Last Name: First Name:

Matriculation Number:

Seat:

Exam
Artificial Intelligence 2

Feb 16, 2023

	To be used for grading, do not write here										
prob.	1.1	1.2	2.1	2.2	3.1	3.2	4.1	4.2	5.1	Sum	grade
total	7	9	8	13	13	10	10	8	7	85	
reached											

The “solutions” to the exam/assignment problems in this document are supplied to give students a starting point for answering questions. While we are striving for helpful “solutions”, they can be incomplete and can even contain errors even after our best efforts.
In any case, grading student’s answers is not a process of simply “comparing with the reference solution”, therefore errors in the “solutions” are not a problem in this case.
If you find “solutions” you do not understand or you find incorrect, discuss this on the course forum and/or with your TA and/notify the instructors. We will – if needed – correct them ASAP.

In the course Artificial Intelligence I/II we award bonus points for the first student who reports a factual error in an old exam. (Please report spelling/formatting errors as well.)

1 Probabilities

Problem 1.1 (Python) 7 pt

Consider the Python program below.

1. Which operation does the function `foo` compute? 4 pt

2. Assume random variables X with domain $\{0, \dots, m-1\}$ and Y with domain $\{0, \dots, n-1\}$. 3 pt
   Assume the Python object C holds their joint probability distribution $P(X, Y)$, i.e., $C[i][j] = P(X = i, Y = j)$.
   Complete the definition of E in the program below in such a way that it holds the probability distribution $P(Y|X = 0)$, i.e., $E[j] = P(Y = j|X = 0)$.
   (Hint: This can be done with relatively little code.)

```
# input: a list of numbers in the interval [0;1]
def foo(a):
    l = len(a)
    s = 0
    for i in range(l):
        s += a[i]
    res = []
    for i in range(l):
        res.append(a[i]/s)
    return res

E =
```

Solution:

1. The normalization αa of a vector a.

2. `E = foo(C[0])`.

Problem 1.2 (Calculations) 9 pt

Assume random variables X, Y both with domain $\{0, 1, 2\}$, whose joint distribution $P(X, Y)$ is given by

x	y	$P(X = x, Y = y)$
0	0	a
0	1	b
0	2	c
1	0	d
1	1	e
1	2	f
2	0	g
2	1	h
2	2	i

1. In terms of $a, b, c, d, e, f, g, h, i$, give $P(X = 0)$. 1 pt
2. In terms of $a, b, c, d, e, f, g, h, i$, give $P(X = 0|Y = 1)$. 2 pt
3. In terms of $a, b, c, d, e, f, g, h, i$, give $P(X \neq 0|Y \neq 1)$. 2 pt
4. The table above is redundant. How can it be stored using less space? 2 pt
5. Now assume X and Y are stochastically independent. How can the information in the table be stored using the least space? 2 pt

---

Solution:

1. $P(X = 0) = P(X = 0, Y = 0) + P(X = 0, Y = 1) + P(X = 0, Y = 2) = a + b + c$
2. $P(X = 0|Y = 1) = P(X = 0, Y = 1)/P(Y = 1) = b/(b + e + h)$.
3. $P(X \neq 0|Y \neq 1) = P(X \in \{1, 2\}, Y \in \{0, 2\})/P(Y \in \{0, 2\}) = (d + f + g + i)/(a + c + d + f + g + i)$.
4. We can remove, e.g., the row for i by using $i = 1 - a - b - c - d - e - f - g - h$.
5. We can store $P(X = 0)$, $P(X = 1)$, $P(Y = 0)$, and $P(Y = 1)$. Then we can compute $P(X = 2) = 1 - P(X = 1) - P(X = 2)$ (and accordingly for Y) and $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$.

---

2 Bayesian Reasoning

Problem 2.1 (Bayes' Rule) 8 pt

Assume you are trying to predict whether a particular topic comes up in an exam. You have collected the following data:

- 30% of all topics come up in the exam.
- 40% of all topics come up in the tutorials.

- If a topic comes up in an exam, it was covered by an assignment 60% of the time.
- If a topic comes up in an exam, it came up in a tutorial 80% of the time.

You model this situation using 3 Boolean random variables X (comes up in exam), S (covered by assignments), and T (came up in a tutorial).

1. By filling in the gaps below, state for each number in the text above, which probability it describes. 2 pt

   (a) $P($_____$) = 0.3$

   (b) $P($_____$) = 0.4$

   (c) $P($_____$) = 0.6$

   (d) $P($_____$) = 0.8$

2. Assume the topic you are interested in **did not** come up in the exam. Argue if and how we can obtain the probability that it was covered by a tutorial. 2 pt

3. The topic you are interested in was covered by a tutorial. Using Bayes' rule, calculate the exact probability that it will come up in the exam. 4 pt

Solution:

1. $P(X = true) = 0.3$, $P(T = true) = 0.4$, $P(S = true|X = true) = 0.6$, $P(T = true|X = true) = 0.8$

2. The joint distribution of X and T contains 4 values: $P(t, x)$, $P(t, \neg x)$, $P(\neg t, x)$ and $P(\neg t, \neg x)$. We know 4 properties about them:
   - $P(x) = P(t, x) + P(\neg t, x) = 0.3$
   - $P(t) = P(t, x) + P(t, \neg x) = 0.4$
   - $P(t|x) = P(t, x)/P(x) = 0.8$
   - $P(t, x) + (t, \neg x) + P(\neg t, x) + P(\neg t, \neg x) = 1$

   From those we can compute the joint distribution as

	t	$\neg t$	Σ
x	0.24	0.06	0.3
$\neg x$	0.16	0.54	0.7
Σ	0.4	0.6	1

   And from that, we can compute $P(t|\neg x) = P(t, \neg x)/P(\neg x) = 0.16/0.7 = 8/35 \approx 0.229$.

   A direct calculation can be done as follows: $P(t|\neg x) = P(t, \neg x)/P(\neg x)$
   $= \big(P(t) - P(t, x)\big)/(1 - P(x))$
   $= \big(P(t) - P(t|x) \cdot P(x)\big)/(1 - P(x))$
   $= (0.4 - 0.8 \cdot 0.3)/(1 - 0.3) = 0.16/0.7 = 8/35 \approx 0.229$

   To get the points, it was sufficient to sketch this procedure. A full calculation was not required.

3. $P(X = true|T = true) = P(T = true|X = true) \cdot P(X = true)/P(T = true) = 0.8 \cdot 0.3/0.4 = 0.6$

Problem 2.2 (Bayesian Networks) 13 pt

Consider the following situation:
- Covid and influenza can cause fever.
- Fever causes stress.
- Tests can detect Covid. But a false-positive Covid-test causes stress as well.
- There are no other causal relationships.

You want to model this situation using Boolean random variables C (Covid infection), I (influenza infection), F (fever), S (stress), and T (positive Covid test).

1. Give a good variable ordering for forming a Bayesian network for this situation. 3 pt

2. Give the resulting network. 3 pt

3. You have a fever and have tested positive for Covid. Now you want to determine if you have influenza. What are the query, evidence, and hidden variables? 2 pt

4. Assume your network is $I \leftarrow C \rightarrow F \leftarrow S \rightarrow T$ (which *may or may not* be a correct solution to the above question). Which probabilities are stored in the conditional probability table of node F? 2 pt

5. Again using the network $I \leftarrow C \rightarrow F \leftarrow S \rightarrow T$, give the formula for

$$P(C = true, I = true, F = true, T = true)$$

in terms of the entries of the conditional probability table of that network. 3 pt

You may abbreviate the event $X = true$ by the lower-case name of the random variable X.

Solution:

1. Causes should occur before effects, so e.g., $CIFTS$ (CI can be swapped, and T must be anywhere between C and S).
2. $C \rightarrow F \leftarrow I$ and $F \rightarrow S$ and $C \rightarrow T \rightarrow S$.
3. Query: I, evidence: F, T, hidden: C, S.
4. The probability distribution $P(F|C,S)$, i.e., $P(F = x|C = y, S = z)$ as a function of Booleans x, y, z.
5. $$P(c,i,f,t) = P(c,i,f,t,s) + P(c,i,f,t,\neg s)$$
$$= P(c) \cdot P(s) \cdot P(t|s) \cdot P(i|c) \cdot P(f|c,s) + P(c) \cdot P(\neg s) \cdot P(t|\neg s) \cdot P(i|c) \cdot P(f|c,\neg s)$$
$$= P(c) \cdot P(i|c) \cdot \big(P(s) \cdot P(t|s) \cdot P(f|c,s) \cdot + P(\neg s) \cdot P(t|\neg s) \cdot P(f|c,\neg s)\big)$$

---

3 Markovian Reasoning

Problem 3.1 (Hidden Markov Models) 13 pt

Consider the following situation:

- You make annual observations about the rainfall at a certain location. Each year the rainfall is high (r_1), medium (r_2), or low (r_3).
- You know this is caused by an atmospheric condition, which is either strong (c_1) or weak (c_2).
- You have previously obtained the following information:
  - when the condition is strong, the rainfall is high 20% and medium 30% of the time,
  - when the condition is weak, the rainfall is high 35% and medium 15% of the time,
  - when the condition is strong, it stays strong next year 70% of the time,

– when the condition is weak, it becomes strong next year 40% of the time,

You want to model this situation as a hidden Markov model with two families of random variables R_a (rainfall) and C_a (condition), each indexed by year number a.

1. Give the state and evidence variables and their domains. 2 pt
2. How can you tell that the model is first-order here? 1 pt
3. Complete the following sentences:
   (a) The transition model T is given by the matrix 2 pt

   $$T = \begin{pmatrix} & \\ & \end{pmatrix} \quad \text{where} \quad T_{ij} = P(C_{a+1} = c_j | C_a = c_i).$$

   (b) The sensor model S is given by the matrix 2 pt

   $$S = \begin{pmatrix} & \\ & \end{pmatrix} \quad \text{where} \quad S_{ij} = P(R_a = r_j | C_a = c_i).$$

4. The atmospheric condition has been strong last year, and the rainfall is low this year. You want to use filtering to obtain the probability distribution of this year's condition.
   You proceed as follows:
   (a) Give the recursive filtering equation for $f_{1:a+1}$. 1 pt
   (b) Give the initial value $f_{1:0}$ to use in this case. 1 pt
   (c) Give the diagonal sensor matrix O_1 to use in this case. 1 pt
   (d) Calculate the resulting distribution. 3 pt
   Fully compute all values including the normalization. (This does not require approximations or a calculator.)

Solution:

1. State variables $C_a \in \{c_1, c_2\}$, evidence variables $R_a \in \{r_1, r_2, r_3\}$.
2. The probability of the atmospheric condition C_a only depends on the previous year C_{a-1}, not earlier years.
3. (a) The transition model T is given by the matrix

$$T = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix} \text{where } T_{ij} = P(C_{a+1} = c_j | C_a = c_i).$$

   (b) The sensor model S is given by the matrix

$$S = \begin{pmatrix} 0.2 & 0.3 & 0.5 \\ 0.35 & 0.15 & 0.5 \end{pmatrix} \text{where } S_{ij} = P(R_a = r_j | C_d = c_i).$$

4. We compute $f_{1:1}$ by applying the filtering equation once.
   (a) $f_{1:a+1} = \alpha(O_{a+1} \cdot T^t f_{1:a})$
   (b) $f_{1:0} = \langle 1, 0 \rangle$
   (c) $O_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$
   (d) $f_{1:1} = \alpha \cdot O_1 \cdot T^t \cdot f_{1:0} = \alpha \langle 0.5 \cdot 0.7, 0.5 \cdot 0.3 \rangle$
   $= 2\langle 0.35, 0.15 \rangle = \langle 0.7, 0.3 \rangle$.

Problem 3.2 (Utility and Decision Processes) 10 pt

Consider an agent moving along a circular arrangement of 8 locations as indicated below.

$$\begin{array}{ccccccc} 0 & \rightarrow & 1 & \rightarrow & 2 & \rightarrow & 3 \\ \uparrow & & & & & & \downarrow \\ 7 & \leftarrow & 6 & \leftarrow & 5 & \leftarrow & 4 \end{array}$$

The agent's movement is as follows:

- It can move forward (in arrow direction) or backward (against arrow direction), and in each case it can move carefully or quickly.
- The careful moves result in moving 1 step in that direction with probability 60% and no move otherwise.
- The quick actions result in moving 1 step in that direction with probability 90% and moving 1 step in the opposite direction otherwise.

The agent's goal is to move to location 0.

1. Choose an appropriate reward function and model this situation as a Markov Decision Process. 4 pt

2. State the Bellman equation relative to your model. 2 pt

3. Give an optimal policy π^*. 2 pt
Hint: This requires deciding whether careful or quick actions lead to the goal faster.

4. Now assume we use a POMDP because the agent is unable to tell what move an action resulted in. Assume we know the agent is initially in location 4. Calculate the belief state after moving backwards carefully twice. 2 pt

---

Solution:

1. One possible model is
   - $S = \{0, \dots, 7\}$
   - $A((i, j)) = \{1, -1\} \times \{c, q\}$
   - $P(s + x|(x, c), s) = 0.6$ and $P(s|(x, c), s) = 0.4$,
     $P(s + x|(x, q), s) = 0.9$ and $P(s - x|(x, q), s) = 0.1$,
     and all other probabilities are 0.
   - A typical choice is any function R that is high for 0 and slightly negative for other states. E.g., $R(0) = 1$ and $R(s) = -0.1$ otherwise.

2. $U(s) = R(s) + \gamma \cdot \max_{a \in A(s)} \Sigma_{s' \in S} U(s') \cdot P(s'|s, a)$

3. Any policy that maps 0 to (x, c) for any x, and $1, 2, 3$ to $(-1, q)$ and $5, 6, 7$ to $(1, q)$ and 4 to (x, q) for any x.

4. The belief state B is a probability distribution over states $s \in S$. The values are $P(B = 2) = 0.6 \cdot 0.6 = 0.36$, $P(B = 3) = 0.6 \cdot 0.4 + 0.4 \cdot 0.6 = 0.48$, $P(4) = 0.4 \cdot 0.4 = 0.16$, and $P(B = s) = 0$ otherwise.

---

4 Learning

Problem 4.1 (Decision Trees and Lists) 10 pt
Consider an unknown 3-letter word $W = L_1, L_2, L_3 \in \{BAT, BET, CAT, CAR\}$. Each letter L_i is currently covered (i.e., it cannot be seen) and can be uncovered individually.

1. Draw the decision tree for W that arises from uncovering the letters in the order L_1, L_2, L_3. (Do not uncover additional letters if the word can already be identified.) 3 pt

2. Which choice would the information gain algorithm make first in this case? Justify your answer. 2 pt

3. Give a decision list in 1-DL for W using only literals of the form $L_i = X$ for characters X. 2 pt

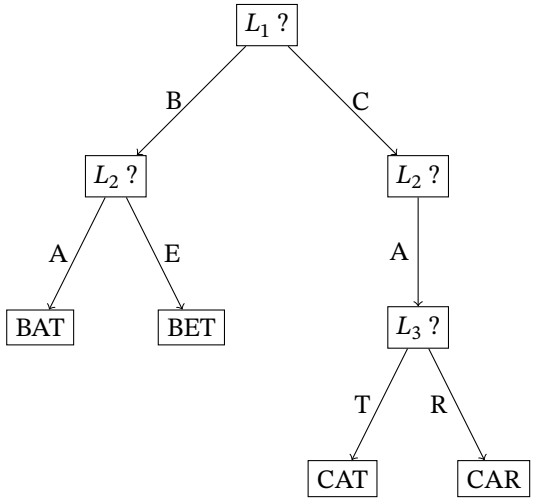
4. Give all minimal sets $A \subseteq \{L_1, L_2, L_3\}$ of letters such that $A \succ W$. 2 pt

5. How would we have to change the set of possible words so that the determination $\{L_1, L_2\} \succ W$ holds? 1 pt

---

Solution:

1. The tree is



2. It would choose L_1. Each letter splits the words into two sets, and only L_1 splits them into equally sized ones, which maximizes information gain. This could be formally justify by computing the information gain.

3. W := if $L_3 = R$ then CAR elif $L_1 = C$ then CAT elif $L_2 = A$ then BAT else BET

4. $\{L_1, L_2, L_3\}$

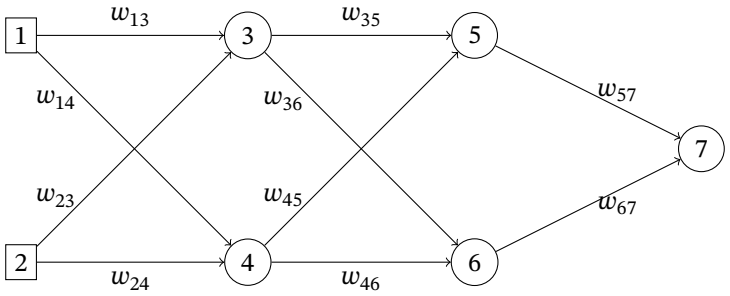
5. Remove or or change CAT or CAR so that there are no two words that agree in L_1 and L_2.

Problem 4.2 (Neural Networks) 8 pt

Consider the neural network **without bias** given below where

- units 1, 2 are inputs,
- unit 7 is output,
- weights are given by the labels on the edges, and

- units $3, 4, 5, 6, 7$ are perceptron units with activation function $T(x) = 1$ for $x > 0.5$ and $T(x) = 0$ otherwise.



1. How many hidden layers are there? 1 pt
2. Assume $w_{ij} = 0.2$ for all weights and $a_1 = a_2 = 1$. What is the resulting output a_7? 1 pt
3. Assume we remove the edge from 5 to 7. Give two reasons why that would be a bad network to use. 2 pt
4. Give the formula for the activation a_6 of unit 6 in terms of the inputs a_1 and a_2. 2 pt
5. Assume the inputs a_1, a_2 are in $\{0, 1\}$. Choose weights such that the output a_7 is always 1 (no matter what the inputs are), or argue why that is impossible? 2 pt

---

Solution:

1. 2
2. 0
3. Node 5 is redundant because its output is not used anymore. Node 7 can be eliminated because it has only one input.
4. $a_6 = T\big(w_{46}T(w_{14}a_1 + w_{24}a_2) + w_{36}T(w_{13}a_1 + w_{23}a_2)\big)$
5. It is impossible. For $a_1 = a_2 = 0$, the output is always $a_7 = 0$, no matter what the weights are (because there is no bias).

---

5 Natural Language Processing

Problem 5.1 (Language Models) 7 pt

1. How many different trigrams does a language with n words have? 1 pt
2. What is a statistical language model? 2 pt
3. Name two applications of statistical language models. 2 pt
4. Why is it work-intensive in practice to build a good statistical language model for a natural language? 2 pt

---

Solution:

1. n^3
2. A probability distribution over words or n-grams occurring in a corpus of the language.
3. Language identification, genre classification, named entity recognition, text generation, spell-checking.
4. Because a representative corpus of texts has to be aggregated that is hard.
