第2回 強化学習ゼミ (8/12)

まさああ

復習

強化学習とは…

$$\begin{tikzcd} \pi \arrow[r, bend left, "\text{方策評価}"] & Q, V \arrow[l, bend left, "\text{方策改善}"] \end{tikzcd}$$

の繰り返し。

今からやること。

- πをどのように定め、QやVをどう改善するのかを見ていく。
- ベルマン方程式は状態遷移確率がわかっている状況なら解けるが、そうでないとき解けない。
(あと計算量が爆発する)

・ 状態遷移確率がわかっている解法

→ モデルベース ← Bellmanが解いた方

〃 わかっていない

→ モデルフリー

→ モデルフリーのアプローチを見ていく。

3.6 モンテカルロ法

- アイデア

たくさん試行錯誤をする。

→ 状態の遷移は近似的に状態遷移確率に従う(大数の法則)

→ 報酬の平均は期待値に収束。

何度も繰り返して得られた経験をもとに
状態価値や状態行動価値を推定する方法を
モンテカルロ法という。

やることの大枠

以下をM回やる. ($m=1 \dots M$)

① 状態列 $S_0, S_1 \cdots S_{T_m}$ をπによって生成する.

② 状態列に対する報酬列 $r_0, r_1 \cdots r_{T_m}$ を観測する.

③ 報酬列に対して状態価値を推定する

④ 状態価値を更新.

③をどうやってするか?

→
- 初回訪問モンテカルロ法.
- 逐一訪問モンテカルロ法

初回訪問モンテカルロ法

状態列 $S_1 \dots S_{T_m}$ について、

状態 S_t が初めての訪問なら、収益 $R(S_t)$ を

$$R(S_t) \leftarrow R(S_t) + \sum_{t'=t+1}^{T_m} r_{t'}$$

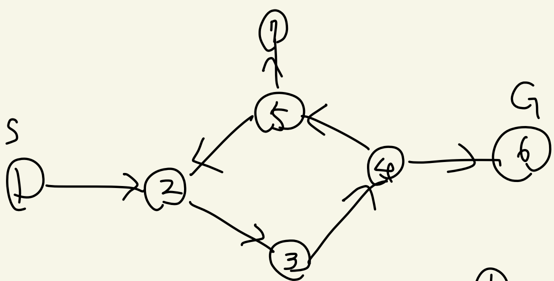
状態価値 $V(S_t)$ を

$$V(S_t) \leftarrow \frac{1}{m} R(S_t)$$

とする。

③の部分。

※ 初回訪問という条件を入れているのは、観測された状態列の中に同じ状態が複数回あらわれて、各状態が不平等な評価になるのをさけるため。



① → ② → ③ → ④ → ⑤ → ②

・逐一モンテカルロ法

初回訪問モンテカルロ法では状態価値を評価した。

モデルが不明なので、とるべき行動がわからない。

→ 状態行動価値を評価する。

状態行動価値最大の行動をすればよい

・問題点

状態行動価値 $Q(s, a)$ について、

(s, a) が観測されないと $Q(s, a)$ が評価できない。

例

$a = \{0, 1, 2, 3, 4\}$ とする。

$\pi(a|s) = \frac{a}{10}$ のとき、

$a=0$ が観測される確率は 0。

~問題の解決~

① すべての行動の確率が0でないと仮定する.

$$
\left\{
\begin{array}{l}
\cdot \quad R(s_t) \leftarrow R(s_t) + \sum_{t'=t+1}^{T_m} r_{t'} \\[2em]
\cdot \quad Q(s_t, a_t) \leftarrow \frac{1}{m} R(s_t)
\end{array}
\right.
$$

とすればOK

「モンテカルロ-ES法」という.

② 方策オン型モンテカルロ制御

確率的な方策を用いることで、任意の状態行動タプル(s, t)

について　$\pi(s, a) > 0$を保証する。

遺伝的アルゴリズムみたい。

1. ε-greedy 法。

εを十分小さな正の値とする $(0 < \varepsilon < 1)$

$$\pi(a_t | S_t) := \begin{cases} \dfrac{1-\varepsilon}{|B(S_t)|} + \dfrac{\varepsilon}{|A(S_t)|} & \left(a_t = \underset{a \in A(S_t)}{\arg\max}\, Q(S_t, a)\right) \\[2ex] \dfrac{\varepsilon}{|A(S_t)|} & \text{otherwise.} \end{cases}$$

ただし、$A(S_t)$: 時刻S_tのときの行動空間

$$B(S_t) = \left\{ a \in A(S_t) \,\middle|\, a = \underset{b \in A(S_t)}{\arg\max}\, Q(S_t, b) \right\}$$

※ 教科書だと $\sum \pi(a_t | S_t) = 1$ にならない。

($\varepsilon = \frac{1}{2^n}$ とかすると良さそう?)

softmax法

温度パラメータ$\tau \in \mathbb{R}_{>0}$を用いて、

$$\pi(a_t | s_t) := \frac{\exp\left(\frac{Q(s_t, a)}{\tau}\right)}{\sum_{b \in A} \exp\left(\frac{Q(s_t, b)}{\tau}\right)}$$

τは学習率をコントロールするパラメータで、

$$\lim_{\varepsilon \to 0} \pi_{greedy}(a_t | s_t) = \lim_{\tau \to 0} \pi_{softmax}(a_t | s_t)$$

(証明)

文字が多いので、$f_k : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$として、

$$f_k(x) = \frac{\exp\left(\frac{b_k}{x}\right)}{\sum_{b_i \in A} \exp\left(\frac{b_i}{x}\right)}$$

の$x=0$付近での挙動を調べる。

$$f_k(x) = \frac{1}{\sum_{b_i \in A} \exp\left(\frac{b_i - b_k}{x}\right)} \quad \left(\text{分母分子}\ \exp\left(\frac{b_k}{x}\right) > 0 \ \text{で割る}\right)$$

である

$$\sum_{b_i \in A} \exp\left(\frac{b_i - b_k}{x}\right)$$ について,

i) $b_i - b_k > 0$ のとき, (b_k が $\max(A)$ でない)

$x > 0$ より,

$$\frac{b_i - b_k}{x} > 0.$$

$$\lim_{x \to 0} \frac{b_i - b_k}{x} = \infty$$

$$\lim_{x \to 0} \exp\left(\frac{b_i - b_k}{x}\right) = \infty$$

$$\lim_{x \to 0} f(x) = 0 \quad \left(\forall i, \exp\left(\frac{b_i - b_k}{x}\right) > 0 \text{ より}\right)$$

ii) $b_i - b_k < 0$ のとき.

$$\frac{b_i - b_k}{x} < 0.$$

$$\lim_{x \to 0} \frac{b_i - b_k}{x} = -\infty$$

$$\lim_{x \to 0} \exp\left(\frac{b_i - b_k}{x}\right) = 0.$$

iii) $b_i = b_k$ のとき,

$$\frac{b_i - b_k}{x} = 0.$$

$$\lim_{x \to 0} \exp\left(\frac{b_i - b_k}{x}\right) = 1.$$

i), ii), iii) をまとめると、

- $b_k = \max(A)$ のとき

$B = \{ b \in A \mid b = \max(A) \}$

として

$$\lim_{x \to 0} f(x) = \frac{1}{|B|}$$

- $b_k < \max(A)$ のとき

$$\lim_{x \to 0} f(x) = 0 , \qquad \text{㊥}$$

③ 方策オフ型モンテカルロ制御

評価・改善される方策π（推定方策）と、
実際の行動する方策π'（行動方策）
という2つの方策を用いる

→ ・推定方策と挙動方策では収益が異なる
・改善した方だけ収益が良くなる

→ 改善した際の方策ごとの差をなくすことを考える

→ 重要度（重み）を定義する

推定方策 π において状態行動タプル列 $(s_0, a_0) \cdots (s_T, a_T)$ と収益 R が観測される確率を P,

挙動方策 π' において、同じ状態行動タプル列と収益が観測される確率を P' とする。

この時、確率的に推定方策では挙動方策の $\frac{P}{P'}$ 倍多く観測している。

- $\frac{P}{P'} = 1$ のとき 推定方策と挙動方策は同じ（方策オン）
- $\frac{P}{P'} > 1$ 推定方策では多く見ている
  → 強い意味をもつ（重要度が高い）
- $\frac{P}{P'} < 1$ 重要度低め。

$\frac{P}{P'}$ 倍の重みをつける（$\frac{P}{P'}$ 回みたことにする）

Algo

$$R_{sum} \leftarrow R_{sum} + \frac{P}{P'} R$$

$$k \leftarrow k + \frac{P}{P'}$$

$$Q^{\pi}(s, a) = \frac{1}{k} R_{sum}$$

(s_t, a_t)を観測し、$(s_{t+1}, a_{t+1}), (s_{t+2}, a_{t+2})\cdots$

を観測する確率 $p^{\pi}(s_t, a_t)$ は以下で表わされる。

$$p^{\pi}(s_t, a_t) = P(s_{t+1} | s_t, a_t) \prod_{k=t+1}^{\infty} \pi(a_k | s_k) P(s_{k+1} | s_k, a_k)$$

重み $w(s_t, a_t)$ について

$$w(s_t, a_t) := \frac{p^{\pi}(s_t, a_t)}{p^{\pi'}(s_t, a_t)}$$

$$= \frac{\cancel{P(s_{t+1} | s_t, a_t)} \prod \pi(a_k, s_k) \cancel{P(s_{k+1} | s_k, a_k)}}{\cancel{P(s_{t+1} | s_t, a_t)} \prod \pi'(a_k, s_k) \cancel{P(s_{k+1} | s_k, a_k)}}$$

$$= \prod_{k=t+1}^{\infty} \frac{\pi(a_k | s_k)}{\pi'(a_k | s_k)} \quad (\text{状態せんい確率に依らない})$$

推定方策が決定論的なものである場合

$$w(s_t, a_t) = \begin{cases} \prod_{k=t+1}^{\infty} \frac{1}{\pi'(s_k, a_k)} & a_k = \arg\max_{a \in A} Q(s_k, a) \\ 0 & \text{otherwise} \end{cases}$$

→最適行動以外は学習する意味がない。

3.7 TD学習

- モンテカルロ法は試行が終了するまで更新できない
  という問題点

→ この欠点を回避したのがTD学習

「1-step TD法」,「k-step TD法」,「TD(λ)法」

があって 今日は 1-stepとk-stepを見る.

1. 1-step TD法

名のとおり 1ステップごとに方策を改善する.

ベルマン方程式の解のひとつは,

$$\sum_{a \in A} \sum_{s_{t+1} \in S} \pi(a|s_t) P(s_{t+1}|s_t, a) = 1 \text{ より},$$

$$V(s_t) = r_t + \gamma V(s_{t+1}) \text{ である.}$$

$Q(s,a)$も同様に(代入して確かめることができる)

$Q(s,a) = r_t + \gamma Q(s_{t+1}, a_{t+1})$ が解のひとつである.

$V(s_t)$ と $Q(s_t, a_t)$ を最適解に近づけることを考える.

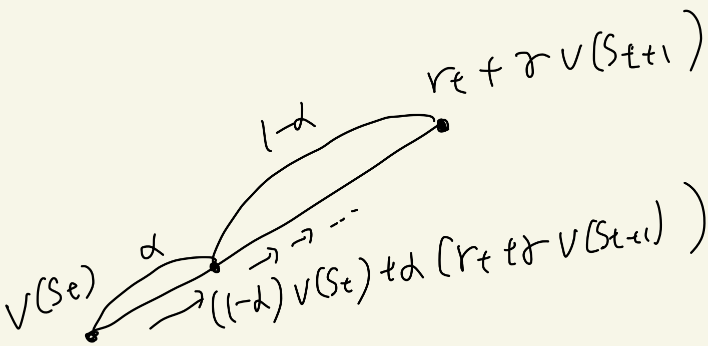
$$
\begin{cases}
V(s_t) \leftarrow V(s_t) + \alpha\left(r_t + \gamma V(s_{t+1}) - V(s_t)\right) \\
Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left(r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right)
\end{cases}
$$

~お気持ち~

右辺を式変形すると,

$$(1-\alpha)V(s_t) + \alpha\left(\underbrace{r_t + \gamma V(s_{t+1})}_{\text{解}}\right)$$

図で書いてみる.



2. k-Step TD法

1-Step TD法だと局所解になる

(例は教科書を参考)

可能性があるので、k個先まで先読みすることを考える。

ベルマン方程式の解は

$$V(S_t) = r_t + \gamma V(S_{t+1})$$

であった。これをもう少し展開して、

$$V(S_t) = r_t + \gamma (r_{t+1} + \gamma V(S_{t+2}))$$

$$\vdots$$

$$= \sum_{n=0}^{k-1} (\gamma^n r_{t+n} + \gamma^k V(S_{t+k}))$$

(証明は数学的帰納法)

これをTD法の更新とする。

$$\boxed{V(S_t) \leftarrow (1-\alpha) V(S_t) + \alpha \sum_{n=0}^{k-1} (\gamma^n r_{t+n} + \gamma^k V(S_{t+k}))}$$

・メリット

k個先まで読むので局所解に陥りにくい。

・デメリット

学習速度が遅い。

トレードオフ。

k→∞でモンテカルロと一致すると書いてるけど

ホントか？って顔になってます。

ここまで。