Name:

Birth Date:

Matriculation Number:

Exam
Artificial Intelligence 2

August 11., 2020

	To be used for grading, do not write here														
prob.	1.1	1.2	1.3	2.1	2.2	2.3	2.4	3.1	3.2	4.1	4.2	4.3	4.4	4.5	Sum	grade
total	6	10	2	6	4	6	9	12	3	10	2	4	8	3	85	
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

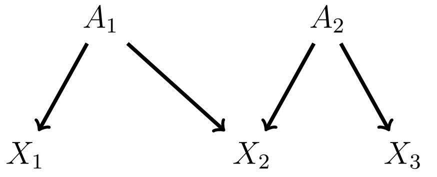
Problem 1.1 (Basic Probability) 6 pt
Let A, B, C be Boolean random variables, and let a, b, c denote the atomic events that A, B, C, respectively, are true. Which of the following equalities are always true? Justify each of your answers in one sentence. 6 min

1. $P(b) = P(a, b) + P(\neg a, b)$
2. $P(a) = P(a|b) + P(a|\neg b)$
3. $P(a, b) = P(a) \cdot P(b)$
4. $P(a, b|c) \cdot P(c) = P(c, a|b) \cdot P(b)$
5. $P(a \vee b) = P(a) + P(b)$
6. $P(a, \neg b) = (1 - P(b|a)) \cdot P(a)$

Solution:

1. True (marginalization over A)
2. Not true (e.g. $P(a|b) = P(a|\neg b) = 0.6$ would result in $P(a) = 1.2$)
3. Not true (only true if A and B are stochastically independent)
4. True (using product rule, both sides become $P(a, b, c)$)
5. Not true (general form is $P(a \vee b) = P(a) + P(b) - P(a, b)$)
6. True ($1 - P(b|a) = P(\neg b|a)$ and via product rule we get $P(a, \neg b)$)

Problem 1.2 (Bayesian Networks) 10 pt
Consider the following Bayesian network with Boolean variables: 10 min



2 pt

1. Give the definition of conditional independence. 3 pt
2. Which nodes in the network are conditionally independent with X_1 given A_1? Explain why. 3 pt
3. Give an example of two nodes from the network above that are stochastically independent. Explain why they are stochastically independent.

2 pt

4. What exactly (formal criterion) does an arrow between two nodes in a Bayesian network mean for the associated events?

Solution:

1. Two events A, B are conditionally independent given C, if $P(A \wedge B|C) = P(A|C)P(B|C)$.
2. Every node in a Bayesian network is conditionally independent of its non-descendants given its parents. This means that X_2, X_3 are conditionally independent of X_1 given A_1
3. The nodes A_1 and A_2 have no parents, so they are in fact stochastically independent of their non-descendants. The non-descendants of A_1 are A_2 and X_3, and the non-descendants of A_2 are A_1 and X_1. This means that the pairs A_1 and A_2, A_1 and X_3, A_2 and X_1.

   Note that it would be also correct to say that X_1 and X_3 are stochastically independent by *D-separation*[1], but this was not covered in the lectures.
4. We draw an arrow from X_j to X_i if X_j is in the smallest set $\texttt{Parents}(X_i)$ with the property $P(X_i|X_{i-1}, ..., X_1) = P(X_i|\texttt{Parents}(X_i))$

Problem 1.3 (Arrows in Bayesian Networks) 2 pt

Suppose that in a Bayesian network N we have variables C_1, E_1, C_2, and E_2, such that C_1 causes E_1 and C_2 causes E_2. 2 min

1. How do we call an arrow going from C_1 to E_1?
2. How do we call an arrow going from E_2 to C_2?

Solution:

1. Causal.
2. Diagnostic

---

2 Decision Theory

Problem 2.1 (Expected Utility) 6 pt

6 min

1. What is the formal(!) definition of *expected utility*? What is the meaning of every variable in the defining equation?
2. How do we use expected utility to make decisions?

Solution:

[1] https://de.wikipedia.org/wiki/D-Separation

1. The expected utility EU is defined as

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

   where

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

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

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

2. The principle of maximum expected utility says that a rational agent should choose the action that maximizes the agent's expected utility.

Problem 2.2 (The Value of Information) 4 pt 4 min

Chef Giordana runs a kitchen that provides food for a large organisation. A salad is sold for €6 and costs €4 to prepare. Therefore, the contribution per salad is €2. At present Giordana must decide in advance how many salads to prepare each day (40 or 60). Actual demand will also be 40 or 60 each day. So Giordana's payoff table looks as follows:

Demand	Probability	40 salads	60 salads
40	0.4	€80	€0
60	0.6	€80	€120

Thus, the expected utility for making 40 salads is 80 and the expected utility for making 60 salads is 72. Based on these expected values without additional information, Giordana would choose to make 40 salads per day with an EU of €80 per day.

She is considering a new ordering system, where the customers must order their salad online the day before. With this new system Giordana will know for certain the daily demand 24 hours in advance. She can adjust production levels on a daily basis. How much is this system worth to her (per day)?

Task: Compute the concrete value in € and explain what you did.

Solution: The value of information is equal to the expected value of best action given the information minus expected value of best action without information. The corresponding formula for the value of perfect information is

$$\mathrm{VPI}_E(E_j) = \sum_k P(E_j = e_{jk}|E) \cdot \mathrm{EU}(\alpha_{e_jk}|E, E_j = e_{jk}) - \mathrm{EU}(\alpha|E)$$

In Giordana's case this amounts to

$$\mathrm{VPI} = (0.4 \cdot 80 + 0.6 \cdot 120) - 80 = (32 + 72) - 80 = 24$$

so the system is worth at most €24 per day.

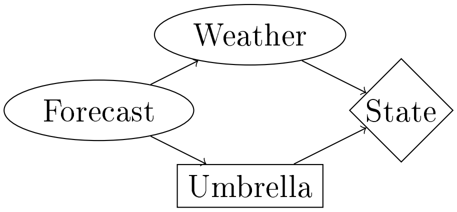
Problem 2.3 (Decision Network) 6 pt 6 min

You try to decide on whether to take an umbrella to Uni. Obviously, it's useful to do so if it rains when you go back home, but it's annoying to carry around if it doesn't even rain. Here are the states you could end up in:

- happy (or relieved) if it *doesn't rain* and you *did not bring an umbrella*,
- annoyed if it *doesn't rain* and you *brought an umbrella*,
- wet if it *rains* and you *did not bring an umbrella*,
- dry if it *rains* and you *brought an umbrella.*

You look at the weather forecast, which has two possible values: sunny and rainy.

You come up with this decision network:



1. Decision networks are extensions of Bayesian networks, which additional kinds of nodes do decision networks have? For each kind give an example from the network above. 2 pt
2. How would you compute whether or not to take an umbrella, assuming you know all of the probabilities $P(\text{state} = s|\text{forecast} = f, \text{umbrella} = u)$ for all 4 pt
    - $s \in \{\text{happy}, \text{annoyed}, \text{wet}, \text{dry}\}$,
    - $u \in$ `Bool`, and
    - $f \in \{\text{sunny}, \text{rainy}\}$.

Solution:

Problem 2.4 (Markov Decision Procedures) 9 pt

8 min

1. How do Markov decision procedures differ from (simple) decision networks? 3 pt
2. What do we use the value iteration and policy iteration for? How do they differ? 4 pt
3. What is the difference between *partially observable Markov decision procedures* and normal MDPs? 2 pt

Solution:

1. In Markov decision procedures, the probabilistic model is a Markov process (i.e. random variables are indexed over time, transitions are subject to the Markov properties).

2. Value iteration iterates over utilities until they converge, obtaining an optimal policy. Policy iteration alternates the steps of policy evaluation (computing utilities given the current policy) and policy improvement (computing a new policy). The policy resulting from value iteration can be stable long before the individual utilities have converged to their precise values.

3. In POMDPs the current state is unknown; instead we have observables and a sensor model $O(s, e) := P(e|s)$ for observables e and states s.

---

3 Markov Models

Problem 3.1 (Stock Market Predictions) 12 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. 12 min

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%.

If the market is in a bull state, the probability that your stock will rise that day is 90%. If it is in a bear state, your stock will more likely fall (with 60% probability).

1. What are the observable and unobservable variables in this model? 1 pt

2. If we consider this as a hidden Markov model, what is its transition matrix T? Remember that we use transition matrices to compute the previous or future states. 1 pt

3. Explain what kind of probabilities *prediction*, *filtering* and *smoothing* compute in this scenario. Do not just give formulas. 6 pt

4. Give the underlying equations for the first two of these algorithms and explain what each variable in the equation represents. 4 pt

---

Solution:

1.

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

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

3. **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 sta stock market was in some state at an earlier point $t_1 < t_0$.

4. We have $P(X_{t+1}|e_{1:t+1}) = \alpha P(e_{t+1}|X_{t+1}) \cdot \sum_{x_t} P(X_{t+1}|x_t, e_{1:t}) \cdot P(x_t|e_{1:t})$ where X represents the state and e the behavior of the stock market.

Problem 3.2 (Stationary) 3 pt

Define what it means for a Markov model to be *stationary*, and why we are interested in stationarity. 3 min

Solution: A Markov process is called stationary if the transition model is independent of time, i.e. $\mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{t-1})$ is the same for all t.

We like stationary Markov processs, since they are finite.

4 Learning

Problem 4.1 (Decision List) 10 pt

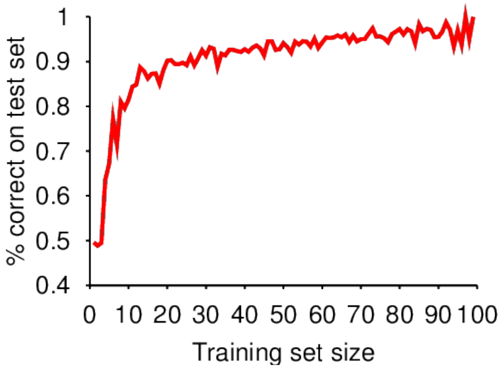
Construct a decision list to classify the data below. The tests should be as small as possible (in terms of attributes), breaking ties among tests with the same number of attributes by selecting the one that classifies the greatest number of examples correctly. If multiple tests have the same number of attributes and classify the same number of examples, then break the tie using attributes with lower index numbers (e.g., select A_1 over A_2). 10 min

Example	A_1	A_2	A_3	A_4	y
x_1	1	0	0	0	1
x_2	1	0	1	1	1
x_3	0	1	0	0	1
x_4	0	1	1	0	0

Solution:

Problem 4.2 (Information Theory) 2 pt

Explain why it is possible (even common) that the learning curve (example given below) never gets to 100% correctness, even for large example sets. 2 min



Solution: If the function f we are approximating is not in the hypothesis space, then even the best approximation can never be 100% correct.

Problem 4.3 (Information Entropy) 4 pt

Explain and define *information entropy.* 4 min

Solution: Information entropy of a (set of) random variable is the average level of "information", "surprise", or "uncertainty" inherent in the variable's possible outcomes.

$$I(\langle P_1, ..., P_n\rangle) = \sum_{i=1}^{n} -P_i \log_2(P_i)$$

Problem 4.4 (Sunbathing) 8 pt

Eight people go sunbathing. Some of them got a sunburn, others didn't: 6 min

Name	Hair	Lotion	Result
Sarah	Light	No	Sunburned
Dana	Light	Yes	None
Alex	Dark	Yes	None
Annie	Light	No	Sunburned
Julie	Light	No	None
Pete	Dark	No	None
John	Dark	No	None
Ruth	Light	No	None

1. Which quantity does the decision tree learning algorithm use to pick the attribute to split on? Write down the formula for it. 2 pt

2. Compute it for for the attributes Hair and Lotion. It is enough to give the formula and insert the correct values for the variables, you do not need to compute the final value. 4 pt

3. Which one would the algorithm pick for the next step? Explain what happens next. 2 pt

Note that *Name* is only an index, not a (meaningful) attribute!

Solution:

1. Information gain.

2.

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

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

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

3. `Hair` has the highest information gain, so we split here. All table entries with `Dark` have result `None`, so we continue with `Hair` = `Light`

Problem 4.5 (Overfitting) 3 pt

Explain what *overfitting* means and why we want to avoid it. 3 min

Solution: Overfitting is a modeling error that occurs when the chosen hypothesis is too closely fit to a sample set of data points. It picks an overly complex hypothesis that also explains idiosyncrasies and errors in the data. A simpler hypothesis that fits the data less exactly is often a better match for the underlying mechanisms.