

# 強化学習

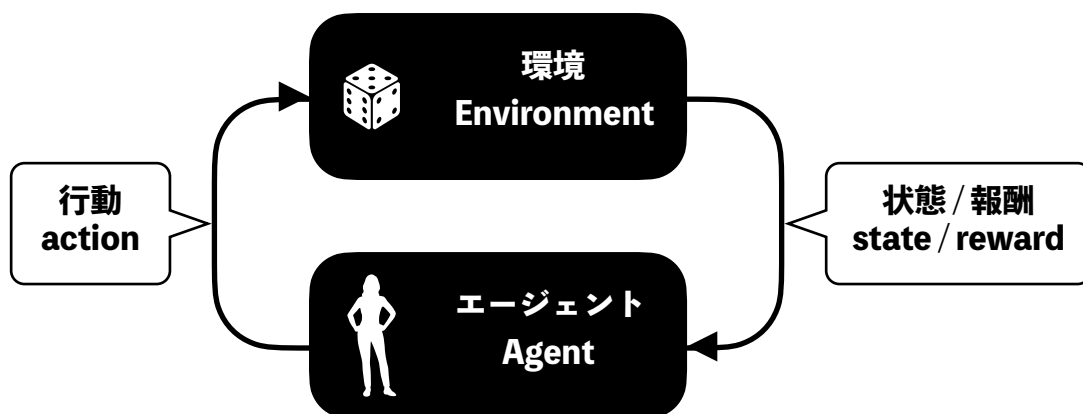
Python 中級

鈴木 敬彦

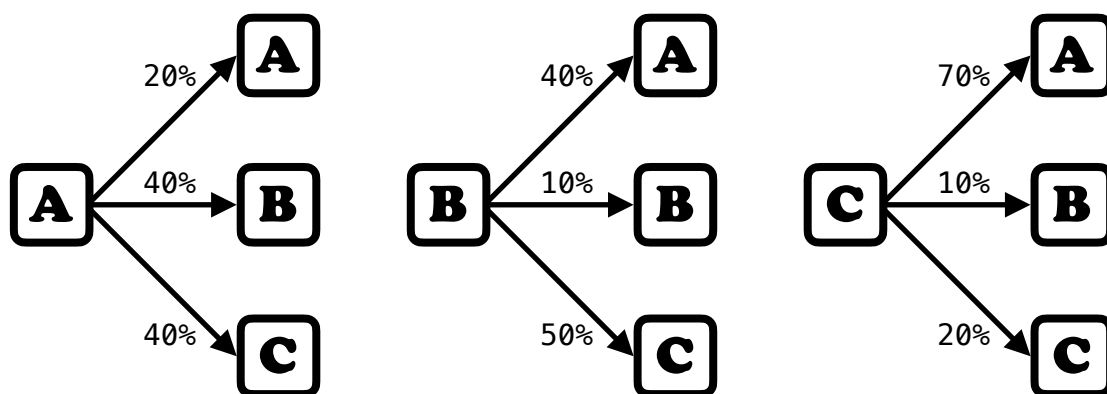
強化学習 (**Reinforcement Learning**) とは、ある環境内において、エージェントが現在の状態を観測し、次を取るべき行動を決定する問題を扱う機械学習アルゴリズムの一つです。

強化学習では、以下の一連のステップを繰り返して、より多くの報酬を得られるような方策 (policy) を学習します。

1. エージェントは環境の状態  $s$  を観測し、方策  $\pi$  に基づいて行動  $a$  を選択する。
2. 環境はエージェントの行動  $a$  と現在の状態  $s$  に基づいて、次の状態  $s'$  に遷移する。
3. エージェントは、状態  $s'$  に基づく報酬  $r$  を受け取る。
4. 次のステップへ  $t \leftarrow t + 1$



方策とは、状態  $s$  において行動  $a$  を選択する確率を意味します。また、確率過程において、確率分布が現在の状態に依存する性質をマルコフ性 (Markov Property) と呼び、そのような確率過程をマルコフ過程 (Markov Process)、マルコフ連鎖 (Markov Chain) と呼びます。強化学習の文脈においては、マルコフ決定過程 (Markov Decision Process) と呼ばれています。



強化学習のアルゴリズムはモデルベースとモデルフリーに大別されます。「モデル」とは、環境の状態遷移と報酬を予測する関数を意味します。(つまり、現在の状態から将来の状態と報酬を予測する。) この「モデル」に基づいてエージェントが行動を決定するアルゴリズムはモデルベースと呼ばれ、代表的な例としては、モンテカルロ木探索 (Monte Carlo Tree Search) が有名です。一方、「モデル」を使用しないアルゴリズムはモデルフリーに分類されます。これは、いわば、

エージェントに試行錯誤を通して学習させるアルゴリズムです。モデルフリーのアルゴリズムは、更に、方策オン（on-policy）型と方策オフ（off-policy）型に分類されます。方策オン型のアルゴリズムは、現在の方策に従った行動から学習を行い、方策オフ型のアルゴリズムは、現在の方策を前提としない行動から学習を行います。

## Q 学習 / Q-learning

Q 学習は、機械学習のアルゴリズムの一種で、モデルフリー、方策オフ型のアルゴリズムです。十分な学習が実施されていれば、有限 MDP において、最適な行動選択方策を特定できることが知られています。

Q 学習では、状態と各状態における各行動に対する価値からなる表を作成します。この表は Q テーブル、表の値は Q 値と呼ばれます。Q 学習では、学習を通してこの Q 値を最適化します。十分な学習を終えた後には、各状態の Q 値は将来の報酬を反映した値となり、最適な行動選択方策を得ることが可能となります。

Q テーブル		行 動							
		a0	a1	a2	a3	a4	a5	a6	a7
状 態	0	Q(0,0)	Q(0,1)	Q(0,2)	Q(0,3)	Q(0,4)	Q(0,5)	Q(0,6)	Q(0,7)
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	255	Q(255,0)	Q(255,1)	Q(255,2)	Q(255,3)	Q(255,4)	Q(255,5)	Q(255,6)	Q(255,7)
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	511	Q(511,0)	Q(511,1)	Q(511,2)	Q(511,3)	Q(511,4)	Q(511,5)	Q(511,6)	Q(511,7)

初期状態で Q テーブルは任意の値（一般的には 0）に初期化されます。次に、エージェントが環境内で行動して得られた報酬から新しい Q 値を計算し Q テーブルを更新していきます。初期状態の方策に意味はないため、行動はランダムな選択により決定します。学習の間を通して一貫してランダムな行動を選択するかは実装者次第ですが、後述の  $\epsilon$ -グリーディー法 ( $\epsilon$ -greedy method) 等を用いるのが一般的です。Q 値の更新には、次のベルマン方程式を用います。

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t))$$

矢印は、矢印の左側の値を右側の値で更新することを表しています。 $t$  はステップを表し、 $t+1$  は次にエージェントが行動を選択する時のステップを表します。 $s$  は状態 (state) を、 $a$  は行動 (action) を、 $r$  は報酬 (reward) を表します。 $s_t$  は時刻  $t$  での状態を、 $a_t$  はその時に選択した行動を意味し、 $Q(s, a)$  は状態  $s$  と行動  $a$  に対する Q 値を表し、 $Q(s_t, a_t)$  は時刻  $t$  における状態  $s_t$  で選択した行動  $a_t$  に対する Q 値を意味します。その結果得られた報酬が  $r_t$  で、遷移した状態が  $s_{t+1}$

となります。 $\max_a Q(s_{t+1}, a)$  は、状態  $s_{t+1}$  において  $Q$  値が最大となる行動  $a$  を選択した時の  $Q$  値を意味します。（つまり、状態  $s_{t+1}$  における各行動に対する  $Q$  値の最大値です。）

$\alpha$  は学習率 (learning rate) と呼ばれ、0 から 1 の間で設定され、学習の度合いを決定します。0 であれば何も学習せず、1 であれば最後に学習した値のみが反映されます。一般的にこの値は定数です。 $\gamma$  は割引率 (discount factor) と呼ばれ、これも 0 から 1 の間で設定されます。0 であれば目の前の報酬にしか反応しないエージェントとなり、1 であれば長期的に大きな報酬を得ようとするエージェントになります。一般的にこの値も定数です。

$$\begin{array}{c}
 \text{更新} \\
 \downarrow \\
 \underbrace{Q(s_t, a_t)}_{\text{新しい値}} \leftarrow \underbrace{Q(s_t, a_t)}_{\text{現在の値}} + \underbrace{\alpha}_{\text{学習率}} \left( \underbrace{r_t}_{\text{割引率}} + \underbrace{\gamma \max_a Q(s_{t+1}, a)}_{\text{将来期待できる最適値}} - \underbrace{Q(s_t, a_t)}_{\text{現在の値}} \right)
 \end{array}$$

この式は以下のようにも書き換えられます。

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha r_t + \alpha \gamma \max_a Q(s_{t+1}, a)$$

$(1 - \alpha)Q(s_t, a_t)$  は、 $(1 - \alpha)$  で重み付けされた現在の  $Q$  値です。 $(1 - \alpha)$  は学習率に対して忘却率と解釈することもできます。 $\alpha r_t$  は学習率  $\alpha$  で重み付けされた報酬です。 $\alpha \gamma \max_a Q(s_{t+1}, a)$  は、学習率  $\alpha$  と割引率  $\gamma$  で重み付けされた状態  $s_{t+1}$  で期待できる報酬の最大値です。

また、全てのエピソード終端の状態に対する  $Q$  値は更新されるべきではなく、その状態から得られる報酬に設定します。

## $\epsilon$ -グリーディー法 / $\epsilon$ -greedy method

学習の段階において、十分な探索を行うためにランダムな選択を取るための手法の一つです。 $\epsilon$  は 0 から 1 の間に設定され、乱数が  $\epsilon$  未満であればランダムな行動を、 $\epsilon$  以上であれば方策に従った行動を選択します。