Last Name: First Name:

Matriculation Number:

Seat:

Final Exam
Künstliche Intelligenz 2

Aug 01., 2017

	To be used for grading, do not write here													
prob.	1.1	1.2	1.3	1.4	2.1	2.2	2.3	3.1	3.2	3.3	4.1	4.2	4.3	Sum	grade
total	8	6	14	8	8	8	20	8	16	8	8	6	20	138	
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

Problem 1.1 (Bayesian Rules) 8 pt

Name four of the basic rules in Bayesian inference and explain each with a short sentence and formula. 8 min

---

Solution:

1. Bayes rule (compute $P(A|B)$ from $P(B|A)$,
2. Normalization (Fixing evidence e, updating the probabilities of all other events using a normalization constant α),
3. Marginalization ($P(A) = \sum_y P(A, y)$),
4. Chain rule ($P(A_1, \ldots, A_n) = P(A_n|A_{n-1}, \ldots, A_1) \cdot P(A_{n-1}|A_{n-2}, \ldots, A_1) \cdot \ldots$)
5. Product rule ($P(A, B) = P(A|B)P(B)$)
6. Conditional Independence (Not really bayesian inference, but rather bayesian networks, but we'll be lenient)

Problem 1.2 (Conditional Independence) 6 pt

Define *conditional independence.* 6 min

Solution: Two events A, B are conditionally independent given C, if $P(A \wedge B|C) = P(A|C)P(B|C)$.

Problem 1.3 (Nuclear Test) 14 pt

Assume it is your responsibility to monitor the Nuclear Test Ban treaty. You receive data from two different stations (seismometers), S_1 and S_2. Each S_i is modeled as a Boolean variable where “true” stands for “I detected a Nuclear test” and “false” stands for “I did not detect a Nuclear test”. The seismometers are not fully reliable, however; they may not detect a Nuclear test even though there was one, and they may mistake an earthquake for a Nuclear test. We model this situation with two additional Boolean variables: N for Nuclear test, and E for Earthquake. 14 min

Use the algorithm from the lecture to construct a Bayesian network for these 4 variables. More precisely:

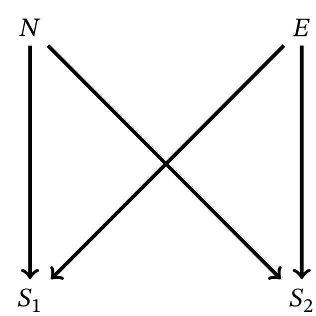
1. State the exact formal condition for when the algorithm inserts an edge between two nodes. 6 pt
2. Execute the algorithm for the variable order $X_1 = N$, $X_2 = E$, $X_3 = S_1$, $X_4 = S_2$. 8 pt

Justify your decisions.

Solution:

(a) $P(X_i|X_{i-1},...,X_1) = P(X_i|\texttt{Parents}(X_i))$

(b) With this variable order, we get the following network:



$X_2 = E$ does not need $X_1 = N$ as a parent because Earthquakes are independent from Nuclear tests. $x_3 = S_1$ needs both $X_1 = N$ and $X_2 = E$ as parents because each of these may influence the measurement; same for $X_4 = S_2$, i.e., here we also need the parents $X_1 = N$ and $X_2 = E$. However, given the values of N and E, the measurements of $X_3 = S_1$ and $x_4 = S_2$ are independent. So $X_4 = S_2$ does not require the parent $X_3 = S_1$.

Problem 1.4 (Causal and Diagnostic) 8 pt

State the difference between *causal* and *diagnostic* edges in (e.g.) a Bayesian network. Use no more than four sentences. 8 min

2 Decision Theory

Problem 2.1 (Decision Preferences) 8 pt

8 min

1. Name and state three of the axioms for *preferences* (i.e. $\succ$).
2. How are preferences related to *value functions*?

Solution:

1. Pick any three of the following:

   Orderability: $A \succ B \vee B \succ A \vee A \sim B$

   Transitivity: $A \succ B \wedge B \succ C \Rightarrow A \succ C$

   Continuity: $A \succ B \succ C \Rightarrow (\exists p.[p, A; 1-p, C] \sim B)$

   Substitutability: $A \sim B \Rightarrow [p, A; 1-p, C] \sim [p, B; 1-p, C]$

   Monotonicity: $A \succ B \Rightarrow (p > q) \Leftrightarrow [p, A; 1-p, B] \succeq [q, A; 1-q, B]$

   Decomposability: $[p, A; 1-p, [q, B; 1-q, C]] \sim [p, A; (1-pq), B; (1-p(1-q)), C]$

2. Ramsey's theorem states that given a set of preferences that obey the constraints above, there is a *value function* U with

$$(U(A) \geq U(B)) \Leftrightarrow A \succeq B \quad \text{and} \quad U([p_1, S_1; \ldots; p_n, S_n]) = \sum_i p_i U(S_i)$$

Problem 2.2 (Expected Utility) 8 pt
What is the formal(!) definition of *expected utility*? Explain every variable in the defining equation. 8 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 2.3 (Textbook Decisions) 20 pt 20 min

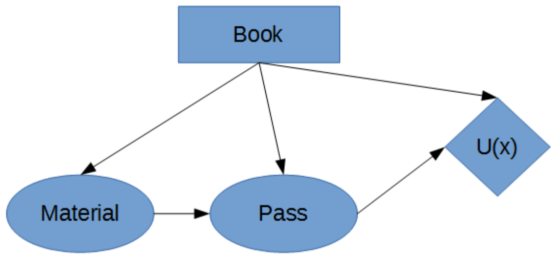
Abby has to decide whether to buy Russell&Norvig for 100$. There are three boolean variables involved in this decision: B indicating whether Abby buys the book, M indicating whether Abby knows the material in the book, and P indicating that Abby passes the course. Additionally, we use a utility node 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 probability tables containing the 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 the decision network for this problem.
2. Explain precisely how to compute the utility of buying the book.

Solution:



P	B	Utility
F	F	0
F	T	-100
T	F	2000
T	T	1900

B	P(B)
F	0.7
T	0.9

B	M	P(P)
F	F	0.3
F	T	0.8
T	F	0.5
T	T	0.9

The above picture has a typo: It should be $P(M)$, not $P(B)$ in the table on the bottom left. The picture includes concrete numbers as examples for how the calculation would be done. But the problem statement requires only to give the formulas and explain that we pick the decision that leads to the higher utility.

We set the decision variable B to each value and in each case:

- calculate the probability of passing/failing the exam
- use those probabilities to calculate the expected utility of the decision

Then we pick the decision with the higher utility.

Case of buying the book $B = T$:

$\mathrm{P(B{=}T,P{=}T,M{=}T)} = \mathrm{P(P{=}T \mid M{=}T,B{=}T) \cdot P(B{=}T) \cdot P(M{=}T \mid B{=}T)} = 0.9 \cdot 1 \cdot 0.9$
$\mathrm{P(B{=}T,P{=}T,M{=}F)} = 0.5 \cdot 1 \cdot 0.1$
$\mathrm{P(B{=}T,P{=}F,M{=}T)} = 0.1 \cdot 1 \cdot 0.9$
$\mathrm{P(B{=}T,P{=}F,M{=}F)} = 0.5 \cdot 1 \cdot 0.1$

$\mathrm{P(B{=}T,P{=}T)} = 0.9 \cdot 1 \cdot 0.9 + 0.5 \cdot 1 \cdot 0.1 = 0.86$
$\mathrm{P(B{=}T,P{=}F)} = 0.1 \cdot 1 \cdot 0.9 + 0.5 \cdot 1 \cdot 0.1 = 0.14$
$\rightarrow \mathrm{U(B{=}T)} = 0.86 \cdot 1900 + 0.14 \cdot -100 = 1620$

Case of not buying the book $B = F$: $\mathrm{P(B{=}F,P{=}T,M{=}T)} = \mathrm{P(P{=}T \mid M{=}T,B{=}F)} \cdot$
$\mathrm{P(B{=}F)}$
$\cdot\, \mathrm{P(M{=}T \mid B{=}F)} = 0.8 \cdot 1 \cdot 0.7$
$\mathrm{P(B{=}F,P{=}T,M{=}F)} = 0.3 \cdot 1 \cdot 0.3$
$\mathrm{P(B{=}F,P{=}F,M{=}T)} = 0.2 \cdot 1 \cdot 0.7$
$\mathrm{P(B{=}F,P{=}F,M{=}F)} = 0.7 \cdot 1 \cdot 0.3$

$\rightarrow \mathrm{P(B{=}F,P{=}T)} = 0.8 \cdot 1 \cdot 0.7 + 0.3 \cdot 1 \cdot 0.3 = 0.65$
$\mathrm{P(B{=}F,P{=}F)} = 0.2 \cdot 1 \cdot 0.7 + 0.7 \cdot 1 \cdot 0.3 = 0.35$
$\rightarrow \mathrm{U(B{=}F)} = 0.65 \cdot 2000 + 0.35 \cdot 0 = 1300$

So in this case, she should buy the book.

3 Markov Models

Problem 3.1 (Prediction, Filtering, Smoothing) 8 pt

Explain the goals of *prediction, filtering* and *smoothing* in terms of conditional probabilities 8 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 3.2 (Markov Mood Detection) 16 pt

16 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

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

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 3.3 (Bellman Equation) 8 pt

State the Bellman Equation and explain every symbol in the equation and what the equation is used for and how. 8 min

---

Solution:

$$U(s) = R(s) + \gamma \cdot \max_{a \in A(s)} \left(\sum_{s'} P(s'|s, a) \cdot U(s') \right)$$

The meaning of the components is as follows:

- $U(s)$: the utility of the state s (long-term, global)
- $R(s)$: the reward at state s (short-term, local)
- $A(s)$: the set of actions available in state s
- $\max_{a \in A(s)}$: take the maximum over all available actions in state s
- $P(s'|s, a)$: the probability that taking action a in state s yields state s'
- $U(s')$: the utility in successor state s'
- $\left(\sum_{s'} P(s'|s, a) \cdot U(s')\right)$: the expected utility of action a by summing over all possible successor states

The equation is used to compute the utility of every state. The algorithm uses the equation as an iteration operator that computes new values for every $U(s)$ by evaluating the right hand side for the current values of U. If this leads to a *fixpoint*, a solution for the utilities has been found.

---

4 Learning

Problem 4.1 (Gradient Descent) 8 pt

Explain a gradient descent algorithm. 8 min

Problem 4.2 (Information Entropy) 6 pt

Explain and define *information entropy*. 6 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 4.3 (Home Decisions) 20 pt

20 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

Apply the decision tree learning algorithm on this table to predict whether people will get sunburned based on the attributes provided.

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`...