公益財団法人
ひろしま産業振興機構

産×学連携

3つのひかり 未来をつくる
広島市立大学
Hiroshima City University

Smart Factory推進Mgr養成 e-Learningコース

強化学習概論

広島市立大学情報科学研究科システム工学専攻
神尾　武司

目次

機械学習入門

機械学習入門

そもそも“機械学習”とは何か？

人間がするような色々な学習や知的作業をコンピュータに実行させるための方法

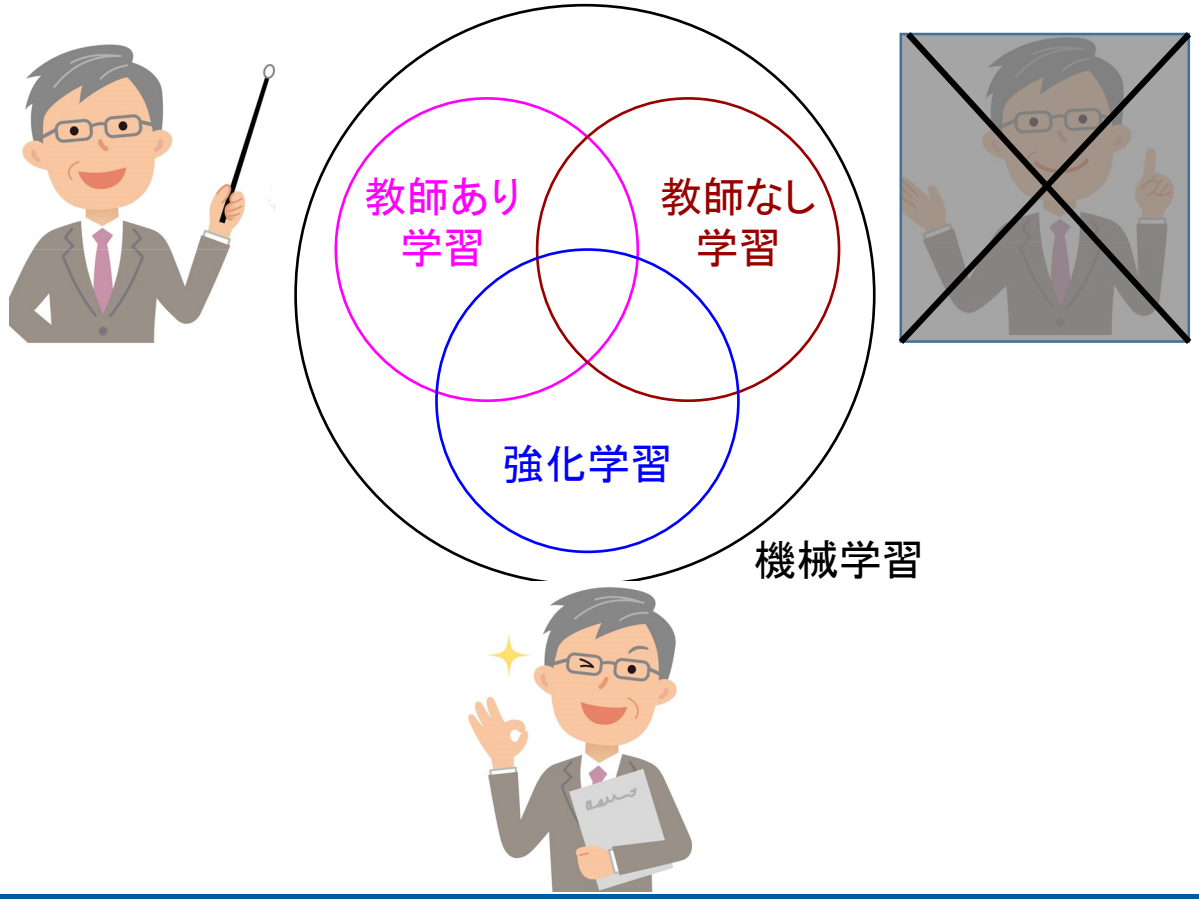
知識を人間が直接アルゴリズムに具体的に書き込んだり教え込んだりせず，データという具体例からコンピュータに自動的に学ばせる方法

機械学習って何でも出来る
夢のプログラムなの？

機械学習入門

“機械学習”の分類

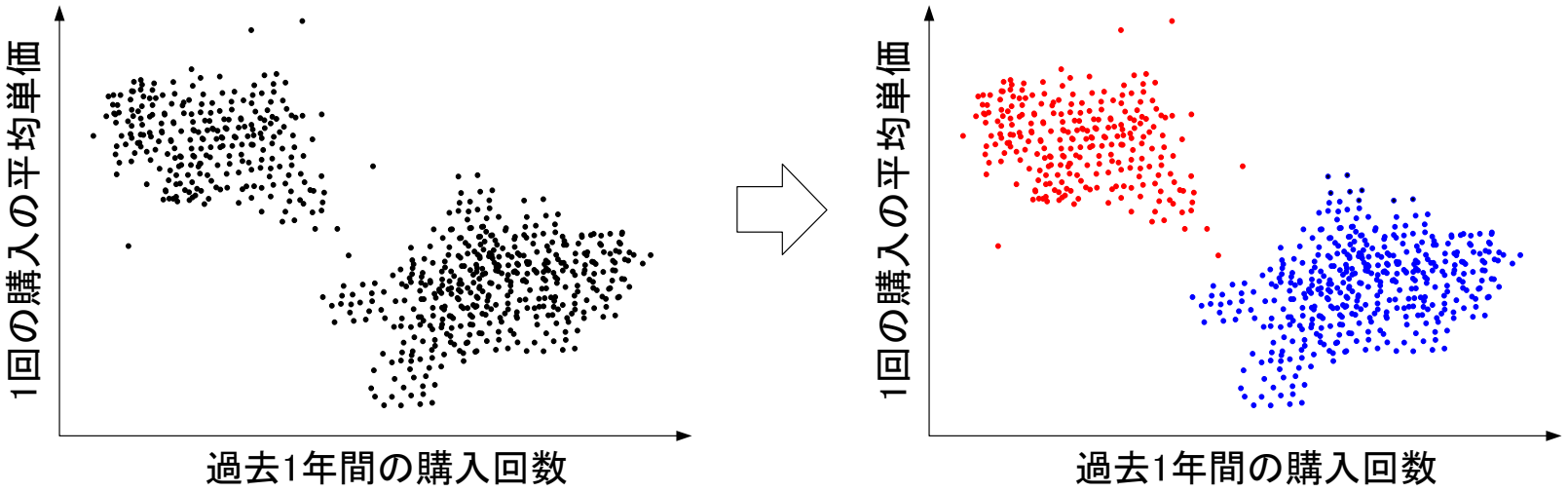
機械学習の分類は“教師”の与え方により３種類



機械学習入門

教師なし学習

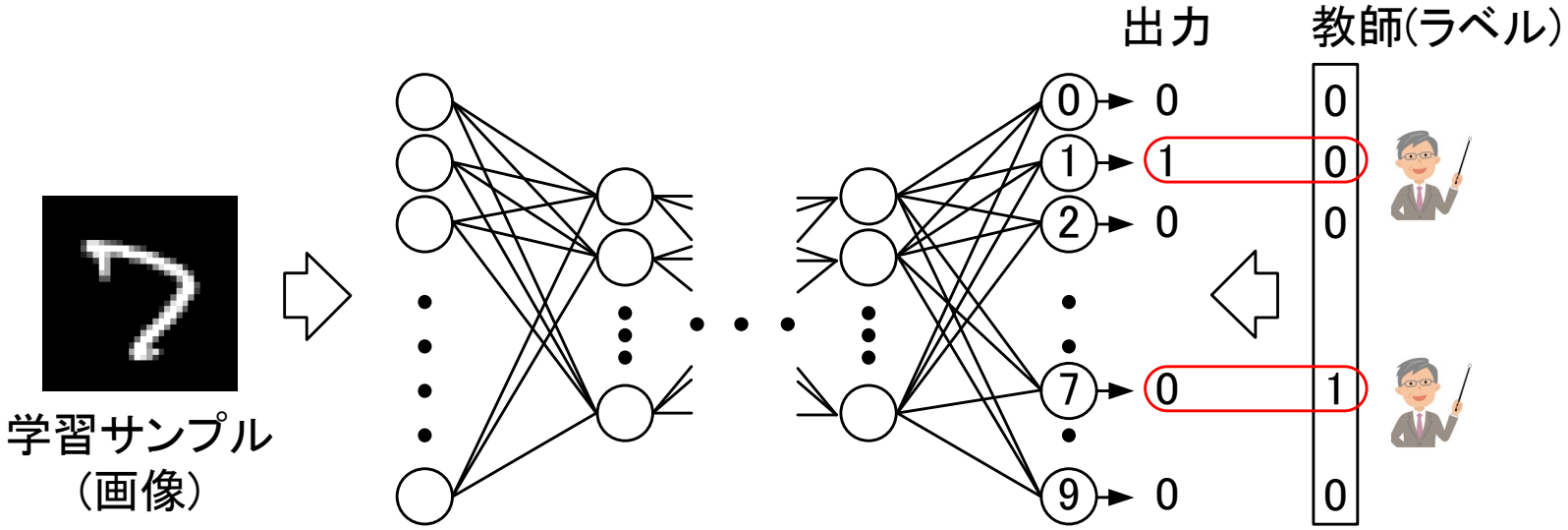
- 入力：学習サンプル，教師：なし
- 学習目標：サンプルの類似性に基づくグループ分け
- 応用：購買データに基づく顧客の分類



機械学習入門

教師あり学習

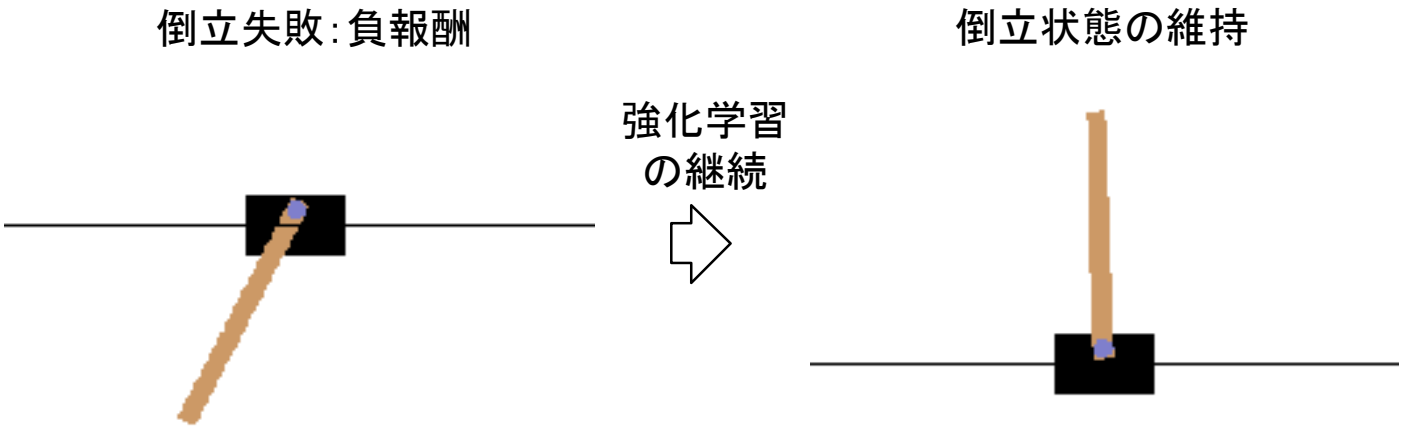
- 入力：学習サンプル，教師：ラベルや数値
- 学習目標：サンプルに対する出力と教師データの一致
- 応用：パターン認識，関数近似



機械学習入門

強化学習

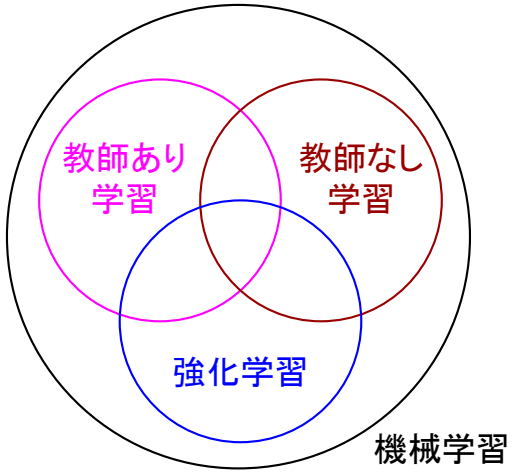
- 入力：知覚(センサ)情報，教師：報酬
- 学習目標：報酬最大化のための行動選択ルールの獲得
- 応用：ロボットの姿勢制御，対戦ゲームの戦略



機械学習入門

機械学習とは何でもできる夢のプログラムか？

- 漠然とした要望に応えるものではない.
- 要望によっては，学習サンプルと教師の準備にコストがかかる.
- 要望に見合う学習システムの選択，構築は人手による.
- データからの有用な知識抽出において人間の能力を超える.



機械学習入門

機械学習の向き／不向き

(機械学習全般)
- 機械学習の種類で対象とする問題が異なるため，それを踏まえて向き／不向きを論じるべき.

(教師あり，教師なし学習)
- 得意：データを分析し，将来を予測すること.
- 不得意：発生回数が少ない事例に対する判断

(強化学習)
- 得意：時系列変化を伴うシステムの制御ルールの構築
- 不得意：制御対象の状態を上手く表現できない問題

強化学習概論

強化学習とは？

根底にある考え方

学習が持つ性質を考えてみると，我々が環境との相互作用を通じて学習しているということが多分最初に思い浮かぶ．幼児が遊んだり，腕を振ったり，あるいは周囲を見まわしたりするときに，そこには教師に相当するものはいないが，その幼児は感覚系と運動系の連携を用いて環境に直接的に働きかける．この連携を用いると原因と結果，動作の結果，そして目標を達成するために何をすべきかについて多くの情報を作り出すことができる．我々が生きていく上においても，このような相互作用が環境と我々自身に関する主要な知識源であることは疑いのないところである．

(中略)

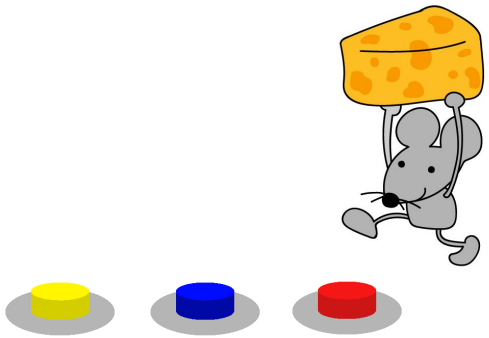
相互作用から学習することは，ほとんどすべての学習理論の根底にある基本的な考え方である．

※強化学習(R. S. Sutton, A. G. Barto著， 三上・皆川 訳， 森北出版， 2000)の序章より抜粋

強化学習とは？

動物実験(スキナーボックス)に見られる強化学習

ラットが箱の中で動き回り，たまたま赤いレバーが見えたので，それを押したらチーズが降ってきた場合を考えよう．ラットにとっては，未知の環境で探索して赤いレバーが見えたという未知の状況に遭遇し，赤いレバーを押すという行為をたまたまとったに過ぎない．しかし，チーズという報酬が与えられたということで，そのような状況での「赤いレバーを押す行動」が強化されたと表現される．青いレバーも押せば報酬がもらえるかもしれないし，黄色いレバーは天敵が出現するかもしれない．さまざまな試行錯誤を繰り返し，報酬や罰の授受により，ラット自身にとって意味のある行動を獲得するようになる．



※ロボットインテリジェンス(浅田稔・國吉康夫 著，岩波書店，2006)の4章より抜粋

強化学習とは？

書籍に見られる強化学習の概要説明

1. 強化学習(R. S. Sutton, A. G. Barto著，三上・皆川 訳，森北出版，2000)

   強化学習では，数値化された報酬信号を最大化するために，なにをすべきか(どのようにして状況に基づく動作選択を行うか)を学習する．

2. 最適化アルゴリズム(長尾智晴 著，昭晃堂，2000)

   強化学習は，自律エージェントの行動を最適化するための一手法である．エージェントは，自分の行った一連の行動に対して環境から与えられる報酬を基にして行動を最適化してゆく．

強化学習とは？

書籍に見られる強化学習の概要説明

3. マルチエージェント学習(高玉圭樹 著，コロナ社，2003)

強化学習とは，さまざまな報酬(即座に得られる報酬や時間的に遅れて得られる報酬など)を手がかりとしながら，単位時間当たりの受け取る報酬値(あるいは最終的に受け取る報酬の総量)を最大化する方策を試行錯誤を通じて獲得するものである.

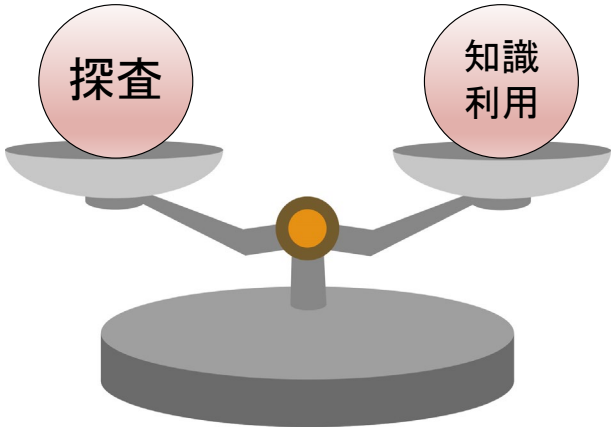
4. ロボットインテリジェンス(浅田稔・國吉康夫 著，岩波書店，2006)

強化学習は，最終結果もしくは，途中経過に対して，どの程度よかったかを示す「報酬信号」に基づき，これらの報酬をなるべく大きくするように探索する.

強化学習とは？

探査と知識利用のトレードオフ

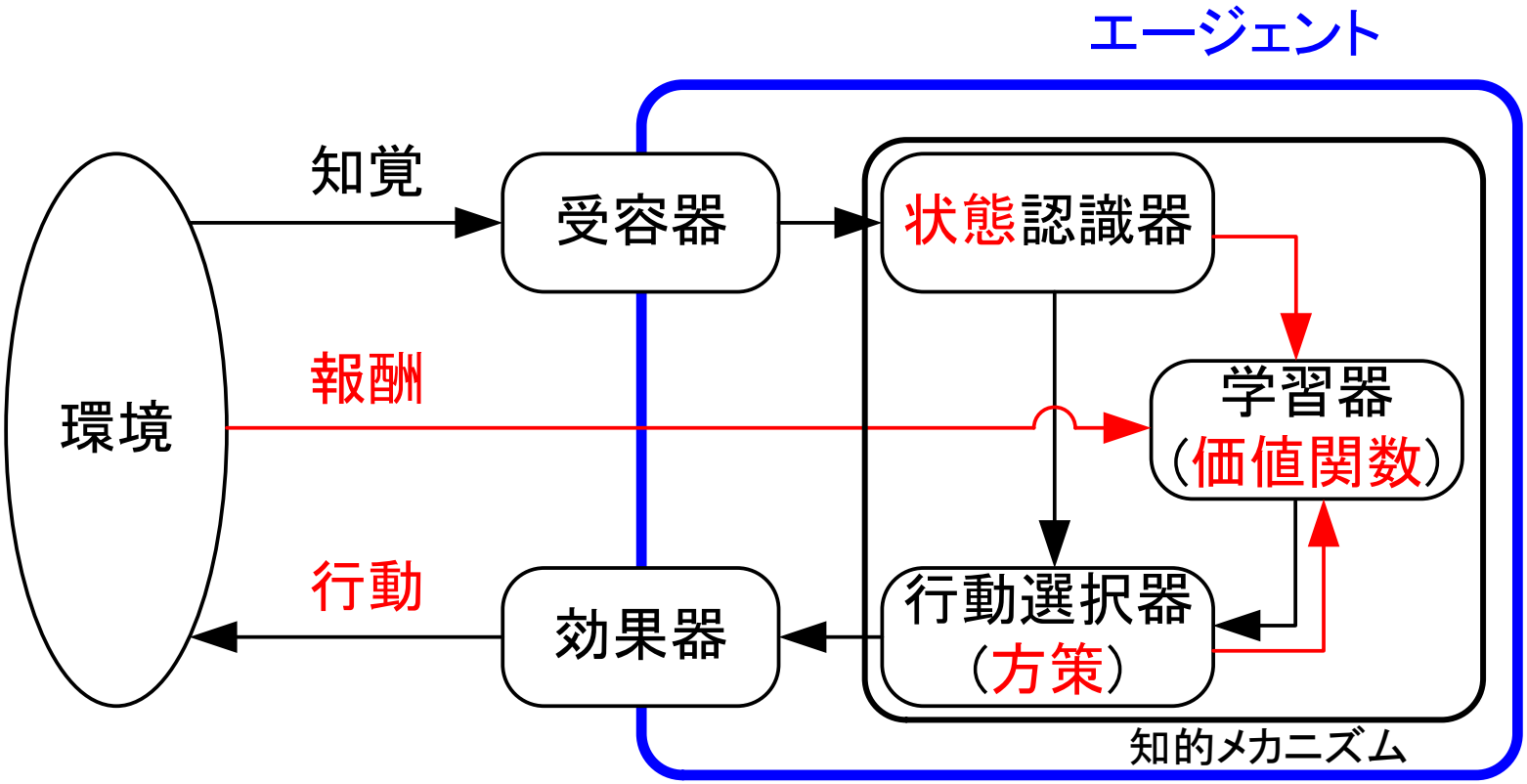
探査と知識利用のトレードオフは強化学習にあって他の学習にはない挑戦的なテーマの1つである．多くの報酬を得るために，強化学習エージェントは過去に試みた行動の中で報酬を得るために効果的なものを優先的に選ばなくてはならない．ところが，このような動作を発見するためには，過去に試みたことのない行動も選択してみなくてはならない．つまり，エージェントは報酬を得るためにすでに持っている知識を利用し，将来的に行動選択を改善するためには探査も行わなくてはならない．ここで生じるジレンマは，探査も知識利用も与えられた作業の失敗なしに独自に遂行されることはないということである．
エージェントは色々な行動を試し，その中で最良と考えられるものを徐々に見出していく必要がある．



※強化学習(R. S. Sutton, A. G. Barto著， 三上・皆川 訳， 森北出版， 2000)の序章より抜粋

強化学習の基本プロセス

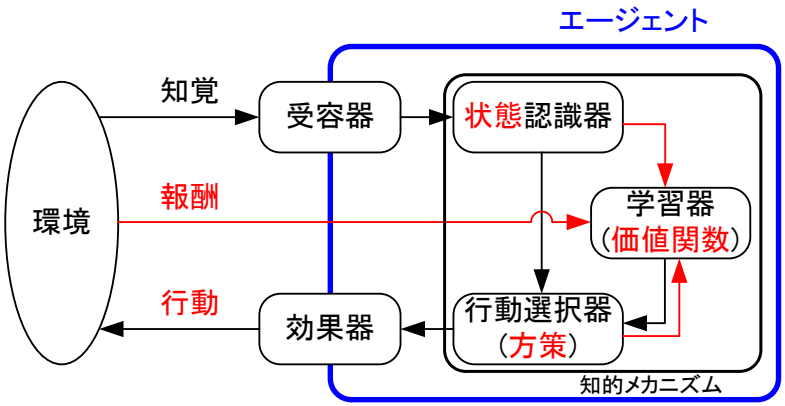
強化学習システムの概略図



強化学習の基本プロセス

用語説明①

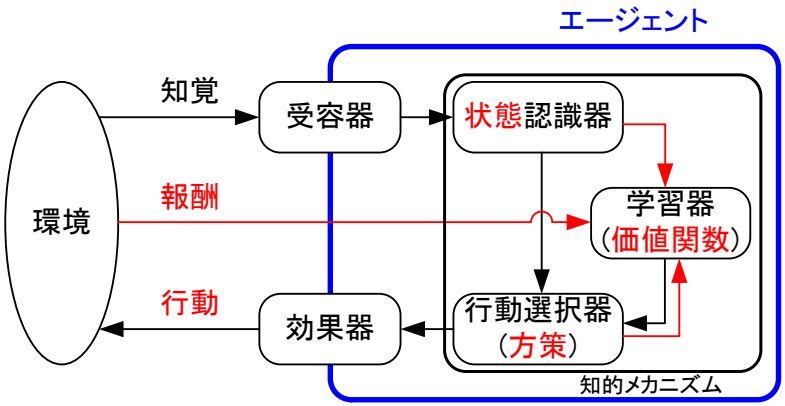
- エージェント：学習と意思決定(行動選択)を行うもの　※制御器
- 環境：エージェントを取り巻く環境　※制御系
- 状態s_t：知覚情報に対する区分．番号で与えられる．
- 価値関数：報酬の獲得に対する価値を表すものであり，状態または状態行動対に対して定義される．前者を状態価値$V^{\pi}(s)$，後者を行動価値$Q^{\pi}(s, a)$という．どちらを使用するかは学習アルゴリズムに依存する．



強化学習の基本プロセス

用語説明②

- 方策π：行動選択の方法
- 行動a_t：エージェントが環境に対して取り得る動作　※制御信号
- (即時)報酬r_t：目的の成否に基づいて数値で与えられる.
- エピソード：エージェントが初期状態から終端状態(目的の成否)に達するまでの期間．エージェントは終端状態に達した後，初期状態に戻る.



強化学習の基本プロセス

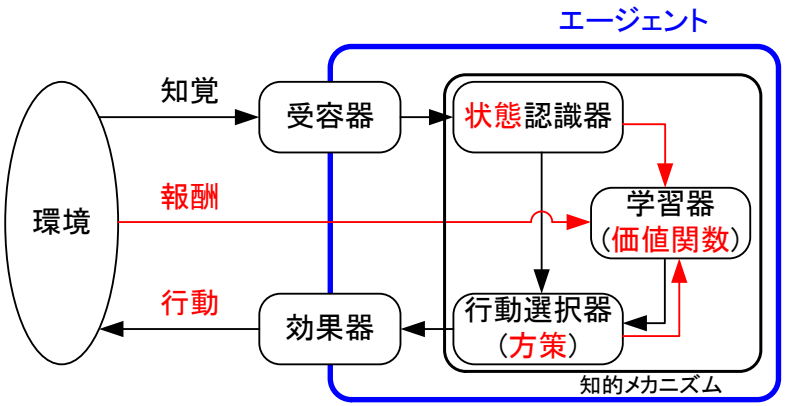
用語説明③

- モデル：正確には，環境のモデル．エージェントが自分の行動に対してどのように応答するかを予測出来る，あらゆる対象を意味するものとする．1個の状態と1個の行動が与えられたとき，結果として生じる次の状態と報酬の予測がモデルによって作り出される．
- (単純)マルコフ性：時刻tの状態と行動のみから時刻t+1の状態が決定される性質のこと．また，その確率過程を(単純)マルコフ過程という．なお，n時刻前まで考慮する場合はn重マルコフ過程という．
- マルコフ決定過程(MDP)：マルコフ性を満たす強化学習タスクのこと．強化学習アルゴリズムの多くはMDPであることを前提としている．
- 非マルコフ決定過程：マルコフ性を満たさない強化学習タスクのこと．

強化学習の基本プロセス

強化学習の基本ルーチン

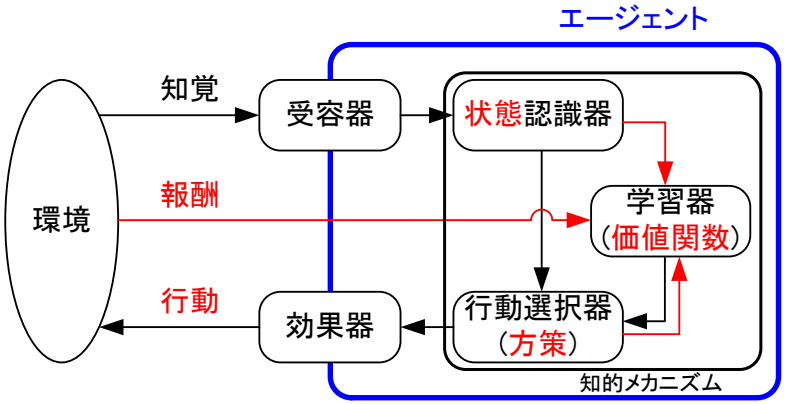
1. エージェントは受容器(センサ)を通して環境から知覚情報を入手し，状態認識器によって状態s_tを知る．
2. 行動選択器は価値関数に基づく方策に従って状態s_tに対する行動a_tを選択する．
3. 効果器(アクチュエータ)により行動a_tが実行され，環境に影響を与える．
4. エージェントは新しい知覚情報を入手し，状態s_{t+1}を知る．同時に報酬r_{t+1}を得る．



強化学習の基本プロセス

強化学習の基本ルーチン

5. s_t, a_t, s_{t+1}, r_{t+1}に基づいて価値関数と方策を更新(つまり，学習)する．
6. エージェントが終端状態(ゴール，失敗)に達したとき，エピソードが終了する．
   - エピソード終了　⇒　エージェントを初期位置に戻した上で，Step7へ．
   - エピソード途中　⇒　Step1に戻る．
7. 学習終了条件(エピソード回数など)を満足していない場合，Step1に戻る．満足した場合は学習を終了する．



強化学習アルゴリズムの基礎理論

1. 仮定

- 環境の変化はマルコフ過程に従う.
- エージェントは環境に関する知識を一切持たない設定から学習する.

2. 累積報酬：R(t)

- 時刻 t から見て，将来的に得られるであろう報酬の総和.
- R(t)を最大化するための方策πを獲得することが強化学習の目的.

$$R(t) = r(t+1) + \gamma r(t+2) + \gamma^2 r(t+2) + \cdots = \sum_{k=0}^{\infty} \gamma^k r(t+k+1)$$

- 割引率 γ ：将来の即時報酬が現在においてどれだけの価値があるかを決定する．未来の報酬であるほど不確定となるため，$0<\gamma<1$とする.

強化学習アルゴリズムの基礎理論

3. 状態価値関数：Vπ(s)

– 定義：時刻tで状態sを観測した後に得られるR(t)の期待値

$$V^{\pi}(s) = \mathrm{E}_{P,\pi}\left[R(t) \mid s(t) = s\right]$$

4. 状態価値関数に対するベルマン方程式

– 導出方法： Vπ(s)を方策πのもとでMDPが与える期待値と見なす.

$$V^{\pi}(s) = R^{\pi}(s) + \gamma \sum_{s'} P^{\pi}(s' \mid s) V^{\pi}(s')$$

5. 行動価値関数Qπ(s, a)とベルマン方程式

$$Q^{\pi}(s,a) = \mathrm{E}_{P}\left[R(t) \mid s(t) = s, a(t) = a\right]$$

$$Q^{\pi}(s,a) = \sum_{s'} P(s' \mid s,a) R(s,a,s') + \gamma \sum_{s'} \sum_{a'} \pi(a' \mid s') P(s' \mid s,a) Q^{\pi}(s',a')$$

強化学習アルゴリズムの基礎理論

6. 最適方策：π^*

- 方策πは方策π'と同等か，より優れた方策である．

$$V^{\pi}(s) \geq V^{\pi'}(s), \quad \forall s \in S$$

- 最適方策π^*に対する状態価値関数

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

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

- 最適方策 $\pi^* = \arg\max_{\pi} V^{\pi}(s)$ に対する行動価値関数

$$Q^*(s,a) = Q^{\pi^*}(s,a) \equiv \max_{\pi} Q^{\pi}(s,a), \quad \forall s, \ \forall a \in A(s)$$

- 状態sで最適方策π^*が選択する行動

$$a_i \in \arg\max_{a} Q^*(s,a)$$

強化学習アルゴリズムの基礎理論

7. 状態価値関数と行動価値関数の関係

– 方策πに対しての関係

$$V^{\pi}(s)=\sum_{a}\mathrm{E}_{P}\left[R(t),a(t)=a\,|\,s(t)=s\right]=\sum_{a}\pi(a\,|\,s)Q^{\pi}(s,a)$$

– 最適方策π^{*}に対しての関係

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

8. ベルマン最適方程式

$$V^{*}(s)=\max_{a\in A(s)}\sum_{s'}P(s'\,|\,s,a)\left(R(s,a,s')+\gamma V^{*}(s')\right)$$

$$Q^{*}(s,a)=\sum_{s'}P(s'\,|\,s,a)\left(R(s,a,s')+\gamma\max_{a'\in A(s)}Q^{*}(s',a')\right)$$

9. 強化学習アルゴリズムの基礎理論

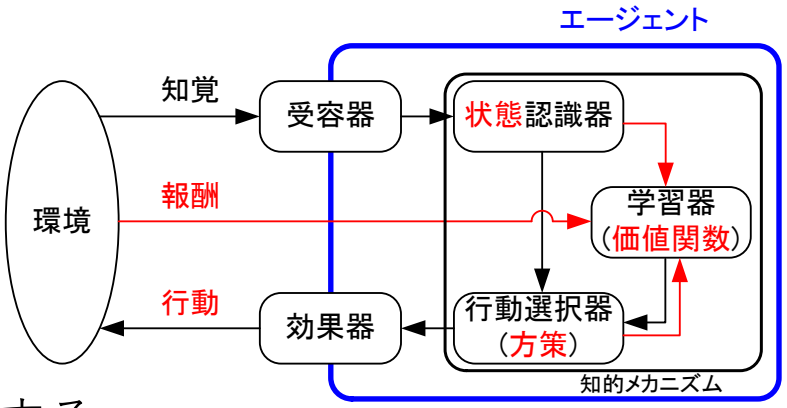
– 強化学習アルゴリズムはベルマン方程式およびベルマン最適方程式に基づいて構築されている.

Q学習

前提条件

- マルコフ決定過程
- 離散状態空間，離散行動空間
- 終端状態：目標状態(ゴール)，失敗
- エピソード：スタート～終端状態
- 報酬：全状態に設定可能
- 学習目標：指定方策を最適化する行動価値関数Q(s, a)の獲得
- Q(s, a)の初期値：零
- 方策：ε-greedy → 確率εでランダム，
  それ以外は最大のQ値を持つ行動を選択.

Q学習



処理フロー

1. 時刻t=0とし，エージェントを
   スタート位置に配置する.
2. 知覚情報から状態s_tを認識する.
3. 方策(ε-greedy)に従い，行動a_tを選択する.
4. 行動a_tを実行後，時刻を更新する(t → t+1).
5. 知覚情報から状態s_{t+1}を認識し，環境から報酬r_{t+1}を受け取る.
6. $Q(s_t, a_t)$の更新

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma \max_{a \in A(s_{t+1})} Q(s_{t+1},a) - Q(s_t,a_t)\right]$$

※ α：学習率，γ：割引率

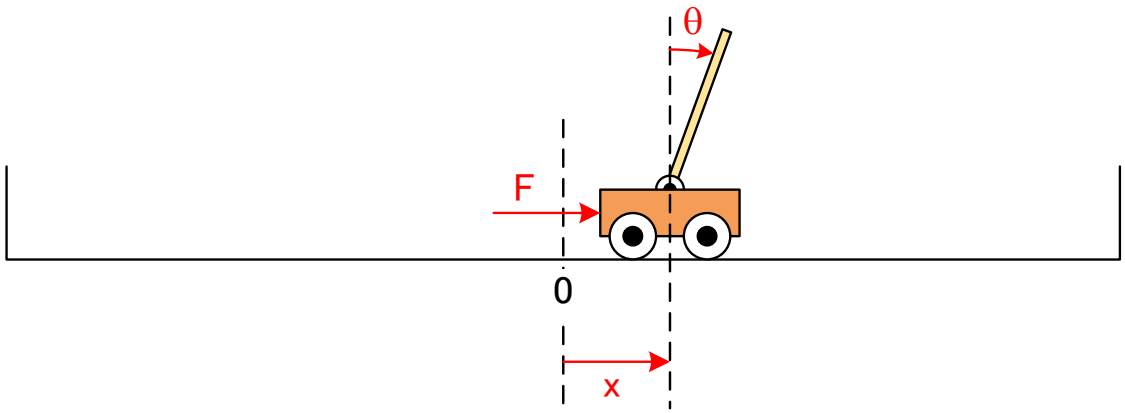
7. 状態s_{t+1}が終端状態であれば，現エピソードを終了し，Step1へ戻る.
   それ以外はt ← t+1 としてStep3へ戻る.

Q学習

実装例：倒立振り子の姿勢制御

(目的)
- 左右に動かせる台車の上に取り付けられたポールを倒立させ続ける.

(条件)
- センサ出力
  台車位置：$x \in [-2.4, 2.4]$(m), ポール角度：$\theta \in [-0.5\pi, 0.5\pi]$(rad),
  台車速度：$v \in [-10, 10]$(m/s), ポール角速度：$\omega \in [-2\pi, 2\pi]$(rad/s)
- 台車に与える力と時間間隔：$F \in \{-10, 10\}$(N), $\tau = 0.02$s
- 初期状態： $|x_0| < 0.1$(m), $|\theta_0| < 0.1$(rad), $v_0 = 0$(m/s), $\omega_0 = 0$(rad/s)



Q学習

実装例：倒立振り子の姿勢制御

(成否の解釈)
- 100sの間，センサ出力が倒立状態を表す範囲内であれば成功．それ以外は失敗．

(強化学習における終端状態)
- 成功：表現できない．
- 失敗：センサ出力が倒立状態を表す範囲を超えた場合

- 台車位置：$x \in [-2.4, 2.4]$ ➡ $x_b \in [-1.6, 1.6]$(m)
- ポール角度：$\theta \in [-0.5\pi, 0.5\pi]$ ➡ $\theta_b \in [-0.4, 0.4]$(rad)
- 台車速度：$v \in [-10, 10]$ ➡ $v_b \in [-2, 2]$(m/s)
- ポール角速度：$\omega \in [-2\pi, 2\pi]$ ➡ $\omega_b \in [-1.4, 1.4]$(rad/s)

センサの出力レンジ　　倒立状態を表す範囲

Q学習

実装例：倒立振り子の姿勢制御

(状態空間の離散化) →表形式表現という

– センサの出力レンジと倒立状態の範囲
  - 台車位置：$x \in [-2.4, 2.4]$ ➡ $x_b \in [-1.6, 1.6]$(m)
  - ポール角度：$\theta \in [-0.5\pi, 0.5\pi]$ ➡ $\theta_b \in [-0.4, 0.4]$(rad)
  - 台車速度：$v \in [-10, 10]$ ➡ $v_b \in [-2, 2]$(m/s)
  - ポール角速度：$\omega \in [-2\pi, 2\pi]$ ➡ $\omega_b \in [-1.4, 1.4]$(rad/s)

– 離散化：倒立状態の範囲を4等分に分割する
  - xの閾値：[–2.4, –1.6, –0.8, 0, 0.8, 1.6, 2.4](m)
  - θの閾値：[–0.5π, –0.4, –0.2, 0, 0.2, 0.4, 0.5π](rad)
  - vの閾値：[–10, –2, –1, 0, 1, 2, 10](m/s)
  - ωの閾値：[–2π, –1.4, –0.7, 0, 0.7, 1.4, 2π](rad/s)

各センサ出力は6分割される.

※赤枠の範囲が失敗を表す.

Q学習

実装例：倒立振り子の姿勢制御

(状態数と状態番号)

- 状態数：N_S

$$N_S = (x\text{の分割数}) \times (\theta\text{の分割数}) \times (v\text{の分割数}) \times (\omega\text{の分割数})$$
$$= 6 \times 6 \times 6 \times 6 = 1296$$

- 状態番号：s = 0～1295

$$s = (x\text{の区分}) \times 6^0 + (\theta\text{の区分}) \times 6^1 + (v\text{の区分}) \times 6^2 + (\omega\text{の区分}) \times 6^3$$

- 状態番号の算出例
「センサ出力がx=0, θ=0, v=0, ω=0であった場合の状態番号は？」

$$s = 3 \times 6^0 + 3 \times 6^1 + 3 \times 6^2 + 3 \times 6^3 = 777$$

※隣り合う閾値a, bに対して，a≤x<bの場合，xは[a,b]の区分にあると考えましょう．

Q学習

実装例：倒立振り子の姿勢制御

(行動数と行動番号)

- 行動数：N_A
  台車に与える力は$F \in \{-10,10\}(N)$の2種類なので，「$N_A=2$」
- 行動番号：$a= 0 \sim 1$

(Qテーブル)

$$Q(s, a)= \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} & \\ & \\ & \\ & \\ & \\ & \end{pmatrix} \end{array} \begin{array}{l} 0 \\ \\ \\ \\ \\ \\ 1295 \end{array}$$

→ 要素：Q値，行列：Qテーブル

Q学習

実装例：倒立振り子の姿勢制御

(報酬設定)

– 倒立成功時：r_A
  成功に対応する状態を表現できないので，「r_Aは未定義」

– 倒立失敗時：r_F
  失敗に対応する状態を指定できるので，「$r_F = -1$」

– それ以外：r_E
  特に，報酬を与える理由もないので，「$r_E = 0$」

(方策設定)

– Q学習での一般的な方策：ε-greedy

$$\varepsilon = \frac{0.5}{1 + \text{episode数}}$$

Q学習

実装例：倒立振り子の姿勢制御

(学習終了条件)

- 総エピソード数の超過
  最大エピソード回数：N_{eps}=20000
- N_{suc}連続エピソードでの倒立成功
  学習成功判定： N_{suc}=100

(その他のパラメータ)

- 学習率：α=0.1
- 割引率：γ=0.99

知名度の高い強化学習アルゴリズム

1. 動的計画法(DP)
   - 前提条件：完全な環境のモデル($R(s, a, s')$, $P(s' | s, a)$)が必要.
   - 強化学習としては実用的ではないが，理論的には重要.
   - 価値関数： $V^{\pi}(s)$
   - 後続の価値に基づいて価値関数を更新 → ブートストラップ

2. モンテカルロ法(MC)
   - 環境のモデルを必要としない.
   - 本質的にDPと同じ方法で最適化を図る.
   - 価値関数：$Q^{\pi}(s, a)$
   - エピソード終了時に価値関数を更新 → 非ブートストラップ

知名度の高い強化学習アルゴリズム

3. Sarsa
   - TD学習 (環境モデル不要，ブートストラップ) ← DPとMCの利点
   - 価値関数：$Q^{\pi}(s, a)$
   - 価値更新で，$s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1}$に関する情報を使用←名前の由来

4. Q学習(QL)
   - TD学習 (環境モデル不要，ブートストラップ)
   - 価値関数：$Q^{\pi}(s, a)$
   - 価値更新で，$\max_a Q(s_{t+1}, a)$を使用．← SarsaではQ(s_{t+1}, a_{t+1})
   - Sarsaよりも高速に最適方策に収束する．

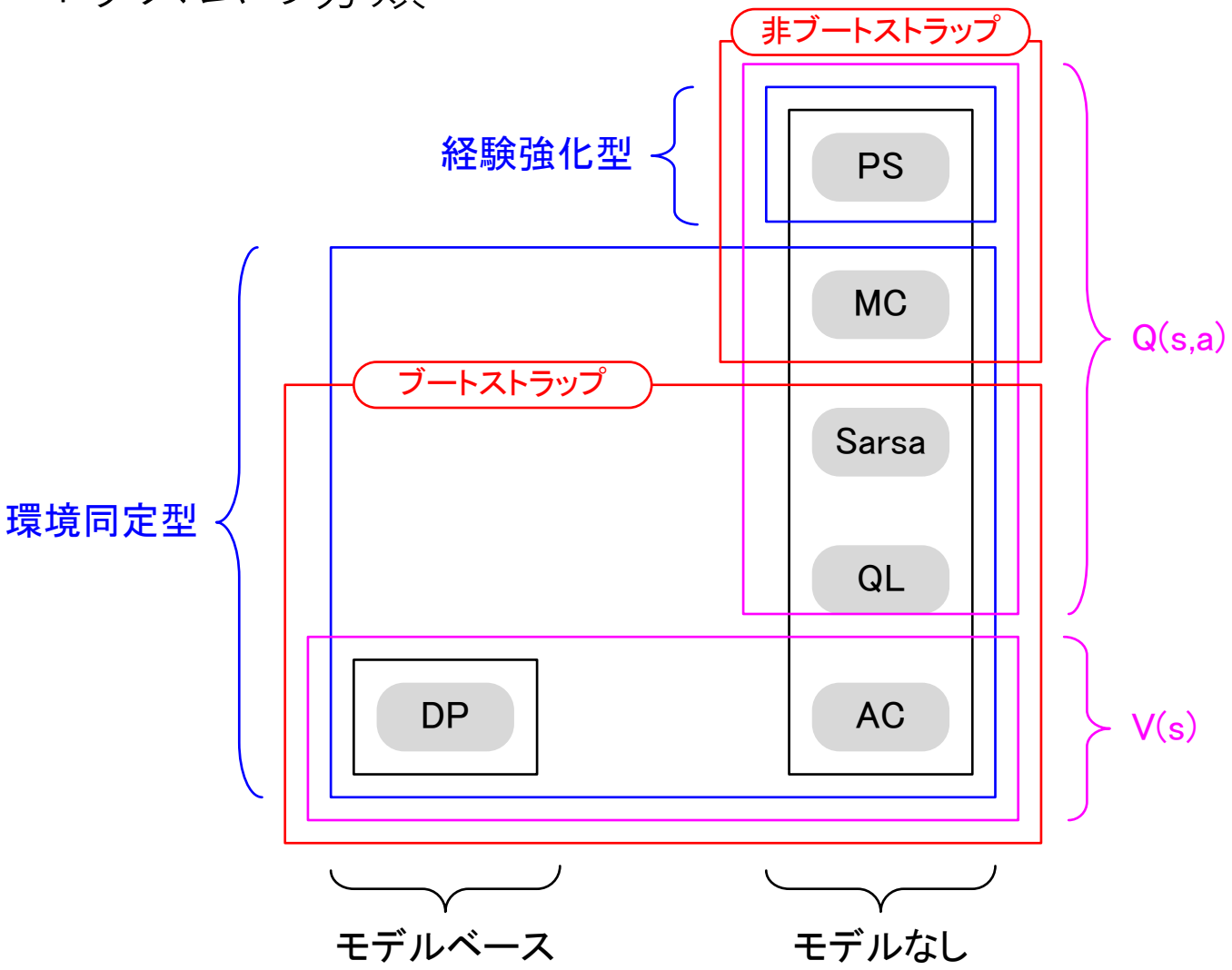
知名度の高い強化学習アルゴリズム

5. Actor-Critic (AC)
   - TD学習 (環境モデル不要，ブートストラップ)
   - 価値関数：$V^{\pi}(s)$
   - Actor (行動器)：方策πに基づく行動選択，TD誤差による方策更新
   - Critic (評価器)：選択行動をTD誤差で評価，TD誤差による$V^{\pi}(s)$更新
   - 連続な状態行動空間への適用がよく試みられた.

6. Profit Sharing (PS)
   - 環境モデル不要，非ブートストラップ，価値関数：$Q^{\pi}(s, a)$
   - 学習効率の重視(経験強化型)　⇔　環境同定型(最適性を重視)
   - 非MDP環境での利用も可能　⇒マルチエージェントへの適用

知名度の高い強化学習アルゴリズム

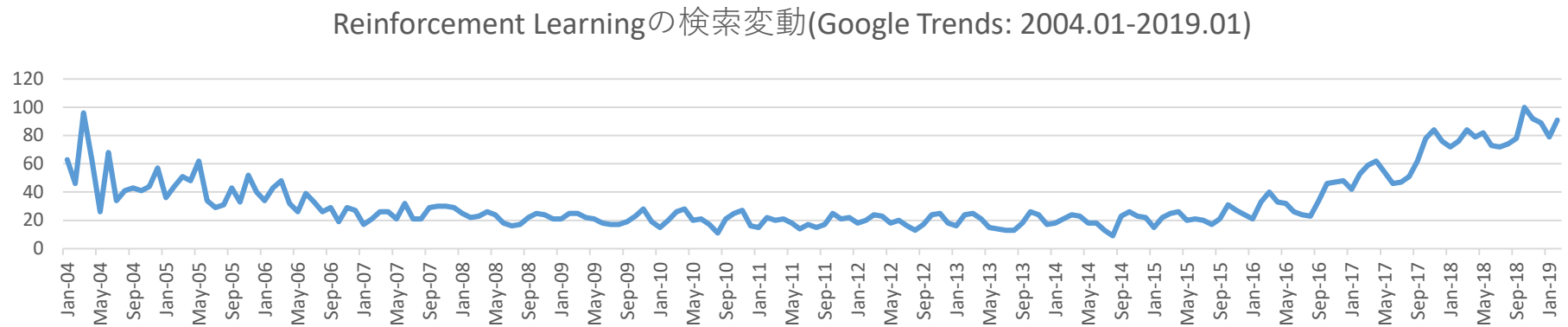
アルゴリズムの分類



強化学習と最近のAIブーム

強化学習の歴史

- 1990年代後半：強化学習と脳の学習メカニズムの類似性が示唆
- 1990年代後半～2000年代前半：強化学習ブーム
- 2000年代後半：状態の縮約表現の問題から強化学習ブームは下火に
- 2012年：ディープラーニング(深層学習)に注目が集まる.
- 2015年～：深層強化学習によるゲーム攻略により強化学習ブームが再燃



強化学習と最近のAIブーム

強化学習(深層強化学習)で何ができる？

- データセンターの冷却効率の改善(DeepMind, 2016.07.20)
  Googleのサーバを集めたデータセンタの冷却効率を強化学習を用いて改善し，消費電力の削減に成功したと発表.
  (https://blog.google/outreach-initiatives/environment/deepmind-ai-reduces-energy-used-for/)
  (https://tech.nikkeibp.co.jp/it/atcl/news/16/072102162/)

- 自動運転に関するデモ展示(Preferred Networks, 2016.01.09)
  車両の幅に対して道路が狭く，車が密集した難易度の高い問題を扱う.
  また，学習時には存在しない，人が操作する車からの回避という困難な問題も扱う.
  (https://research.preferred.jp/2016/01/ces2016/)

- 深層強化学習で超高層ビルの地震に備える(NTTファシリティーズ, 2017.08)
  (https://inforium.nttdata.com/foresight/ai-vibration-control.html)

小テスト

問題1：倒立振り子の状態空間の離散化

倒立状態を表す各変数の範囲について(※一部変更した)，x_bを8等分，θ_bを6等分，v_bを6等分，ω_bを4等分して状態空間の離散化を行う場合，各センサ出力(x, θ, v, ω)の閾値を示せ．

(センサの出力レンジと倒立状態の範囲)

- 台車位置：$x \in [-2.4, 2.4]$ ➡ $x_b \in [-1.6, 1.6]$(m)
- ポール角度：$\theta \in [-0.5\pi, 0.5\pi]$ ➡ $\theta_b \in [-0.6, 0.6]$(rad)
- 台車速度：$v \in [-10, 10]$ ➡ $v_b \in [-3, 3]$(m/s)
- ポール角速度：$\omega \in [-2\pi, 2\pi]$ ➡ $\omega_b \in [-1.4, 1.4]$(rad/s)

センサの出力レンジ　　倒立状態を表す範囲

解答1：倒立振り子の状態空間の離散化

(センサの出力レンジと倒立状態の範囲)

- 台車位置：$x \in [-2.4, 2.4]$ ➡ $x_b \in [-1.6, 1.6]$(m) ➡8等分
- ポール角度：$\theta \in [-0.5\pi, 0.5\pi]$ ➡ $\theta_b \in [-0.6, 0.6]$(rad) ➡6等分
- 台車速度：$v \in [-10, 10]$ ➡ $v_b \in [-3, 3]$(m/s) ➡6等分
- ポール角速度：$\omega \in [-2\pi, 2\pi]$ ➡ $\omega_b \in [-1.4, 1.4]$(rad/s) ➡4等分

センサの出力レンジ　　倒立状態を表す範囲

(答え)

- xの閾値：[–2.4, –1.6, –1.2, –0.8, –0.4, 0, 0.4, 0.8, 1.2, 1.6, 2.4](m)
- θの閾値：[-0.5π, –0.6, –0.4, –0.2, 0, 0.2, 0.4, 0.6, 0.5π](rad)
- vの閾値： [–10, –3, –2, –1, 0, 1, 2, 3, 10](m/s)
- ωの閾値：[-2π, –1.4, –0.7, 0, 0.7, 1.4, 2π](rad/s)

問題2：倒立振り子の状態番号の算出

問題1で求めたセンサ出力の閾値を使う場合，状態数N_Sと状態番号の範囲を求めよ．さらに，センサ出力が以下の通りに与えられた場合，その状態番号を算出せよ．

(センサ出力の閾値)

- xの閾値：[–2.4, –1.6, –1.2, –0.8, –0.4, 0, 0.4, 0.8, 1.2, 1.6, 2.4](m)
- θの閾値：[–0.5π, –0.6, –0.4, –0.2, 0, 0.2, 0.4, 0.6, 0.5π](rad)
- vの閾値： [–10, –3, –2, –1, 0, 1, 2, 3, 10](m/s)
- ωの閾値：[–2π, –1.4, –0.7, 0, 0.7, 1.4, 2π](rad/s)

(センサ出力)

- x = 0.9 (m)
- θ = –0.3 (rad)
- v = –2.5 (m/s)
- ω = 0.5 (rad/s)

(ヒント)
3月2日10時5分は3月1日0時0分ら見て何番目の分？
ただし，3月1日0時0分を0番目としましょう．

5 × (1) + 10 × (1 × 60) + 2 × (1 × 60 × 24) = 3485番目

解答2：倒立振り子の状態番号の算出

(センサ出力の閾値)

xの閾値：[–2.4, –1.6, –1.2, –0.8, –0.4, 0, 0.4, 0.8, 1.2, 1.6, 2.4](m) ➡10分割

θの閾値：[–0.5π, –0.6, –0.4, –0.2, 0, 0.2, 0.4, 0.6, 0.5π](rad) ➡8分割

vの閾値： [–10, –3, –2, –1, 0, 1, 2, 3, 10](m/s) ➡8分割

ωの閾値：[–2π, –1.4, –0.7, 0, 0.7, 1.4, 2π](rad/s) ➡6分割

(答え)

- 状態数： $N_S = 10 \times 8 \times 8 \times 6 = 3840$
- 状態番号の範囲：0 ～ 3839
- 状態番号：s = 2027　※ xを1桁目と考えた場合

x = 0.9 (m) ➡第7区分

θ = –0.3 (rad) ➡第2区分

v = –2.5 (m/s) ➡第1区分

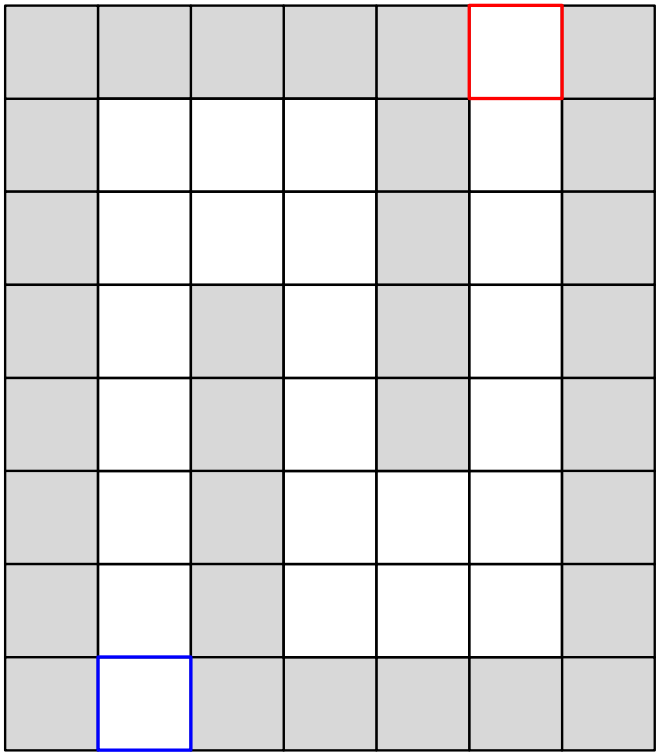
ω = 0.5 (rad/s) ➡第3区分

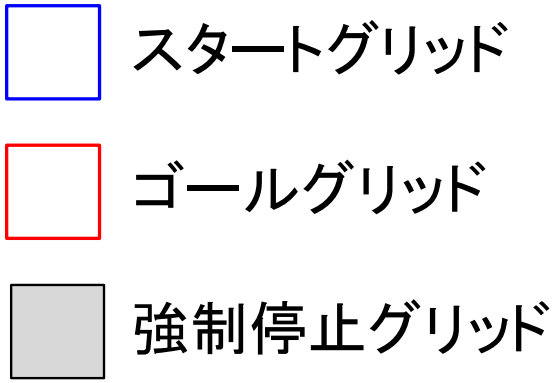
$$s = 7 \times (1) + 2 \times (1 \times 10) + 1 \times (1 \times 10 \times 8) + 3 \times (1 \times 10 \times 8 \times 8) = 2027$$ →xを1桁目と考えた場合

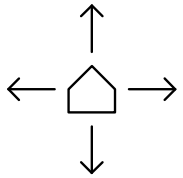
$$\left(s = 3 \times (1) + 1 \times (1 \times 6) + 2 \times (1 \times 6 \times 8) + 7 \times (1 \times 6 \times 8 \times 8) = 2793\right)$$ →ωを1桁目と考えた場合

問題3：迷路問題の状態と行動設定

下図のようなS字迷路があります．ロボットをスタート(S)に上向きに配置し，強制停止グリッドに侵入せずに最短経路でゴール(G)を目指すようにQ学習を実装したいです．**状態番号と行動番号を具体的に設定してみましょう．**ただし，ロボットは1回の行動で上下左右いずれかのグリッドに移動可能であり，センサによりどのグリッドにいるか認識できると仮定します．





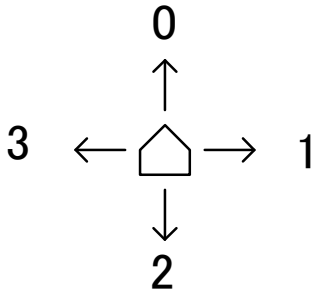


移動ロボット
（上下左右）

解答3：迷路問題の状態と行動設定

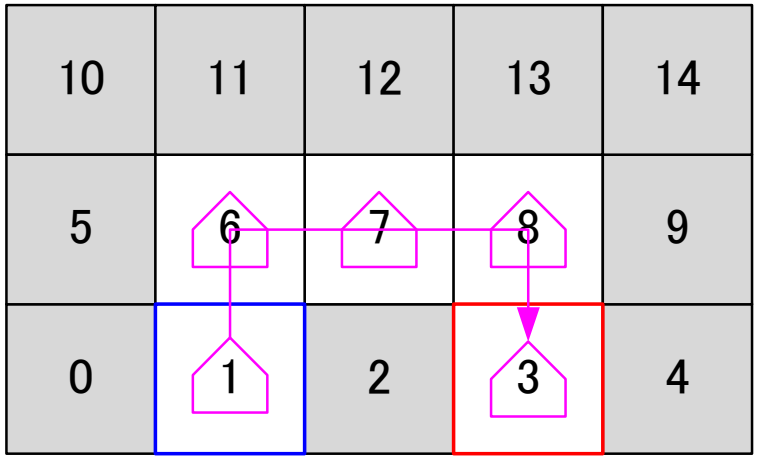
(答え)

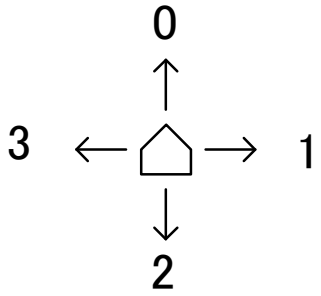
49	50	51	52	53	54	55
42	43	44	45	46	47	48
35	36	37	38	39	40	41
28	29	30	31	32	33	34
21	22	23	24	25	26	27
14	15	16	17	18	19	20
7	8	9	10	11	12	13
0	1	2	3	4	5	6



問題4：Q値の計算

下図に示すように，迷路に対して，ロボットがゴールまでの経路を辿ったとする．今回が初回エピソードである場合，ロボットが経験した状態と行動に対するQ値を計算せよ．ただし，ゴール到達に対する報酬を+1，強制停止グリッド到達に対する報酬を–1，それ以外のグリッド到達に対する報酬を0とする．さらに，ロボットは常に上を向いて移動するものとする．また，パラメータとして，学習率α=0.1，割引率γ=1.0を使用せよ．

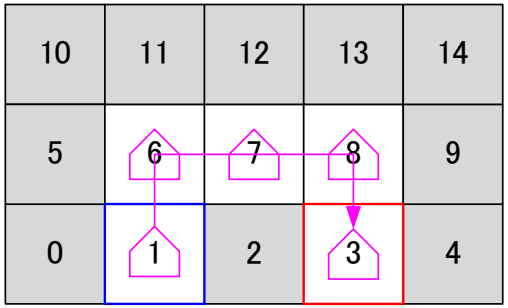


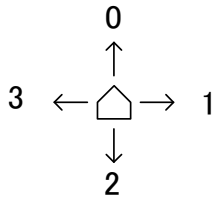


解答4：Q値の計算

(答え)

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \max_{a \in A(s_{t+1})} Q(s_{t+1}, a) - Q(s_t, a_t)\right]$$





※重要
終端状態にQ値は存在しない

各時刻tにおけるs_t, a_t, s_{t+1}, r_{t+1}

① t=0: s_t=1, a_t=0, s_{t+1}=6, r_{t+1}=0 ➡ $Q(1,0) \leftarrow 0 + 0.1\left[0 + 1 \times 0 - 0\right] = 0$

② t=1: s_t=6, a_t=1, s_{t+1}=7, r_{t+1}=0 ➡ $Q(6,1) \leftarrow 0 + 0.1\left[0 + 1 \times 0 - 0\right] = 0$

③ t=2: s_t=7, a_t=1, s_{t+1}=8, r_{t+1}=0 ➡ $Q(7,1) \leftarrow 0 + 0.1\left[0 + 1 \times 0 - 0\right] = 0$

④ t=3: s_t=8, a_t=2, s_{t+1}=3, r_{t+1}=1 ➡ $Q(8,2) \leftarrow 0 + 0.1\left[1 + 1 \times 0 - 0\right] = 0.1$