Université Lyon 1

Master Automatique

Artificial Intelligence & Data Analysis

04 juin 2024

Examen final
Partie 2: Apprentissage par renforcement ("Reinforcement Learning")

Durée : 45 minutes. **Aucun document autorisé.**
Répondre aux questions suivantes directement sur la feuille en cochant la bonne réponse. Un seule réponse doit être sélectionnée à chaque question, les autres sont de fausses affirmations.

---

1. Qu'est-ce que le "Reinforcement Learning" (RL) ? (1 point)
   - ☐ Une méthode d'apprentissage profond ("deep learning") supervisée
   - ☐ Une méthode basée sur l'imitation d'une loi de commande issue de l'automatique
   - ☐ Une méthode basée sur l'interaction avec un système inconnu

2. Pour un agent, que représente l'environnement ? (1 point)
   - ☐ L'ensemble des actions qu'il peut effectuer
   - ☐ Le monde extérieur à l'agent
   - ☐ Le tableau qui stocke les valeurs des états
   - ☐ Le réseau de neurones qui permet d'estimer les valeurs des actions
   - ☐ Les données utilisées pour l'entraînement ("replay buffer")

3. Quel est l'objectif du RL ? (1 point)
   - ☐ Entraîner un agent à accomplir une tâche
   - ☐ Minimiser la somme des récompenses
   - ☐ Explorer les états

4. Que désigne le "discount factor" ? (1 point)
   - ☐ Un coefficient favorisant l'exploration : ϵ
   - ☐ Un coefficient favorisant les "rewards" rapides : γ
   - ☐ Le temps maximal pour prendre une décision : β
   - ☐ C'est un synonyme de "learning rate" : α

5. Supposons qu'on se trouve dans l'état s et qu'on ait une estimation des Q-valeurs $Q(s, a)$ pour chaque action a. Qu'est-ce qu'une sélection ϵ-gourmande ("ϵ-greedy") ? (1 point)
   - ☐ choisir la meilleure action (par rapport à Q, i.e. $\arg\max_a Q(s, a)$)
   - ☐ choisir la moins bonne action
   - ☐ choisir avec une probabibilité de ϵ une action au hasard et sinon la meilleure action
   - ☐ choisir une action qui vaut moins que ϵ (i.e. telle que $Q(s, a) < \epsilon$)

6. Qu'est qu'une politique ? (1 point)
   - ☐ Un ensemble d'états

☐ Une séquence d'actions

☐ Une fonction qui à chaque état fait correspondre une action

☐ Une fonction qui à chaque couple (état, action) fait correspondre une valeur

☐ Une séquence de "rewards"

7. Quelle notation désigne habituellement une politique ? (1 point)
☐ α ☐ γ ☐ s ☐ π ☐ ϵ

8. On utilise le RL pour commander un robot. La position du robot est : (1 point)
☐ une action ☐ une politique ☐ un état ☐ une récompense ("reward")

9. On cherche à commander un robot pour la découverte de la planète Mars. L'état 1 correspond à une découverte scientifique importante, l'état 2 à une petite découverte et l'état 3 à un robot détruit. Quelle fonction reward R correspond le mieux ? (Remarque: en général une fonction reward dépend de l'état et de l'action; ici on considère une fonction qui ne dépend que de l'état) (1 point)

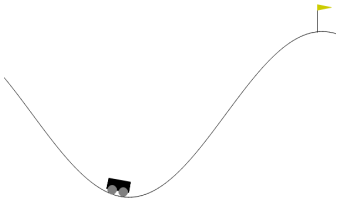
☐ $R(1) < R(2) < R(3)$ où $R(1)$ et $R(2)$ sont négatives et $R(3)$ positive

☐ $R(1) > R(2) > R(3)$ où $R(1)$, $R(2)$ et $R(3)$ sont négatives

☐ $R(1) > R(2) > R(3)$ où $R(1)$ et $R(2)$ sont positives et $R(3)$ négatives

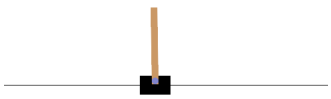
☐ $R(1) > R(2) > R(3)$ où $R(1)$, $R(2)$ et $R(3)$ sont positives

10. On souhaite apprendre à un agent (un véhicule) à s'arrêter au niveau d'un drapeau (à droite) : (1 point)



Soit x la position du véhicule et x_d la position du drapeau. Une fonction "reward" possible est $r(x) = 1$ si $x = x_d$ et $r(x) = 0$ sinon. Mais le problème est alors difficile à résoudre car les "rewards" sont presque toujours nuls ("sparse rewards").
Quelle autre fonction reward peut-on choisir ?
☐ $r(x) = x$ ☐ $r(x) = -x$ ☐ $r(x) = (x_d - x)^2$ ☐ $r(x) = -(x_d - x)^2$

11. Un pendule inverse est posé sur un chariot : (1 point)



Soit x la position du chariot ($x = 0$ étant au centre) et θ l'angle du pendule avec la verticale. En choisissant une fonction "reward" qui vaut 1 si $|x|$ et $|\theta|$ sont petits et qui vaut 0 sinon, on peut apprendre au chariot à ne pas s'éloigner du centre avec le pendule près de la verticale.
Pour apprendre à rester le plus près possible du centre avec le pendule le plus près possible de la verticale (en automatique, on parle de stabilisation), quelle autre fonction reward peut-on choisir ?
☐ $r(x, \theta) = x$ ☐ $r(x, \theta) = -\theta$ ☐ $r(x, \theta) = -x^2 - \theta^2$ ☐ $r(x, \theta) = \frac{1}{x} + \frac{1}{\theta}$

12. Lequel de ces algorithmes est un algorithme de RL adapté aux petits environnements ? (1 point)
☐ "Tabular Q-learning" ☐ "Monte Carlo Tree Search (MCTS)" ☐ "Deep Q-learning"
☐ "Policy gradient"

13. On utilise les notations standard: s désigne un état, a une action, s' l'état suivant (à partir de s quand on choisi a), A l'ensemble des actions, etc. (1 point)
Laquelle de ces équations correspond à une équation de Bellman ?

☐ $V^*(s) = \max_{a \in A} r(s, a)$

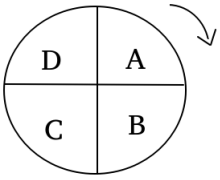
☐ $V^*(s) = \max_{a \in A} r(s,a) + V^*(s')$
☐ $V^*(s) = \max_{a \in A} r(s,a) + \gamma V^*(s')$
☐ $V^*(s) = \max_{a \in A} r(s,a) + \sum_{t=0}^{+\infty} r(s_t, a_t)$

14. Quelle relation est exacte ? (1 point)
☐ $V^*(s) = \max_{a \in A} r(s,a)$ ☐ $Q^*(s,a) = \max_{a \in A} r(s,a)$
☐ $Q^*(s,a) = r(s,a) + \gamma V^*(s')$ ☐ $V^*(s) = \arg\max_{a \in A} Q^*(s,a)$

15. On considère le système suivant : $\begin{matrix} & A(-1) & \\ B(-3) & C(?) & D(-4) \\ & E(-2) & \end{matrix}$ (1 point)
On se trouve en C. Les actions haut, bas, gauche et droite permettent respectivement d'aller en A, E, B et D avec un "reward" de -1. La valeur optimale de ces états est écrit entre parenthèses (par exemple $V^*(B) = -3$).
Quelle est la meilleure action ?
☐ haut ☐ bas ☐ gauche ☐ droite

16. On reprend le système de la question précédente. Sachant que le "discount factor" est de 0.9, que vaut $V^*(C)$? (1 point)
☐ 0 ☐ -0.9 ☐ -1 ☐ -1.9 ☐ -2.9

17. On considère le système suivant, composé de 4 états (A,B,C,D) : (1 point)



Pour chaque état, 2 actions sont possibles (avancer, attendre).
L'agent démarre dans l'état A : si il attend il reste dans l'état A, si il avance il va en B.
En B si il attend il reste dans l'état B, si il avance il va en C.
En C si il attend il reste dans l'état C, si il avance il va en D et l'épisode est fini (D est un état terminal).
La fonction reward R est défini ainsi : $R(s, attendre) = 0 \ \forall s$ (le reward est nul quel que soit l'état); $R(A, avancer) = R(B, avancer) = -1$ et $R(C, avancer) = 10$.
Sachant que le "discount factor" est de 0.5, que vaut $V^*(A)$ (i.e. quelle est la valeur de l'état A si on suit la politique optimale)?
☐ 0 ☐ 0.5 ☐ 1 ☐ 2 ☐ 5 ☐ 10

18. On considère le même système qu'à la question précédente (avec le même "discount factor"). (1 point)
Que vaut $Q^*(B, attendre)$, i.e. que vaut la Q-valeur de la paire état-action (B,attendre) ?
☐ 0 ☐ 0.5 ☐ 1 ☐ 2 ☐ 5 ☐ 10

19. Qu'affiche le programme python ci-dessous ? (1 point)
```
import numpy as np
y = np.max([-1.81, -2.782, -1.609])
print("%.2f " % (y,))
```
☐ -1.8 ☐ -2.782 ☐ -2.8 ☐ -1.609 ☐ -1.61

20. Qu'affiche le programme python ci-dessous ? (1 point)
```
import numpy as np
y = np.argmax([-1,2,3,0])
print(y)
```
☐ -1 ☐ 0 ☐ 1 ☐ 2 ☐ 3