Cours d'apprentissage par renforcement

2. Processus de décision Markovien finis

Pierre-Louis Guhur

11 septembre 2023

Table des matières

1 Introduction

Quand on découvre un nouveau domaine

- Se faire une image mentale des notions clés.
- Comprendre le sens des équations plutôt que de les apprendre par coeur.
- Chercher des exemples supplémentaires.
- Tracer un mindmap récapitulatif.
- Inutile d'aller plus loin si les notions ne sont pas claires

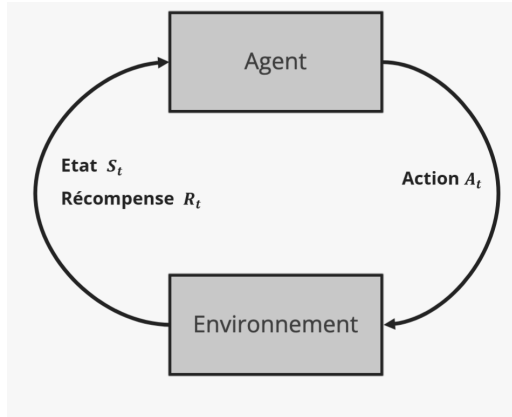
Table des matières

2 Processus de décision Markovien

Agent et environnement

2 Processus de décision Markovien

- Un agent est un système capable de percevoir son environnement et d'agir sur celui-ci.
- L'environnement est tout ce qui n'est pas l'agent.
- L'agent et l'environnement interagissent à chaque étape.
- L'agent perçoit l'état de l'environnement et choisit une action.
- L'environnement reçoit l'action de l'agent et renvoie un nouvel état et une récompense.



Exemples

2 Processus de décision Markovien

Trois exemples d'agents :

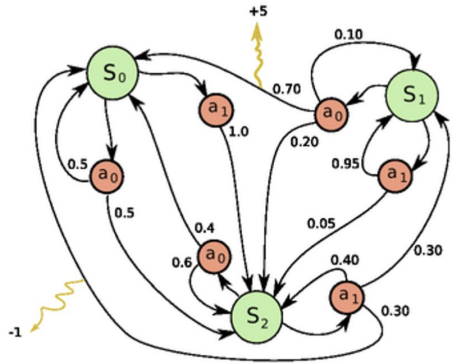
- Atari Breakout ;
- Robot aspirateur ;
- Joueur dans Starcraft.

Fonction de transition

2 Processus de décision Markovien

- L'état de l'environnement est un vecteur d'informations.
- L'action de l'agent est un vecteur de commandes.
- La fonction de transition est une fonction qui prend en entrée l'état de l'environnement et l'action de l'agent et renvoie l'état suivant.

$$f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$$



- La politique est une fonction qui prend en entrée l'état de l'environnement et renvoie l'action de l'agent.
- La politique est déterministe ou stochastique.

$$\pi : \mathcal{S} \to \mathcal{A}$$

Récompense

- La récompense est une fonction qui prend en entrée l'état de l'environnement et l'action de l'agent et renvoie un nombre réel.
- La récompense est immédiate ou cumulative.

$$r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

Attention aux effets secondaires

Les récompenses sont souvent mal définies et peuvent avoir des effets secondaires non désirés. Pour plus d'infos : `https://openai.com/research/faulty-reward-functions`.

- Un processus de décision Markovien est un processus de décision où l'état suivant ne dépend que de l'état actuel et de l'action choisie.
- Un processus de décision Markovien est défini par un tuple $(\mathcal{S}, \mathcal{A}, f, r)$.

$$f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$$

$$r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

Épisodique ou continu

- Un processus de décision Markovien est épisodique si il se termine après un nombre fini d'étapes.
- Un processus de décision Markovien est continu si il ne se termine jamais.

Table des matières

3 Équations de Bellman

Gain

3 Équations de Bellman

- Le gain est la somme des récompenses reçues par l'agent.
- Le gain est une fonction de la politique.

$$G(\pi) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

Que signifie γ ?

- $\gamma \in [0, 1]$ est le facteur d'actualisation.
- $\gamma = 0$: l'agent ne prend en compte que la récompense immédiate.
- $\gamma = 1$: l'agent prend en compte toutes les récompenses.

Lien entre le gain actuel et le gain futur

3 Équations de Bellman

$$G(\pi) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

$$G(\pi) = r(s_0, a_0) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t)$$

$$G(\pi) = r(s_0, a_0) + \gamma G(\pi')$$

Remarques :

- On peut définir le gain actuel en fonction du gain futur.
- C'est une équation récursive, appelée équation de Bellman.

Fonction de valeur

3 Équations de Bellman

- La fonction de valeur est une fonction qui prend en entrée l'état de l'environnement et renvoie le gain attendu.
- La fonction de valeur est une fonction de la politique.

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[G(\pi) \mid s_0 = s\right]$$

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[r(s_0, a_0) + \gamma V^{\pi}(s_1) \mid s_0 = s\right]$$

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} p(s' \mid s, a)\left[r(s, a) + \gamma V^{\pi}(s')\right]$$

Fonction de qualité

3 Équations de Bellman

- La fonction de qualité est une fonction qui prend en entrée l'état de l'environnement et l'action de l'agent et renvoie le gain attendu.
- La fonction de qualité est une fonction de la politique.

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[G(\pi) \mid s_0 = s, a_0 = a\right]$$

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[r(s_0,a_0) + \gamma V^{\pi}(s_1) \mid s_0 = s, a_0 = a\right]$$

$$Q^{\pi}(s,a) = \sum_{s' \in \mathcal{S}} p(s' \mid s,a)\left[r(s,a) + \gamma V^{\pi}(s')\right]$$

Lien entre la fonction de valeur et la fonction de qualité

3 Équations de Bellman

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) Q^{\pi}(s, a)$$

$$Q^{\pi}(s, a) = \sum_{s' \in \mathcal{S}} p(s' \mid s, a) \left[r(s, a) + \gamma \sum_{a' \in \mathcal{A}} \pi(a' \mid s') Q^{\pi}(s', a') \right]$$

Table des matières

4 Fonction de valeur optimale

Fonction de valeur optimale

- La fonction de valeur optimale est une fonction qui prend en entrée l'état de l'environnement et renvoie le gain maximal.
- La fonction de valeur optimale est indépendante de la politique.

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$$

Fonction de qualité optimale

4 Fonction de valeur optimale

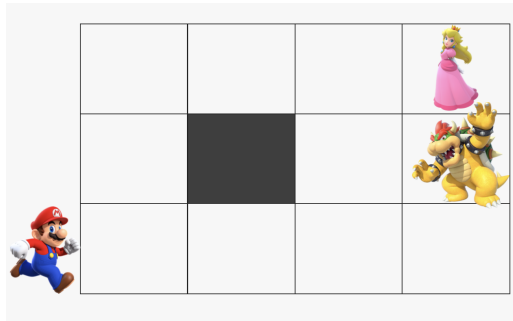
- La fonction de qualité optimale est une fonction qui prend en entrée l'état de l'environnement et l'action de l'agent et renvoie le gain maximal.
- La fonction de qualité optimale est indépendante de la politique.

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

$$Q^*(s,a) = \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \left[r(s,a) + \gamma V^*(s') \right]$$

Exemple de fonction de valeur

- Dans cette grille 3×4, l'agent peut se déplacer dans les 4 directions.
- La case Peach apporte une récompense de 1 point.
- La case Bowser apporte une récompense de -1 point.
- Les autres cases n'apportent aucune récompense.
- L'agent ne peut pas sortir de la grille.



Exemple de fonction de valeur

4 Fonction de valeur optimale

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[G(\pi) \mid s_0 = s\right]$$

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[r(s_0, a_0) + \gamma V^{\pi}(s_1) \mid s_0 = s\right]$$

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} p(s' \mid s, a)\left[r(s, a) + \gamma V^{\pi}(s')\right]$$

Cours d'apprentissage par renforcement

Merci pour votre attention