

目次

はじめに	III
謝辞	IV
本書の読み方	V
サンプルコードの見方について	VII

Chapter 1 強化学習のゴールと課題 1

1.1 強化学習の考え方	2
1.2 実行環境のセットアップ	5
1.3 バンディットアルゴリズム (基本編)	11
1.3.1 多腕バンディット問題	11
1.3.2 平均値計算の効率化	18
1.3.3 『活用』と『探索』の組み合わせ	19
1.3.4 ハイパーパラメーター・チューニング	25
1.4 バンディットアルゴリズム (応用編)	28
1.4.1 非定常状態への対応	28
1.4.2 初期値を用いた探索のトリック	36

Chapter 2 環境モデルを用いた強化学習の枠組み 45

2.1 マルコフ決定過程による環境のモデル化	46
2.1.1 状態遷移図と報酬設計	46
① エピソード的タスク	49
② 非エピソード的タスク (報酬の割引率)	51
2.1.2 確率的な状態変化とバックアップ図	55
2.2 エージェントの行動ポリシーと状態価値関数	60
2.2.1 状態価値関数の定義と計算例	60
2.2.2 状態価値関数の比較による最善の行動ポリシーの発見	65
2.3 動的計画法による状態価値関数の決定	66
2.3.1 ベルマン方程式と動的計画法	67
2.3.2 動的計画法による計算例	71
① 1次元のグリッドワールド	71
② 2次元のグリッドワールド	86
③ 確率的な状態遷移を含む場合	93

Chapter 3 行動ポリシーの改善アルゴリズム 101

3.1 ポリシー反復法 102

3.1.1 『一手先読み』による行動ポリシーの改善 102

3.1.2 ポリシー反復法の適用例 109

① 1次元のグリッドワールド 109

② 2次元のグリッドワールド 115

3.2 価値反復法 124

3.2.1 状態価値関数と行動ポリシーの並列更新 124

3.2.2 価値反復法の適用例 126

3.3 より実践的な実装例 131

3.3.1 三目並べ 131

3.3.2 レンタカー問題 145

Chapter 4 サンプルングデータを用いた学習法 163

4.1 モンテカルロ法 164

4.1.1 シミュレーションによるデータ収集 164

4.1.2 サンプルングによる状態価値関数の評価 170

4.1.3 サンプルングを用いた価値反復法 180

4.1.4 オフポリシーでのデータ収集 188

4.2 TD (Temporal-Difference) 法 201

4.2.1 オフポリシーでのTD法 Q-Learning 202

4.2.2 オンポリシーでのTD法 SARSA 210

Chapter 5 ニューラルネットワークによる関数近似 219

5.1 ニューラルネットワークによる状態価値関数の計算 220

5.1.1 関数近似の考え方 220

5.1.2 強化学習におけるニューラルネットワークの学習方法 222

① 1次元のグリッドワールド 一次関数による近似例 227

② 2次元のグリッドワールド ニューラルネットワークによる近似例 236

5.2 ニューラルネットワークを用いたQ-Learning 245

5.2.1 CNN (畳み込みニューラルネットワーク) による特徴抽出 246

5.2.2 DQNの実装例 251

5.2.3 実行時の先読みによる性能向上 267

おわりに	276
参考書籍	277
索引	278

Column

Colaboratoryのランタイムについて	10
データフレームを用いたグラフ描画	17
NumPyのarrayオブジェクト	31
無限級数の公式	53
ポアソン分布による確率の計算	147