1 強化学習の位置づけ

1.1 強化学習における問題設定

強化学習では、遷移先の状態が、直前の状態とそこでの行動にのみ依存することと、報酬は、直前の状態と遷移先の状態にのみ依存するという性質をもつ環境を取り扱う。このような性質をマルコフ性 (Markov property)、そして環境をマルコフ決定過程 (Markov Decision Process, MDP) という。MDP の構成要素は次の 4 つである。

- 状態 (State): $s \in \mathcal{S}$
- 行動 (Action): $a \in \mathcal{A}$
- 遷移関数 (Transition function): $T(s,a) : \mathcal{S} \times \mathcal{A} \to \mathcal{S} \times [0,1]$
  状態と行動を引数に取り、遷移先 (次の状態) と遷移確率を出力する関数
- 報酬関数 (Reward function): $R(s,s') : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$
  状態と遷移先 (次の状態) を引数に取り、報酬を出力する関数 (行動を引数に加える場合もある)

状態を引数に取り、行動を出力する関数を戦略 (Policy)$\pi(a|s)$、戦略に従って動く行動主体をエージェント (Agent) という。報酬は直前の状態と遷移先の状態に応じて決まるが、これを即時報酬 (Immediate reward)r_t という。エージェントは即時報酬の合計を最大化するように戦略 $\pi(a|s)$ を学習する。即時報酬の総和が最大になるような行動をとりたいが、未来の即時報酬はエピソードが終了するまでは分からないので、実際には即時報酬の総和を見積もる (予想を立てる) ことになる。見積もりは不確かな値であるため、未来の報酬は、割引率 γ を用いて割り引く。未来の報酬ほど、不確かさが増すので割引を強く働かせる。これより、ある時刻 t における報酬の総和の見積もりは、次のように書ける (エピソードは時刻 T で終了する)。

$$G_t \equiv r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots \gamma^{T-t-1} r_T = \sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1} \tag{1}$$

G_t は次のように、再帰的に書くことができる。

$$\begin{aligned} G_t &\equiv r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots \gamma^{T-t-1} r_T && (2)\\ &= r_{t+1} + \gamma \left(r_{t+2} + \gamma r_{t+3} + \cdots \gamma^{T-t-2} r_T \right) && (3)\\ &= r_{t+1} + \gamma \sum_{k=0}^{T-t-2} \gamma^k r_{t+k+2} && (4)\\ &= r_{t+1} + \gamma G_{t+1} && (5) \end{aligned}$$

1.2 価値の定義とベルマン方程式

報酬の総和を見積もった値 G_t の問題は、将来の即時報酬の値が判明している必要がある点と、その即時報酬の値が必ず得られると予想している点の 2 つある。即時報酬がどのような値になるかは行動によって異なるほか、その行動による遷移先の状態は確率的に決まるため、即時報酬の値がどのようになるかは分からない。将来の即時報酬が判明していなければならないという問題は、G_t を再帰的に表現することで回避できる。そして、2 つ目の問題については G_t の期待値を取ることで解決できる。

戦略 π に基づいて行動するとき、状態 s において行動 a を選択する確率は $\pi(a|s)$ であるほか、遷移先の状態 s' は遷移確率 $T(s'|s,a)$ によって決まる。戦略 π に基づき行動した結果得られる価値を $V^\pi(s)$ と表すとき、この $V^\pi(s)$ は G_t の期待値として次のように定義できる。

$$V^\pi(s_t) = \mathbb{E}_\pi \left[r_{t+1} + \gamma V^\pi(s_{t+1}) \right] \tag{6}$$

報酬 r_{t+1} を、状態 s から s' に遷移したときに得られる報酬を表す、報酬関数 $R(s,s')$ で表すことにすれば、上記は次のように書ける (これをベルマン方程式という)。

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} T(s'|s,a) \left(R(s,s') + \gamma V^\pi(s') \right) \tag{7}$$

2 動的計画法

2.1 価値反復法

以下で表される価値 $V(s)$ を、収束するまで繰り返し計算する。

$$V^{(t+1)}(s) \leftarrow \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} T(s'|s,a) \left(R(s,s') + \gamma V^{(t)}(s') \right) \tag{8}$$

方策 $\pi(a|s)$ に ϵ-greedy を使う場合は次の計算を繰り返す。$\mathcal{A}(s) \subseteq \mathcal{A}$ は、状態 s の下で取ることのできる行動の集合 ($\mathcal{A}$ の部分集合) である。

$$V^{(t+1)}(s) \leftarrow \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} T(s'|s,a) \left(R(s,s') + \gamma V^{(t)}(s') \right) \tag{9}$$

$$\pi(a|s) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}(s)|} & (a = \arg\max_{a' \in \mathcal{A}} Q(s,a') \text{ のとき}) \\ \frac{\epsilon}{|\mathcal{A}(s)|} & (\text{それ以外}) \end{cases} \tag{10}$$

2.2 動的計画法

戦略 $\pi(a|s)$ と価値 $V(s)$ を交互に最適化していく。

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} T(s'|s,a) \left(R(s,s') + \gamma V^\pi(s') \right) \tag{11}$$

$$\pi(a|s) = \begin{cases} 1 & (a = \arg\max_{a' \in \mathcal{A}} Q(s,a') \text{ のとき}) \\ 0 & (\text{それ以外}) \end{cases} \tag{12}$$