INFO8006 Introduction to Artificial Intelligence

Exam of January 2021

Instructions

- *The exam starts at 8:30 AM.*
- *Questions must be answered on paper.*
- *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.*
- *Your answers must be scanned and submitted on eCampus by 12:45 AM.*
- *Your answers must be submitted in PDF in 5 distinct files named as*
  - *LASTNAME_Firstname_Q1.pdf*
  - *LASTNAME_Firstname_Q2.pdf*
  - *LASTNAME_Firstname_Q3.pdf*
  - *LASTNAME_Firstname_Q4.pdf*
  - *LASTNAME_Firstname_Q5.pdf*

Question 1 [4 points]

Multiple choice questions. Choose one of the four choices. Correct answers are graded $+\frac{4}{10}$*, wrong answers are graded* $-\frac{2}{15}$ *and the absence of answers is graded* 0*. The total of your grade for Question 1 is bounded below at* $0/4$*.*

1. An approximate Q-Learning agent based on a linear model of the Q-table is an example of
   (a) learning simple reflex agent.
   (b) learning model-based reflex agent.
   (c) learning goal-based planning agent.
   (d) learning utility-based planning agent.
2. A^* necessarily reduces to uniform-cost search when
   (a) the heuristic h is always null.
   (b) the heuristic h is admissible.
   (c) the heuristic h is consistent.
   (d) the heuristic h is random.
3. In adversarial search,
   (a) if not looked deep enough, bad moves may appear as good moves, because their consequences are hidden beyond the search horizon.
   (b) the deeper in the tree the evaluation function is buried, the more the quality of the evaluation matters.
   (c) the horizon effect arises when the evaluation function is perfect.
   (d) the horizon effect is negligible unless TD learning's learning rate α tends to 1.
4. Given that $A \perp B|C$, $\mathbf{P}(A,B|C)$ is equivalent to
   (a) $\frac{\mathbf{P}(A,C)\mathbf{P}(B|C)}{\mathbf{P}(C)}$.
   (b) $\frac{\mathbf{P}(C|A)\mathbf{P}(A|B)\mathbf{P}(B)}{\mathbf{P}(C)}$.
   (c) $\frac{\mathbf{P}(B,C|A)\mathbf{P}(A)}{\mathbf{P}(B,C)}$.
   (d) $\frac{\mathbf{P}(C|A)\mathbf{P}(C|B)}{\mathbf{P}(C)}$.
5. In the Kalman filter, the prior $\mathbf{P}(X_0)$

(a) must be a uniform distribution over the discrete state values $1, \dots, S$.

(b) is arbitrary, since it depends on one's initial beliefs.

(c) must be a Gaussian distribution over the state values.

(d) can be any continuous distribution, as long as its mean is zero.

6. Which of following is wrong? Convolutional neural networks...

   (a) are usually trained by solving a maximum likelihood estimation problem.

   (b) are fit for processing spatially structured data, such as images or sequences.

   (c) usually count thousands to millions of parameters.

   (d) have a topology that encodes conditional independence assumptions between random variables.

7. In a Markov Decision Process, if rewards r are associated to transition tuples (s, a, s'), i.e. $r = R(s, a, s')$, then the Bellman equations should be expressed as

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

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

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

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

8. Which of the following is true? In reinforcement learning, ...

   (a) direct utility estimation is an efficient algorithm for estimating V^π since its time complexity (for reaching a desired level of precision of the estimate) grows sub-linearly with the size of the state-action space.

   (b) ϵ-greedy is an exploration policy that avoids visiting states s for which $V(s) < \epsilon$.

   (c) the state-action-value $Q(s, a)$ of the q-state (s, a) is the maximum utility starting out having taken action a from s and thereafter acting optimally.

   (d) the Q-table has size equal to $|\mathcal{S} \times \mathcal{A}|$.

9. In DQN (Mnih et al, 2015), the Q-table is approximated with a so-called Q-network. This network is

   (a) a Bayesian network.

   (b) a dynamic Bayesian network.

   (c) a convolutional neural network.

   (d) a decision network.

10. In DQN (Mnih et al, 2015), the Q-network is trained using a variant of

    (a) direct utility estimation.

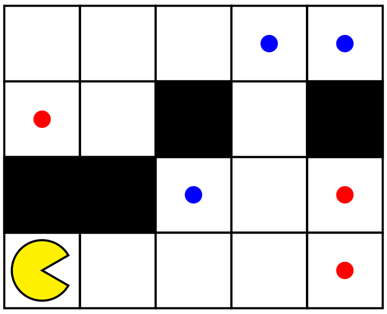
    (b) temporal difference learning.

    (c) value iteration.

    (d) policy iteration.

Question 2 [4 points]

Let us assume a grid-world in which there are two kinds of food pellets, each with a different color (blue and red). Mrs Pacman is only interested in tasting the two different kinds of food: the game ends when she has eaten 1 blue pellet and 1 red pellet. Mrs Pacman has four actions: moving north, south, east, or west. There are K blue pellets and K red pellets and the dimensions of the grid-world are $N \times M$. Mrs Pacman cannot move into any of the B walls.



(a) Formally define an efficient search problem for Mrs Pacman. Assume she starts in location (i_0, j_0).

- A state s is described by the current location (i, j) and two booleans b and r indicating whether Mrs Pacman has eaten (or not) a blue and red pellet, respectively. The initial state is $s_0 = ((i_0, j_0), 0, 0)$.
- The available actions are $actions(s) = \{N, S, E, W\}$.
- For the transition model $s' = result(s, a)$, if the action leads Mrs Pacman out of the maze or within a wall, $s' = s$. Otherwise, either i or j is incremented/decremented (e.g. $(i', j') = (i, j + 1)$ for $a = N$) and if the new position corresponds to a blue or red pellet, then $b' = 1$ or $r' = 1$, respectively.
- The state is terminal if $b = r = 1$.
- The step cost is $c(s, a, s') = 1$.

(b) Provide a tight upper bound on the size of the state space.

$M \times N \times 4$.

(c) Provide a tight upper bound on the branching factor of the search problem.

4.

(d) For each of the following heuristics, indicate (yes/no) and motivate briefly (1-2 sentences) whether or not it is admissible.

- h_1: the number of remaining pellets.

  No, it could be larger than the remaining cost h^*. Increasing the number of pellets increases h_1 but not h^*.

- h_2: the Euclidean distance to the closest remaining pellet.

  Yes, if the state isn't terminal, Mrs Pacman needs to eat at least one extra pellet to reach a terminal state and the cost of moving to that extra pellet is greater or equal than the Euclidean distance to the closest pellet. However, the heuristic should be zero for terminal states.

- h_3: the maximum Euclidean distance between any two remaining pellets.

  No, it could be larger than the remaining cost h^*. Adding far away pellets increases h_3 but not h^*.

- h_3: the minimum Euclidean distance between any two remaining pellets of opposite colors.

  No, if we have already eaten one pellet, this heuristic could be larger than the remaining cost.

Question 3 [4 points]

Pinky the ghost escaped from Pacman and tries to hide from him. Pinky randomly moves around floors A, B, C, and D. Pinky's location at time t is X_t. At the end of each timestep, Pinky stays on the same floor with probability 0.5, goes upstairs with probability 0.3 , and goes downstairs with probability 0.2. If Pinky is on floor A, he goes down with probability 0.2 and stays put with probability 0.8. If Pinky is on floor D, he goes upstairs with probability 0.3 and stays put with probability 0.7.

A
B
C
D

(a) Given the prior $\mathbf{P}(X_0)$ below, determine the distribution of Pinky's location at time $t = 1$.

X_0	$P(X_0)$
A	0.1
B	0.2
C	0.3
D	0.4

$$P(X_1) = \sum_{x_0} P(X_1, x_0) = \sum_{x_0} P(X_1|x_0)P(x_0)$$

$$= T\,P(X_0) = \begin{pmatrix} 0.8 & 0.3 & 0 & 0 \\ 0.2 & 0.5 & 0.3 & 0 \\ 0 & 0.2 & 0.5 & 0.3 \\ 0 & 0 & 0.2 & 0.7 \end{pmatrix} \begin{pmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{pmatrix} = \begin{pmatrix} 0.14 \\ 0.21 \\ 0.31 \\ 0.34 \end{pmatrix}$$

(b) Determine the distribution of Pinky's location at time $t = \infty$.

We have to find the stationary distribution

$$\begin{aligned} P(X_\infty) &= T\,P(X_\infty) \\ \Leftrightarrow \qquad 0 &= (T - \mathbb{I})P(X_\infty) \end{aligned}$$

such that $\sum_x P(X_\infty = x) = 1$, which amounts to solving the following system of linear equations

$$\begin{cases} 0 = -0.2P(X_\infty = A) + 0.3P(X_\infty = B) \\ 0 = 0.2P(X_\infty = A) - 0.5P(X_\infty = B) + 0.3P(X_\infty = C) \\ 0 = 0.2P(X_\infty = B) - 0.5P(X_\infty = C) + 0.3P(X_\infty = D) \\ P(X_\infty = D) = 1 - P(X_\infty = A) - P(X_\infty = B) - P(X_\infty = C) \end{cases} .$$

Using a calculator,

$$P(X_\infty) \approx \begin{pmatrix} 0.4154 \\ 0.2769 \\ 0.1846 \\ 0.1231 \end{pmatrix} .$$

To aid the search, a sensor S^r is installed on the roof and a sensor S^b is put in the basement. Both sensors detect either a ghost ($+g$) or no ghost ($-g$). The distribution of sensor measurements is determined by d, the number of floors between Pinky and the sensor. For example, if Pinky is on floor B, then $d_b = 2$ because there are two floors (C and D) between floor B and the basement, and $d_r = 1$ because there is one floor (A) between floor B and the roof. Pinky will not go onto the roof nor into the basement.

S^r	$P(S^r \vert d_r)$
$+g$	$1 - 0.3d_r$
$-g$	$0.3d_r$

S^b	$P(S^b \vert d_b)$
$+g$	$1 - 0.2d_b$
$-g$	$0.2d_b$

(c) You decide to track Pinky by particle filtering with 3 particles. At time t, the particles (1), (2) and (3) are at positions $X_t^{(1)} = A$, $X_t^{(2)} = B$ and $X_t^{(3)} = C$. In Step 1 of particle filtering, without incorporating any sensory information, what is the (joint) probability that the particles will be resampled respectively to B, B and C when projected through the transition model?

In step 1 of particle filtering, the particles are projected *independently* through the transition model.

$$\begin{aligned}&P(X_{t+1}^{(1:3)} = (B,B,C)|X_t^{(1:3)} = (A,B,C))\\&= P(X_{t+1}^{(1)} = B|X_t^{(1)} = A)P(X_{t+1}^{(2)} = B|X_t^{(2)} = B)P(X_{t+1}^{(3)} = C|X_t^{(3)} = C)\\&= 0.2 \times 0.5 \times 0.5 = 0.05\end{aligned}$$

(d) Assume the particles have been resampled to B, B and C in Step 1. At $t+1$, the sensors observe $S^r_{t+1} = +g$ and $S^b_{t+1} = -g$. In Step 3 of particle filtering, now taking this sensory information into account, what is the (joint) probability that the particles will be such that $X_{t+1}^{(1)} = B$, $X_{t+1}^{(2)} = B$ and $X_{t+1}^{(3)} = B$?

In step 2 of particle filtering, the particles are weighted by the evidence likeliness

$$\begin{aligned}w_i &= P(S^r_{t+1} = +g, S^r_{t+1} = -g|X_{t+1}^{(i)})\\&= P(S^r_{t+1} = +g|X_{t+1}^{(i)})P(S^r_{t+1} = -g|X_{t+1}^{(i)})\end{aligned}$$

and, then in step 3, re-sampled with respect to the weights. In our case,

$$\begin{aligned}w_1 = w_2 &= (1 - 0.3 \times 1)(0.2 \times 2) = 0.28\\w_3 &= (1 - 0.3 \times 2)(0.2 \times 1) = 0.08\end{aligned}$$

and

$$P(X_{t+1}^{(1:3)} = (B,B,B)) = P(X_{t+1}^{(i)} = B)^3 = \left(\frac{w_1 + w_2}{w_1 + w_2 + w_3}\right)^3 = 0.875^3 \approx 0.6699$$

(e) You find out that particle filtering with 3 particles is not very reliable and decide instead to make use of a Bayes filter. Express recursively the probability of X_t given all the measurements from the two sensors.

We are asked to express

$$\begin{aligned}P(X_t|e_{1:t}) &= \alpha P(e_t|X_t)P(X_t|e_{1:t-1})\\&= \alpha P(e_t|X_t)\sum_{x_{t-1}} P(X_t, x_{t-1}|e_{1:t-1})\\&= \alpha P(e_t|X_t)\sum_{x_{t-1}} P(X_t|x_{t-1})P(x_{t-1}|e_{1:t-1})\end{aligned}$$

In our case, $E_t = (S^r_t, S^b_t)$ and

$$P(e_t|X_t) = P(s^r_t|X_t)P(s^b_t|X_t).$$

Question 4 [4 points]

You observe a grandmaster playing Pacman and wish to learn from her games. To this end, you write down in a table all the state-action pairs (s,a) played by the grandmaster, together with their corresponding Q-values $Q(s,a)$. You describe each state-action pair with six features: the horizontal and vertical position of Pacman (x_P, y_P) and of the ghost (x_G, y_G); the distance d to the closest food pellet; and an action feature. Unfortunately, you did not sleep too well the night before and make random errors when computing and reporting the Q-values. Assuming Gaussian errors (of zero mean and unit variance), how would you learn a model of the Q-function using your data?

(a) Describe formally the learning problem you would have to solve in the case of a linear model of the Q-function (i.e., the data, the model, and its parameters). Write down the optimization problem to estimate the model parameters.

Let $x \in \mathbb{R}^6 \times \{1\}$ denote the six provided features (plus a constant 1) and $y \in \mathbb{R}$ the Q-value (with noise). Our data is a set $d = \{(x_i, y_i)\}_{i=1}^N$. The assumption is that y follows a linear Gaussian distribution $\mathcal{N}(w^T x, 1^2)$, where $w \in \mathbb{R}^7$ are the model parameters. The optimization problem is to estimate the parameters that maximize the likelihood, i.e.

maximum likelihood estimation (MLE).

$$\begin{aligned} w^* &= \arg\max_w P(d|w) \\ &= \arg\max_w \prod_i P(x_i, y_i|w) \\ &= \arg\max_w \prod_i P(y_i|w, x_i)P(x_i) \\ &= \arg\max_w \sum_i \log P(y_i|w, x_i) \\ &= \arg\max_w \sum_i \log\left[\frac{1}{\sqrt{2\pi 1^2}} \exp\left(-\frac{(w^Tx_i - y_i)^2}{2 1^2}\right)\right] \\ &= \arg\min_w \sum_i (w^Tx_i - y_i)^2 \end{aligned}$$

(b) Assuming a Gaussian prior (of zero mean and unit variance) on the model parameters, revise your optimization problem above. What is the name of the resulting estimator?

We now assume that parameters follow a Gaussian prior

$$P(w) = \mathcal{N}(0, \mathbb{I}) = \frac{1}{\sqrt{2\pi}^7} \exp\left(-\frac{1}{2}\sum_{j=1}^{7} w_j^2\right).$$

The optimization problem is now to estimate the parameters that are the most likely, given the data, i.e. maximum a posteriori (MAP) estimation.

$$\begin{aligned} w^* &= \arg\max_w P(w|d) \\ &= \arg\max_w P(d|w)P(w) \\ &= \arg\max_w \log P(w) + \sum_i \log P(y_i|w, x_i) \\ &= \dots \\ &= \arg\min_w \sum_{j=1}^{7} w_j^2 + \sum_i (w^Tx_i - y_i)^2. \end{aligned}$$

(c) Derive a closed-form formula for computing the parameters the solution of this optimization problem.

The optimal parameters are the ones that minimize the objective

$$\begin{aligned} L &= \sum_{j=1}^{7} w_j^2 + \sum_i (w^Tx_i - y_i)^2 \\ &= \sum_{j=1}^{7} w_j^2 + \sum_i (w^Tx_i)^2 + y_i^2 - 2w^Tx_iy_i \\ &= w^Tw + w^TX^TXw + Y^TY - 2w^TX^TY, \end{aligned}$$

where $X = (x_i^T) \in \mathbb{R}^{N\times 7}$ and $Y = (y_i) \in \mathbb{R}^N$. As this objective is convex (sum of squares), we can find w^* by canceling the gradient of L, i.e.

$$\begin{aligned} 0 &= \nabla_w \left(w^Tw + w^TX^TXw + Y^TY - 2w^TX^TY\right)\Big|_{w^*} \\ &= 2w^* + 2X^TXw^* - 2X^TY \\ \Leftrightarrow \quad w^* &= (X^TX - \mathbb{I})^{-1}X^TY. \end{aligned}$$

(d) For the data in the table below, compute the parameters of the linear model.

x_P	y_P	x_G	y_G	d	a	Q
2	4	4	2	4	1	-1
2	2	2	2	-2	10	1
0	0	4	4	3	3	1

Using a calculator,

$$w^* \approx \begin{pmatrix} -0.151 \\ -0.296 \\ 0.025 \\ 0.171 \\ -0.043 \\ 0.141 \\ 0.005 \end{pmatrix}.$$

Question 5 [4 points]

Let us consider a simplified version of Blackjack where the deck is infinite and the dealer does not play. The deck contains cards 2 through 10, J, Q, K, and A, which are all equally likely to be drawn. Each card is worth the number of points shown on it, except for the cards J, Q, and K that are worth 10 points, and A that is worth 11. At each turn, you may either draw or stop. If you choose to draw, you are dealt with an additional card and move to the next turn. You don't receive any immediate reward. If you stop, you receive a reward of 0 if the total number of points for the cards you hold is exactly 15, 10 if it is higher than 15 but not higher than 21, and -10 otherwise (i.e., lower than 15, or larger than 21). After taking the stop action, the game ends. If your total number of points reaches 22 or higher, then you have failed: you may only choose the stop action, in which case you lose and receive a reward of -10.

Let us now assume the state space $\mathcal{S}$ to be the set $\{0, 2, ..., 21, \text{fail}, \text{end}\}$, such that each state indicates either the ongoing total number of points of the player, whether the player has failed ("fail"), or if the game has ended ("end").

(a) Assume you have already performed j iterations of Value Iteration. Compute $V_{j+1}(12)$ given the table below for $V_j(s)$. The discount factor is $\gamma = 0.5$.

s	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	fail	end
$V_j(s)$	0	0	0	0	0	0	1	1	1	1	2	2	10	10	10	10	10	10	10	10	-10	0

We are asked to perform a single step of the value iteration for the state $s = 12$, that is calculating

$$\begin{aligned} V_{j+1}(s=12) &= R(s=12) + \gamma \max_a \sum_{s'} P(s'|s=12,a)V_j(s') \\ &= 0 + 0.5 \max\{V_j(fail), \frac{1}{13}(V_j(14) + V_j(15) + \cdots + V_j(21) + 5 \times V_j(fail))\} \\ &= 0.5\frac{30}{13} \approx 1.154 \end{aligned}$$

(b) You are informed that the cards do not actually appear with equal probability and decide to use Q-learning instead of value iteration. Explain the Q-Learning algorithm as precisely as possible.

The state-action-value $Q(s, a)$ is the expected utility of acting optimally after taking the action a in state s, i.e.

$$\begin{aligned} Q(s,a) &= R(s) + \gamma \sum_{s'} P(s'|s,a)V(s') \\ &= R(s) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a'). \end{aligned}$$

The Q-learning algorithm allows to estimate the Q-values *without knowing the transition model* $P(s'|s,a)$. The principle is to apply iteratively the temporal difference update

$$\begin{aligned} Q(s,a) &\leftarrow Q(s,a) - \alpha(Q(s,a) - r - \gamma \max_{a'} Q(s',a')) \\ &\leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a')) \end{aligned}$$

from observed trial/episode transition tuples (s, a, r, s'). This update can be seen/formalized as a stochastic gradient descent step. Additionally, in some settings, the agent also needs to generate the transition tuples by exploring the environment itself. In the state s, the exploring Q-learning agent

i. takes the action $a = \arg\max_a f(Q(s,a), N(s,a))$;

ii. collects the reward $r = R(s)$ and the next state $s' \sim P(s'|s,a)$;

iii. updates $Q(s,a)$ accordingly (see TD update);

iv. increments $N(s,a)$ and

v. starts again with $s \leftarrow s'$.

The exploration function $f(q,n)$ defines the trade-off between exploitation (of good actions) and exploration (of uncertain actions). It should be increasing in q and decreasing in n (e.g. $f(q,n) = q + \frac{C}{n}$).

(c) Given the incomplete table of initial Q-values below, update the Q-values after the following episode occurred. Assume a learning rate of $\alpha = 0.5$ and a discount factor of $\gamma = 0.5$. Do not consider the values for which Q-learning does not update.

s	a	$Q(s,a)$
19	draw	-2
19	stop	5
20	draw	-4
20	stop	7
21	draw	-6
21	stop	8
fail	stop	-8

(a) Initial values.

s	a	r	s	a	r	s	a	r
19	draw	0	21	draw	0	fail	stop	-10

(b) Episode.

$$\begin{aligned} Q(19, draw) &\leftarrow (1-\alpha)Q(19, draw) + \alpha(0 + \gamma \max\{Q(21, draw), Q(21, stop)\}) = 1 \\ Q(21, draw) &\leftarrow (1-\alpha)Q(21, draw) + \alpha(0 + \gamma \max\{Q(fail, stop)\}) = -5 \\ Q(fail, stop) &\leftarrow (1-\alpha)Q(fail, stop) + \alpha(-10 + 0) = -9 \end{aligned}$$

(d) Dissatisfied with tabular Q-learning, you decide to model your Q-values with a linear model, representing them as $\sum_i w_i f_i(s,a)$. First consider the two feature functions

$$f_1(s,a) = \begin{cases} 0, \text{if a} = \text{stop} \\ 1, \text{if a} = \text{draw and } s \geq 15 \\ -1, \text{if a} = \text{draw and } s < 15 \end{cases} \qquad f_2(s,a) = \begin{cases} 0, \text{if a} = \text{stop} \\ 1, \text{if a} = \text{draw and } s \geq 18 \\ -1, \text{if a} = \text{draw and } s < 18 \end{cases} . \tag{1}$$

For which of the following partial policy tables is it possible to represent Q-values in the form $w_1 f_1(s,a) + w_2 f_2(s,a)$ while implying that policy unambiguously? Explain in one sentence.

s	$\pi(s)$
14	draw
15	draw
16	draw
17	draw
18	draw
19	draw

(a)

s	$\pi(s)$
14	stop
15	draw
16	draw
17	draw
18	stop
19	stop

(b)

s	$\pi(s)$
14	draw
15	draw
16	draw
17	draw
18	stop
19	stop

(c)

s	$\pi(s)$
14	draw
15	draw
16	draw
17	draw
18	draw
19	stop

(d)

s	$\pi(s)$
14	draw
15	draw
16	draw
17	stop
18	draw
19	stop

(e)

For any state s, if the policy selects the action a unambiguously, then $Q(s,a) > Q(s,a')$ for any action $a' \neq a$. For the policies a, b, d and e, writing these inequalities for the states 17, 18 and 19 introduces contradictions.