Last Name: First Name

Matriculation Number: Birth Date:

Seat:

Exam
Artificial Intelligence 2

Feb 17, 2022

	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	Sum	grade
total	10	14	15	10	15	15	5	10	94	
reached										

Exam Grade: Bonus Points: Final Grade:

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 5 bonus points for the first student who reports a factual error (please report spelling/formatting errors as well) in an assignment or old exam and 10 bonus points for an alternative solution (formatted in LaTeX) that is usefully different from the existing ones.

1 Bayesian Reasoning

Note: When working with an upper case **Boolean** random variable X, you may abbreviate the event $X = true$ by the corresponding lower-case letter x. If you do that, make sure the distinction between upper and lower case letters is clear in your writing.

Problem 1.1 (Bayesian Rules) 10 pt

5 pt

1. Assume a prevalence of SARS-CoV-2 infections in your area of 1/100 (share of people who are infected) and a positive-test rate of 19% (share of tests that return positive). Moreover, assume you you are using a test with a false-positive rate of 1% (share of uninfected people who test positive) and a false-negative rate of 5% (share of infected people who test negative). You have tested positive. Apply Bayes' rule to determine the probability that you are infected.

5 pt

2. Assume three random variables A, B, C such that A and B are conditionally independent given C. You know
   - the probability distribution of C,
   - the conditional probability distribution of A given C,
   - the conditional probability distribution of B given C.

   In terms of the above, give the formula for the probability distribution of C given the event $A = a, B = b$.

Solution:

1. We use Boolean random variables I for an infection and P for testing positive. We have $P(i|p) = P(p|i) \cdot P(i)/P(p) = 95/100 \cdot (1/100)/(19/100) = 5/100$.

2. $$P(C|A = a, B = b) = \alpha \cdot P(C, A = a, B = b) = \alpha \cdot P(C)P(B = b|C)P(A = a|B = b, C) =$$

   $$\alpha \cdot P(C)P(B = b|C)P(A = a|C)$$

   where α is a constant factor that normalizes the distribution.

Problem 1.2 (Bayesian Networks) 14 pt

Consider the following situation:

- To attend university, people need academic qualification and interest.
- Both of those are more likely for people with university-educated fathers.
- Interest is also generated by special schooling programs.
- There are no other causal relationships.

You want to model this situation using Boolean random variables U (attend university), Q (qualified), I (interested), F (father attended university), and S (covered by a school program).

1. You do not know if a person attends university and want to determine that by asking about their father's education. Which variables are the evidence, query, and hidden variables? 2 pt

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

3. Give the resulting network. 3 pt

4. Now assume your network is $Q \to F \to I \leftarrow U \to S$ (which *may or may not* be a correct solution to 3. above) and assume that I is deterministic.

   (a) What does being deterministic mean for the probability distribution of I? 2 pt
   (b) Which entries in the conditional probability table do we save by exploiting that I is deterministic? 2 pt
   (c) Give the formula for

   $$P(I|F = true, U = true, S = false)$$

   in terms of the entries of the conditional probability table of the network. 2 pt

---

Solution:

1. Evidence F, query U, hidden I, Q, S
2. Causes should occur before effects, so e.g., $SFIQU$.
3. $F \to Q \to U$ and $F \to I \to U$ and $S \to I$.
4. (a) The value of I is a fixed boolean function $I(f, u)$ in terms of the values of F and U. Thus, $P(i) = \sum_{f,u,I(f,u)=i} P(F = f, U = u)$.
   (b) We can save all entries for I, i.e., $P(I|f, u)$, $P(I|\neg f, u)$, $P(I|f, \neg u)$, and $P(I|\neg f, \neg u)$. The corresponding entries for $\neg I$ can also be saved, but they could already by computed from the others anyway.
   (c) We have $P(I|F = true, U = true, S = false) = P(I|F = true, U = true)$ (This question deceptively looked like marginalization might be necessary.) This could be further simplified by using the fact that I is deterministic.

---

2 Markovian Reasoning

Problem 2.1 (Hidden Markov Models) 15 pt

Consider the following situation:

- You record daily how your mood was. Each day you feel good or bad.
- You know this is caused by your daily workload, which is either high, medium, or low.
- You have previously assessed the following information:
  - high workload makes you feel bad 80% of the time,
  - medium workload makes you feel bad 60% of the time,
  - low workload makes you feel bad 10% of the time,

- from day to day, your workload never jumps from high to low or vice versa,
- 40% of the time, your workload stays the same from one day to the next,
- if your workload is medium, an increased and a decreased workload on the next day are equally likely,
- 2 days ago your workload was high.

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

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

2. How can you tell whether this model is stationary? 1 pt

3. How can you tell whether this model has Markov order 1? 1 pt

4. Complete the following sentences:

   (a) The transition model is given by the matrix 3 pt

   $$T = \begin{pmatrix} \quad & \quad & \quad \\ & & \\ & & \end{pmatrix} \qquad \text{where} \qquad T_{ij} = P(\quad = j | \quad = i).$$

   (b) The sensor model is given by the matrix 3 pt

   $$M = \begin{pmatrix} \quad & \quad \\ & \\ & \end{pmatrix} \qquad \text{where} \qquad M_{ij} = P(\quad = j | \quad = i).$$

   where we map values to matrix indices via *good* $= 1$ and *bad* $= 2$ as well as *low* $= 1$, *medium* $= 2$, and *high* $= 3$.

5. Your mood was good yesterday ($d = 1$) and bad today ($d = 2$). Give the matrix form of the recursive filtering equation and state precisely which concrete values to plug in to obtain the probability distribution of your workload at $d = 2$. (You do not have to actually compute the distribution.) 5 pt

---

Solution:

1. Evidence variables: $M_d \in \{g, b\}$ (for good/bad mood); state variables: $W_d \in \{l, m, w\}$ (for low/medium/high workload).

2. The probabilities do not depend on the day d.

3. The probabilities at $d + 1$ only depend on the values at d.

4. (a) The transition model is given by the matrix

   $$T = \begin{pmatrix} 0.4 & 0.6 & 0 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.6 & 0.4 \end{pmatrix} \qquad \text{where} \qquad T_{ij} = P(W_d = j | W_{d-1} = i).$$

(b) The sensor model is given by the matrix

$$M = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \\ 0.2 & 0.8 \end{pmatrix} \qquad \text{where} \qquad M_{ij} = P(M_d = j | W_d = i).$$

5. We compute $f_{1:2}$ by applying the filtering equation $f_{1:w+1} = \alpha(O_{w+1} \cdot T^t f_{1:w})$ twice where
   - $O_1 = \begin{pmatrix} 0.9 & 0 & 0 \\ 0 & 0.4 & 0 \\ 0 & 0 & 0.2 \end{pmatrix}$ and $O_2 = \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0.6 & 0 \\ 0 & 0 & 0.8 \end{pmatrix}$ are the diagonal sensor matrices for the observation of good mood at $d = 1$ and bad mood at $d = 2$,
   - $f_{1:0} = \langle 0, 0, 1 \rangle$ is the prior probability,
   - α is a constant factor to normalize the distribution.

Problem 2.2 (Markov Decision Processes) 10 pt

1. Give an optimal policy π^* for the following MDP: 4 pt
   - set S of states: integers $\mathbb{Z}$ with initial state 0
   - set of actions for $s \in S$: $A(s) = \{-1, 0, 1\}$
   - transition model for $s, s' \in S$ and $a \in A(s)$: $P(s'|s, a)$ is such that
     - $s' = s + a$ with probability $2/3$,
     - $s' = s - a$ with probability $1/3$.
   - reward function: $R(10) = 1$ and $R(s) = -0.1$ for all other states s
2. Assume that all utilities are initialized as $U(s) = 0$. Perform one round of value iteration using $\gamma = 1$. 2 pt
3. Now assume that utilities are initialized as $U(s) = R(s)$. Give the value of $U(0)$ after one round of value iteration using $\gamma = 1$. 4 pt

---

Solution:

1. $\pi^*(s) = 1$ if $s < 10$, and $\pi^*(s) = -1$ if $s > 10$, and $\pi^*(s) = 0$ if $s = 10$
2. We need to apply $U(s) := R(s) + \gamma \max_{a \in A(s)} \sum_{s' \in S} U(s')P(s'|s, a)$ with $\gamma = 1$ and current values of $U(s) = 0$. We obtain $U(s) = R(s)$ for all states.
3. We need to apply $U(0) := R(0) + \gamma \max_{a \in \{-1,0,1\}} \sum_{s' \in S} U(s')P(s'|s, a)$ with $\gamma = 1$ and current values of $U(s) = R(s)$. This yields $U(0) := -0.1 + \max\{U(-1) \cdot 2/3 + U(1) \cdot 1/3, U(0) \cdot 1, U(-1) \cdot 1/3 + U(1) \cdot 2/3\} = -0.2$.

---

3 Learning

Problem 3.1 (Decision Trees) 15 pt

You observe the values below for 6 different football games of your favorite team. You want to construct a decision tree that predicts the result.

#	Day	Weather	Location	Opponent	Result
1	Sunday	Cloudy	Away	Strong	Draw
2	Wednesday	Rainy	Away	Strong	Win
3	Monday	Sunny	Home	Weak	Loss
4	Monday	Rainy	Home	Weak	Loss
5	Monday	Cloudy	Home	Weak	Loss
6	Sunday	Sunny	Away	Strong	Draw

1. Assume you choose attributes in the order $Opponent, Location, Weather, Day$. Give the resulting decision tree. 4 pt

2. How does the information-theoretic algorithm choose an attribute? 1 pt

3. Without using the above observations, give the formula for the information gain of the attribute $Weather$. 3 pt

4. Using the above observations, give the results of 2 pt

   - $I(P(Result)) =$
   - $P(Result = Loss|Weather = Rainy) =$

   You do not have to compute irrational logarithms.

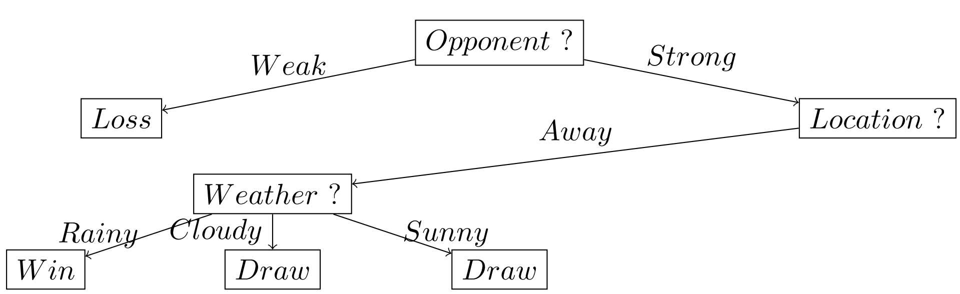
5. Give all minimal sets A of attributes such that $A \succ Result$ hold for the above observations. 3 pt

6. Explain *why or why not* the determination $Result, Location \succ Opponent$ holds for the above observations. 2 pt

---

Solution:

1. The tree is



2. The algorithm chooses the attribute with the highest information gain.

3. $Gain(Weather) = I(P(Result))$
   $-P(Weather = Cloudy) \cdot I(P(Result|Weather = Cloudy))$
   $-P(Weather = Sunny) \cdot I(P(Result|Weather = Sunny))$
   $-P(Weather = Rainy) \cdot I(P(Result|Weather = Rainy))$

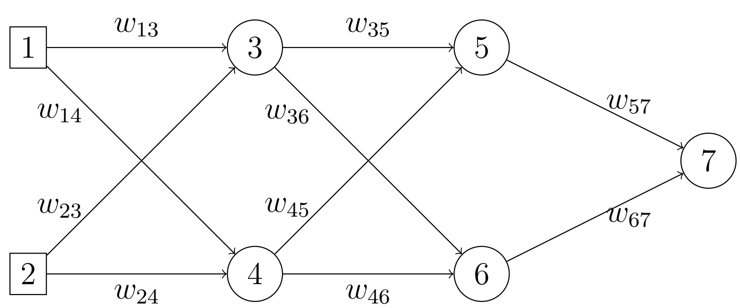
4. $I(P(Result)) = -1/2 \log_2 1/2 - 1/3 \log_2 1/3 - 1/6 \log_2 1/6$ and $P(Result = Loss|Weather = Rainy) = 1/2$.

5. $A = \{Weather, Opponent\}$ and $A = \{Day\}$ and $A = \{Weather, Location\}$

6. It holds. Games 1 and 6 as well as Games 3-5 agree in *Result* and *Location*; in both cases they also agree on *Opponent*.

Problem 3.2 (Neural Networks) 15 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. Which nodes are part of a hidden layer? 1 pt

2. Which important property of the network changes if we add an edge from 5 to 1? 1 pt

3. Give the formula for the activation a_5 of unit 5 in terms of the activations a_3 and a_4 and the weights w_{ij}. 3 pt

4. Assume $w_{ij} = 1$ for all weights w_{ij} and $a_1 = a_2 = 1$. What is the resulting output a_7? 2 pt

5. Assume $a_1, a_2 \in \{0, 1\}$ and $w_{13} = w_{23} = 1$ and $w_{14} = w_{24} = 0.3$ and $w_{35} = w_{46} = 1$ and $w_{57} = 1$. Choose appropriate values for the other weights such that the network implements the XOR function, i.e., $a_7 = a_1 \, XOR \, a_2$. 4 pt

6. Complete the high-level description of the back-propagation algorithm *on the next page*. To learn a target function $a_7 = f(a_1, a_2)$, do the following for each input to f: 4 pt

- compute the

- determine the error between

  and propagate

- use the propagation results to update

Solution:

1. 3 and 4 as well as 5 and 6
2. The network becomes recurrent.
3. $a_5 = T(w_{35}a_3 + w_{45}a_4)$
4. 1
5. The intended solution was $w_{36} = w_{45} = 0$, $w_{67} = -1$. This copies a_3 and a_4 over to a_5 and a_6 and then builds XOR at a_7 in the usual way. But other solutions were also correct.
6. - compute the activations of all units,
   - determine the error between network output and target result and propagate it back to all inner units,
   - use the propagated values to update all weights.

---

4 Natural Language Processing

Problem 4.1 (Language Models) 5 pt

1. How can we obtain a trigram model for a language? Explain the probability distribution involved. 2 pt
2. Explain informally how we can use trigram models to identify the language of a document D. 2 pt

3. Explain briefly what named entity recognition is. 1 pt

Solution:

1. We need a corpus of words over L. Then we count how often each trigram occurs in it and use that to estimate the probability distribution $P(T = t)$ of trigrams t.

2. We build a trigram model for each candidate language. Then we use each model to compute the probability of D occurring in that language. We choose the language with the highest probability.

3. The task of finding, in a text, names of things and deciding what class they belong to.

Problem 4.2 (Information Retrieval) 10 pt

Consider the following two texts

- d_1: "The air is cold."
- d_2: "The sun warms the air."
- d_3: "The day is over."

Let $D = \{d_1, d_2, d_3\}$.

Below we use alphabetical order for the vector components:

air, cold, day, is, over, sun, the, warms

Simplify all results as much as possible but without introducing approximate values.

1. Give the vector $tf(\_, d_2)$. 3 pt

2. Give the vector $idf(\_, D)$. 3 pt

3. Let w be the word "is". Give the value $tfidf(w, d_1, D)$. 2 pt

4. Now assume we have computed $tfidf(x, d, D)$ for every word x and every $d \in D$. How do we use those values to rank the texts in D for a query q? 2 pt

---

Solution:

1. $tf(\_, d_2) = \langle 1/5, 0, 0, 0, 0, 1/5, 2/5, 1/5 \rangle$.

2. $idf(\_, \{d_1, d_2\}) = \log_{10}(3/\langle 2, 1, 1, 2, 1, 1, 3, 1 \rangle) = \langle l-k, l, l, l-k, l, l, 0, l \rangle$ with $k = \log_{10} 2$ and $l = \log_{10} 3$.

3. $tfidf(w, d_2, D) = 1/4 \cdot idf(w, D) = (l-k)/4 = 1/4 \log_{10}(3/2)$.

4. We additionally compute $tfidf(x, q, D)$. Then we compute the angles between the vectors $tfidf(\_, d_i, D)$ and $tfidf(\_, q, D)$ and rank documents with lower angles higher.
