Last Name: First Name:

Matriculation Number:

Seat:

Exam
Artificial Intelligence 2

July 20, 2021

	To be used for grading, do not write here									grade
prob.	1.1	1.2	2.1	2.2	3.1	3.2	4.1	4.2	Sum	grade
total	10	15	15	10	15	15	5	10	95	
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 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

0 min

1. Assume that, in some area, the prevalence of SARS-CoV-2 infections is $1/10,000$. 5 pt
   Moreover, assume that such an infection causes a cough half the time, and that in general on any given day 1 person out of $1,000$ is coughing. Apply Bayes' rule to determine the probability that someone who coughs is infected.
2. Assume three random variables A, B, C such that A and B are conditionally independent given C. You know 5 pt
   - 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 C for coughing. We have $P(i|c) = P(c|i) \cdot P(i)/P(c) = 1/2 \cdot (1/10,000)/(1/1000) = 1/20$.
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.

Grading:

1. usually correct
2. partial credit: 1 point per derivation step; −0.5 for minor mistakes

Problem 1.2 (Bayesian Networks) 15 pt

0 min

Consider the following situation:

- You have a rock in your yard, which can feel wet or not.
- Rain may cause humidity in the air.
- Any one of rain, humidity, and whether the lawn was sprinkled may cause the rock to be wet.
- There are no other causal relationships.

You want to model this situation using Boolean random variables W (wet rock), R (rain), H (humidity), and S (sprinkling).

1. You do not know if it rained today and want to determine that by touching the rock. 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. Which of the variables are deterministic? 2 pt
5. Now assume your network is $W \to R \to H \to S$ (which *may or may not* be a correct solution to 3. above). How many entries do the conditional probability tables of that network have in total? 2 pt
6. Now assume a correct network is $W \to R \leftarrow H \leftarrow S$. Give the formula for 3 pt

$$P(R|W = true, S = true)$$

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

Solution:

1. Evidence W, query R, hidden H, S
2. Causes should occur before effects, so e.g., $RHSW$, $RSHW$, or $SRHW$.
3. $R \to H \to W \leftarrow S$ and $R \to W$
4. None
5. We need 2 for W: $P(W)$ and $P(\neg W)$. Then 4 for R: $P(R|W)$, $P(R|\neg W)$, $P(\neg R|W)$, and $P(\neg R|\neg W)$. Then accordingly for 4 each for H and S. So 14 in total. Because we can compute the negative probabilities from the positive ones, 7 is also an acceptable solution.
6.

$$P(R|w,s) =$$

$$\alpha\big(P(R,h,w,s) + P(R,\neg h,w,s)\big) =$$

$$\alpha\big(P(R|h,w,s)P(h|w,s)P(w|s)P(s) + P(R|\neg h,w,s)P(\neg h|w,s)P(w|s)P(s)\big) =$$

$$\alpha\big(P(R|h,w)P(h|s)P(w)P(s) + P(R|\neg h,w)P(\neg h|s)P(w)P(s)\big)$$

$$\alpha'\big(P(R|h,w)P(h|s) + P(R|\neg h,w)P(\neg h|s)\big)$$

where α, α' are constant factors that normalize the distribution. The penultimate line is an acceptable solution too.

Grading:

1. 0.5 points per variable
2. 2 points if minor mistake, 1 point if completely wrong but a variable ordering

3. Relative to solution of 2. 2.5 points if one edge missing or too much, 2 points if more mistakes, 1 point if not similar to solution but a Bayesian network.
4. No partial credit
5. Full credit for 14, 7, or any explained answer that partially drops the entries for negative probabilities. Partial credit if explanation for wrong answer: 1.5 or 1 depending on severity of mistake.
6. 2.5 points if minor mistake, 2 points if recognizable as the solution but wrong, 1 if recognizable, 0.5 mercy points

---

2 Markovian Reasoning

Problem 2.1 (Hidden Markov Models) 15 pt

0 min

Consider the following situation:

- You make weekly observations about your business with a client. Each week business is either good or bad.
- You know this is caused by the mood of your client, who feels either optimistic or pessimistic about the economy.
- You have previously obtained the following information:
  - when your client is optimistic, they remain optimistic next week 90% of the time,
  - when your client is optimistic, your business is good 70% of the time,
  - when your client is pessimistic, they remain pessimistic next week 25% of the time,
  - when your client is pessimistic, your business is bad 80% of the time,
  - your client was optimistic two weeks ago with probability 60%.

You want to model this situation as a hidden Markov model with Boolean random variables indexed by week number w.

1. Give the state and evidence variables. 2 pt
2. Is the model stationary? 1 pt
3. What order does the model have? 1 pt
4. Complete the following sentences:
   (a) The transition model is given by the matrix 3 pt
   $$T = \begin{pmatrix} \quad & \quad \\ \quad & \quad \end{pmatrix} \qquad \text{where} \qquad T_{ij} = P(\qquad = j \mid \qquad = i).$$
   (b) The sensor model is given by the matrix 3 pt
   $$M = \begin{pmatrix} \quad & \quad \\ \quad & \quad \end{pmatrix} \qquad \text{where} \qquad M_{ij} = P(\qquad = j \mid \qquad = i).$$

   To map Boolean values to matrix indices i, j, we use $true = 1$ and $false = 2$.

5 pt

5. Your business was good last week ($w = 1$) and bad this week ($w = 2$). 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 client's current mood at $w = 2$. (You do not have to actually compute the distribution.)

Solution:

1. State variables C_w (true if client optimistic), evidence variables B_w (true if business is good)

2. Yes

3. 1

4. (a) $T = \begin{pmatrix} 0.9 & 0.1 \\ 0.75 & 0.25 \end{pmatrix}$ where $T_{ij} = P(C_w = j | C_{w-1} = i)$

   (b) $M = \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix}$ where $M_{ij} = P(B_w = j | C_w = i)$

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
   - $O_1 = \begin{pmatrix} 0.7 & 0 \\ 0 & 0.2 \end{pmatrix}$ and $O_2 = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.8 \end{pmatrix}$ are the diagonal sensor matrices for the observation of good business at $w = 1$ and bad business at $w = 2$,
   - $f_{1:0} = \langle 0.6, 0.4 \rangle$ is the prior probability,
   - α is a constant factor to normalize the distribution.

Grading:

1. 1 point each; -0.5 for minor mistakes
2. no partial credit
3. no partial credit
4. -0.5 per minor mistake
5. 2 points for the filtering equation, 3 points for plugging in values

---

Problem 2.2 (Markov Decision Processes) 10 pt

0 min

1. Give an optimal policy π^* for the following MDP:
   - set of states: $S = \{0, 1, 2, 3, 4, 5\}$ with initial state 0
   - set of actions for $s \in S$: $A(s) = \{-1, 1\}$
   - transition model for $s, s' \in S$ and $a \in A(s)$: $P(s'|s, a)$ is such that
     - $s' = (s + a) \mod 6$ with probability $2/3$,
     - $s' = (s + 3) \mod 6$ with probability $1/3$.

- reward function: $R(5) = 1$ and $R(s) = -0.1$ for $s \in S \setminus \{5\}$

2. State the Bellman equation.

3. Complete the following high-level description of the value iteration algorithm:
   - The algorithm keeps a table $U(s)$ for $s \in S$, that is initialized with
   - In each iteration, it uses the

     in order to
   - $U(s)$ will converge to the

Solution:

1. $\pi^*(s) = 1$ if $s \in \{3, 4\}$ and $\pi^*(s) = -1$ if $s \in \{0, 1\}$ and arbitrary for $s \in \{2, 5\}$
2. $U(s) = R(s) + \gamma \max_{a \in A(s)} \sum_{s' \in S} U(s')P(s'|s, a)$
3. - The algorithm keeps a table $U(s)$ for $s \in S$, that is initialized with arbitrary values, e.g. all 0 or the rewards.
   - In each iteration, it uses the Bellman equation in order to update $U(s)$.
   - $U(s)$ will converge to the expected utility of s.

---

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. 0 min

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

4 pt

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

   1 pt

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

   3 pt

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

   3 pt

4. Using the above observations, give the results of

   - $I(P(Result)) =$
   - $P(Result = Loss|Opponent = Strong) =$

   You do not have to compute irrational logarithms.

   2 pt

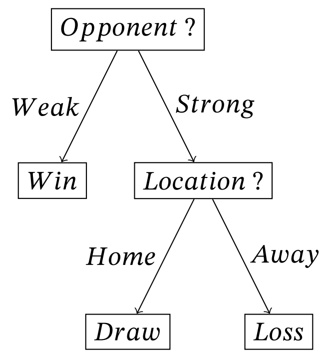
5. Give a minimal set A of attributes such that $A \succ Result$ holds for the above observations.

   2 pt

6. Explain *why or why not* the determination $Day, Weather \succ Result$ holds for the above observations.

Solution:

1. The tree is



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

3. $Gain(Opponent) = I(P(Result)) - P(Opponent = Strong) \cdot I(P(Result|Opponent = Strong)) - P(Opponent = Weak) \cdot I(P(Result|Opponent = Weak))$

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

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

6. It does not hold. Games 4 and 6 agree on Day and $Weather$ but not on $Result$.

Grading:

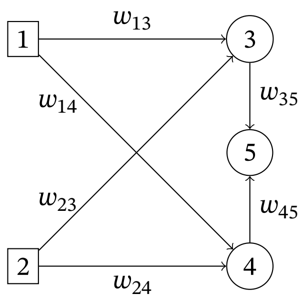
1. −0.5 for a decision list, −0.5 if for missing leaf node labels, or similar representation errors, −1 for including the redundant attributes, −0.5 for other mistakes; at least 3 points for an essentially correct solution, at least 2 points if recognizably similar to solution, 1 point if some decision tree
2. 0.5 points if only information gain is mentioned but not highest.
3. 2.5 points if minor mistake, 2 points if clearly recognizable as the solution, 1 point if some overlap with solution but very wrong, 0.5 mercy points if anything correct is written.
4. 1.5 points each. −0.5 per minor mistake
5. 1 point if off by one attribute
6. 1 point for the answer (no partial credit); 1 point for the explanation, −0.5 for mistakes

---

Problem 3.2 (Neural Networks) 15 pt

0 min

Consider the neural network below where units 1, 2 are inputs, unit 5 is output, weights are given by the labels on the edges, and units 3, 4, 5 are perceptron units with activation function $T(x) = 1$ for $x > 0.5$ and $T(x) = 0$ otherwise.



1 pt

1. Is it recurrent? 1 pt
2. How many hidden layers does it have? 3 pt
3. Give the formula for the activation a_3 of unit 3 in terms of the inputs a_1 and a_2 and the weights w_{ij}. 2 pt
4. Assume $w_{ij} = 1$ for all weights w_{ij} and $a_1 = a_2 = 1$. What is the resulting output a_5? 4 pt
5. Assume $a_1, a_2 \in \{0, 1\}$ and $w_{13} = w_{23} = 1$. Choose appropriate values for the other weights such that the network implements the XOR function, i.e., $a_5 = a_1 \, XOR \, a_2$. 4 pt
6. Complete the high-level description of the back-propagation algorithm *on the next page*. To learn a target function $a_5 = f(a_1, a_2)$, do the following for each input to f:
   - compute the
   - determine the error between

     and propagate
   - use the propagation results to update

Solution:

1. No
2. 1
3. $a_3 = T(w_{13}a_1 + w_{23}a_2)$
4. 1
5. $w_{14} = w_{24} = 0.3$ (range: each ≤ 0.5 and sum > 0.5), $w_{35} = 1$ (range: > 0.5), $w_{45} = -1$ (range $w_{35} + w_{45} \leq 0.5$)
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.

Grading:

1. no partial credit
2. no partial credit; 2 accepted if they list the 2 nodes in the hidden layer
3. -0.5 for minor mistakes, -1 for major mistakes; up to 1.5 points for wrong but vaguely recognizable solutions
4. solutions with bias accepted if used consistently; 1 point for wrong solution with good computation
5. solutions with bias accepted if used consistently; -1 for mistakes, at least 2 points if recognizable/correct intention, 1 point for well-typed solution
6. 1 point per blank; -0.5 for mistakes (in particular: missing computation of activations for all units)

4 Natural Language Processing

Problem 4.1 (Trigram Models) 5 pt

0 min

1. How many trigrams does a language with 10 words have? 1 pt
2. Explain informally how we can obtain a trigram model for a language L. 2 pt
3. Name two applications of trigram models. 2 pt

Solution:

1. $10^3 = 1000$
2. 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 of trigrams.
3. Language identification, genre classification, named entity recognition

Grading:

1. no partial credit
2. 1 point each for mentioning the corpus and the probability distribution of trigram occurrences
3. 1 point per correct answer (up to 2)

Problem 4.2 (Information Retrieval) 10 pt

0 min

Consider the following two texts

- d_1: "Information retrieval is hard."
- d_2: "Machine learning is very, very hard."

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

Below we use alphabetical order for the vector components:

hard, information, is, learning, machine, retrieval, very

Simplify all results 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 q be the query consisting of the word "retrieval". Give the value $tfidf(\text{retrieval}, q, D)$. 2 pt
4. How can we use $tfidf$ for choosing how to rank the texts in D for the query q? 2 pt

Solution:

1. $tf(\_, d_2) = \langle 1/6, 0, 1/6, 1/6, 1/6, 0, 1/3 \rangle$.
2. $idf(\_, \{d_1, d_2\}) = \log_{10}(2/\langle 2, 1, 2, 1, 1, 1, 1 \rangle) = \langle 0, k, 0, k, k, k, k \rangle$ with $k = \log_{10} 2$.
3. $tfidf(\text{retrieval}, q, D) = 1/1 \cdot idf(\text{retrieval}, D) = k$.
4. We compute the angle between the vectors $tfidf(\_, d_i, D)$ and $tfidf(\_, q, D)$. Lower values are ranked higher.

Grading:

1. -0.5 per wrong entry
2. -0.5 per wrong entry
3. graded relative to solutions to subproblem 2; -0.5 for minor mistakes, 1 point for roughly correct attempt
4. 1 point for mentioning the angle between the correct vector, 1 point for the ranking criterion

---