RITSUMEIKAN

機械学習 第１５回 強化学習

立命館大学 情報理工学部

福森 隆寛

Beyond Borders

講義スケジュール

担当教員１：福森（第1回～第15回）

1	機械学習とは、機械学習の分類
2	機械学習の基本的な手順
3	識別（１）
4	識別（２）
5	識別（３）
6	回帰
7	サポートベクトルマシン
8	ニューラルネットワーク

9	深層学習
10	アンサンブル学習
11	モデル推定
12	パターンマイニング
13	系列データの識別
14	半教師あり学習
15	**強化学習**

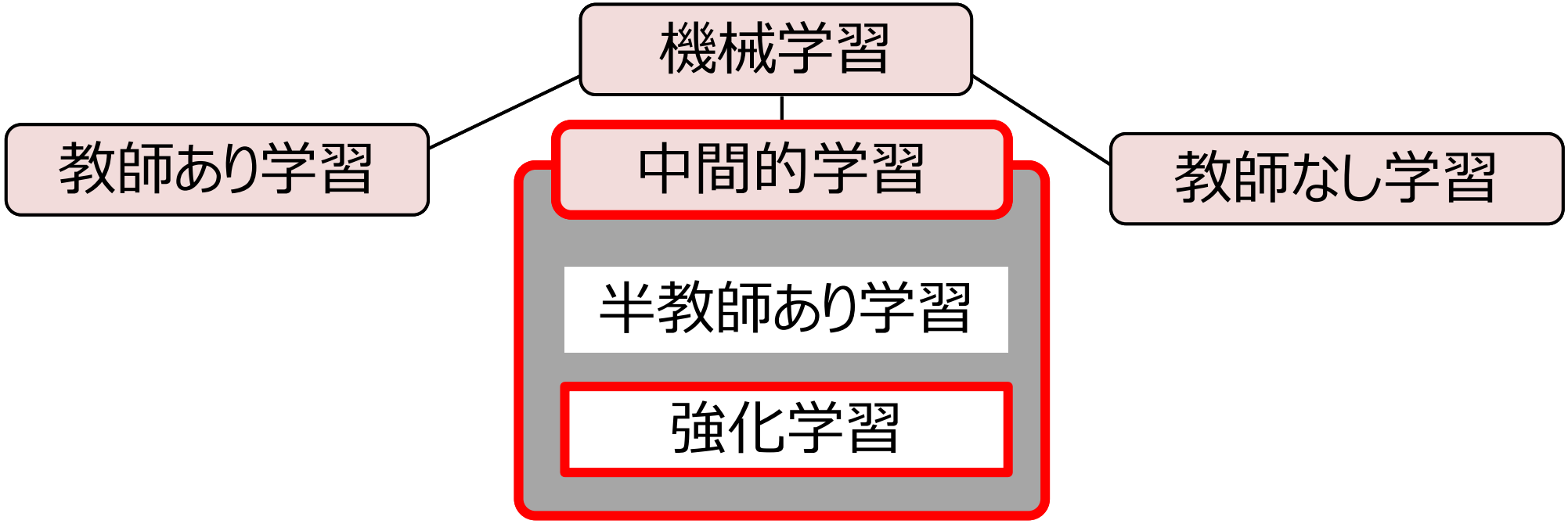
担当教員２：叶昕辰先生（第16回の講義を担当）

今回の講義内容

- 取り扱う問題の定義
- 強化学習 報酬を得るために、環境に対して何らかの行為を行う意思決定エージェントの学習
- マルコフ決定過程による定式化 マルコフ性をもつ確率過程における意思決定問題
  - K-armed bandit 問題
- Q値の推定方法
  - モデルベース 環境をモデル化する知識（状態遷移確率と報酬の確率分布）が与えられている場合に、動的計画法の考えを用いて Value iterationアルゴリズム
  - モデルフリー（TD学習）
- 演習問題 環境のモデルを持っていない場合（状態遷移確率と報酬の 確率分布が未知の場合）、試行錯誤を通じて環境と相互作用をした結果を使って学習する
- 定期試験について TD (Temporal Difference) 学習
  モデルが未知なので、環境の探索が必要になる
  報酬と遷移は未知だが決定的に定まる場合の TD学習を考える

取り扱う問題の定義：強化学習

- **教師信号に準(じゅん)ずる情報が、一部の学習データのみに与えられる状況で、各状態における最適な出力を学習**
  - 教師あり/教師なし学習の中間的な設定
    - 教師時々(ときどき)あり学習という位置づけ



強化学習

□ 強化学習

■ 報酬を得るために、環境に対して
何らかの行為を行う意思決定エージェントの学習

- 行為を行う意思決定エージェントの例 執行操作的决策代理示例
  - ロボット、将棋や囲碁などを行うプログラムなど
- エージェントには、環境に関する情報が与えられる 代理获得有关环境的信息
  - ロボットの場合：センサ・カメラ・マイクなどからの入力が環境

■ エージェントがなるべく多くの報酬を得ることを目的として
状態（カテゴリ）や状態の確率分布（連続値）を入力として、
行為（カテゴリ）を出力する関数を学習

- 学習過程の定式化に**マルコフ決定過程**が用いられる

強化学習：マルコフ決定過程

ロ マルコフ決定過程（Markov Decision Process; MDP）

具有马尔可夫性质的随机过程中的决策问题

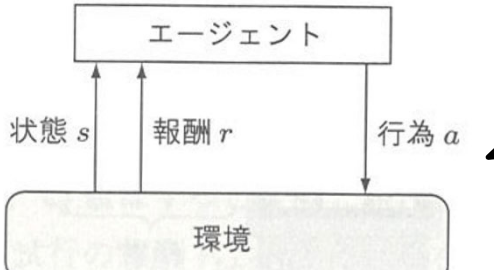
■ **マルコフ性**をもつ確率過程における意思決定問題

马尔可夫性质 次の状態において、ある事象の起こる確率は現在の状態だけから決まる（過去の状態には依存しない）という性質

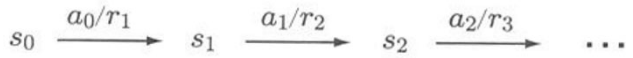
■ マルコフ決定過程は、以下の条件を仮定

1. 環境を離散的な状態の集合 $S = \{s|s \in S\}$ でモデル化
2. 時刻 t で、ある状態 s_t において、エージェントが行為 a_t を行うと報酬 r_{t+1} が得られ、状態 s_{t+1} に遷移
3. 状態遷移は確率的で、その確率は遷移前の状態にのみ依存

状态转换是随机的，其概率仅取决于转换前的状态。



マルコフ決定過程



報酬rは、たまにしか与えられない

将棋やチェスなどのゲームを考えると、「個々の手が良いか？悪いか？」はその手だけでは判断できず、最終的に勝ったときに報酬が与えられる

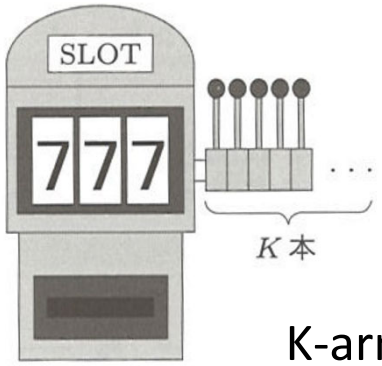
将棋等棋牌游戏只有在胜利的时候才会给酬劳

1状態問題の定式化 –K-armed bandit 問題-

有k个不同的选择（或者说动作）摆在你的面前，你需要不断重复地选择其中一个，每次你选择其中一个之后，会根据你选择的动作给你一个数值奖励，这个数值奖励符合正态分布。

K-armed bandit

- K本のアームをもつスロットマシン
- マルコフ決定過程のもとで最も単純な例
  - **1状態**：1台のスロットマシン
  - K種の行為：K本の内、どのアームを引くか？
  - 報酬：即時に与えられる
    - K本のアームは、それぞれ賞金の期待値が異なる（とする）
  - 学習結果：スロットマシンで、最大の報酬を得る行為



K-armed bandit

1状態問題の定式化 –K-armed bandit 問題-

- 報酬が**決定的**な状況での定式化
  - 全ての行為を順に試(こころ)みて
    最も報酬の高い行為を学習結果とすれば良い
  - Q値を最大にする行為を考える
    - Q値：行為aによって得られる報酬の推定値 $Q(a)$
      动作 a 获得的奖励估计值为 Q (a)
  - 定式化
    1. 行為aによって得られる報酬量が不明(ふめい)なので、
       全てのaについて$Q(a) = 0$とする
    2. 可能なaを順番(じゅんばん)に行い、そのときの報酬r_aを得る → $Q(a) = r_a$
    3. Q値が最大のaが最終的に得られる行為

1状態問題の定式化 –K-armed bandit 問題-

報酬が**非決定的**な状況での定式化

- 行為aに対応する報酬rが確率分布$p(r|a)$に従うと仮定
  - 各アームを1回だけ引くのではなく、何度も引いて、平均的な報酬が多いアームを選ぶことになる
    - 何度も試行して確率分布 $p(r|a)$を推定することと同じ 与多次尝试估计概率分布相同
  - 下式に従って、試行を繰り返して行為aの報酬の推定値$Q(a)$を収束させれば良い

$$Q_{t+1}(a) = Q_t(a) + \eta\{r_{t+1}(a) - Q_t(a)\}$$

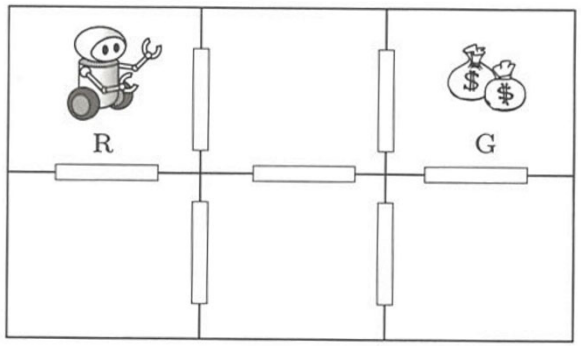
- $Q_{t+1}(a)$：時刻$t+1$における行為aの報酬（推定値）
- $Q_t(a)$：時刻tにおける行為aの報酬（推定値）
- η：学習係数
- $r_{t+1}(a) - Q_t(a)$：変動幅 波动范围 時刻tにおける行為aによる試行の報酬$r_{t+1}(a)$と、現在のQ値の差

※ 学習係数 η：Q値が収束するように時刻tの増加に従って減少（初期値：1以下の適当な値）

マルコフ決定過程による定式化

複数の状態をもつ問題に拡張

- ロボットRが迷路を移動して、ゴールGに到着すれば報酬が与えられる状況を考える



状態遷移を伴う問題

假设奖励和转换是概率性的

報酬や遷移が確率的であると想定

例えば、ロボットのゴールを探知するセンサがノイズで誤作動をしたり、路面状況でスリップが生じるなどの不確定（確率的）な要因で行為が成功しない状況が考えられる

机器人会因为噪音、路面打滑等不确定因素导致动作不成功

- この問題を以下の状況でのマルコフ決定過程として定式化
  - 報酬と次状態への遷移の確率：現在の状態と行為のみに依存

- 時刻tにおける状態 $s_t \in S$
- 時刻tにおける行為 $a_t \in A(s_t)$
- 報酬 $r_{t+1} \in \mathbb{R}$（実数）、確率分布 $p(r_{t+1}|s_t, a_t)$
- 次状態 $s_{t+1} \in S$、確率分布 $p(s_{t+1}|s_t, a_t)$

マルコフ決定過程による定式化

マルコフ決定過程における学習 基于马尔可夫决策过程中学习

- 「各状態でどの行為をとれば良いのか？」という意思決定規則（政策π）を獲得していくプロセス
- 政策πの良さは、その政策に従って行動したときの累積報酬の期待値で評価

  判断政策的好坏：根据政策行事时积累的报酬的期待值进行评价

  - 状態s_tから政策πに従って行動した時に得られる累積報酬の期待値 $V^{\pi}(s_t)$

$$V^{\pi}(s_t) = E(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots) = E\left(\sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i}\right)$$

– γ：割引率 $(0 \leq \gamma < 1)$

» あとに得られる報酬ほど割引いて計算するための係数

» 同じ報酬に辿り着けるなら、より短い手順を優先

マルコフ決定過程による定式化

- 学習の目標は、最適政策 π^*を獲得すること
  - **最適政策 π^***
    - 累積報酬の期待値が全ての状態に対して最大となる政策

$$\pi^* \equiv \underset{\pi}{\operatorname{argmax}} V^{\pi}(s_t), \forall s_t$$

  - 最適政策 π^*に従ったときの累積報酬の期待値 $V^{\pi^*}(s_t)$
    - 状態s_tで行為a_tを行った後、最適政策に従ったときの期待累積報酬の見積もり $Q^*(s_t, a_t)$が最大となる行為a_tを選択

$$Q^*(s_t, a_t) = E(r_{t+1}) + \gamma \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})$$

（※ 式の導出：次スライドの補足資料）

    - 状態s_tでの最適政策 $\pi^*(s_t)$

$$\pi^*(s_t): \text{Choose } a_t^* \quad \text{if} \quad Q^*(s_t, a_t^*) = \max_{a_t} Q^*(s_t, a_t) = V^{\pi^*}(s_t)$$

どのようにしてQ値を推定するか？

補足資料

- マルコフ決定過程による定式化
  - 状態s_tで行為a_tを行った後、最適政策に従ったときの期待累積報酬の見積もり $Q^*(s_t, a_t)$ の算出方法

$$V^{\pi^*}(s_t) = \max_{a_t} Q^*(s_t, a_t) = \max_{a_t} E\left(\sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i}\right)$$

最適政策π^*に従ったときの累積報酬の期待値

$$= \max_{a_t} E\left(\sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i}\right) = \max_{a_t} E\left(r_{t+1} + \gamma \sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i+1}\right)$$

$$= \max_{a_t} E\left(r_{t+1} + \gamma V^{\pi^*}(s_{t+1})\right)$$

遵循最优策略，积累报酬

$V^{\pi^*}(s_{t+1})$
状態s_{t+1}以降も最適政策π^*に従ったときの累積報酬

無限時刻の和で表現される状態評価関数を、隣接時刻間の再帰方程式で表現

補足資料

マルコフ決定過程による定式化（つづき）

前のスライドでは
無限時刻の和の状態評価関数を、隣接時刻間の再帰方程式で表現
※この再帰方程式をベルマン方程式 (Bellman equation)**と呼ぶ**

$$V^{\pi^*}(s_t) = \max_{a_t} E\left(\sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i}\right) = \max_{a_t} E\left(r_{t+1} + \gamma V^{\pi^*}(s_{t+1})\right)$$

無限時刻の和の状態評価関数

隣接時刻間の再帰方程式

状態遷移確率を明示的にすると…

$$V^{\pi^*}(s_t) = \max_{a_t} \left\{ E(r_{t+1}) + \gamma \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) V^{\pi^*}(s_{t+1}) \right\}$$

状態遷移確率

Q値を用いて書き換えると…

$$V^{\pi^*}(s_{t+1}) = \max_{a_t} Q^*(s_{t+1}, a_{t+1})$$

$$Q^*(s_t, a_t) = E(r_{t+1}) + \gamma \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1})$$

Q値の推定手法

- Q値の推定手法は
  モデルに関する知識の前提によって分類
  - **モデルベースの手法**
    - 環境をモデル化する知識（状態遷移確率と報酬の確率分布）が与えられている場合に、動的計画法（どうてきけいかくほう）の考えを用いて
      Q値を求める
  - **モデルフリーの手法**
    - 環境のモデルを持っていない場合（状態遷移確率と報酬の確率分布が未知の場合）、試行錯誤を通じて環境と相互作用（そうごさよう）をした結果を使って学習する 通过反复试验学习使用与环境交互的结果

Q値の推定手法：モデルベースの学習

□ モデルベースの手法

- 以下の2つの情報が与えられているものとする
  - 状態遷移確率 $P(s_{t+1}|s_t, a_t)$
  - 報酬の確率分布 $P(r_{t+1}|s_t, a_t)$

- Value iterationアルゴリズムによって、
  状態評価関数 $V(s)$の最適値を求める
  - それぞれの状態でQ値を最大とする行為（最適政策）が求まる
  - 次スライドでValue iterationアルゴリズムを説明

Q値の推定手法：モデルベースの学習

価値迭代算法

□ Value iterationアルゴリズム

$V(s)$を任意の値で初期化
repeat
　　for all $s \in S$ **do**
　　　　for all $a \in A$ **do**
　　　　　　$Q(s,a) \leftarrow E(r|s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V(s')$
　　　　end for
　　　　$V(s) \leftarrow \max_a Q(s,a)$
　　end for
until $V(s)$が収束

※ $V(s)$：状態価値(かち)関数、$E(r|s,a)$：報酬の期待値、$P(r_{t+1}|s_t,a_t)$：報酬の確率分布

※ 報酬がもらえる状態（例：ゴール）が1つだけある場合
ゴール状態の1つ手前(てまえ)での最適行為が得られ、次にその一つ手前、さらにその一つ手前…と
繰り返しを重ねるごとに正しい最適値が得られる状態がゴールを中心に広がっていくイメージ

Q値の推定手法：モデルフリーの学習

TD (Temporal Difference) 学習

- モデルが未知なので、環境の探索が必要になる
- 探索戦略(せんりゃく)としてϵ-greedy法を用いる
  - 確率$1-\epsilon$ $(0<\epsilon<1)$で最適な行為、
    確率ϵで、それ以外の行為を実行する探索手法
  - 実際は、Q値を確率に変換した下式を基準に行為を選択

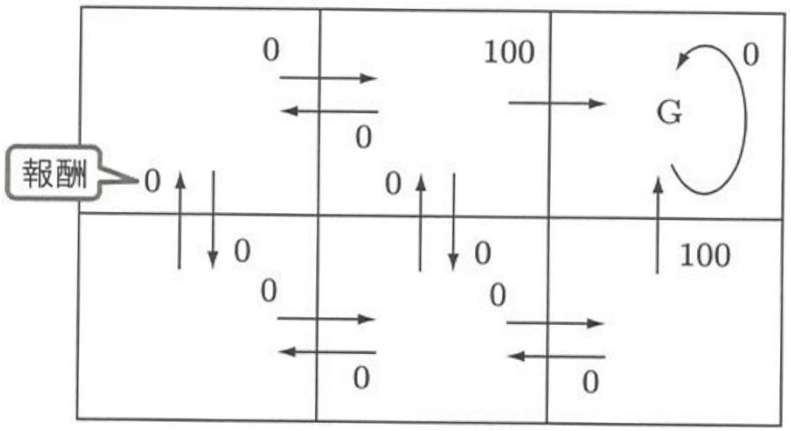
$$P(a|s)=\frac{\exp\{Q(s,a)/T\}}{\sum_{a\in A}\exp\{Q(s,a)/T\}}$$

    - 探索の初期は色々な行為を試(ため)し、落ち着いてくると最適な行為を多く選ぶように温度Tの概念を導入
      - » 学習が進むにつれて、Tを小さくすることで、学習結果が安定
    - 温度Tが高ければ全ての行為を等確率に近い確率で選択し、低(ひく)ければ最適なものに偏(かたよ)る

Q値の推定手法：決定的なTD学習

- **報酬と遷移は未知だが決定的に定まる場合**のTD学習を考える
  - 例：迷路での最適行為の獲得
    - この場合のベルマン方程式は、確率的な要素を取り除いて表現

$$Q(s_t, a_t) = r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})$$



迷路の例（迷路での最適行為の獲得）

Q値の推定手法：決定的なTD学習

- TD学習のアルゴリズム
  - 報酬と遷移が決定的な場合

$Q(s,a)$を0に初期化
for all エピソード **do**
　repeat
　　探索基準に基づき行為aを選択
　　行為aを実行し、報酬rと次状態s'を観測
　　/* 以下の式でQ値を更新 */
　　$Q(s,a) \leftarrow r + \gamma \max_{a'} Q(s',a')$
　　$s \leftarrow s'$
　until sが終了状態
end for

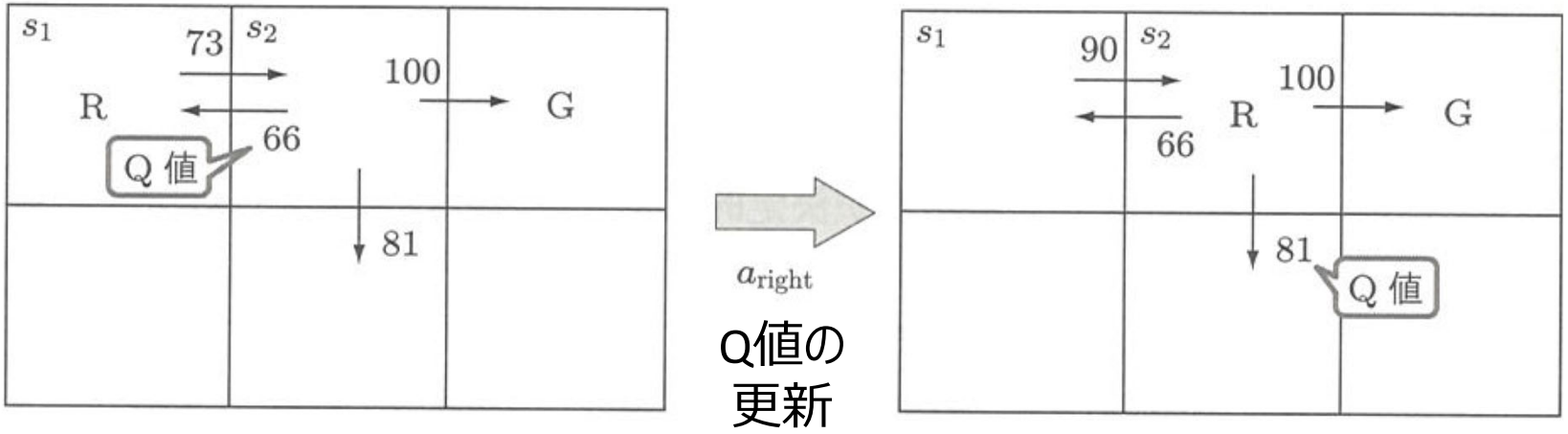
※ エピソード：1回の試行（スタートからゴールに着くか、ある移動回数に達するまでの行為系列）
※ 学習データ：エピソードの集合

Q値の推定手法：決定的なTD学習

- TD学習（Q値の更新）の例
  - 状態s_1にロボットRがいるときのQ値が左図であったとする
  - 右に移動する行為a_{right}をとると、報酬は0、状態s_2になる
    - Q値は以下のように更新

$$Q(s_t, a_{\text{right}}) \leftarrow r + \gamma \max_{a'} Q(s_2, a') \leftarrow 0 + 0.9 \max\{66,81,100\} \leftarrow 90 \quad (※\ \gamma = 0.9)$$

    - これを可能な全ての遷移系列について繰り返せば、ゴールGの報酬が末端（まったん）まで伝播して、全状態での最適行動が求まる



Q値の推定手法：確率的なTD学習

- **報酬と遷移は非決定的な場合**のTD学習を考える
  - 現在のQ値に一定割合の更新分を加えて、その割合を時間とともに減らす更新式を用いる
    - 1状態・非決定性の問題と同様

$$Q(s,a) \leftarrow Q(s,a) + \eta\{r + \gamma \max_{a'} Q(s',a') - Q(s,a)\}$$

  - 学習係数ηを適切に設定し、各状態で全ての行為を十分な回数行えれば、Q値が収束することが証明
    - あくまで理論上の話で、実際にロボットを動かして強化学習を行わせるようなケースは少なく、パラメータを変えてシミュレーション結果を評価することが多い

演習問題15-1（5分間）

- 「強化学習」と「教師あり/教師なし学習」の違いを考えなさい

定期試験について

◘ 機械学習の講義内容

1	機械学習とは、機械学習の分類
2	機械学習の基本的な手順
3	識別（１）
4	識別（２）
5	識別（３）
6	回帰
7	サポートベクトルマシン
8	ニューラルネットワーク

9	深層学習
10	アンサンブル学習
11	モデル推定
12	パターンマイニング
13	系列データの識別
14	半教師あり学習
15	強化学習

再度（さいど）、シラバスを読んで、講義内容や到達目標などを確認してください

定期試験について

- 後日(ごじつ)、以下の資料を配布(はいふ)します
  - 講義資料
    - これまでの資料と大幅な変更はありません
  - 講義中の演習問題の解答例
- 成績評価方法
  - 定期試験　60%
    - 選択問題（理論や手法に関する語句を選択肢から選ぶ）
    - 記述問題（機械学習の用語を回答）
    - 計算問題
  - 日常点評価　40%
    - 今回はオンライン講義のため、出席点のみで評価

定期試験について

- **選択問題**の例 **（注意：以下の問題は定期試験に出題されません）**
  - 「　　」に入る言葉を、以下の選択肢から1つ選択せよ。
    - アルゴリズムとして明示的に解法が与えられないタスクに対して、そのタスクを遂行するためのモデルを学習データから構築する処理を「　　」と言う。
    - **選択肢**

      標準化、機械学習、DNN、尤度

    - **答え：機械学習**

定期試験について

- **記述問題**の例 **（注意：以下の問題は定期試験に出題されません）**
  - 「　　」に入る言葉を答えよ。
    - 統計的識別では事後確率を直接的に求めることが困難であることが多い。そこで、「　　」に基づいて事後確率を尤度と事前確率の積に変換して、事後確率を間接的に求める。
    - **答え：ベイズの定理**

定期試験について

- **計算問題**の例 **（注意：以下の問題は定期試験に出題されません）**
  - ある箱に白玉が1個・黒玉が3個入っている。この箱から白玉が取り出される確率を求めよ。

    - **答え：0.25**
      - $P(\text{白玉}) = \frac{1}{4} = 0.25$

定期試験について

- 全15回で取り扱った問題の定義を再確認すること
  - どのようなデータを用いて、どのような問題を解いたのか？
- 専門用語(ようご)とその意味を理解しておくこと
  - 特に講義資料の「**赤い文字(あか)**で記載されている用語」は必ず確認・理解しておくこと
- 演習問題を理解しておくこと
  - 演習問題よりも難易度(なんいど)の高い問題を出題する予定(よてい)はありません
  - 【注意】演習問題を丸暗記(まるあんき)しないこと