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

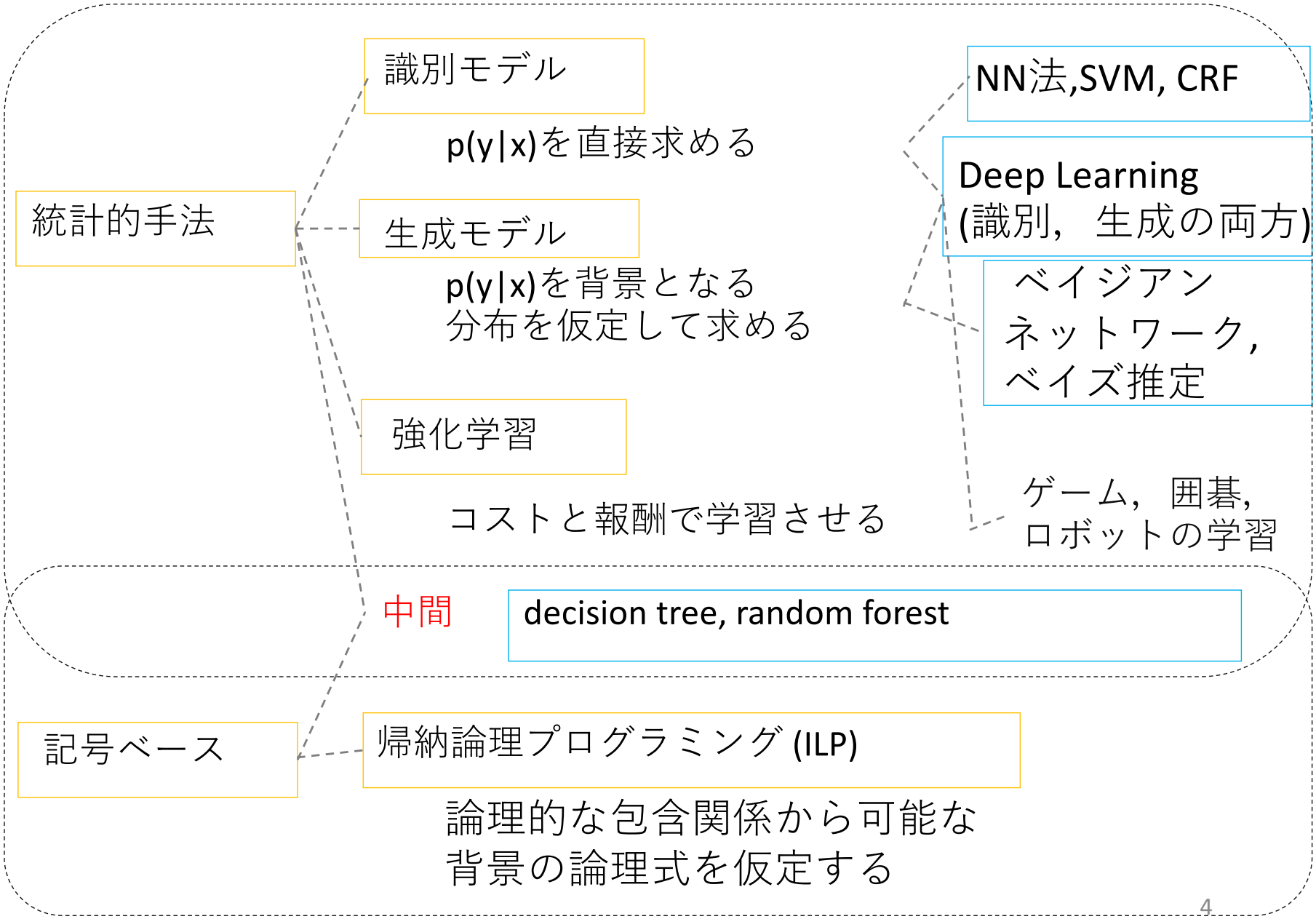
使う必要がある場合とは?

- ■学習データが使えないとき
  - ■人手による正解が作れない

例)

- ■ロボットの部品を掴む動作/ゲームの動作
  - ■状況によって腕のどの角度にするかは1通りでは無い
  - ■1つの正解例(人間が教える)としても違ってても，目標さえ達成すれば腕の角度やスピードに関して幅がある
  - ■人手で作成する学習データとして可能な全ての組合せでうまくいく例をつくるのは不可能
- ■強化学習では
  - ■正解(手順を教える)の代わりに，うまくいったときに報酬を与えることで手順はモデルに生成してもらう

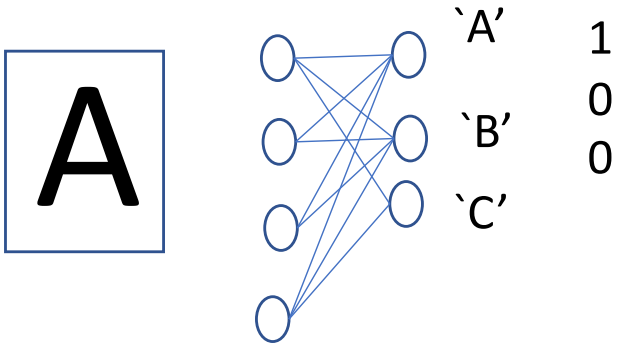
(統計的)パターン認識まわりの学習法



教師あり/教師なし(報酬あり)

教師あり

各データに対して1つの正解が与えられる (例)文字認識



深層学習など使うと良い

強化学習不要

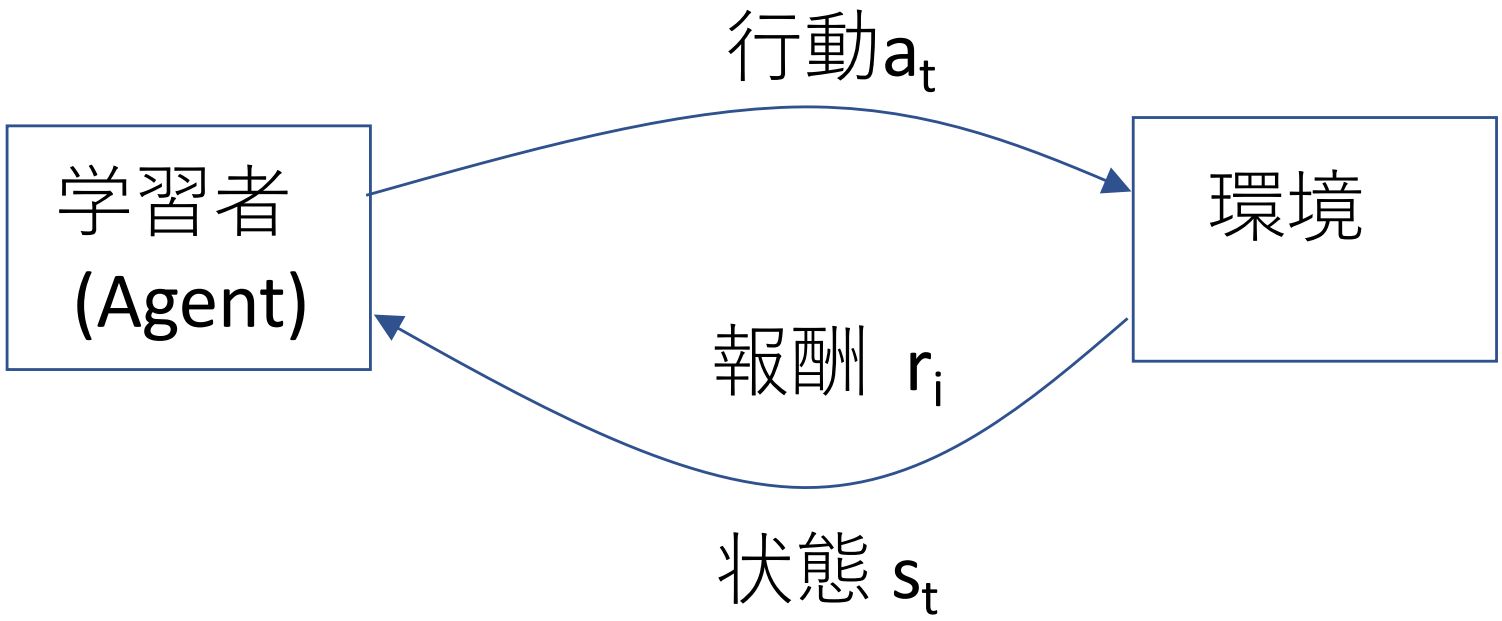
ある状態だけ良い

とり得る値(x)に対して正解は無いが，時系列後に正解がある
(例)迷路を解く

S	→				
	↓	障			
			障		G

強化学習が使える

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

練習

下記のように時刻t=1のとき状態 s_t = (1,1)にいるとする．t=2のときに s_2= (1,2) に移動して，報酬を2もらったとする．次にs_3=(2,3)に移動して報酬1をもらったとする．

問1　s_1からs_3に至る行動を示せ．ここでエージェントが取れる行動は上下左右とする

問2　s_2で得た報酬はいくらか．

問3　s_1からs_3まで移動したときの収益はいくらか

問3　次にs4に移動したときは報酬がなく，s5で報酬3を得た．s1からs5までの行動と収益を求めよ

	1	**2**	**3**
1	s_1	s_2	
2		s_3	
3		s_4	s_5

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

練習

- 下記の移動問題で各状態sの価値V(s)が計算できたとする．政策πを「価値が最大の状態に進む」とするとき，SからGに向けてどういう状態遷移をするか状態列を示せ．
  - ここで状態は(1,1)など座標で表すとする
  - 行動は上下左右のみとする

座標	1	2	3
1	S 0.0	5.8	8.1
2	3.5	9.6	G 20.5