強化学習　第3章
「探索と活用のトレードオフ」

1　方策モデル

環境が未知である場合，エージェントは環境に対し行動を入力し報酬や次状態を観測することでデータを収集し，データから方策を学習する．この「環境に対し行動を入力」するための方策が求める最適方策とは別に定義されている必要がある．この時の行動選択の基準には，データ収集のための「探索」とリターンを最大化しようとする「活用」があり，両者のバランスを取ることが重要となる．(探索と活用のトレードオフ)

- 効用関数に従い間接的に方策を規定するアプローチ
  - 貪欲方策
  - ϵ 貪欲方策
  - ソフトマックス方策
- 数理モデルによって直接的に方策を規定するアプローチ
  - 関数近似を用いた強化学習法

1.1　効用関数にもとづく方策モデル

リターン：

$$C_0 = \lim_{T\to\infty} \sum_{t=0}^{T} \gamma^t R_t$$

を用いて効用関数を行動価値関数 Q^π の推定値 $q(s,a)$ を用いる．

$$Q^\pi(s,a) \coloneqq \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|A_t, S_t\sim p_T}[C_0|S_0 = s, A_0 = a]$$

- 貪欲方策モデル：効用関数が最大になる行動を選択する決定的方策．データ活用に特化した方策．

$$\pi_{\text{greedy}}(s;q) \coloneqq \arg\max_{a\in\mathcal{A}} q(s,a)$$

- ϵ 貪欲方策モデル：確率 $\epsilon \in [0,1]$ でランダムに行動を選択して，それ以外は貪欲方策に従う．

$$\pi_e(a|s;q,\epsilon) = \begin{cases} 1-\epsilon+\frac{\epsilon}{|\mathcal{A}|}, & (\text{if } a = \arg\max_{b\in\mathcal{A}} q(s,b)) \\ \frac{\epsilon}{|\mathcal{A}|}, & (\text{otherwise}) \end{cases}$$

貪欲方策にランダム性を取り入れることでデータの探索と活用のトレードオフを考慮している．

- ソフトマックス方策モデル：状態 s にある時，行動 a を選択する確率を次のように定める．

$$\pi_s(a|s;q,\beta) = \frac{\exp(\beta q(s,a))}{\sum_{b\in\mathcal{A}}\exp(\beta q(s,b))}$$

$0 \leq \beta$ が大きければ効用 q が高い行動を高い確率で選択し，β が小ならばランダムに行動を選択するようになる，ソフトマックス関数は「マックス関数」のなだらかなバージョンとして解釈することができ，βq の大小関係を保存しつつ行動 a の確率を与える．

ソフトマックス方策 π_s を $q(s,b)$ で偏微分する．$b \neq a$ の時，

$$\begin{aligned} \frac{\partial \pi_s(a|s;q,\beta)}{\partial q(s,b)} &= \frac{-\beta\exp(\beta q(s,a))\exp(\beta q(s,b))}{\{\sum_{b\in\mathcal{A}}\exp(\beta q(s,b))\}^2} \\ &= -\beta\pi_s(a|s,q,\beta)\pi_s(b|s,q,\beta) \end{aligned}$$

$b = a$ の時，

$$\begin{aligned} \frac{\partial \pi_s(a|s;q,\beta)}{\partial q(s,b)} &= \frac{\beta\exp(\beta q(s,a))\{\sum_{b\in\mathcal{A}}\exp(\beta q(s,b))\} - \beta\exp(\beta q(s,a))\exp(\beta q(s,b))\}}{\{\sum_{b\in\mathcal{A}}\exp(\beta q(s,b))\}^2} \\ &= \beta\pi_s(a|s,q,\beta)(1-\pi_s(a|s,q,\beta)) \end{aligned}$$

2　楽観的方策

状態が変化しない単純な問題である多腕 Bandit 問題を考える．すなわち，エージェントは複数の Bandit から一つを選び，$0 \leq r \leq 1$ の報酬を得る．また問題は即時報酬の最大化であるとする (割引率 $\gamma = 0$)．例えば，各 Bandit から今までの行動で得られた報酬の標本平均が次の様だったとする．

- Bandit1… 0.7 (0.6)
- Bandit2… 0.3 (0.8)
- Bandit3… 0.5 (0.3)

ただし，括弧内は真の報酬の期待値である．現在は Bandit1,Bandit3 を過大評価していて Bandit2 を過小評価している状況である．この時，前述の方策モデルによって行動を選択していくと，推定効用の高いものほど選択されやすいので，1,3 の経験は今後多く得られ，効用の推定値を修正できる可能性がある．しかし，Bandit2 は選ばれづらく 1,3 よりも評価を正す機会が少ない，という問題がある．
そこで,「楽観的方策」というヒューリスティックがよく用いられる．すなわち，不確実度 $u : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ なる指標を導入して，効用関数を u で補正した $\tilde{q}$ によって行動を選択する．

$$\tilde{q}(s, a) \coloneqq q(s, a) + u(s, a)$$

例：UCB1 法
時点 t において，これまでに行動 a を選択した回数を $n_t(a)$ と置く．

$$n_t(a) \coloneqq \sum_{k=0}^{t-1} \mathbb{I}(a_k = a)$$

また，時間ステップ t における行動 a の効用 q を報酬の標本平均：

$$q(a) = \frac{1}{n_t(a)} \sum_{k=0}^{t-1} r_k \mathbb{I}(a_k = a)$$

とする．これが上記の 0.7,0.3,0.5 に対応する．さらに不確実度 u を

$$u(a) \coloneqq \sqrt{\frac{2 \log t}{n_t(a)}}$$

と定義し，$\tilde{q}(a) = q(a) + u(a)$ をもとに行動を選択する．