards with a numerical value of 7 from different suits, those cards are considered to be of the same rank. Similarly, if you have two cards, a jack of hearts and a jack of spades, those cards are considered to be of the same rank.

Choose the rank of the pair: ${}^{13}C_1$ = 13

Choose the pair from that rank, i.e., pick 2 of 4 cards: ${}^{4}C_2$ = 6

Choose the rank of the triple (from the remaining 12 ranks): ${}^{12}C_1$ = 12

Choose the triple from that rank: ${}^{4}C_3$ = 4

Number of ways to get a full house: 13 * 6 * 12 * 4 = 3744 ways

Number of ways to pick any 5 cards out of 52: 2598960 ways

Prob = (3744/2598960) = 0.00144058

Correct Answer: **0.001440** or **0.0014406**

Questions 3 to 6 depend on the following.

Two six-sided dice are rolled.

A = ‘sum of two dice equals 3’

B = ‘sum of two dice equals 7’

C = ‘Atleast one of the dice shows a 1’

3. [2%] What is P(A|C)?

There are 11 configurations in which C is true. Of those, 2 sum to three, so P(A|C) = 2/11 = 0.182

Correct Answer: **0.182** or **0.1818**

4. [2%] What is P(B|C)?

Correct Answer: **0.182** or **0.1818**

5. [1%] Are A and C independent?

Correct Answer: **No**

6. [1%] Are B and C independent?

Correct Answer: **No**

P(A) = { (1,2), (2,1) } = 2/36

P(B) = { (1,6), (6,1), (2,5), (5,2), (3,4), (4,3) } = 6/36

P(C) = { (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (3,1), (4,1), (5,1), (6,1) } = 11/36

P(A ∩ C) = { (1,2), (2,1) } = 2/36

P(A|C) = P(A ∩ C)/P(C) = 2/11

P(B|C) = P(B ∩ C)/P(C) = 2/11

Total possibilities of rolling two six-sided dice:

(1,1) (2,1) (3,1) (4,1) (5,1) (6,1)

(1,2) (2,2) (3,2) (4,2) (5,2) (6,2)

(1,3) (2,3) (3,3) (4,3) (5,3) (6,3)

(1,4) (2,4) (3,4) (4,4) (5,4) (6,4)

(1,5) (2,5) (3,5) (4,5) (5,5) (6,5)

(1,6) (2,6) (3,6) (4,6) (5,6) (6,6)

6. HMM, Temporal Model [15%]

A POMDP environment is defined as follows:

States = {S0, S1, S2, S3}

Observation = {a, b, c}

Initial State Probabilities: π (<state>)

	π
S0	0.325
S1	0.449
S2	0.0
S3	0.226

Transition probabilities: T (<next state> | <initial state>)

This table contains the transition probabilities to move from one state to the next state. For example, the probability of moving from state S1 to state S3 will be P(S3 | S1) = 0.139.

	S0	S1	S2	S3
S0	0.082	0.135	0.363	0.420
S1	0.315	0.262	0.284	0.139
S2	0.497	0.141	0.173	0.189
S3	0.252	0.248	0.369	0.131

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

This table contains the emission probabilities of seeing an observation from a state. For example, the probability of observing b from state S0 will be P(b | S0) = 0.34.

	a	b	c
S0	0.26	0.34	0.4
S1	0.22	0.48	0.3
S2	0.25	0.3	0.45
S3	0	0.75	0.25

Answer the below questions based on the information above:

1. [5%] What is the probability of seeing the observation sequence P(O) where O=<a,a>?

   Hint 1: P(O) = Σ_{s} P (O, s) where s is a possible state sequence.

   Hint 2: There are only six possible state sequences where P(<a, a>, s) ≠ 0.0

P(<a,a>, <S0,S0>) = 0.325 * (0.26) * 0.082 * (0.26) = 0.00018015 or 0.0002

P(<a,a>, <S0,S1>) = 0.325 * (0.26) * 0.135 * (0.22) = 0.00250965 or 0.0025

P(<a,a>, <S0,S2>) = 0.325 * (0.26) * 0.363 * (0.25) = 0.00766838 or 0.0077

P(<a,a>, <S1,S0>) **= 0.449 * (0.22) * 0.315 * (0.26) = 0.00809008 or 0.0081**

P(<a,a>, <S1,S1>) = 0.449 * (0.22) * 0.262 * (0.22) = 0.00569368 or 0.0057

P(<a,a>, <S1,S2>) = 0.449 * (0.22) * 0.284 * (0.25) = 0.00701338 or 0.0070

All remaining state sequences have a probability of 0.

P(O) = Σ P (O, s) = 0.00018015 + 0.00250965 + 0.00766838 + 0.00809008 + 0.00569368 + 0.00701338 = 0.03115532

Or = 0.0002 + 0.0025 + 0.0077 + 0.0081 + 0.0057 + 0.0070 = 0.0312

Correct Answer: **0.031, 0.0312**

2. [3%] Given the Observation sequence (a, a), which is the most probable state sequence? (Write the answer in the box below, as "S? -> S?" where ? is a number).

S1 -> S0

3. [4%] Given the Observation sequence (c, a, b) and the initial two states as (S0 -> S2), which is the most probable next state? (Write the answer in the box below, as "S?" where ? is a number).

S0 -> S2 -> S?

Ans: **S0**

Let the probability of seeing the first two states P(<c, a>, <S0, S2>) be x.

P(<c, a, b>, <S0, S2, S0>) = x * 0.497 * (0.34) = 0.16898 * x

P(<c, a, b>, <S0, S2, S1>) = x * 0.141 * (0.48) = 0.06768 * x

P(<c, a, b>, <S0, S2, S2>) = x * 0.173 * (0.3) = 0.0519 * x

P(<c, a, b>, <S0, S2, S3>) = x * 0.189 * (0.75) = 0.14175 * x

So, the most probable next state is **S0**.

4. [3%] Given the Observation sequence (c, a, b) and the initial two states as (S0 -> S2), what is the probability of the most probable state sequence?

S0 -> S2 -> S0 = 0.325 * (20/50) * 0.363 * (15/60) * 0.497 * (17/50) = 0.00199354

Correct Answer: **0.002, 0.0019**

7. Naive Bayes [15%]

Consider the following GOOD and BAD reviews of XYZ restaurant.

GOOD

- Best Pizza EVER! Not just on the lower Cape…. Anywhere!
- DELICIOUS BREAKFAST! We ordered the eggs mulligan, corn beef hash, and the California Focaccia. All items were great. Excellent service.
- Friendly staff and probably the best cheese pizza I've had!. It's so sad to hear that the owner died recently.
- The food is excellent, with generous portions and great prices. The service was fast and friendly. I highly recommend it if you're in the Wellfleet area
- One of the best places I've eaten on Cape Cod. Great cocktails, awesome

BAD

- Nobody should pay this much to be sad. Worst food ever
- Simply unpalatable. So angry and sad about visiting this place.
- Worst slice and customer service. Asked them to reheat the COLD SLICE that they gave us to eat. GARBAGE.
- Overhyped restaurant. Not the best place to visit on weekends.

These reviews are analyzed according to the presence of specific keywords.

A) Fill in the blanks. (If any of the blanks is zero, fill just zero) (2 marks)
(P(W | GOOD) is the (number of sentences that contain the word 'W' in GOOD reviews/number of GOOD reviews))

W (word in review)	P(W | GOOD)	P(W | BAD)	P(W)
best	0.600	1/4	0.444
service	2/5	1/4	1/3
sad	1/5	0.500	1/3

B) Compute the probabilities *P(GOOD | X)* and *P(BAD | X)*, where *X* = *"service"* (2 marks)

Ans:

P(GOOD | service) = P(service/GOOD) * P(GOOD) / P(service)
= ((2/5)*(5/9))/(1/3) = **2/3 =0.666**

P(BAD | service) = P(service/BAD) * P(BAD) / P(service)
= ((1/4)*(4/9))/(1/3) = **1/3 =0.333**

C) Consider the following new review just received by the restaurant:
"This restaurant provides the best service, but it's so sad to see many tables empty"

Using the keywords (best, service, sad) from the review, compute P(GOOD/ best, service, sad), and P(BAD/best, service, sad). Also, state the predicted class of the Naive Bayes classifier for the new review.

Ans:

P(best, service, sad) = P(best, service, sad/GOOD) * P(GOOD) + P(best, service, sad/BAD)* P(BAD)
= (3/5 * 2/5 * 1/5) * (5/9) + (1/4 * 1/4 * 1/2) * (4/9)
= (2/75) + (1/72)

a. P(GOOD/ best, service, sad) = P(best, service, sad/GOOD) * P(GOOD) / P(best, service, sad) (2 marks)
= (2/75) / ((2/75) + (1/72))
= 0.658

A) 0.761
B) 0.658
C) 0.813
D) 0.692

b. P(BAD/ best, service, sad) = P(best, service, sad/BAD)* P(BAD) / P(best, service, sad) (2 marks)
= 0.342

A) 0.239
B) 0.187
C) 0.308
D) 0.342

c. GOOD (1 mark)

D) State whether the following statement is True or False (1 mark)
P(GOOD/best, service) < (BAD/best, service)

Ans: False

8. Multiple Choice [5%]

Answer	Reference	Description	A	B	C	D
BCD	wk 18, sl 10	A Partially Observable MDP (POMDP) is defined by:	A set of fully observable states.	A set of actions for each state	A transition model and a sensor model	A reward function
A	wk 17, sl 55	Which of these is a way to write Bayes Rule in simple English (non-mathematically)?	The posterior is proportional to the prior * likelihood	The prior is proportional to the likelihood * posterior	The likelihood is proportional to the prior * posterior	The posterior is proportional to the prior / likelihood
ACD	sk18 Learning, sl 19-25	Which of the following are true about decision trees?	They can be used to perform supervised learning	A decision tree is designed to split it's decisions based on the features that give the lowest information gain	Entropy is a way to measure information gain and a common non-leaf node splitting criteria.	Leaf nodes display the class of an instance that reaches them.
ABD	wk23, 6-22	Which of the following are true about neural networks?	The perceptron is their fundamental building block	Layers of perceptrons can be strung together to learn more complicated functions such as XOR	Back propagation maximizes the error between predicted outputs and the actual target values.	Gradient descent is useful in the process because it calculates the magnitude the weights should change
ABCD	wk19, 31	Which of the following are true about temporal models?	Localization/filtering/ monitoring are the tasks for tracking the belief state over time	Markov chains have no explicit actions, states transition randomly and have no explicit actions	Calculating the probability at a previous state is known as smoothing	Quantifying how correct a model is at time t is known as model learning