Name:

Birth Date:

Matriculation Number:

Exam
Artificial Intelligence 2

July 30., 2019

	To be used for grading, do not write here													
prob.	1.1	1.2	2.1	2.2	3.1	3.2	3.3	3.4	4.1	4.2	5.1	5.2	Sum	grade
total	7	5	4	10	10	4	8	3	10	8	5	6	80	
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 2) 7 pt

Both smoking and living in a city with high air pollution can cause lung cancer, which can be indicated by a patient coughing up blood. We consider the following random variables for a given patient: 7 min

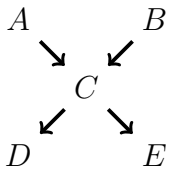
- *Smoke*: The patient is a smoker.
- *Smog*: The patient lives in a polluted city.
- *Blood*: The patient is coughing up blood.
- *LC*: The patient has lung cancer.

1. Draw the corresponding Bayesian network for the above data using the algorithm presented in the lecture, assuming the variable order $Smoke, Smog, Blood, LC$. Explain rigorously(!) the exact criterion for whether to insert an arrow between two nodes.
2. Which arrows are causal and which are diagnostic? Which order of variables would be better suited for constructing the network?
3. How do we compute the probability the patient is a smoker, given that they have lung cancer? State the query variables, hidden variables and evidence and write down the equation for the probability we are interested in.

Solution:

Problem 1.2 (Stochastic and Conditional independence) 5 pt

Consider the following Bayesian network: 5 min



Find

1. two variables that are stochastically independent, but not conditionally independent for some condition, and
2. two variables that are **not** stochastically independent, but conditionally independent for some condition.

For both pairs, state the condition explicitly and justify why they are correct answers. Assume that the conditional probabilities modelled by this network are not degenerate or exceptional in any way (e.g. no 0 or 1 entries).

Solution:

1. A, B are independent (no arrow between them), but not conditionally independent given C. For example, suppose that you are flipping two coins. Let A be the event that your first coin flip is heads and B the event that your second coin flip is heads. Let C be the event that your first two flips were the same. Then A and B here are independent. However, A and B are conditionally dependent given C, since if you know C then your first coin flip will inform the other one.

2. D, E are not independent (both are influenced by C), but conditionally independent given C. Now suppose that D and E are again coin flips (first coin flip is heads and second coin flip is heads), and C the event that I gave you a biased coin. D and E here are dependent, since flipping a coin once will give you evidence on whether or not the coin is biased. This evidence will help inform your next flip. However the flips are conditionally independent given C.

2 Decision Theory

Problem 2.1 (Expected Utility) 4 pt

What is the formal(!) definition of *expected utility*? Explain every variable in the defining equation. 4 min

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) 10 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. 10 min

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) 10 pt

10 min

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 (Prediction, Filtering, Smoothing) 4 pt

Explain the goals of *prediction, filtering* and *smoothing* in terms of conditional probabilities 4 min

Solution:

Prediction $P(X_{t+k}|e_{1:t})$

Filtering $P(X_t|e_{1:t})$

Smoothing $P(X_k|e_{1:t})$ for $0 \leq k < t$

Problem 3.3 (Markov Mood Detection) 8 pt

8 min

On any given day d, your roommate Moody is in one of two states – either he is happy (H_d) or he is in a bad mood (B_d). Usually when he's in a bad mood, it's because he had a fight with his boyfriend and those tend to go on for a couple of days, so $P(B_{d+1}|B_d) = 0.7$, but aside from that he's a cheery guy, so ($P(H_{d+1}|H_d) = 0.85$).

Of course you try to avoid talking to people, but you can hear his music blasting all day which tends to shift depending on his mood. On a good day he usually listens to Jazz (i.e. $P(J_d|H_d) = 0.7$), on a bad day he slightly prefers Death Metal ($P(DM_d|B_d) = 0.6$). He has a limited taste in music, so it's always one of the two.

You know that he was in a good mood at day d_0. Assume he's been listening to death metal for n days straight since then. Explain how to compute the probability that he is in a bad mood on day d_{n+1}. State the equations underlying this algorithm explicitly.

Solution: We have $P(H_0) = 1$ and

$$\langle P(H_d), P(B_d)\rangle = \langle P(H_d|H_{d-1}) + P(H_d|B_{d-1}), P(B_d|H_{d-1}) + P(B_d|B_{d-1})\rangle$$

which allows us to update using the information DM_d:

$$\langle P(H_d|DM_d), P(B_d|DM_d)\rangle = \alpha\langle P(DM_d|H_d)P(H_d), P(DM_d|B_d)P(B_d)\rangle$$

Problem 3.4 (Stationary) 3 pt

3 min

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

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 (Home Decisions) 10 pt

10 min

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

Name	Hair	Height	Weight	Lotion	Result
Sarah	Blonde	Average	Light	No	Sunburned
Dana	Blonde	Tall	Average	Yes	None
Alex	Brown	Short	Average	Yes	None
Annie	Blonde	Short	Average	No	Sunburned
Julie	Blonde	Average	Light	No	None
Pete	Brown	Tall	Heavy	No	None
John	Brown	Average	Heavy	No	None
Ruth	Blonde	Average	Light	No	None

Explain how the information-theoretic decision tree learning algorithm would proceed on this table (up to two iterations). Explicitly state how to compute the information gain (and what that means).

Note that you do not need to compute any actual values; if it is helpful for your explanation, you may guess any values you might want to use.

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

Solution:

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

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

$$\texttt{Gain(Height)} = E_0 - \underbrace{\frac{4}{8}I(\langle\frac{1}{4}, \frac{3}{4}\rangle)}_{\texttt{Average}} - \underbrace{\frac{2}{8}I(\langle 0, 1\rangle)}_{\texttt{Tall}} - \underbrace{\frac{2}{8}I(\langle\frac{1}{2}, \frac{1}{2}\rangle)}_{\texttt{Short}} \approx 0.16$$

$$\texttt{Gain(Weight)} = E_0 - \underbrace{\frac{3}{8}I(\langle\frac{1}{3}, \frac{2}{3}\rangle)}_{\texttt{Average}} - \underbrace{\frac{3}{8}I(\langle\frac{1}{3}, \frac{2}{3}\rangle)}_{\texttt{Light}} - \underbrace{\frac{2}{8}I(\langle 0, 1\rangle)}_{\texttt{Heavy}} \approx 0.12$$

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

`Hair` has the highest information gain, so we split here. All table entries with `Brown` have result `None`, so we continue with `Hair` = `Blonde`:

$$E_1 := I(\langle\frac{2}{5}, \frac{3}{5}\rangle) \approx 0.97$$

$$\texttt{Gain(Height)} = E_1 - \underbrace{\frac{3}{5}I(\langle\frac{1}{3}, \frac{2}{3}\rangle)}_{\texttt{Average}} - \underbrace{\frac{1}{5}I(\langle 0, 1\rangle)}_{\texttt{Tall}} - \underbrace{\frac{1}{5}I(\langle 1, 0\rangle)}_{\texttt{Short}} \approx 0.42$$

$$\texttt{Gain(Weight)} = E_1 - \underbrace{\frac{2}{5}I(\langle\frac{1}{2}, \frac{1}{2}\rangle)}_{\texttt{Average}} - \underbrace{\frac{3}{5}I(\langle\frac{1}{3}, \frac{2}{3}\rangle)}_{\texttt{Light}} - \underbrace{0}_{\texttt{Heavy}} \approx 0.02$$

$$\texttt{Gain(Lotion)} = E_1 - \underbrace{\frac{1}{5}I(\langle 0, 1\rangle)}_{\texttt{Yes}} - \underbrace{\frac{4}{5}I(\langle\frac{2}{4}, \frac{2}{4}\rangle)}_{\texttt{No}} \approx 0.17$$

`Height` has the highest information gain, so we proceed here. All short blondes are sunburned, all tall blondes are not, hence we only need consider `Average`...

Problem 4.2 (Backpropagation)

8 pt

8 min

Explain what *Backpropagation* means in the context of Neural Networks, when and why we need it, and how to do it using an example.

Solution: A possible answer:

Backpropagation is an algorithm for training feedforward neural networks for supervised learning. It computes the gradient of the loss function with respect to the weights of the network for a single input–output example.

5 Communication with Natural Language

Problem 5.1 (Ambiguity) 5 pt

5 min

1. Explain the concept of ambiguity of natural languages.
2. Give two examples of different kinds of ambiguity and explain the readings.

Solution:

1. Ambiguity is the phenomenon that in natural languages a single utterance can have multiple readings.
2. Here are some examples
   - *bank* can be a financial institution or a geographical feature.
   - In *I saw her duck* the word *duck* can be a verb or a noun.
   - *Time flies like an arrow* could be about the preferences of special insects (*time flies*) or about the fact that time passes quickly – e.g. in an exam.
   - In *Peter saw the man with binoculars*, it could be Peter who is using binoculars, or it could be that Peter saw *the man* who had *binoculars*.

Problem 5.2 (Language Identification) 6 pt

6 min

You are given an English, a German, a Spanisch, and a French text corpus of considerable size, and you want to build a language identification algorithm A for the EU administration. Concretely A takes a string as input and classifies it into one of the four languages $\ell^* \in \{English, German, Spanish, French\}$. The prior probability distribution for the strings being English/German/Spanisch/French, is $\langle 0.4, 0.2, 0.15, 0.15\rangle$.

How would you proceed to build algorithm A? Specify the general steps and give/derive the formula for computing ℓ given a string $\mathbf{c}_{1:N}$.

Solution:

1. Build a trigram language model $\mathbf{P}(\mathrm{c}_i \mid \mathbf{c}_{i-2:i-1}, \ell)$ for each candidate language ℓ by counting trigrams in a ℓ-corpus.

2. Apply Bayes' rule and the Markov property to get the most likely language:

$$\begin{aligned}
\ell^* &= \operatorname*{argmax}_{\ell}(P(\ell \mid \mathbf{c}_{1:N})) \\
&= \operatorname*{argmax}_{\ell}(P(\ell) \cdot P(\mathbf{c}_{1:N} \mid \ell)) \\
&= \operatorname*{argmax}_{\ell}(P(\ell) \cdot \prod_{i=1}^{N} P(\mathrm{c}_i \mid \mathbf{c}_{i-2:i-1}, \ell))
\end{aligned}$$
