L’apprentissage par renforcement (RL)
(Markov Decision Process)

Markov Decision Process (MDP):

C'est la fondation théorique de l'apprentissage par renforcement, elle est la base de tout processus de ce dernier.

Il est un généralement un Framework utilisé pour modéliser un problème en apprentissage par renforcement.

Le mot « Markov » ici fait référence à la propriété de Markov (Markov Property) qui est un principe fondamental de la chaine de Markov (Markov Process). La propriété de Markov sera expliquée avec la chaine de Markov dans les parties suivantes, et aussi expliquer comment transformer la chaine de Markov en MDP.

Markov Property (la propriété de Markov) :

La propriété de Markov est satisfaite lorsque l'état actuel du processus est suffisant pour prédire l'état futur du processus.

Exemple : si nous avons deux états (ensoleillé ou pluvieux) et si nous sommes dans l'état « ensoleillé », **on n’utilise pas la séquence des états précédents** pour les transitions des états futures.

Dans une autre définition. Un processus stochastique possède la propriété de Markov si la distribution de probabilité conditionnelle des états futurs du processus (conditionnée par les valeurs passées et présentes) dépend uniquement de l'état présent.

Pourquoi la propriété de Markov ?

La propriété de Markov est importante dans l'apprentissage par renforcement car les décisions et les valeurs sont supposées être uniquement en fonction de l'état actuel. Pour que celles-ci soient efficaces et informatives.

Markov Process (Markov Chain) :

La définition formelle du processus de Markov est la suivante :

- Un ensemble d'états (S) dans lesquels un système peut se trouver.
- Une matrice de transition (T), avec des probabilités de transition, qui définit la dynamique du système.

Remarque : Les observations forment une séquence d'états ou une chaîne, c'est pour ça Markov Process est aussi appelée « Markov Chain ».

La séquence d'observation dans le temps forme une chaine d'état, cette séquence est appelée l'historique.

Important : Pour appeler un système comme MP (Markov Process) **il doit remplir la propriété de Markov (Markov Property)**

Markov Reward Process (MRP) :

Maintenant nous devons élargir un peu notre modèle de processus de Markov en ajoutant une valeur pour chaque transition qui s'appelle la valeur de récompense (Reward), qui est un concept clé dans l'apprentissage par renforcement, elle peut être négative ou positif, et aussi on ajoute une valeur qui s'appelle le facteur de réduction (discount factor) qui est une valeur comprise entre 0 et 1 (généralement 0.9 ou 0.99).

La route vers MDP :

Maintenant pour transformer notre MRP en MDP et englober tous les entités de l'apprentissage par renforcement, on ajoute un ensemble finis d'actions.

Résumé:

Markov Chain (Markov Process) + Rewards = **Markov Reward Process (MRP)**

Markov Reward Process + Actions = **MDP (Markov Decision Process)**

Donc:

Markov Chain (Markov Process) + Rewards + Actions
⇨ **Markov Decision Process (MDP)**.

On peut aussi voir la relation dans la figure suivante :