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
  - 最適価値関数
  - Q学習

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/方策 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

最適政策

■最適政策

■状態sにおける期待収益が高い政策

$$V^{\pi}(s) \geq V^{\pi'}(s)$$

■最適状態価値関数

$$V^{*}(s) = \max_{\pi} V^{\pi}(s)$$

$$V^{*}(s) = \max_{a} Q^{\pi *}(s, a)$$

■最適行動価値関数 (ある状態での行動価値)

$$Q^{*}(s, a) = \max_{\pi} Q^{\pi}(s, a)$$

$$Q^{*}(s, a) = E\{r_{t+1} + \gamma V^{*}(s_{t+1}) | s_t, a_t = a\}$$

Q(s,a) 行動価値の学習

- Q(s,a)の行動価値学習
  - 方策オン型 SARSA
    - 方策に従った行動価値
  - 方策オフ型 Q-learning
    - 最適政策に従った行動価値
- Q-learning
  - ある状態である行動を取るときの価値を行動価値最大で計算．方策と無関係に学習

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

αは学習率

Q(s,a) 行動価値の学習

■SARSA

- ■ある状態である行動を取るときの行動価値
- ■方策で行動aを決定
- ■$s_t, a_t, r_t, s_{t+1}, a_{t+1}$の頭文字

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

注意: 教科書にあわせて r_{t+1} で記述している

結局TD学習の形と同じ

$$V(s) \leftarrow V(s) + \alpha\{r_{t+1} + \gamma V(s') - V(s)\}$$