INFO8006 Introduction to Artificial Intelligence

Exam of January 2019 – Solutions

Instructions

- *Duration: 4 hours.*
- *Answer the questions on separate sheets, labeled with the question number, your first name, last name and student id.*
- *Answer in English or in French.*
- *Follow the same mathematical notation conventions as in the course, or properly define your conventions otherwise.*
- *Non-programmable calculators are allowed.*
- *Notes or documents of any kind are forbidden.*

Question 1 [3 points]

Multiple choice questions. Choose one of the four choices. Correct answers are graded $+\frac{3}{10}$*, wrong answers are graded* $-\frac{1}{10}$ *and the absence of answers is graded* 0*. The total of your grade for Question 1 is bounded below at* 0/3. Solutions are given in red.

1. In a partially observable and stochastic environment, a rational agent is ...

   (a) an agent that chooses whichever action that minimizes the expected value of the performance measure, given its percept history.

   (b) an agent that chooses whichever action that maximizes the value of the performance measure, given its percept history.

   (c) an agent that chooses whichever action that maximizes the value of the performance measure, regardless of its percept history.

   (d) an agent that chooses whichever action that maximizes the expected value of the performance measure, given its percept history.

2. Consider the game of Pacman in the absence of ghosts and with a single food dot in the maze. In state s, Pacman is located at position (i_s, j_s) while the food dot is located at (x_s, y_s). At timestep t, the score of the game is $500 - t$. The game ends when the food is eaten. Which of the following heuristics is not admissible?

   (a) $h(s) = |i_s - x_s| + |j_s - y_s|$

   (b) $h(s) = \sqrt{(i_s - x_s)^2 + (j_s - y_s)^2}$

   (c) $h(s) = (i_s - x_s)^2 + (j_s - y_s)^2$

   (d) $h(s) = \min(|i_s - x_s|, |j_s - y_s|)$

3. Which of the following propositions is not equivalent to the others?

   (a) The sentence α entails the sentence β.

   (b) α is true in all models where β is true.

   (c) β follows logically from α.

   (d) $\alpha \vDash \beta$.

4. In Monte Carlo Tree Search, at a node n during the selection step, the UCB1 policy picks the child node n' of n that maximizes

$$\frac{Q(n', p)}{N(n')} + c\sqrt{\frac{2 \log N(n)}{N(n')}}.$$

   Which of the following is true?

   (a) The first term encourages the exploitation of higher-reward nodes, while the second encourages the exploration of less-visited nodes.

   (b) The first term encourages the exploration of less-visited nodes, while the second term encourages the exploitation of higher-reward nodes.

(c) The first term encourages the exploitation of highly-visited nodes, while the second term encourages the exploration of lesser-rewarding nodes.

(d) The first term encourages the exploration of lesser-rewarding nodes, while the second term encourages the exploitation of highly-visited nodes.

5. The Bayes' rule states that ...

   (a) $P(x|y) = P(y|x)P(x) \,/\, P(y)$.

   (b) $P(x|y) = P(y|x)P(y) \,/\, P(x)$.

   (c) $P(x|y) = P(y) \,/\, P(y|x)P(x)$.

   (d) $P(y|x) = P(x|y)P(x) \,/\, P(y)$.

6. Which of the following is false?

   (a) Variable elimination is an exact inference algorithm that can be used both for discrete and continuous variables (under appropriate assumptions).

   (b) In variable elimination, the elimination ordering can greatly affect the computational complexity.

   (c) In likelihood weighting sampling, the evidence variables are all taken into account by the sampling distribution.

   (d) Gibbs sampling settles into a dynamic equilibrium in which the long-run fraction of time spent in each state is exactly proportional to its posterior probability.

7. In a first-order Markov process, as time passes ...

   (a) the state distribution always converges to a fixed point, called the stationary distribution.

   (b) the state distribution always converges to a fixed point, called the stationary distribution. This distribution is always of maximum uncertainty.

   (c) the state distribution sometimes converges to a fixed point, called the stationary distribution.

   (d) the state distribution sometimes converges to a fixed point, called the stationary distribution. This distribution is always of maximum uncertainty.

8. In a Markov decision process, the goal is ...

   (a) to find the optimal next action to take.

   (b) to find the action that maximizes the reward in the next state.

   (c) to find an optimal policy that maps states to actions.

   (d) to find an optimal plan, or sequence of actions, from start to goal.

9. Logistic regression models $P(Y = 1|\mathbf{x})$ as ...

   (a) $P(Y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b)$, where $\sigma(x) = \frac{1}{1+\exp(-x)}$.

   (b) $P(Y = 1|\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$.

   (c) $P(Y = 1|\mathbf{x}) = \tanh(\mathbf{w}^T\mathbf{x} + b)$.

   (d) $P(Y = 1|\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$.

10. State-of-the-art approaches for speech recognition are based on ...

    (a) A*.

    (b) hidden Markov models.

    (c) neural networks.

    (d) the Kalman filter.

Question 2 [3 points]

The game "21" is played as a misère game with any number of players who take turns saying a number. The first player says "1" and each player in turn increases the number by 1, 2, or 3, but may not exceed 21; the player who says "21" or a larger number loses.

1. Define the search problem associated with the 2-player version of the "21" game.

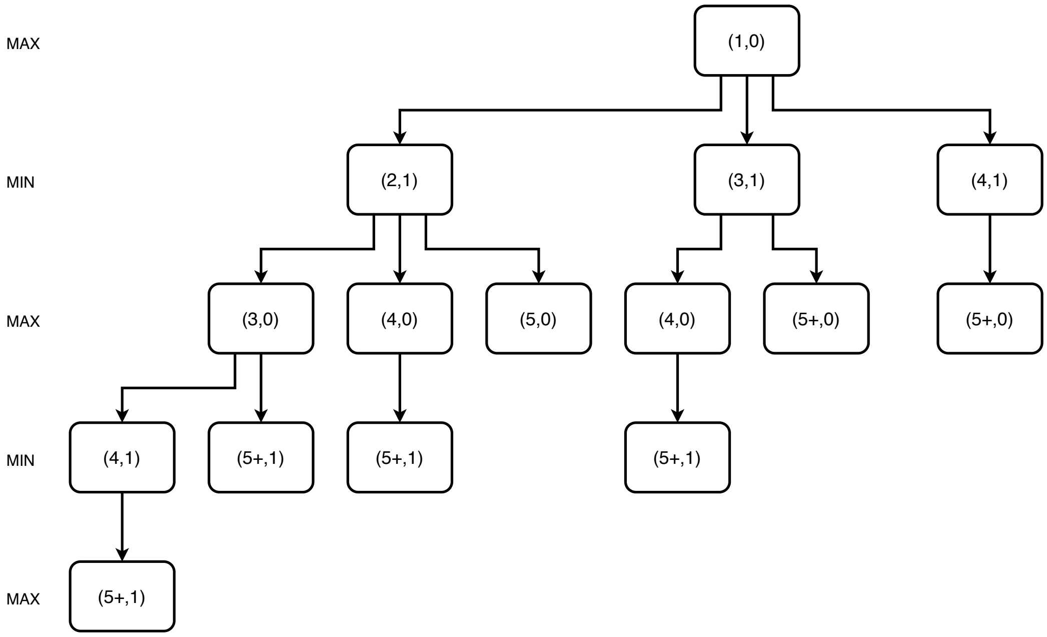
   Solution: We define a state s as a pair $s = (s^0, s^1) \in \mathbb{Z} \times \{0, 1\}$ such that the first element s^0 is the value played by the last player and the second element s^1 corresponds to the player to play next. The first player has id $p = 1$ and the second player has id $p = 0$.

   - Initial state: $s_0 := (s_0^0, s_0^1) := (1, 0)$.

- Player function: $player(\cdot) : \mathbb{Z} \times \{0,1\} \to \{0,1\}$, such that $player(s = (s^0, s^1)) := s^1$.
- Action function: actions are denoted by the integer values that can be played by the current player. We define the action function $action(\cdot) : \mathbb{Z} \times \{0,1\} \to \mathbb{Z}^3$ such that $action(s = (s^0, s^1)) := \{s^0 + 1, s^0 + 2, s^0 + 3\}$.
- Transition function: $result(\cdot, \cdot) : (\mathbb{Z} \times \{0,1\}) \times \mathbb{Z} \to \mathbb{Z} \times \{0,1\}$ such that $result(s = (s^0, s^1), a) := (a, (s^0 + 1) \bmod 2)$.
- Terminal function: $terminal(\cdot) : \mathbb{Z} \times \{0,1\} \to \{0,1\}$ such that $terminal(s = (s^0, s^1)) := s^1 \geq 21$.
- Utility function: $utility(\cdot, \cdot) : (\mathbb{Z} \times \{0,1\}) \times \{0,1\} \to \{-1,1\}$ such that $utility(s = (s^0, s^1), p) := 1 - 2|p - s^0|$.

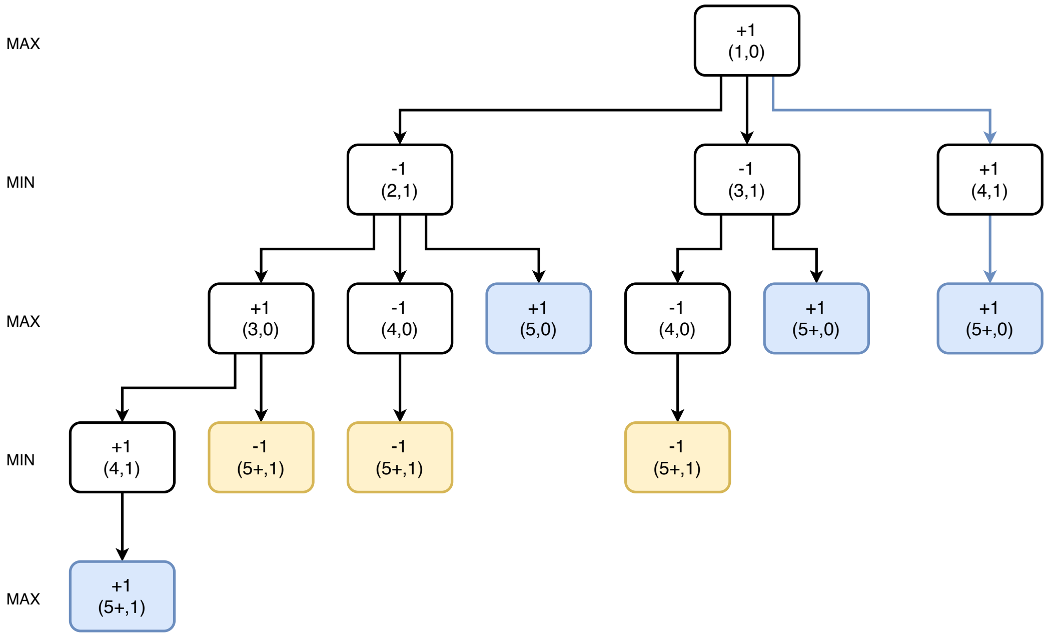
2. For this subquestion and the following, consider the game of "5" (still in its 2-player version) which has the same rule except that you should not say 5 or more. Show the whole game tree.

Solution:



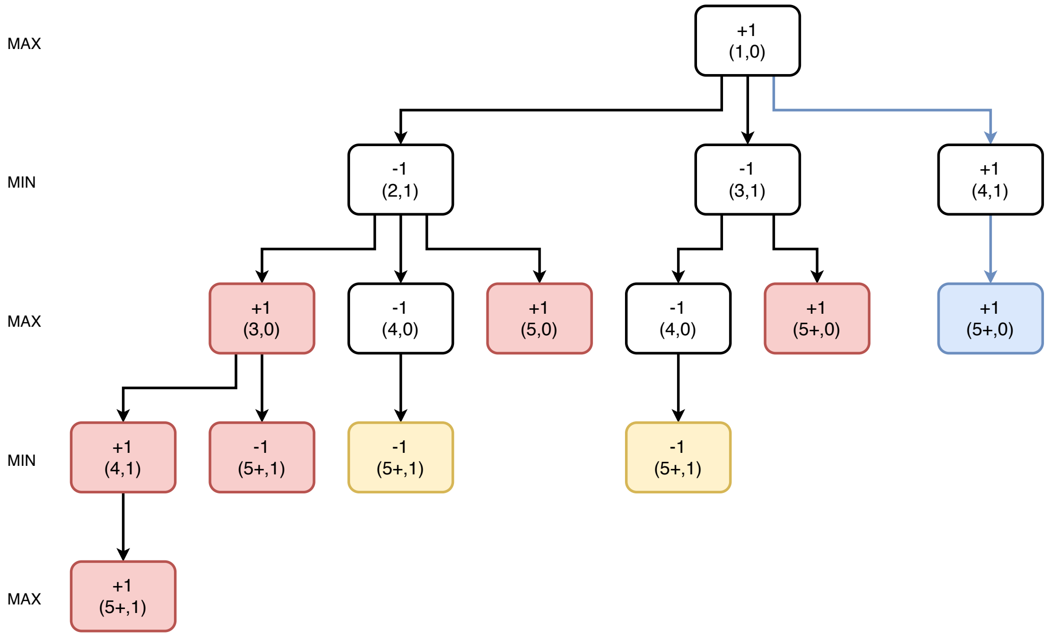
3. (a) Using the minimax algorithm, mark on your tree the backed-up values, and use those values to choose the best starting move

Solution: From the game tree we conclude that the best move is to play 4. NB: The winning strategy for the game of 21 is to always say a multiple of 4; it is then guaranteed that the other player will ultimately have to say "21" – so in the standard version where the first player opens with "1", they start with a losing move.



(b) Assume alpha–beta pruning were applied in optimal order. Draw the game tree containing only the nodes that would be evaluated.

Solution: Nodes that are not visited are colored in red.



Question 3 [3 points]

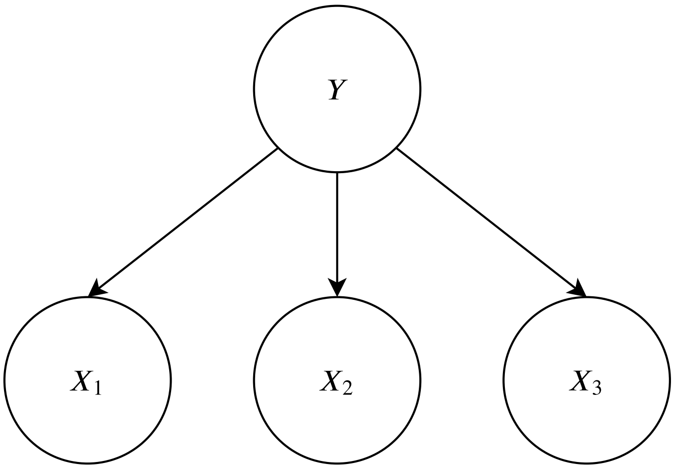
In order to quickly estimate the exam results, the teaching assistant wants to predict whether or not a student will pass the exam (Y) based on the following evidence:

- X_1: The student got a grade greater or equal than 10/20 at the first exercise (Yes/No).
- X_2: The student got a grade greater or equal than 10/20 in average for the projects (Yes/No).
- X_3: The student liked the course (Yes/No).

To simplify his model, the teaching assistant assumes that X_1, X_2 and X_3 are pairwise conditionally independent given Y. Answer the following questions:

1. Draw a Bayesian network that represents the joint distribution of Y, X_1, X_2 and X_3, and that incorporates the independence assumptions listed above. What is the name of this model?

   Solution: The corresponding Bayesian network is shown below. It is a Naive Bayes model.



2. Use Table 1 to compute an estimate of the conditional probability tables in the Bayesian network.

   Solution: The four conditional probability tables are as follows:

y	$P(Y=y)$
T	.78
F	.22

x_1	$P(X_1=x_1\|Y=\text{T})$	$P(X_1=x_1\|Y=\text{F})$
T	$\frac{54}{78}$	$\frac{10}{22}$
F	$\frac{78-54}{78}$	$\frac{22-10}{22}$

x_2	$P(X_2=x_2\|Y=\text{T})$	$P(X_2=x_2\|Y=\text{F})$
T	$\frac{70}{78}$	$\frac{15}{22}$
F	$\frac{78-70}{78}$	$\frac{22-15}{22}$

x_3	$P(X_3=x_3\|Y=\text{T})$	$P(X_3=x_3\|Y=\text{F})$
T	$\frac{45}{78}$	$\frac{4}{22}$
F	$\frac{78-45}{78}$	$\frac{22-4}{22}$

3. Give the prediction rule of Y for this model. What are the parameters of the model? Predict whether or not Pierre will pass the exam, knowing that he got 15/20 at the first question, liked the course and got 14/20 in average for the projects.

Solution: The prediction rule can be expressed as follows:

$$\begin{aligned} y^\star(x_1,x_2,x_3) &= \operatorname*{arg\,max}_{y=\{T,F\}} P(Y=y|X_1=x_1,X_2=x_2,X_3=x_3) \\ &= \operatorname*{arg\,max}_{y=\{T,F\}} \frac{P(Y=y,X_1=x_1,X_2=x_2,X_3=x_3)}{P(X_1=x_1,X_2=x_2,X_3=x_3)} \\ &= \operatorname*{arg\,max}_{y=\{T,F\}} P(Y=y,X_1=x_1,X_2=x_2,X_3=x_3) \\ &= \operatorname*{arg\,max}_{y=\{T,F\}} P(Y=y)P(X_1=x_1|Y=y)P(X_2=x_2|Y=y)P(X_3=x_3|Y=y) \end{aligned}$$

For Pierre, we have:

$$\begin{aligned} y^\star(x_1=T,x_2=T,x_3=T) &= \operatorname*{arg\,max}_{y=\{T,F\}}\{(y=T): 0.78\times\frac{54}{78}\times\frac{70}{78}\times\frac{45}{78}=0.28, (y=F): 0.22\times\frac{10}{22}\times\frac{15}{22}\times\frac{4}{22}=0.012\} \\ &= T. \end{aligned}$$

Pierre is predicted to pass the exam.

Y	Passed	Failed
Number students	78	22
Number of students who got $\geq 10/20$ at the first exercise	54	10
Number of students who got $\geq 10/20$ at the projects	70	15
Number of students who liked the course	45	4

Table 1: Data collected for 100 students.

Question 4 [4 points]

1. Define mathematically i) first-order Markov processes and ii) the inference tasks of prediction and filtering. Discuss how the latter can be useful to an agent.

Solution:

i. A Markov process is a stochastic process in which the (unobserved) state $\mathbf{X}_t$ at time t only depends on a subset of the previous states $\mathbf{X}_{1:t-1}$. In particular, in a first-order Markov process, we have

$$P(\mathbf{X}_t|X_{1:t-1}) = P(\mathbf{X}_t|\mathbf{X}_{t-1}),$$

i.e. the state $\mathbf{X}_t$ is conditionally independent of all previous states given X_{t-1}.

ii. Prediction is the inference task of computing $P(\mathbf{X}_{t+k}|\mathbf{e}_{1:t})$ for $k > 0$. Filtering is the inference task of computing $P(\mathbf{X}_t|\mathbf{e}_{1:t})$. The latter is helpful for maintaining a belief state about a partially observable environment, as we collect new evidence.

2. Derive the recursive update equation of the Bayes filter, assuming discrete variables. Does the Bayes filter generalize to continuous variables? If yes, outline how? If not, why?

   Solution:

$$\begin{aligned}
P(\mathbf{X}_{t+1}|\mathbf{e}_{1:t+1}) &= P(\mathbf{X}_{t+1}|\mathbf{e}_{1:t}, \mathbf{e}_{t+1}) \\
&= \alpha P(\mathbf{e}_{t+1}|\mathbf{X}_{t+1}, \mathbf{e}_{1:t})P(\mathbf{X}_{t+1}|\mathbf{e}_{1:t}) \\
&= \alpha P(\mathbf{e}_{t+1}|\mathbf{X}_{t+1})P(\mathbf{X}_{t+1}|\mathbf{e}_{1:t}) \\
&= \alpha P(\mathbf{e}_{t+1}|\mathbf{X}_{t+1}) \sum_{\mathbf{x}_t} P(\mathbf{X}_{t+1}|\mathbf{x}_t, \mathbf{e}_{1:t})P(\mathbf{x}_t|\mathbf{e}_{1:t}) \\
&= \alpha P(\mathbf{e}_{t+1}|\mathbf{X}_{t+1}) \sum_{\mathbf{x}_t} P(\mathbf{X}_{t+1}|\mathbf{x}_t)P(\mathbf{x}_t|\mathbf{e}_{1:t})
\end{aligned}$$

   where the normalization constant

$$\alpha = \frac{1}{P(\mathbf{e}_{t+1}|\mathbf{e}_{1:t})} = 1/\sum_{\mathbf{x}_{t+1}} P(\mathbf{e}_{t+1}|\mathbf{x}_{t+1})P(\mathbf{x}_{t+1}|\mathbf{e}_{1:t})$$

   is used to make probabilities sum to 1. Therefore, if we write $\mathbf{f}_t = P(\mathbf{X}_t|\mathbf{e}_{1:t})$, we can define the recursive relation $\mathbf{f}_{t+1} = \alpha\text{forward}(\mathbf{e}_{t+1}, \mathbf{f}_t)$.

   The Bayes filter generalizes to continuous variables: replace probability distributions P with probability density functions p and summations with integrals:

$$p(\mathbf{x}_{t+1}|\mathbf{e}_{1:t+1}) = \alpha\, p(\mathbf{e}_{t+1}|\mathbf{x}_{t+1}) \int p(\mathbf{x}_{t+1}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{e}_{1:t})d\mathbf{x}_t.$$

3. Let us consider a continuous Pacman world in which Pacman cannot directly observe ghosts. However, Pacman is equipped with a device that yields noisy estimates of the ghost positions. Assuming

   - a one-dimensional world with a single ghost,
   - a Gaussian prior of constant variance σ_0^2 for the ghost position,
   - a Gaussian transition model that nudges the ghost with random perturbations of fixed variance σ_x^2,
   - a sensor model that yields measurements with Gaussian noise of fixed variance σ_z^2 of the ghost position,

   derive the update equations from timestep t to $t+1$ of the parameters of the belief distribution of the ghost position, including its mean μ_{t+1} and its variance σ_{t+1}^2. You are free to use identities from Appendix A if needed.

   Solution: This exercise is adapted from section 15.4.2 and exercise 15.11 of the textbook.

   The belief distribution about the ghost position can be maintained with a Kalman filter for which we have:

   - a Gaussian prior $p(\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0|\mu_0, \sigma_0^2)$,
   - a linear Gaussian transition model $p(\mathbf{x}_{t+1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t+1}|\mathbf{x}_t, \sigma_x^2)$,
   - a linear Gaussian sensor model $p(\mathbf{e}_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{e}_t|\mathbf{x}_t, \sigma_z^2)$.

   Given the properties of the Gaussian distribution, the belief distribution $p(\mathbf{x}_t|\mathbf{e}_{1:t})$ remains Gaussian for all t, with $p(\mathbf{x}_t|\mathbf{e}_{1:t}) = \mathcal{N}(\mathbf{x}_t|\mu_t, \sigma_t^2)$.

   From t to $t+1$, using identities (1), (2) and (3) of Appendix A for $\mathbf{x} := \mathbf{x}_t$ and $\mathbf{y} := \mathbf{x}_{t+1}$, we have

$$\begin{aligned}
p(\mathbf{x}_{t+1}|\mathbf{e}_{1:t}) &= \mathcal{N}(\mathbf{x}_{t+1}|\mathbf{A}\mu + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \\
&= \mathcal{N}(\mathbf{x}_{t+1}|\mu_t, \sigma_x^2 + \sigma_t^2).
\end{aligned}$$

   Conditioning on the new evidence $\mathbf{e}_{t+1}$, we find using identities (1), (2) and (4) of Appendix A for $\mathbf{x} := \mathbf{x}_{t+1}$ and $\mathbf{y} := \mathbf{e}_{t+1}$ that

$$p(\mathbf{x}_{t+1}|\mathbf{e}_{1:t}, \mathbf{e}_{t+1}) = \mathcal{N}(\mathbf{x}_{t+1}|\underbrace{\boldsymbol{\Sigma}\left(\mathbf{A}^T\mathbf{L}(\mathbf{e}_{t+1} - \mathbf{b}) + \boldsymbol{\Lambda}\mu\right)}_{\mu_{t+1}}, \underbrace{\boldsymbol{\Sigma}}_{\sigma_{t+1}^2})$$

   where

$$\begin{aligned}
\sigma_{t+1}^2 &= \frac{\sigma_z^2(\sigma_x^2 + \sigma_t^2)}{\sigma_z^2 + \sigma_x^2 + \sigma_t^2} \\
\mu_{t+1} &= \frac{\mathbf{e}_{t+1}(\sigma_x^2 + \sigma_t^2) + \mu_t\sigma_z^2}{\sigma_z^2 + \sigma_x^2 + \sigma_t^2}.
\end{aligned}$$

4. Let us examine the behavior of the variance update.
   - As $t \to \infty$, σ_t^2 converges to a fixed point σ^2. Calculate the value of σ^2.
   - Give a qualitative explanation for what happens as i) $\sigma_x^2 \to 0$ and ii) as $\sigma_z^2 \to 0$.

   Solution: This exercise is adapted from exercise 15.12 of the textbook.

   As $t \to \infty$, $\sigma^2_\infty = \sigma^2_{\infty+1} \Leftrightarrow \frac{\sigma_z^2(\sigma_x^2+\sigma^2)}{\sigma_z^2+\sigma_x^2+\sigma^2} = \sigma^2$. Solving for σ^2 and given that the variance is positive, we find the fixed point

$$\begin{aligned} \sigma^4 + \sigma_x^2\sigma^2 - \sigma_x^2\sigma_z^2 = 0 \\ \Leftrightarrow \sigma^2 = \frac{-\sigma_x^2 + \sqrt{\sigma_x^4 + 4\sigma_x^2\sigma_z^2}}{2}. \end{aligned}$$

   When the variance of the transition model and the variance of the sensor model tends to 0, we have

   i. $\sigma_x^2 \to 0 \Rightarrow \sigma^2 \to 0$,

   ii. $\sigma_z^2 \to 0 \Rightarrow \sigma^2 \to 0$.

   In the first case, this means that the ghost does not move and we eventually infer its exact position as infinitely many evidences are collected. In the second case, the sensor directly returns the exact ghost position, hence the absence of uncertainty.

Question 5 [4 points]

1. Formally define what is i) a Markov Decision Process (MDP) and ii) an optimal policy.

   Solution:

   i. A Markov decision process (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, P, R)$ such that:
      - $\mathcal{S}$ is a set of states s;
      - $\mathcal{A}$ is a set of actions a;
      - P is a (stationary) transition model such that $P(s'|s,a)$ denotes the probability of reaching state s' if action a is done in state s;
      - R is a reward function that maps immediate (finite) reward values $R(s)$ obtained in states s.

   ii. The expected utility obtained by executing a policy π starting in s is given by

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t)\right]\Bigg|_{s_0=s}$$

   where the expectation is with respect to the probability distribution over state sequences determined by s and π. Among all policies the agent could execute, the optimal policy is the policy π_s^* that maximizes the expected utility:

$$\pi_s^* = \arg\max_\pi V^\pi(s)$$

   Because of discounted utilities, the optimal policy is independent of the starting state s. Therefore we write π^*.

2. Sometimes MDPs are formulated with a reward function $R(s,a)$ that depends on the action taken or with a reward function $R(s,a,s')$ that also depends on the outcome state.

   (a) Write the Bellman equations for these formulations.

   (b) Show how an MDP with reward function $R(s,a,s')$ can be transformed into a different MDP with reward function $R(s,a)$, such that optimal policies in the new MDP correspond exactly to optimal policies in the original MDP. Then do the same to convert MDPs with $R(s,a)$ into MDPs with $R(s)$.

   Solution: This exercise is adapted from exercise 17.4 of the textbook.

   (a) For $R(s,a)$ we have

$$V(s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')\right].$$

   For $R(s,a,s')$, we have

$$V(s) = \max_a \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V(s')\right].$$

(b) There are a variety of solutions here. One is to create a pre-state $\text{pre}(s,a,s')$ for every s,a,s', such that executing a in s leads not to s' but to $\text{pre}(s,a,s')$. In this state is encoded the fact that the agent came from s and did a to get here. From the pre-state, there is just one action b that always leads to s'. Let the new MDP have transition P', reward R' and discount γ'. Then,

$$\begin{aligned}
&P'(\text{pre}(s,a,s')|s,a) = P(s'|s,a)\\
&P'(s'|\text{pre}(s,a,s'),b) = 1\\
&R'(s,a) = 0\\
&R'(\text{pre}(s,a,s'),b) = \gamma^{\frac{1}{2}}R(s,a,s')\\
&\gamma' = \gamma^{\frac{1}{2}}.
\end{aligned}$$

Similarly, for converting MDPs with $R(s,a)$ into MDPs with $R(s)$, we can create states $\text{post}(s,a)$ for every s,a such that

$$\begin{aligned}
&P'(\text{post}(s,a)|s,a) = 1\\
&P'(s'|\text{post}(s,a),b) = P(s'|s,a)\\
&R'(s) = 0\\
&R'(\text{post}(s,a)) = \gamma^{\frac{1}{2}}R(s,a)\\
&\gamma' = \gamma^{\frac{1}{2}}.
\end{aligned}$$

3. Discuss whether an agent taking actions in the real world can be modeled as a MDP. If yes, how? If not, why?

   Solution: In general, it is difficult to model the real world as an MDP because the environment is i) only partially observable and ii) known only up to the limits of our knowledge of the laws of physics. This implies that all four elements $\mathcal{S},\mathcal{A},P,R$ cannot be defined exactly, nor be evaluated. Under simplifying assumptions, we may consider the real world as a known environment, but a realistic model will remain only partially observable.

   In this case, the real world would correspond to POMDP, which can be reduced to an MDP over belief states $b(s)$ with transition model $p(b'|b,a)$ and reward function $\rho(b)$. The resulting model is an MDP, although its continuous state space makes it intractable to solve exactly.

Question 6 [3 points]

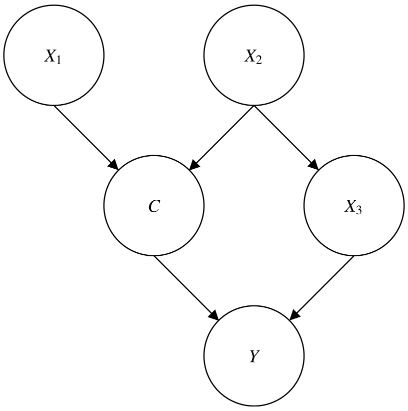


Figure 1: Bayesian network of the final grade calculation.

The Bayesian network shown in Figure 1 models the quantity of salt in a chocolate cake. In this model, X_1, X_2, X_3 respectively denote the quantity of salt in the chocolate before cooking, the quantity of salt in the eggs and the quantity of salt in the egg white beaten in snow. To cook a chocolate cake, you first have to mix chocolate with egg yellow, together with sugar and butter. The quantity C of salt in the resulting dough is modeled as $C = a_2X_1 + a_3X_2 + \mathcal{N}(\mu_C, \sigma_C^2)$. When the egg white is beaten in snow, a bit of salt is usually added, which we model as $X_3 = a_1X_2 + \mathcal{N}(\mu_3, \sigma_3^2)$. Finally, the quantity of salt in the chocolate cake is modeled as $Y = a_4C + a_5X_3 + \mathcal{N}(\mu_y, \sigma_y^2)$. Answer the following questions about this model:

1. Assume the parameters of the model are known, give the prediction rule of Y given X_1 and X_2. Simplify this rule and give a minimal set of parameters that are required to predict Y given X_1, X_2. What is the name of the resulting predictive model?

   Solution: We first express the random variable Y in terms of the input random variables X_1, X_2, X_3:

$$\begin{aligned}
Y &= a_4C + a_5X_3 + \mathcal{N}(\mu_y, \sigma_y^2)\\
&= a_4(a_2X_1 + a_3X_2 + \mathcal{N}(\mu_C, \sigma_C^2)) + a_5X_3 + \mathcal{N}(\mu_y, \sigma_y^2)\\
&= a_4(a_2X_1 + a_3X_2 + \mathcal{N}(\mu_C, \sigma_C^2)) + a_5a_1X_2 + a_5\mathcal{N}(\mu_3, \sigma_3^2) + \mathcal{N}(\mu_y, \sigma_y^2).
\end{aligned}$$

Due to the properties of the Gaussian distribution, this implies that

$$p(y|x_1, x_2) = \mathcal{N}(y|a_2a_4x_1 + (a_3a_4 + a_5a_1)x_2 + a_4\mu_C + a_5\mu_3 + \mu_y, a_4^2\sigma_C^2 + a_5^2\sigma_3^2 + \sigma_y^2).$$

The distribution can further be simplified to a Gaussian $\mathcal{N}(w_1x_1 + w_2x_2 + b, \sigma)$, which effectively depends on the four parameters w_1, w_2, b and σ. The resulting model is a linear Gaussian model.

A maximum likelihood prediction rule $y^*(x_1, x_2)$ for Y is to predict the mean of this linear Gaussian model:

$$\begin{aligned} y^*(x_1, x_2) &= \arg\max_y p(y|x_1, x_2) \\ &= w_1x_1 + w_2x_2 + b. \end{aligned}$$

2. From a data set of N points $d = \{(x_{1,1}, x_{2,1}, y_1), \dots, (x_{1,N}, x_{2,N}, y_N)\}$ (that you assume to be iid) explain mathematically how you would compute the parameters of the simplest model.

   Solution: We apply the maximum likelihood estimation principle, which leads to the following optimization problem:

$$\begin{aligned} W &= \arg\max_W p(d|W) \\ &= \arg\max_W p(y_{1:N}, X_{1:N}|W) \\ &= \arg\max_W \log(p(y_{1:N}, X_{1:N}|W)) \\ &= \arg\max_W \sum_{i=1}^N \log(p(y_i|X_i, W) + \log(p(X_i)) \\ &= \arg\max_W \sum_{i=1}^N \log(p(y_i|X_i, W) \\ &= \arg\min_W \sum_{i=1}^N \frac{1}{2}\log(2\pi\sigma) - \frac{(W^TX_i - y_i)^2}{2\sigma^2} \\ &= \arg\min_W \sum_{i=1}^N (W^TX_i - y_i)^2 \end{aligned}$$

   where $W = [w_1, w_2, b]^T$ and $X = [x_1, x_2, 1]$. Since the last expression is convex, it can be minimized by finding the value of W which cancels the gradient:

$$\begin{aligned} &\nabla_W \sum_{i=1}^N (W^TX_i - y_i)^2 = 0 \\ \Leftrightarrow &\nabla_W \left[(XW - y)^T(XW - y)\right] = 0 \\ \Leftrightarrow &2X^T(XW - y) = 0 \\ \Leftrightarrow &X^T(XW - y) = 0 \\ \Leftrightarrow &X^TXW - X^Ty = 0 \\ \Leftrightarrow &W = (X^TX)^{-1}X^Ty \end{aligned}$$

   where $X = [X_1, \dots, X_N]^T \in \mathbb{R}^{N\times 4}$ and $y = [y_1, \dots, y_N]^T \in \mathbb{R}^N$.

3. Is this model always realistic? Given a large amount of data, what other model could you use instead? What if you only have a small amount of data?

   Solution: A linear model is often satisfactory for many situations, but fails to represent complex (non-linear) interactions between variables. When we have access to a large amount of data, we can train instead a more complicated model, such as a neural network which is the best way (to date) to represent arbitrarily complex relations. If we only have a small amount of data, simpler models should be preferred. In this situation, a neural network cannot be fit as it would typically lead to overfitting.

A Cheat sheet for Gaussian models (Bishop, 2006)

Given a marginal Gaussian distribution for $\mathbf{x}$ and a linear Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \mathbf{\Lambda}^{-1}) \tag{1}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2}$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mu + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T) \tag{3}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\mathbf{\Sigma}\left(\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda}\mu\right), \mathbf{\Sigma}) \tag{4}$$

where

$$\mathbf{\Sigma} = (\mathbf{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}. \tag{5}$$