Last Name: First Name:

Matriculation Number:

Retake Exam
Artificial Intelligence 2

April 8, 2024

Please ignore the QR codes; do not write on them, they are for grading support

	To be used for grading, do not write here												
prob.	1.1	1.2	2.1	2.2	2.3	3.1	3.2	4.1	4.2	5.1	5.2	Sum	grade
total	7	8	7	3	10	13	10	11	8	7	8	92	
reached													

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)

1. We use Python objects p to hold the *joint probability distribution* of *random variables* X and Y, i.e, $p[i][j] = P(X = i, Y = j)$. 4 pt
   Consider the Python *program* below:

```
def bar(p):
    res = []
    for i in range(len(p)):
        s = 0
        q = p[i]
        for j in range(len(q)):
            s += q[j]
        res.append(s)
    return res
```

   Which probability-related operation does `bar` implement?

   Solution: The *probability distribution* of X, i.e., `bar(p)[i]` is $P(X = i)$.

2. Possibly using `bar` from above, write a *function* `cond` such that `cond(p,i,j)` returns the *conditional probability* $P(Y = j \mid X = i)$. 3 pt

```
def cond(p,i,j):
```

   Solution:

```
def cond(p,i,j):
  return p[i][j]/bar(p)[i]
```

Problem 1.2 (Calculations)

Assume *random variables* X and Y, both with *domain* $\{0, 1, 2\}$ with the following *conditional probability distribution*:

x	y	$P(X = x \mid Y = y)$
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. Give all *subsets* of $\{a, b, c, d, e, f, g, h, i\}$ whose elements *sum* to 1. 2 pt

Solution: $\{a, d, g\}, \{b, e, h\}, \{c, f, i\}$

2. In terms of the *values* $a, b, c, d, e, f, g, h, i$, give $P(X \neq 0 \mid Y = 0)$. 2 pt

Solution: $d + g$

3. Which of the following *value* can be computed from the *values* $a, b, c, d, e, f, g, h, i$? 2 pt
   - ☐ $P(X = 0)$
   - ☐ $P(Y = 0)$
   - ☐ $P(Y = 0 \mid X = 0)$
   - ☑ $P(X = 0 \mid Y = 0)$
4. Which property of the values $a, b, c, d, e, f, g, h, i$ holds iff X and Y are *independent*? 2 pt

Solution: $a = b = c \wedge d = e = f \wedge g = h = i$

2 Bayesian Reasoning

Problem 2.1 (Bayesian Calculations)

Consider a disease with a prevalence of 1/1000, i.e., 1 in 1000 people have it. You are using a test that gives a yes/no answer for whether a person has the disease. However, the test randomly returns the wrong result 1% of the time.

1. Model this situation using *random variables*. State all *probabilities* whose *values* are given in the text. 3 pt

Solution: Boolean *RVs* D (for whether someone has the disease) and T (for the test result).
$P(D^+) = 1/1000$
$P(T^+ \mid D^+) = 99/100$ (equivalently: $P(T^- \mid D^+) = 1/100$)
$P(T^- \mid D^-) = 99/100$ (equivalently: $P(T^+ \mid D^-) = 1/100$).

2. Calculate the *probability* of a test returning yes. 2 pt

Solution: Marginalization: $P(T^+) = P(T^+ \mid D^+) \cdot P(D^+) + P(T^+ \mid D^-) \cdot P(D^-) = 0.00099 + 0.00999 = 0.01098$.

3. You are using the test on a person, and it returns yes. Calculate the *probability* that she has the disease. 2 pt

Solution: Bayes rule: $P(D^+ \mid T^+) = P(T^+ \mid D^+) \cdot P(D^+)/P(T^+) = 99/1098 = 11/122$

Problem 2.2 (Conditional Bayes Rule)

1. Consider 3 *Boolean random variables* X, Y, C. We write x, y, c for the *events* where the corresponding *variable* is *true*. Prove that 3 pt

$$P(x \mid y, c) = P(y \mid x, c) \cdot P(x \mid c)/P(y \mid c)$$

Solution: This follows immediately from

$$P(x \mid y, c) \cdot P(y \mid c) \cdot P(c) = P(x, y, c) = P(y \mid x, c) \cdot P(x \mid c) \cdot P(c)$$

after canceling $P(c)$.

Problem 2.3 (Bayesian Networks)

Consider the following situation about a car:

- Your car is unusable if it is out of gas or if it is broken. These two are the only causes.
- You might be late for work if your car does not work or if you oversleep. These two are the only causes.

You want to model this situation as a *Bayesian network* using *Boolean random variables*.

1. Give an appropriate *set* of *random variables* and their meaning. Give a good *variable ordering* and draw the resulting *Bayesian network*. 3 pt

Solution: *Random variables*variable: C (car unusable), G (out of gas), B (broken), L (late for work), S (overslept).
Order: $\{B, G, \}, C, L$ with S anywhere except at the end
Network: $G \to C \leftarrow B$ and $C \to L \leftarrow S$

2. Give the *probability* of the car being unusable in terms of the entries of the *conditional probability table* of your *network*. 2 pt

Solution: $P(C^+) = \Sigma_{b,g \in \{true, false\}} P(C^+ \mid B = b, G = g) \cdot P(B = b) \cdot P(G = g)$

3. Now you decide to make the car-unusable node *deterministic*. Explain (in about 2 sentences) why that choice is justified based on the description above, and how it affects the *conditional probability table* of that *node*. 2 pt

Solution: The description says that the car **is** (rather than e.g., "might be") unusable if is a broken or without gas, i.e., that the relation is *deterministic* and not governed by *probability*. Formally: $P(C^+ \mid G^+ \vee B^+) = 1$. Thus, we do not have to store a *CPT* for C and only need to store the *function* $C = G||B$.

4. Now you decide to make the late-for-work node a *noisy disjunction node*. Explain (in about 2 sentences) which two properties must hold about its *probability distribution* for this decision to be justified. Judge if these are backed by the description. 3 pt

---

Solution: Firstly, the two *causes* must be the only *causes*, i.e., $P(L^+ \mid C^-, S^-) = 0$. This is explicitly stated in the description.
Secondly, the two *causal relationships* must be *independent* of each other. Formally, if both *causes* are present, the *probability* of non-lateness must be the product of the two *inhibition factors*: $P(L^+ \mid C^+, S^+) = 1 - P(L^- \mid C^+, S^-) \cdot P(L^- \mid C^-, S^+)$. This is not commented on by the description. Common sense background knowledge indicates that the probably of being late is even higher if both *causes* are present, e.g., if oversleeping prevents catching a bus.

---

3 Markovian Reasoning

Problem 3.1 (Hidden Markov Models)
Consider the following situation, which you want to model as a *hidden Markov model*:

- You make daily observations about your business (B), which can go well (b_1), average (b_2), or badly (b_3).
- This is caused by the weather (W), which can be good (w_1) or bad (w_2). Over a period of days d, you have collected the following *probabilities* for this *causal relationship*:

$$S_{ij}^d = P(B_d = b_j \mid W_d = w_i) = \begin{pmatrix} 0.4 & 0.4 & 0.2 \\ 0.5 - 1/(2d) & 0.5 & 1/(2d) \end{pmatrix}$$

- The weather is influenced by the previous day's weather as follows:
  - If the weather is good, it stays good 60% of the time.
  - If the weather is bad, it stays bad 30% of the time.

1. Give the *state* and *evidence variables* and their *domains*. 2 pt

---

Solution: *evidence variables* $B_d \in \{b_1, b_2, b_3\}$, *state variables* $W_d \in \{w_1, w_2\}$.

---

2. Fill in the following sentence: 2 pt
The *transition matrix* is given by

$$T_{ij} = P(\qquad\qquad) = \begin{pmatrix} & \\ & \end{pmatrix}$$

Solution:

$$T_{ij} = P(W_{d+1} = w_j \mid W_d = w_i) = \begin{pmatrix} 0.6 & 0.4 \\ 0.7 & 0.3 \end{pmatrix}$$

---

3. Explain (in about one sentence) why this *model* is or is not *first-order*. 1 pt

---

Solution: It is *first-order* because the *probability distribution* of W_d only depends on W_{d-1}.

4. Explain (in about 2 sentences) whether the *transition model* and the *sensor model* are *stationary*. 2 pt

Solution: The *transition model* is *stationary* because T does not depend on d. The *sensor model* is not *stationary* because S^d depends on d.

5. There is a 75% chance for the weather to be good at day $d = 1$. Calculate the resulting *probability distribution* for the business at day 1. 2 pt

Solution: $P(B_1) = \langle 0.75, 0.25\rangle \cdot S^1 = \langle 0.3 + 0, 0.3 + 0.125, 0.15 + 0.125\rangle = \langle 0.3, 0.425, 0.275\rangle$.

This problem is continued on the next page

Now assume we have made *observations* e_d of B_d for a sequence of days.
Consider the *filtering algorithm* in *matrix* form:

$$f_{1:d+1} = \alpha \cdot O_{d+1} \cdot T^t \cdot f_{1:d}$$

where α is a *normalization constant*.

6. State the definition of O_d in terms of e_d. 2 pt

Solution: If $e_d = w_i$, then O_d is the *diagonal matrix* formed from the i-th column of S^d.

7. If $f_{1:1}$ is the *distribution* of W_1, which *distribution* is given by $f_{1:d}$? 2 pt

Solution: $P(W_d \mid B_1 = e_1, \ldots, B_d = e_d)$

Problem 3.2 (Decision Processes and Utility)

Consider an *agent* moving along 8 locations as indicated below.

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

The *agent*'s movement is as follows:

- In every *state*, it can make moves called -1, 0, and 1.
- Additionally, in *states* 1 and 5, it can make a move called 5.
- Each move n, made in location l moves the *agent* to location $l \oplus n$ (where $\oplus$ is addition modulo 8). However, 10% of the time, a move fails, and the *agent*'s location does not change.

The *agent*'s *goal* is to move to location 7.

1. Model this situation as a *Markov Decision Process*. 4 pt

Solution: One possible model is

- $\mathcal{S} = \{0, \ldots, 7\}$
- $\mathrm{Act}(s) = \{-1, 0, 1\} \cup \{5 | s = 5 \vee s = 1\}$

- The *transition model* is given by
  - $P(s \mid s, 0) = 1$
  - for $a \neq 0$: $P(s \oplus a \mid a, s) = 0.9$ and $P(s \mid a, s) = 0.1$.

  All other *probabilities* are 0.
- A typical choice is any *function* R that is high for the *goal* and slightly negative for other *states*. E.g., $R(7) = 1$ and $R(s) = -0.1$ otherwise.

From now on, consider the policy π defined by $\pi(7) = 0$ and $\pi(s) = 1$ otherwise.

2. Calculate the *probability distribution* of the *agent*'s location after starting in location 0 and making 2 moves according to π. 2 pt

---

Solution: $P(s_0) = \langle 1, 0, 0, 0, 0, 0, 0, 0\rangle$
$P(s_1) = \langle 0.1, 0.9, 0, 0, 0, 0, 0, 0\rangle$
$P(s_2) = \langle 0.01, 0.18, 0.81, 0, 0, 0, 0, 0\rangle$
where the vectors contains the probabilities for the values $0, \ldots, 7$.

---

3. State the *equation* for evaluating the *policy* π, which we can use to iteratively calculate the *utility* of each *state* under *policy* π. Include the initial values of the *utilities*. 2 pt

---

Solution: $U(s) = R(s) + \gamma \Sigma_{s' \in S} U(s') \cdot P(s' \mid s, \pi(s))$
Initially $U(s) = 0$.

---

4. Assume we have already calculated the *utility* $U(s)$ of each *state* s. Explain (in about 2 sentences, including the relevant formulas) how can we determine if π is *optimal*? 2 pt

---

Solution: For each s, we check if $\pi(s)$ *maximizes* the *expected utility* $EU(s, a)$. That is given by $\Sigma_{s' \in S} P(s' \mid s, a) \cdot U(s')$.

---

4 Learning

Problem 4.1 (Decision Trees and Lists)

Consider a *word* W chosen uniformly from $\{bad, bed, bend, pend, pad, ped\}$. You are allowed to ask the following questions about W:

A length of the word
B first letter of the word
C second letter of the word
D last letter of the word

1. Show that there is no *decision tree* for W of depth 2. 2 pt

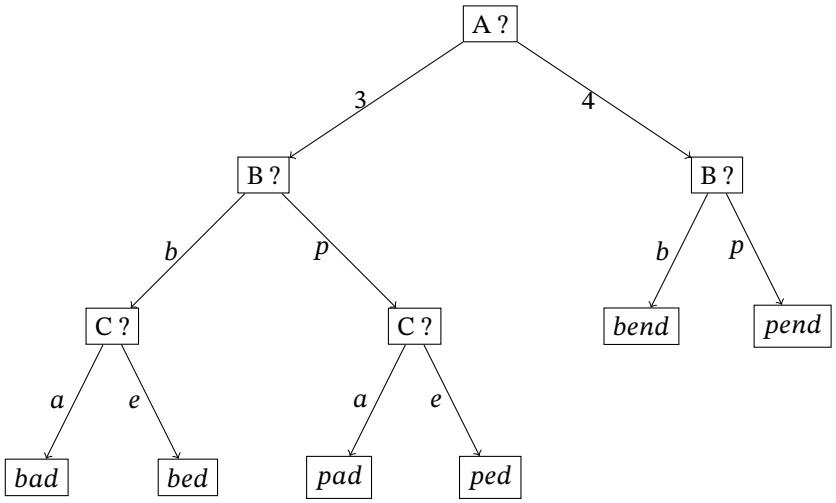
---

Solution: All questions have at most 2 possible answers, so a *decision tree* of *depth* 2 has at most 4 *leaves*. But we need at least 5 *leaves* to cover all options for W.

---

2. Draw the *decision tree* for W that arises from asking the questions in the order A,B,C,D. (Do not ask additional questions if the *word* can already be identified.) 3 pt

Solution: The tree is



3. Calculate the *information gain* for all 4 questions. 2 pt

Solution: A, C: $-1/3\log_2 1/3 - 2/3\log_2 2/3$
B: $-1/2\log_2 1/2 - 1/2\log_2 1/2 = 1$
D: $-1\log_2 1 = 0$

4. Which question would the *information gain algorithm* ask first? 1 pt

Solution: B

5. Give all minimal sets $Q \subseteq \{A, B, C, D\}$ of questions for which the *determination* $Q \succ W$ holds? 1 pt

Solution: $\{A, B, C\}$

6. Give two *words* such that removing them from the choices for W makes the *determination* $\{A, B\} \succ W$ hold. 2 pt

Solution: Any pair out of $\{bad, bed\} \times \{pad, ped\}$

Problem 4.2 (Statistical Learning)

You observe the values below for 50 games of a tennis player. You want to predict the result based on time of day and opponent.

Time	Opponent	Number of wins	Number of losses
Morning	Weaker	5	1
Afternoon	Weaker	6	2
Evening	Weaker	3	0
Morning	Similar	3	3
Afternoon	Similar	2	3
Evening	Similar	4	5
Morning	Stronger	2	2
Afternoon	Stronger	1	3
Evening	Stronger	1	4

1. What is the *hypothesis space* for this situation, seen as a *decision tree learning* problem? 2 pt

   Solution: The set of *functions*

   $$\{Morning, Afternoon, Evening\} \times \{Weaker, Similar, Stronger\} \to \{Win, Loss\}$$

2. Explain (in about 2 sentences) the key characteristic of this data that makes *decision tree learning* inapplicable, and under what circumstances it would be applicable. 2 pt

   Solution: The results are not uniquely determined by the input. So no *decision tree* can exist. If each row had a 0 in one of the Win/Loss columns, it would be applicable.

3. Now instead, consider this as a *statistical learning* problem. As *hypotheses*, we use the *probability distributions* $P(Result \mid Time, Opponent)$. 2 pt
   Relative to the observed data, give the *likelihood* of the following *hypothesis*: The player's *probability* to win is 80% if the game is against a weaker player, and it is 10% otherwise.

   Solution: The *likelihood* is the *probability* of the data under the condition that the *hypothesis* holds. This is $0.8^{14} \cdot 0.2^3 \cdot 0.1^{13} \cdot 0.9^{20}$

4. To learn a *hypothesis* via *Bayesian learning*, we model this situation as a *Bayesian network* $Time \to Result \leftarrow Opponent$. Give the resulting entries of the *conditional probability table* for 2 pt
   1. $P(Opponent = Weaker) =$
   2. $P(Result = win \mid Time = Afternoon, Opponent = Weaker) =$

   Solution: $P(Opponent = Weaker) = 0.34$
   $P(Result = win \mid Time = Afternoon, Opponent = Weaker) = 0.75$

5 Natural Language Processing

Problem 5.1 (Grammars)

Consider the following *probabilistic grammar*:

S	$\rightarrow$	$NP\ VP$[1]
NP	$\rightarrow$	$Article\ Noun$[0.6] | $Name$[0.4]
VP	$\rightarrow$	$Verb$[0.5] | $TransVerb\ NP$[0.5]
$Article$	$\rightarrow$	the[0.7] | a[0.3]
$Noun$	$\rightarrow$	stench[0.2] | breeze[0.3] | wumpus[0.5]
$Name$	$\rightarrow$	John[0.3] | Mary[0.7]
$Verb$	$\rightarrow$	smells[1]
$TransVerb$	$\rightarrow$	sees[0.6] | shoots[0.4]

1. Which of the *production* above comprise the *lexicon*? 1 pt

   Solution: $Article$ to $TransVerb$; $Noun$ to $TransVerb$ was also accepted.

2. Explain (in about 2 sentences) why it is practical to separate the *lexicon* from the other *productions*. 2 pt

   Solution: The *production* in the *lexicon* are usually much more numerous and much less standardized. Moreover, they only occur as *leaves* of the *tree* and are thus not essential for the grammatical structure.

3. Give the *probability* of the *sentence* 2 pt

   Mary shoots the breeze.

   Solution: $0.4 \cdot 0.7 \cdot 0.5 \cdot 0.4 \cdot 0.6 \cdot 0.7 \cdot 0.3 = 0.007056$

4. Explain (in about 2 sentences) how we can use a *treebank* to learn the *probabilities* of the *production*. 2 pt

   Solution: For every *production* p for *non-terminal* L, we count how often it occurs in a *subtree* in the *treebank*, say n_L. Then we count how many of those *subtrees* use the *production* p, say n_p. We learn the *probability* n_p/n_L.

Problem 5.2 (Information Retrieval)

Consider the *corpus* $D = \{d_1, d_2, d_3\}$ where

- d_1: "The man is tall."
- d_2: "The tall man sees the woman."
- d_3: "The woman shouts at the tall man."

Below we use alphabetical order for the *vector components*:

at, is, man, sees, shouts, tall, the, woman

1. Give the *vector* $\mathrm{tf}(\_, d_3)$ 2 pt

   Solution: $\mathrm{tf}(\_, d_3) = \langle 1/7, 0, 1/7, 0, 1/7, 1/7, 2/7, 1/7 \rangle$.

2. Give the *vector* $\mathrm{idf}(\_, D)$. 2 pt

Solution: $\mathrm{idf}(\_, D) = \log_{10}(3/\langle 1,1,3,1,1,3,3,2\rangle) = \langle k,k,0,k,k,0,0,l\rangle$ with $k = \log_{10} 3$ and $l = \log_{10} 1.5$.

3. State the definition of tfidf. 2 pt

Solution: $\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$

4. Explain (in about 2 sentences) the point of using the *inverse document frequency*inverse document frequency in the definition of tfidf. Use the word *the* and the *corpus D* as an example. 2 pt

Solution: $\mathrm{idf}(t,D)$ can be used as a measure of the relevance of a *word* for characterizing a *document* — *words* with low idf score occur in many documents and are thus less distinctive. For example, $\mathrm{idf}(the,D) = 0$ and thus occurrences of the *word the* are ignored when calculating the tfidf *vectors*.