Last Name: First Name:

Matriculation Number:

Exam
Artificial Intelligence 2

September 24, 2025

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	3.1	3.2	4.1	4.2	4.3	5.1	5.2	Sum	grade
total	9	8	10	10	9	8	9	9	8	6	4	90	
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. Consider the Python program below where J holds the joint probability distribution of two random variables such that $J[x][y] = P(X = x, Y = y)$. Which probability-related operation does the function `foo` compute? 4 Points

```
def foo(J):
   res = 0
   for x in range(len(J)):
     for y in range(len(J[x])):
       res += (x+y)*J[x][y]
   return res
```

Solution: It computes the expected value of $X + Y$.

---

2. Consider a state space $\{0, \ldots, n-1\}$, a transition model such that $T[i][j]$ is the probability of state i transitioning into state j, and a probability distribution of the current state, i.e., $S[i]$ is the probability of currently being in state i. 5 Points
Complete the function definition below to return the probability distribution of the state after one transition.

```
def next(T,S):
```

Solution:

```
def next(T,S):
  n = len(S)
  res = []
  for j in range(n):
     pj = 0
     for i in range(n):
        pj += S[i]*T[i][j]
     res.append(pj)
  return res
```

(This is just the matrix product $S \cdot T$ with S seen as a $1 \times n$ matrix.)

Problem 1.2 (Working with Distributions)
Assume Boolean random variables X, Y, Z. Their joint probability distribution is given as follows:

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

1. What is the sample space Ω here? 2 Points

Solution: $\{0,1\}^3$

2. Give all subsets of the probabilities $\{a, \dots, h\}$ that must sum to 1. 2 Points

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

3. In terms of $a, \dots, h$, give $P(X = 0|Y = Z)$. 2 Points

Solution: $(a + d)/(a + d + e + h)$

4. If $e = f = g = h = 0$, under what conditions are X and Y independent? 2 Points

Solution: They are always independent in that case (because X is always 0 anyway).

2 Bayesian Reasoning

Problem 2.1 (Basic Rules)
Let A and B be Boolean random variables (using values 0 and 1) expressing that *Alice* and *Bob*, respectively, will pass a test. Let H be a Boolean random variable expressing that the test will be hard.

You estimate that Alice has a 60% chance to pass in general but only a 30% chance if the test is hard. 8 of the 10 most recent tests were hard, and you expect this to accurately predict the difficulty of the next test.

1. Give the following probabilities: 2 Points

$$P(H = 1) = \boxed{0.8}$$

$$P(A = 1|H = 1) = \boxed{0.3}$$

Solution: filled in above

2. Calculate the updated probability that the test was hard after finding out that Alice has passed. 3 Points

Solution: Bayes' rule: $P(H=1|A=1) = P(A=1|H=1)\cdot P(H=1)/P(A=1) = 0.3\cdot 0.8/0.6 = 0.4$

3. Now assume you know nothing about Alice, but you know that $P(B=1,H=1) = 30\%$ and $P(B=0,H=0) = 5\%$. 3 Points
Calculate the probability that Bob passes.

Solution: Marginalization in two ways:

$$0.2 = P(H=0) = P(B=1,H=0) + P(B=0,H=0)$$

$$P(B=1) = P(B=1,H=1) + P(B=1,H=0) = 0.3 + (0.2 - 0.05) = 0.45$$

4. Assume Alice and Bob have only studied together all semester. Explain in about 2 sentences, how that knowledge affects the probability analysis in this situation. 2 Points

Solution: For two arbitrary students, it is not guaranteed but reasonable to assume that A and B are conditionally independent given H. But now we should not assume that: If they studied together, there is a higher-than-usual chance that their test results will be the same.

Problem 2.2 (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 Points

Solution: random variablesvariable: 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 Points

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 Points

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 Points

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 probability 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:

- You make daily observations about your business B. Each day business is either good (b_1) or bad (b_2).
- You know this is caused by the general economic situation G, which you cannot easily observe, and which can be getting worse (g_1), be stable (g_2), or getting better (g_3).
- You have previously obtained the following information:
  - when the economy gets worse, your business is good 36% of the time,
  - when the economy is stable, your business is good 84% of the time,
  - when the economy gets better, your business is good 90% of the time,
  - half the time, the economy is the same as on the previous day,
  - when the economy changes from one day to the next, each change is equally likely.

You want to model this situation as a hidden Markov model with two families of random variables indexed by day number d.

1. Give the state and evidence variables and their domains. 2 Points

Solution: State variables $G_d \in \{g_1, g_2, g_3\}$, evidence variables $B_d \in \{b_1, b_2\}$

2. How can you tell that the sensor model is stationary here? 1 Points

Solution: The business-economy relation is the same for each day.

3. What order does the model have? 1 Points

Solution: first-order

4. Complete the following sentence: The transition model T is given by the matrix 2 Points

$$T = \left(\quad \right) \qquad \text{where} \qquad T_{ij} = P(G_{d+1} = g_j \mid G_d = g_i).$$

Solution: $T = \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$

5. Complete the following sentence: The sensor model S is given by the matrix 2 Points

$$S = \left(\quad \right) \qquad \text{where} \qquad S_{ij} = P(B_d = b_j \mid G_d = g_i).$$

Solution: $S = \begin{pmatrix} 0.36 & 0.64 \\ 0.84 & 0.16 \\ 0.9 & 0.1 \end{pmatrix}$

6. Assume you want to apply filtering after observing good business at $t = 1$. Give the diagonal sensor matrix O_1 to use in this case. 1 Points

Solution: $O_1 = \begin{pmatrix} 0.36 & 0 & 0 \\ 0 & 0.84 & 0 \\ 0 & 0 & 0.9 \end{pmatrix}$

Problem 3.2 (Decisions and Utility)
You want to calculate the utilities of state sequence $\vec{s} = s_0, s_1, \ldots$ experienced by an agent. You want to use a reward function R from states to real numbers.

1. In about 1 sentence, explain the problem of naively computing the total reward as $R(\vec{s}) = \Sigma_i R(s_i)$. 2 Points

Solution: The reward can become unreasonably large for long state sequences and even diverge to infinity for infinite sequences.

2. Finite horizon and reward discounting are two possible solutions for this problem. Explain each in about 1 sentence or 1 formula. 2 Points

Solution: Finite horizon only considers the next h states: $R(\vec{s}) = \Sigma_{i=1}^{h} R(s_i)$. Reward discounting assigns decreasing rewards to future states: $R(\vec{s}) = \Sigma_i \gamma^i R(s_i)$ for some $0 < \gamma < 1$.

3. Fill in the gaps in the following text: 4 Points

Both the value iteration and the policy evaluation algorithms use update rules derived from the Bellman equation to iteratively recompute the utility of all states.
In each iteration, the utility of each state s is recomputed by applying one action and adding the reward of s and the expected utility of the next state (computed using the current utility values) multiplied by a discount factor γ .
They differ in which actions are chosen:
The latter considers only one action per state , which is given by a fixed policy .
The former picks the action that maximizes expected utility .

Solution: filled in above

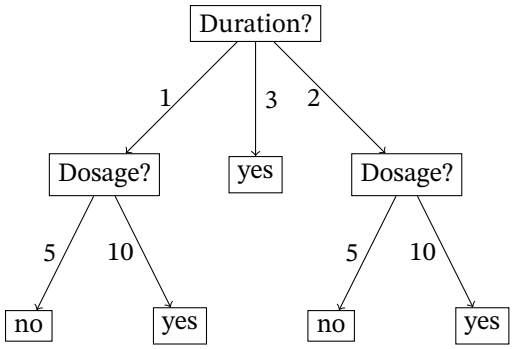
4 Learning

Problem 4.1 (Learning for Data with Attributes)
Consider the following data observed for some patients during a medical trial of a new drug:

Duration of Treatment 1, 2, or 3 weeks	Dosage 5 or 10 mg/week	Success yes/no
1	5	no
1	10	yes
2	5	no
2	10	yes
3	5	yes
3	10	yes

1. Give the decision tree for the attribute order "duration, dosage" that determines the patient's condition given the duration and dosage of the treatment. 3 Points

Solution:



2. Give the entropy of the attribute "dosage". 1 Points

Solution: $-3/6 \log 3/6 - 3/6 \log_2 3/6 = 1$

3. Give the entropy of the attribute "duration". 1 Points

Solution: $-2/6\log_2 2/6 - 2/6\log_2 2/6 - 2/6\log_2 2/6 = \log_2 3$

To give a logical formulation, we use binary predicate symbols *Dur* and *Dos* and a unary predicate symbol *Suc*. For example, $Dur(p,2) \wedge Suc(p)$ represents that patient p was successfully treated for 2 weeks.

4. Give the shortest first-order formula that correctly represents the treatment results for all patients p. 2 Points

Solution: $\forall p.Suc(p) \Leftrightarrow (Dur(p,3) \vee Dos(p,10))$

5. Which clause(s) would inductive logic programming learn to represent the predicate *Suc*? 2 Points

Solution: $Dur(p,3) \Rightarrow Suc(p)$ and $Dos(p,10) \Rightarrow Suc(p)$.

Problem 4.2 (Loss)
Our goal is to find a linear approximation $h_a(x) = ax$ for the series of squares $0,1,4,9,16$ of the numbers $0,1,2,3,4$.

1. Model this situation as an inductive learning problem. 2 Points

Solution: The inductive learning problem is $(\mathcal{H}, f)$ where

- the hypothesis space $\mathcal{H}$ is the set containing all functions $h_a(x) = ax$ with $\mathrm{dom}(h_a) = \{0,\ldots,4\}$ for $a \in \mathbb{R}$
- the target function is $f(x) = x^2$ with $\mathrm{dom}(f) = \{0,1,\ldots,4\}$

2. Assuming all 5 possible examples are equally probable, compute the generalized loss using the squared error loss function. 3 Points

Solution: Each example (x, x^2) has probability $1/5$. For each x, the loss is $L_2(x^2, ax) = (x^2 - ax)^2$. Thus for each $h(x) = ax$, we have

$$GenLoss(h_a) = \sum_{x=0,\ldots,4} (x^2 - ax)^2 \cdot 1/5$$

$$= ((1-a)^2 + (4-2a)^2 + (9-3a)^2 + (16-4a)^2)/5 = (354 - 200a + 30a^2)/5$$

3. Explain in about 2 sentences how you would continue to eventually determine that $h^*(x) = 10x/3$. 2 Points

Solution: We need to find the a that minimizes the loss. The derivative of *GenLoss* for a is

$(60a - 200)/5$. So the minimum is at $a = 10/3$.

4. What is the error rate of h^*? 2 Points

Solution: The error rate is $4/5$ because $h^*(x) = 10x/3$ predicts 4 out of 5 examples incorrectly. (E.g., $h_a(x) = x$ would have better error rate $3/5$ despite having higher generalized loss.)

Problem 4.3 (Bayesian Learning)
Consider an experiment with 2 different results 0 and 1. You repeat the experiment 3 times obtaining results $d = (d_1, d_2, d_3) = (1, 1, 0)$.
Your hypothesis space contains the functions h_β given by $h_\beta(i) = \beta^i$ for $0 \leq \beta \leq 1$ where $h_\beta(i)$ is the probability that the i-th repetition (starting at $i = 1$) will yield 1.

1. Calculate $P(d|h_{1/2})$, i.e., the likelihood of the data d under hypothesis $h_{1/2}$. 2 Points

Solution: $h_{1/2}(1) \cdot h_{1/2}(2) \cdot (1 - h_{1/2}(3)) = 1/2 \cdot 1/4 \cdot (7/8) = 7/64$

2. Why does it make sense to exclude hypotheses h_β for $\beta > 1$? 2 Points

Solution: Because we would have $h_\beta(i) > 1$, which cannot be a probability.

3. Before performing the experiment, you suspected β to be large and judged the probability of h_β to be $1 - \beta$. 2 Points
Now that you've obtained the data d, state the formula obtained from Bayes' rule that updates the probability of h_β.

Solution: $P(h_\beta|d) = \alpha(P(d|h_\beta) \cdot P(h_\beta)) = \alpha(\beta\beta^2(1 - \beta^3)(1 - \beta))$

4. Let us call the probability obtained in the previous problem $P(\beta)$. Explain in about 2 sentences how you can choose a hypothesis using the Maximum a Posteriori approximation. 2 Points

Solution: Find the β_{MAP} that maximizes $P(\beta)$ (e.g., by setting the derivative of $P(\beta)$ to 0 and solving for β). The learned hypothesis is $h_{\beta_{MAP}}$.

5 Natural Language Processing

Problem 5.1 (Language Models)
Consider a language L over the alphabet with characters x, y, and z and a corpus for it consisting of the four words

$$x\,x\,x\,y\,z \qquad x\,x\,x\,y\,y \qquad x\,x\,x\,y\,z\,y \qquad x\,x\,x\,z\,y\,z$$

You want to build a 3-gram model using random variables $c_1, ...$

1. Seen as a stationary Markov process, what is the meaning of the c_i and the order of such a model? 2 Points

   Solution: The random variables are the characters in a word $c_1, \ldots,$ over the language. The order is 2.

2. Give the value of $P(c_i = z | c_{i-2} = x, c_{i-1} = y)$ in the resulting model. 2 Points

   Solution: $2/3$ (3 occurrences of xy, 2 of which are followed by z).

3. Consider the word prefix xxx. By applying your model twice, which two characters would be predicted to follow? 2 Points

   Solution: xx (x is the most frequent next character after xx.)

Problem 5.2 (Information Retrieval)
Consider a corpus of n documents. You have already computed the `tfidf` vector for each document.

1. Informally, explain in about 2 sentences how can you use the `tfidf` vectors to choose the 3 most relevant document for the query q. 2 Points

   Solution: Compute the `tfidf`-vector of q. Then choose the 3 documents with the highest cosine similarity to it.

2. Let TP, TN, FP, FN be the number of true/false positive/negative results to a query. 2 Points
   In terms of those, give the definitions of
   1. precision: $\boxed{TP/(TP+FP)}$
   2. recall: $\boxed{TP/(TP+FN)}$

   Solution: filled in above