人工知能概論　7章用語集

- 方策(policy)　　ある状態にいたときに，どういう行動をどれほどの確率で選択するか
  $\pi(s,a) = P(a_t = a \mid s_t = s)$
- 価値関数(value function)　　状態や行動の価値の評価
- マルコフ決定過程(MDP，Markov Decision Process)　p.84
  状態 s_t と行動 a_t に依存して次状態 s_{t+1} が決まる確率システム
  状態遷移確率　$P(s_{t+1} \mid s_t , a_t)$　と　報酬関数　$r(s_t , a_t)$
- 割引累積報酬(discounted return)　　$Rt = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$
  将来にわたって得られる報酬の和。遠い未来であればあるほど，割り引いて換算される。ただし γ=1 では T→∞で発散
- 割引率(discount rate) γ $(0 \le \gamma < 1)$
- 状態価値関数 $V_\pi(s)$ 状態 s からスタートし方策 π に従う場合に得られる割引累積報酬の期待値
  $$V_\pi(s) = E_\pi[R_t|s_t = s] = E_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s]$$
- 行動価値関数(action-value function) 状態 s で行動 a を取った後に方策 π に従って得られる割引累積報酬の期待値 $Q_{\pi(s,a)}$
  $$Q_\pi(s,a) = E_\pi[R_t|s_t = s, a_t = a] = E_\pi[\sum^{\infty} \gamma^k r_{t+k+1}|s_t = s, a_t = a]$$
  状態価値関数 V_π との関係　$$V_\pi(s) = \sum_a \pi(s,a) Q_\pi(s,a)$$
- Q 値 (Q-value)　　行動価値関数 Q_π の値
- 最適行動価値関数（π*を、行動価値を最大にする最適方策とする）
  $$Q^*(s,a) \equiv Q_{\pi^*}(s,a) = \max_\pi Q_\pi(s,a)$$
- 状態価値関数 V_π に対するベルマン方程式
  $$V_\pi(s) = \sum_a \pi(s,a) \sum_{s'} P(s_{t+1} = s'|s_t = s, a_t = a)[r_{t+1} + \gamma V_\pi(s')]$$
  マルコフ決定過程において価値関数はこの性質を満たす（現状態の状態価値は**次の報酬**と次**状態の価値**だけで決まる）
  行動価値関数のベルマン方程式
  $$Q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} V_\pi(s') P(s'|s,a)$$
  $$V_\pi(s') = \sum_{a'} \pi(s',a') Q_\pi(s',a')$$
  時刻 t の価値関数の値が、その次に得られる報酬 r_{t+1} と次状態の価値関数の値で決まることを表す（注釈: s_t=s, r_{t+1}=$r(s_t,a_t)$）
- Q 学習（Q-learning）　　最適行動価値関数 Q*(s,a)の Q 値を推定することで強化学習を実現する学習　　$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\, \delta_t$　(α:学習係数，学習率　$0 < \alpha \leqq 1$)
- 行動学習の TD 誤差(Tmporal Difference error) δ_t　　最適行動価値関数 Q 値と学習中の Q 値との差　　$\delta_t = r_{t+1} + \gamma \max_{a_{t+1} \in A} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$
- 行動選択の方策

- ランダム法　　全ての行動を等確率で選択する.
- グリーディ法　　各状態においてその時に最適と思われる行動を選択する.
- ε –グリーディ法　　確率 ε でランダムに行動を選択肢，確率(1– ε)でグリーディ法を行う.

$$\pi(s_t, a_t) = P(a_t|s_t) = (1-\varepsilon)\delta(a_t, \underset{a}{\operatorname{argmax}}\, Q(s_t, a)) + \frac{\varepsilon}{\#(A)}$$

- ボルツマン選択

パラメータ T により $\exp(Q(s,a)/T)$に比例した確率で行動選択を行う．Tが大きくなればランダム法へ，Tが小さくなればグリーディ法に近づく.