G検定対策講座　第 9 回

強化学習

「正解」ではなく「報酬」で育てる

今日学ぶこと

1. **強化学習の基本**
   エージェントと環境のやり取り
2. **マルコフ決定過程と割引率**
   数学的な土台
3. **探索と活用のジレンマ**
   バンディット問題・ε-greedy・UCB
4. **価値関数の学習**
   Q学習 と SARSA
5. **方策の学習**
   方策勾配法・REINFORCE・アクター・クリティック

ゲームAIは
どうやって強くなる？

「これが正解」と教えたわけではない。勝てば報酬をもらう、それだけ

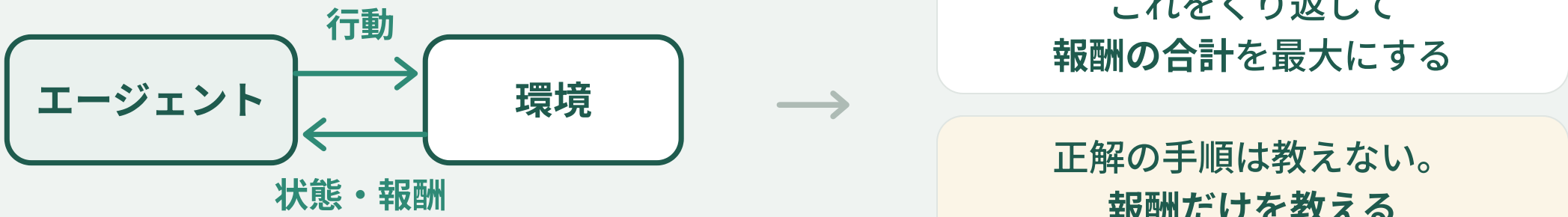
報酬から学ぶ

強化学習とは

環境とやり取りしながら、
報酬を最大化する行動を学ぶ仕組み

> **方策**＝状態から行動への対応関係
> 「敵が近ければ逃げる」など

具体例 学習のループ



これをくり返して
報酬の合計を最大にする

正解の手順は教えない。
報酬だけを教える

POINT 最良の方策は、AIが**試行錯誤で自分で見つける**

3つの学習の違い

何を手がかりに学ぶかが、それぞれ違う

教師あり学習

入力と**正解**のペアから学ぶ
画像と「これは猫」というラベル

教師なし学習

ラベルのないデータから
構造を発見する
クラスタリングなど

強化学習

報酬から試行錯誤で学ぶ
「答え」ではなく「目的」を教える

POINT 正解の手順を人間が知らなくても、**報酬さえ設計できれば**学習させられるのが強み

どんな問題に効くか

強化学習が向いている問題

正解は分からないが、ゴールは明確な問題

何度も試せるシミュレータがあれば、失敗のコストはゼロ

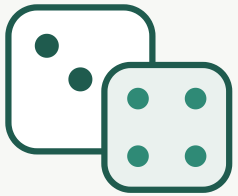
具体例 3つの条件と、具体例

- 正解は不明でもゴールは決められる
- 一手ごとに状況が変わる連続した意思決定
- 何度も試行錯誤できる環境がある

→

- 囲碁・将棋
- ロボットの制御
- 広告配信
- 在庫管理

POINT 囲碁や将棋は手の組み合わせが膨大で、**最善手は人間も知らない**



前提となる性質

マルコフ性

次の状態は、**いまの状態と行動だけ**で決まる

いまだけ覚えていればよいので
計算が現実的になる

対比 マルコフ性がある ／ ない

◯ ある：サイコロ

3回続けて1が出ても

次はまた **どの目も 1/6**

過去は関係ない

× ない：株価の予測

過去5日間の値動きが影響するなら

現在の状態だけでは決まらない

POINT **「過去にどう来たかは関係ない」**——これがマルコフ性

MDP

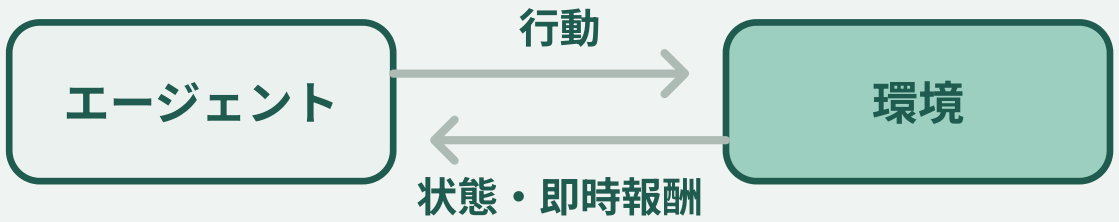
マルコフ決定過程

強化学習を**数学的に表す枠組み**

前提はマルコフ性。
ほぼすべての手法がこの枠組みで書ける

具体例 4つの要素でできている

この4つで「強化学習の問題」を書き表す



- **状態** いまの状況
- **行動** 取れる選択肢
- **遷移確率** 次に移る確率
- **即時報酬** その場の報酬

エージェントが行動し
環境が**次の状態と報酬**を返す

この形に書ければ
強化学習の道具が全部使える

POINT Q学習・SARSA・方策勾配法――すべてこの枠組みの上で定式化される**共通言語**

γ（ガンマ）

割引率

未来の報酬を割り引くための係数（0〜1）

実用では **0.9〜0.99** あたりが
よく使われる

具体例 なぜ割り引くのか

先の報酬ほど
γを掛けた分だけ小さく評価する

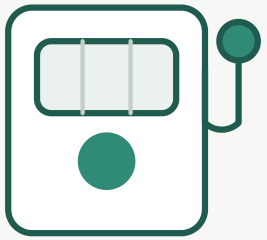
γ＝0.9 なら
10ステップ先は**およそ1/3**まで縮む

→

理由

未来は遠いほど不確実
いまの100円は確実でも
10年後の100円は分からない

POINT γが小さいと**目先重視**、大きいと**長期視点**になる



探索と活用

バンディットアルゴリズム

探索と**活用**のバランスを取る手法の総称

広告配信・推薦システム・ABテストで
広く使われている

対比 探索 ／ 活用

実績は確か
＝ 活用したい

当たりかもしれない
＝ 探索したい

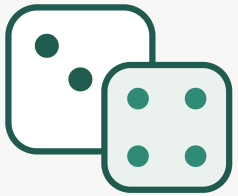
どちらに賭けるか、が探索と活用のジレンマ

活用だけだと
もっと良い台を見逃す

探索だけだと
確実な報酬を捨てることになる

ε-greedy 方策 ／ UCB 方策

POINT どちらかに偏ると損をする。代表的な2つの方策を、次のスライドから見ていく



いちばん単純

ε-greedy 方策

ふだんはベストを選び、**たまに違うものを試す**

εを徐々に減らしていく
ε-decay という工夫もある

具体例 **ε＝10% の場合**

10回のうち 1回だけ、あえて別のものを試す

90%＝これまでで
最良の選択肢（活用）

10%＝ランダムに
別のものを試す（探索）

ふだんはベスト
たまに一割だけ冒険

POINT いちばんシンプルな探索戦略。実装が簡単で、まず試される

理論的に強い

UCB 方策

期待報酬に**不確実性ボーナス**を足して選ぶ

試した回数が少ない選択肢ほど
ボーナスが大きくなる

具体例 **序盤と後半で、ふるまいが変わる**

序盤：まだ試していない選択肢を
順に試していく

後半：有望なものに集中する

→

「不確実なものを優先して試す」発想
ランダムを混ぜる ε-greedy より
無駄な試行が少ない

POINT ε-greedy＝ランダムに冒険／UCB＝**不確実さを狙って冒険**

V と Q

価値関数

ある状況の**良さを数値にした**関数

行動を選ぶときは**Qの値を直接**
比べればよい

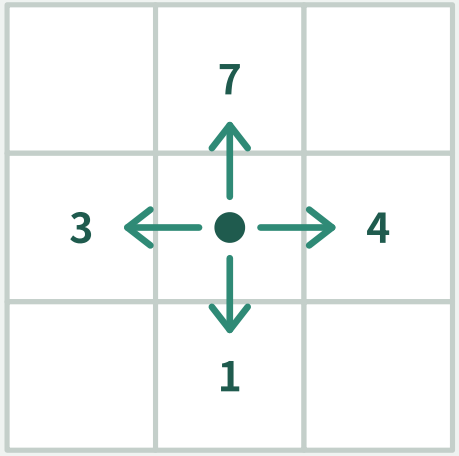
対比 **状態価値関数 V ／ 行動価値関数 Q**

V（状態価値）

2	5	9
1	4	7
0	2	3

マスそのものの良さ

Q（行動価値）

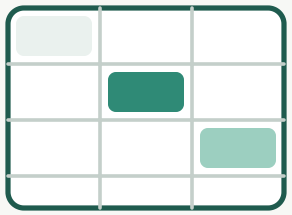


その向きに進むことの良さ

V＝状態だけを評価
「ここからだと、これくらい」

Q＝状態と行動のセットを評価
「上に進むと、これくらい」

POINT 次に見る**Q学習とSARSAは、このQを学習する**



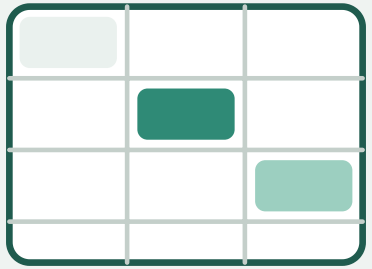
オフポリシー

Q学習

行動価値関数 Q を学ぶ、価値ベースの代表

深層化したものが **DQN**
（ディープQネットワーク）

具体例 **学ぶものと、その性質**



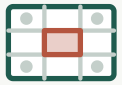
状態 × 行動 ごとの
「将来もらえる報酬の
予想値」

→

オフポリシー

動いている方策と
学んでいる方策が別でよい
試験で頻出のキーワード

POINT **Q学習＝オフポリシー**。必ず押さえる



オンポリシー

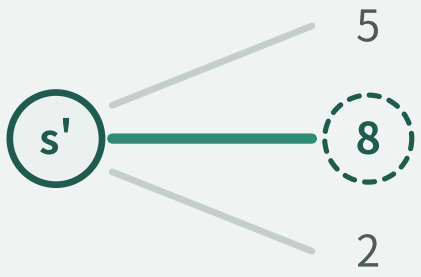
SARSA（サーサ）

学ぶものはQ学習とまったく同じ。違いは1点だけ

予想と現実のズレを使って
Qの値を補正していく

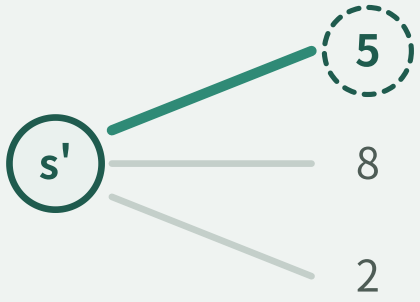
唯一の違い　ズレを計算するときの「次の行動」の仮定

Q学習（方策オフ）



いちばん大きい値を使う

SARSA（方策オン）



実際に選んだ行動の値を使う

Q学習：次に最良の行動を取ると仮定 ＝ オフポリシー

SARSA：次に実際に取る行動を使う ＝ オンポリシー

POINT　「**Q学習はオフポリシー、SARSAはオンポリシー**」――この対比が最頻出

方策ベース

方策勾配法

うまくいった行動の確率を上げ、
ダメだった行動の確率を下げる

代表的な手法が **REINFORCE**
（レインフォース）

対比 価値ベース ／ 方策ベース

価 価値ベース

価値を学んで行動を選ぶ

Q学習・SARSA

方 方策ベース

方策そのものを直接学ぶ

価値関数を介さない

連続的な行動にも対応できる
自動運転のハンドル角度など

POINT REINFORCEは、1エピソードを最後まで走らせて**報酬が大きかった行動の確率を上げる**

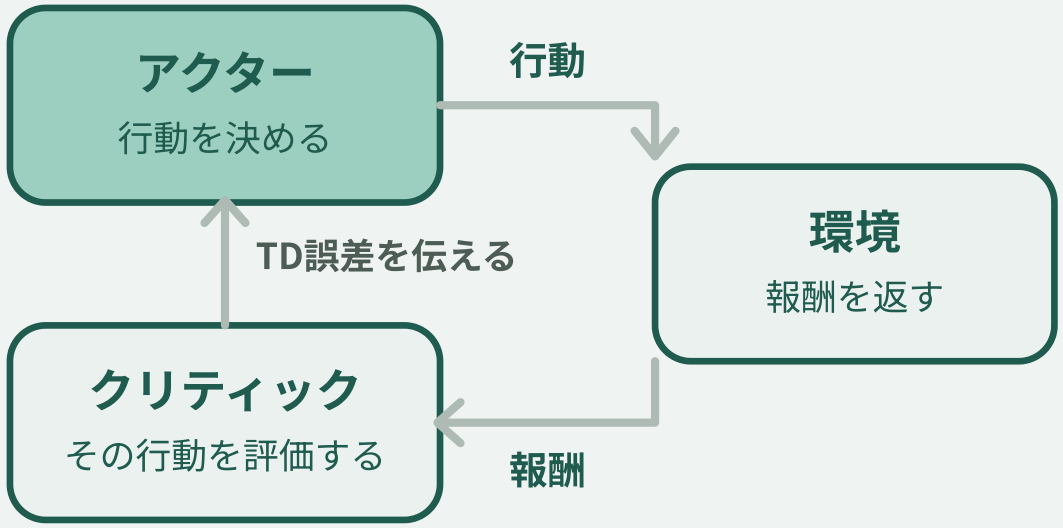
ハイブリッド

アクター・クリティック

価値ベースと方策ベースを**組み合わせた**手法

現代の深層強化学習**A3C・PPO**はすべてこの発展形

具体例 2つの役割を分担する



アクター
方策に従って行動を選ぶ

クリティック
その行動を評価する

評価をもとに
方策を改善していく

POINT クリティックが期待報酬を見積もり、アクターがそれを元に方策を良くしていく

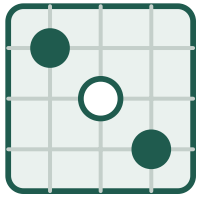
実世界での応用

強化学習が使われている場面

正解は分からないが目的は明確な
問題で力を発揮する

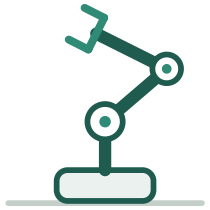
ChatGPTでも使われている
RLHF（人間フィードバック強化学習）

具体例 代表的な4つ



ゲーム

AlphaGo が
世界王者を倒した



ロボット制御

歩行の学習
物をつかむ

自動運転

経路の選択
ハンドル操作

対話AI

RLHF
ChatGPTなど

POINT ロボットは**シミュレータで試行を重ねてから**実機に移す

キーワードを分類して振り返る

枠組み
マルコフ性
マルコフ決定過程
割引率 γ

探索戦略
バンディット問題
ε-greedy
UCB

価値関数
状態価値関数 V
行動価値関数 Q

価値ベース
Q学習（オフポリシー）
SARSA（オンポリシー）
DQN

方策ベース
方策勾配法
REINFORCE

融合と応用
アクター・クリティック
AlphaGo・ロボット
自動運転・RLHF

POINT シラバスの**12個のキーワード**を、これですべて押さえた