Court terme vs long terme

1 Définitions

1.1 Episodes

1.2 Récompenses

Le choix des actions par l'agent doit être guidée par la récompense espérée, mais pas uniquement la récompense à court terme qui n'est pas suffisante. L'apprentissage d'un agent doit tenir compte de l'écart qu'il peut y avoir entre une récompense à court terme et une récompense à long terme. Rechercher la récompense à court terme peut nous éloigner d'une récompense à long terme. Ainsi, lorsque face à une montagne on choisit le contournement de celle -ci, le chemin peut s'avérer plus long. Pour effectuer la balance entre ces deux horizons il faut définir une mesure (car seuls les chiffres sont pertinents). La définition courrament admise est la suivante :

$$G_t = R_{t+1} + R_{t+2} + ...R_\infty \tag{1}$$

Elle établit un lien entre la récompense espérée à l'instant t comme la somme des récompenses futures.

Cependant, plusieurs problèmes apparaîssent immédiatement :

- la récompense peut être infini si l'épisode n'a pas de fin.
- une récompense très lointaine à le même poids qu'une récompense à court terme. Or on souhaite que l'agent aille le plus vite possible vers les meilleurs récompenses

Pour corriger ces points, il est nécessaire d'ajouter un facteur de "discount" γ qui va atténuer l'effet des récompenses à long terme vis à vis des récompenses à court terme.

$$G_t = \sum_{n=0}^{\infty} \gamma^n R_{t+n+1} \tag{2}$$

où $\gamma \in [0, 1]$. Avec ce facteur, les récompenses à long terme participent de moins en moins à la récompenses espérée.

Reste à construire des algorithmes autour de cette définition pour correctement paramétré l'agent.

1.3 La fonction de valeur

L'agent suite à une action va atteindre un certain état d l'environnement. Mais comment choisir sa destination et donc l'action ? Il faut pour cela déterminer à quel point il peut être intéressant d'être dans un certain état. L'agent choisira une action parce que elle l'emmènera vers un état qui lui sera favorable en terme de récompense. Reste à quantifier cette idée d'état favorable. Il s'agit de la récompense espérée en étant dans l'état s en suivant une politique π

$$v_\pi(S) = E_\pi\{G_t|s\} \tag{3}$$

$$= \sum_a \pi(s,a) \sum_{s'} p(s,r|s',a)[r + \gamma v(s')] \tag{4}$$

C'est une équation récursive ($\propto v_\pi$) , elle décrit donc la dynamique d'un épisode complet en partant de l'état s. Elle inclue la référence à la récompense à long terme. s' est l'état précédent l'état s et la transition entre s' et s se fait par l'action a. $p(s,r|s',a)$ décrit la possibilité de passer de l'état s' à l'état s à l'aide de l'action a en obtenant une récompense r. La probabilité pouvant bien sur être nulle. La politique $\pi(s,a)$ fournie la liste des actions disponible en étant dans l'état s.

2 références

https://medium.com/@m.alzantot/deep-reinforcement-learning-demysitifed-episode-2-policy-iteration-value-iteration-and-q-978f9e89ddaa