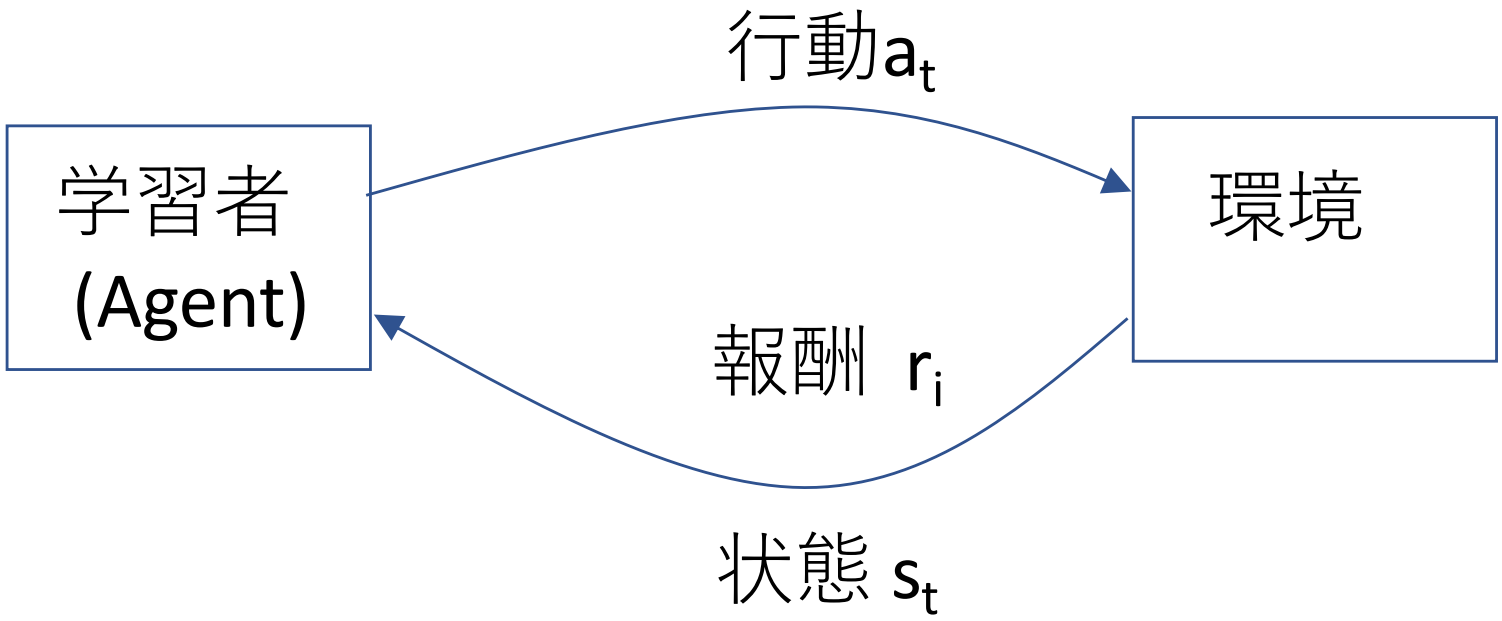
知識工学

岡山大学大学院
講師　竹内孔一

本日の内容

- 強化学習
  - 状態価値関数
  - 政策
  - TD学習

強化学習の枠組



学習者は環境から状態 s_t を受け取り，ある行動a_t をとる．
それにより報酬を得る．

学習者は環境から状態 s_t を受け取り，ある行動a_t をとる．
それにより環境がs_{t+1}に遷移．報酬r_{t+1}を得る．

強化学習の枠組

- 状態 s_t (state)
  - エージェントが時刻tで取る状態
- 行動a_t (action)
  - エージェントが時刻tで取る行動
- 報酬 r_t (reward) ・収益 R_t (return)
  - 報酬: エージェントが行動により得る値
  - 収益: 時刻 t (またはt+1)以降に得られる報酬の総量
- 政策π(s,a) (policy)
  - 状態sのときに行動aをとる関数
- 価値関数 V(s) (state-value function)
  - 状態sで将来得られる報酬の総量(=収益)の期待値 (つまり予測値)
- 行動価値関数Q(s,a) (action-value function)
  - 状態sで行動aを取るとき将来得られる報酬の総量(=収益)の期待値
- 環境モデル
  - 状態sで行動aを取ったとき，次にどういう状態に行くか，報酬はあるかあるとしたらいくらか，エージェントに与える

強化学習の基本枠組(これで全部)

- マルコフ決定過程
  - 状態遷移モデル
- エージェントの行動
  - 状態 s_tをで行動a_tを選択
  - 環境から次状態$_{st+1}$と報酬$_{rt+1}$を得る
  - (注) 行動a_tで報酬r_{t+1} (教科書などによるので注意)
- 収益
  - これから得られる報酬の総量 (t はTまで)
  - 将来の収益は割り引いて考える γ (割引率)
- 状態価値V(s) vs. 行動価値Q(s, a)
  - (ある状態sでの価値)　vs.　(状態sで行動aの時の価値)
- 政策πによって違う値をとる
  - 状態価値　$V^{\pi}(s)$　vs.　行動価値 $Q^{\pi}(s, a)$
    - 期待値を求めて，行動選択の指針にする
    - 状態価値を使うか行動価値を使うかはユーザが選択
- 学習方法 (状態V(s)か行動Q(s,a)の学習か2種)
  - V(s)の学習: TD 学習 (Temporal difference learning)
    - 各状態sでの価値V(s)が数値として求まる
  - Q(s,a)の学習(1): 方策オン型学習
    - SARSA (方策πに従った学習法)
  - Q(s,a)の学習(2): 方策オフ型学習
    - Q-learning (方策は関係無く価値最大の行動をとると固定) 簡単に利用できる

強化学習で行動を決める (政策π)

- 先に$V^{\pi}(s)$ が求まった時に，どうエージェントが行動するかを考えてみよう
  - 強化学習で状態価値関数V(s)が求まったとする(下記)

例えば
$V^{\pi}(s=(1,1)) = 2.1$

スタートsから
どこに行くか

座標	1	2	3	4
1	S 2.1	5.8	8.1	9.5
2	3.5	9.6	20.5	20.8
3	4.2	10.8	22.1	30.3

ある状態s で行動 aを選ぶのは政策 π (s, a) の仕事.
もし，グリーディ手法 (価値の高いものを選ぶ) ならば

S= (1,1) -> (1,2) -> (2,2) -> (2,3) -> (3,3) -> (3,4) と進む

ただし行動aは上下左右のみ

政策の例

■ランダム手法

■グリーディ手法 (greedy)

■e-グリーディー手法(e-greedy)

■ソフトマックス手法

状態価値関数を求める

■収益の期待値の最大化

■状態価値の再帰的関係とBellman方程式

$$V^{\pi}(s_t) = E_{\pi}\{R_t \mid s_t\} = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t\right\}$$

$$= E_{\pi}\left\{\gamma_t + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t\right\}$$

$$= \sum_{a} \pi(s,a) \sum_{s'} P_{ss'}^{a}[R_{ss'}^{a} + \gamma V^{\pi}(s')]$$

TD学習

■状態価値V(s) を更新する考え方

$$\mathrm{V}(s) \leftarrow V(s) + \alpha \Delta \mathrm{V}(s)$$

■前のスライドの差分ΔV(s)

$$\mathrm{V}(s) \leftarrow r_t + \gamma \mathrm{V}(s')$$

$$\Delta\, \mathrm{V}(s) \leftarrow r_t + \gamma V(s') - \mathrm{V}(s)$$

■更新式

$$\mathrm{V}(s) \leftarrow \mathrm{V}(s) + \alpha\{r_t + \gamma \mathrm{V}(s') - \mathrm{V}(s)\}$$

$$\mathrm{V}(s) \leftarrow (1-\alpha)\mathrm{V}(s) + \alpha\{r_t + \gamma \mathrm{V}(s')\}$$

➔V(s)の更新式．Q学習でも同様の式が出てくる

Vπ(s)を計算する

■練習問題(1)を解いてみよう