Name:

Birth Date:

Matriculation Number:

Exam
Artificial Intelligence 2

Feb. 18., 2021

	To be used for grading, do not write here												
prob.	1.1	1.2	1.3	2.1	2.2	3.1	3.2	4.1	4.2	4.3	4.4	Sum	grade
total	12	6	5	8	15	12	10	10	8	4	4	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.

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.

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.

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

Problem 1.1 (Medical Bayesian Network) 10 pt

Both Malaria and Meningitis can cause a fever, which can be measured by checking for a high body temperature. Of course you may also have a high body temperature for other reasons. We consider the following random variables for a given patient:

- Mal: The patient has malaria.
- Men: The patient has meningitis.
- HBT: The patient has a high body temperature.
- Fev: The patient has a fever.

Consider the following Bayesian network for this situation:

$$\begin{tikzcd}
 & Fev & \\
 & HBT \arrow[u] & \\
Mal \arrow[ur] & & Men \arrow[ul]
\end{tikzcd}$$

1. Explain the purpose of the edges in the network regarding the conditional probability table.
2. What would have happened if we had constructed the network using the variable order Mal, Men, Fev, HBT? Would that have led to a better network?
3. How do we compute the probability the patient has malaria, given that he has a fever? State the query variables, hidden variables and evidence and write down the equation for the probability we are interested in.

Solution:

1. The parents (i.e, nodes from which there are incoming edges) of X are the variables that X may depend on. The conditional probability table for X must take all of those as additional inputs.
2. We would have obtained the same network, but with a causal instead of a diagnostic one. That would be better. That would be a worse network because more edges increase the complexity.
3. Query variable: Mal. Evidence: Fev. Hidden variables: Men, HBT. We get:

$$P(Mal|Fev) = \alpha \sum_{v_{HBT}, v_{Men}} P(Mal) \cdot P(v_{Men}) \cdot P(v_{HBT}|Mal, v_{Men}) \cdot P(F|v_{HBT})$$

Problem 1.2 (Bayesian Rules) 8 pt

For each of the following principles, state the key formula and explain the variables occurring in it:

1. Bayes rule

2. Marginalization

Solution: See lecture notes.

Problem 1.3 (Stochastic and Conditional independence) 6 pt

Consider the following random variables:

- three flips C_1, C_2, and C_3 of the same fair coin, which can be heads or tails
- the variable E which is 1 if both C_1 and C_2 are heads and 0 otherwise
- the variable F which is 1 if both C_2 and C_3 are heads and 0 otherwise

Out of the above 5 random variables,

1. give three random variables X, Y, Z such that X and Y are stochastically independent but not conditionally independent given Z,

2. give three random variables X, Y, Z such that X and Y are not stochastically independent but conditionally independent given Z.

Solution:

1. E.g., C_1 and C_2 with $Z = E$.
2. E.g., E and F with $Z = C_2$.

2 Decision Theory

Problem 2.1 (Expected Utility) 6 pt

What is the formal(!) definition of *expected utility*? Explain every variable in the defining equation.

Solution: The expected utility EU is defined as

$$EU(a|e) = \sum_{s'} P(R(a) = s'|a, e) \cdot U(s')$$

where

1. a is the action for which we want to find out the expected utility, given the evidence e.

2. $U(s')$ is the utility of a state s'.

3. $R(a)$ is the result of the action a.

Problem 2.2 (Textbook Decisions) 15 pt

Abby has to decide whether to buy Russell&Norvig for 100\$. There are three boolean variables involved in this decision: B indicating whether Abby buys the book, M indicating whether Abby knows the material in the book perfectly anyway and P indicating that Abby passes the course. Additionally, we use a utility node U.

Abby's utility function is additive, so $U(B) = -100$. Furthermore, she evaluates passing the course with a utility of $U(P) = 2000$. The course has an open book final exam, so B and P are not independent given M.

Assume the conditional probabilities $P(P|B, M)$, $P(P|B, \neg M)$, $P(P|\neg B, M)$, $P(P|\neg B, \neg M)$, $P(M|B)$, $P(M|\neg B)$ are given.

1. Draw a good decision network for this problem.
2. Explain precisely how to compute the utility of buying the book.

Solution:

3 Markov Models

Problem 3.1 (Markov Decision Procedures) 12 pt

1. What are the mathematical components of an unambiguous Markov decision procedure?
2. What is the Bellman equation and what algorithm is it used for? How does that algorithm work?
3. What is the difference between *partially observable* MDPs and normal MDPs?

Solution:

1. A set S of states, a set A_s of actions for each state $s \in S$, a transition model $T(s_1, a, s_2) := P(s_2|s_1, a)$ for $a \in A_{s_1}$, and a reward function $R : S \to \mathbb{R}$.
2. Value iteration: We assign a random utility to each state and update them using the Bellman equation:
$$U(s) = R(s) + \gamma \cdot \max_a \left(\sum_{s'} U(s') \cdot T(s, a, s') \right)$$
Once this iteration has converged, we can compute the "best" action for each state by considering the expected utilities of all possible actions.
3. Current state is unknown; instead we have observables and a sensor model $O(s, e) := P(e|s)$ for observables e and states s.

Problem 3.2 (Stock Market Predictions) 10 pt

You bought SpaceY stock recently and try to predict whether to buy more or sell. The stock market is in one of two possible states; bull state or bear state. In a bull state, it will (in the long term) be advantageous to buy stock; in a bear state it will be more advantageous to sell.

If the market is in a bull state, the probability it will still be in a bull state tomorrow is 60%. If it is in a bear state, the probability it will remain so tomorrow is 80%.

1. If we consider this as a hidden Markov model, what are the random variable, its domain, and its transition matrix T?
2. In terms of the stock market example, explain what probabilities are computed by *prediction*, *filtering* and *smoothing*. You do not need to give the formulas.

Solution:

1. We take X_t to be a discrete random variables with domain $\{\text{bull}, \text{bear}\}$.

$$T = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix}$$

2. **Prediction** Given the *behavior* of the stock market up to time t_0, compute the probability of the state the stock market will be in at time $t_1 > t_0$

   Filtering Given the behavior of the stock market up to now, compute the probability of the state the stock market is in right now

   Smoothing Given the behavior of the stock market up to t_0, compute the probability that the stock market was in some state at an earlier point $t_1 < t_0$.

4 Learning

Problem 4.1 (Sunbathing) 12 pt

Eight people go sunbathing. They are categorized by the attributes Hair and Lotion and the result of whether they got sunburned.

Name	Hair	Lotion	Result: Sunburned
Sarah	Light	No	Yes
Dana	Light	Yes	No
Alex	Dark	Yes	No
Annie	Light	No	Yes
Julie	Light	No	No
Pete	Dark	No	No
John	Dark	No	No
Ruth	Light	No	No

1. Which quantity does the information theoretic decision tree learning algorithm use to pick the attribute to split on?
2. Compute that quantity for the attributes Hair and Lotion. (Simplify as much as you can without computing logarithms.)
3. Assuming the logarithms are computed, how does the algorithm pick the attribute?

Solution:

1. Information gain.
2. $$E_0 := I(\langle\frac{2}{8}, \frac{6}{8}\rangle) = -\frac{2}{8}\log_2(\frac{2}{8}) - \frac{6}{8}\log_2(\frac{6}{8}) \approx 0.81$$

$$\mathtt{Gain(Hair)} = E_0 - \underbrace{\frac{5}{8}I(\langle\frac{2}{5}, \frac{3}{5}\rangle)}_{\mathtt{Light}} - \underbrace{\frac{3}{8}I(\langle 0, 1\rangle)}_{\mathtt{Dark}} \qquad \approx 0.20$$

$$\mathtt{Gain(Lotion)} = E_0 - \underbrace{\frac{2}{8}I(\langle 0, 1\rangle)}_{\mathtt{Yes}} - \underbrace{\frac{6}{8}I(\langle\frac{2}{6}, \frac{4}{6}\rangle)}_{\mathtt{No}} \qquad \approx 0.12$$

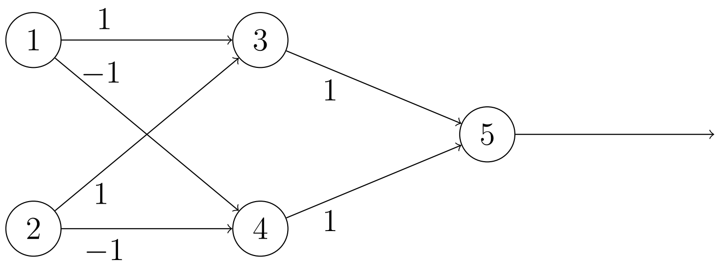
3. It picks the one with the highest information gain (in this case Hair).

Problem 4.2 (XOR Neural Network) 8 pt

Consider the following neural network with

- inputs a_1 and a_2
- units $3, 4, 5$ with activation functions such that $a_i \leftarrow \begin{cases} 1 & \text{if } \Sigma_j w_{ji} a_j > b_i \\ 0 & \text{otherwise} \end{cases}$

- weights w_{ij} as given by the labels on the edges



5 pt

1. Assume $b_1 = b_2 = b_3 = 0$ and inputs $a_1 = a_2 = 1$. What are the resulting activations a_3, a_4, and a_5? 3 pt
2. Choose appropriate values for b_1, b_2, and b_3 such that the network implements the XOR function.

Solution:

1. $a_3 = 1$, $a_4 = 0$, $a_5 = 1$
2. $b_1 = 0.5$, $b_2 = -1.5$, $b_3 = 1.5$

Problem 4.3 (Overfitting) 3 pt

Explain what *overfitting* means and why we want to avoid it.

Solution: Overfitting is a modeling error that occurs when a function is too closely fit to a limited set of data points. Overfitting the model generally takes the form of making an overly complex model to explain idiosyncrasies in the data under study rather than the underlying mechanisms.

Problem 4.4 (Information Theory) 5 pt

Consider the learning curve, which gives the percentage of correct answers on the test set in terms of the size of the training set.

1. Informally describe the typical shape of this curve.
2. In practice, does the correctness commonly reach 100%? Briefly justify your answer.

Solution:

1. The curve starts at low values and increases towards approximating 100%. It does not monotonely but jitters upwards.
2. No. If the function f we are approximating is not in the hypothesis space, then even the best approximation can never be 100% correct.