

УДК 004.85
ББК 32.971.3
С21

Саттон Р. С., Барто Э. Дж.

С21 Обучение с подкреплением: Введение. 2-е изд. / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2020. – 552 с.: ил.

ISBN 978-5-97060-097-9

Идея обучения с подкреплением возникла десятки лет назад, но этой дисциплине предстояло пройти долгий путь, прежде чем она стала одним из самых активных направлений исследований в области машинного обучения и нейронных сетей. Сегодня это предмет интереса ученых, занимающихся психологией, теорией управления, искусственным интеллектом и многими другими отраслями знаний.

Подход, принятый авторами книги, ставит акцент на практическое использование обучения с подкреплением. В первой части читатель знакомится с базовыми его аспектами. Во второй части представлены приближенные методы решения в условиях ограниченных вычислительных ресурсов. В третьей части книги обсуждается важность обучения с подкреплением для психологии и нейронаук.

Издание предназначено для студентов технических вузов, разработчиков, специализирующихся на машинном обучении и искусственном интеллекте, а также представителей нетехнических профессий, которые могут использовать описанные методики в своей работе.

УДК 004.85
ББК 32.971.3

Original English language edition published by The MIT Press Cambridge, MA. Copyright © 2018 Richard S. Sutton and Andrew G. Barto. Russian-language edition copyright © 2020 by DMK Press. All rights reserved.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN 978-0-262-03924-6 (англ.)
ISBN 978-5-97060-097-9 (рус.)

Copyright © 2018 Richard S. Sutton and Andrew G. Barto
© Оформление, издание, перевод, ДМК Пресс, 2020

Содержание

Вступительное слово	11
Предисловие ко второму изданию	12
Предисловие к первому изданию	17
Обозначения	20
От издательства	25
Глава 1. Введение	26
1.1. Обучение с подкреплением.....	26
1.2. Примеры	30
1.3. Элементы обучения с подкреплением	31
1.4. Ограничения и круг вопросов	33
1.5. Развернутый пример: игра в крестики-нолики.....	34
1.6. Резюме	39
1.7. История ранних этапов обучения с подкреплением	39
Библиографические замечания.....	49
Часть I. ТАБЛИЧНЫЕ МЕТОДЫ РЕШЕНИЯ	50
Глава 2. Многорукие бандиты	51
2.1. Задача о k -руком бандите.....	51
2.2. Методы ценности действий	53
2.3. 10-рукий испытательный стенд.....	54
2.4. Инкрементная реализация.....	57
2.5. Нестационарная задача	59
2.6. Оптимистические начальные значения.....	60
2.7. Выбор действия, дающего верхнюю доверительную границу.....	62
2.8. Градиентные алгоритмы бандита.....	64
2.9. Ассоциативный поиск (контекстуальные бандиты).....	68
2.10. Резюме	69
Библиографические и исторические замечания.....	71
Глава 3. Конечные марковские процессы принятия решений	74
3.1. Интерфейс между агентом и окружающей средой.....	74
3.2. Цели и вознаграждения	80
3.3. Доход и эпизоды	82
3.4. Унифицированная нотация для эпизодических и непрерывных задач	84
3.5. Стратегии и функции ценности	86
3.6. Оптимальные стратегии и оптимальные функции ценности	91
3.7. Оптимальность и аппроксимация	96
3.8. Резюме	97
Библиографические и исторические замечания.....	99

Глава 4. Динамическое программирование	102
4.1. Оценивание стратегии (предсказание)	103
4.2. Улучшение стратегии	107
4.3. Итерация по стратегиям	109
4.4. Итерация по ценности	112
4.5. Асинхронное динамическое программирование	115
4.6. Обобщенная итерация по стратегиям	116
4.7. Эффективность динамического программирования	117
4.8. Резюме	118
Библиографические и исторические замечания	119

Глава 5. Методы Монте-Карло	122
5.1. Предсказание методами Монте-Карло	123
5.2. Оценивание ценности действий методом Монте-Карло	127
5.3. Управление методом Монте-Карло	129
5.4. Управление методом Монте-Карло без исследовательских стартов	132
5.5. Предсказание с разделенной стратегией посредством выборки по значимости	135
5.6. Инкрементная реализация	142
5.7. Управление методом Монте-Карло с разделенной стратегией	143
5.8. *Выборка по значимости с учетом обесценивания	146
5.9. *Приведенная выборка по значимости	147
5.10. Резюме	149
Библиографические и исторические замечания	150

Глава 6. Обучение на основе временных различий	152
6.1. Предсказание TD-методами	152
6.2. Преимущества TD-методов предсказания	157
6.3. Оптимальность TD(0)	159
6.4. Sarsa: TD-управление с единой стратегией	162
6.5. Q-обучение: TD-управление с разделенной стратегией	165
6.6. Expected Sarsa	167
6.7. Смещение максимизации и двойное обучение	169
6.8. Игры, послесостояния и другие специальные случаи	171
6.9. Резюме	173
Библиографические и исторические замечания	174

Глава 7. n-шаговый бутстрэппинг	176
7.1. n -шаговое TD-предсказание	176
7.2. n -шаговый алгоритм Sarsa	181
7.3. n -шаговое обучение с разделенной стратегией	184
7.4. *Приведенные методы с переменным управлением	186
7.5. Обучение с разделенной стратегией без выборки по значимости: n -шаговый алгоритм обновления по дереву	188
7.6. *Унифицированный алгоритм: n -шаговый Q(σ)	190
7.7. Резюме	193
Библиографические и исторические замечания	194

Глава 8. Планирование и обучение табличными методами	195
8.1. Модели и планирование	195
8.2. Дупа: объединение планирования, исполнения и обучения	198

8.3. Когда модель неверна	203
8.4. Приоритетный проход	206
8.5. Сравнение выборочного и полного обновлений	210
8.6. Траекторная выборка	213
8.7. Динамическое программирование в реальном времени	216
8.8. Планирование в момент принятия решений	220
8.9. Эвристический поиск	221
8.10. Разыгрывающие алгоритмы	224
8.11. Поиск по дереву методом Монте-Карло	226
8.12. Резюме главы	229
8.13. Резюме части I: оси	230
Библиографические и исторические замечания	233

Часть II. ПРИБЛИЖЕННЫЕ МЕТОДЫ РЕШЕНИЯ 236

Глава 9. Предсказание с единой стратегией и аппроксимацией 238

9.1. Аппроксимация функции ценности	239
9.2. Целевая функция предсказания (\bar{V}_E)	240
9.3. Стохастические градиентные и полуградиентные методы	242
9.4. Линейные методы	246
9.5. Конструирование признаков для линейных методов	252
9.5.1. Полиномы	252
9.5.2. Базис Фурье	254
9.5.3. Грубое кодирование	257
9.5.4. Плиточное кодирование	260
9.5.5. Радиально-базисные функции	265
9.6. Выбор размера шага вручную	266
9.7. Нелинейная аппроксимация функций: искусственные нейронные сети	267
9.8. Алгоритм TD наименьших квадратов	272
9.9. Аппроксимация функций с запоминанием	274
9.10. Аппроксимация с помощью ядерных функций	276
9.11. Более глубокий взгляд на обучение с единой стратегией: заинтересованность и значимость	278
9.12. Резюме	280
Библиографические и исторические замечания	281

Глава 10. Управление с единой стратегией и аппроксимацией 288

10.1. Эпизодическое полуградиентное управление	288
10.2. Полуградиентный n -шаговый Sarsa	292
10.3. Среднее вознаграждение: новая постановка непрерывных задач	294
10.4. Возражения против постановки с обесцениванием	299
10.5. Дифференциальный полуградиентный n -шаговый Sarsa	301
10.6. Резюме	302
Библиографические и исторические замечания	303

Глава 11. *Методы с разделенной стратегией и аппроксимацией 304

11.1. Полуградиентные методы	305
11.2. Примеры расходимости в случае с разделенной стратегией	307
11.3. Смертельная триада	312

11.4. Геометрия линейной аппроксимации функций ценности	314
11.5. Градиентный спуск по беллмановской ошибке	318
11.6. Беллмановская ошибка необучаема	322
11.7. Градиентные TD-методы	327
11.8. Эмфатические TD-методы	330
11.9. Уменьшение дисперсии.....	332
11.10. Резюме	334
Библиографические и исторические замечания.....	335

Глава 12. Следы приемлемости 337

12.1. λ -доход	338
12.2. TD(λ)	342
12.3. n -шаговые усеченные λ -доходные методы	346
12.4. Пересчет обновлений: онлайнный λ -доходный алгоритм	348
12.5. Истинно онлайнный TD(λ)	350
12.6. *Голландские следы в обучении методами Монте-Карло	352
12.7. Sarsa(λ)	354
12.8. Переменные λ и γ	359
12.9. Следы с разделенной стратегией и переменным управлением	361
12.10. От Q(λ) Уоткинса к Tree-Backup(λ)	364
12.11. Устойчивые методы с разделенной стратегией со следами приемлемости	367
12.12. Вопросы реализации.....	368
12.13. Выводы.....	369
Библиографические и исторические замечания.....	371

Глава 13. Методы градиента стратегии..... 373

13.1. Аппроксимация стратегии и ее преимущества	374
13.2. Теорема о градиенте стратегии	376
13.3. REINFORCE: метод Монте-Карло на основе градиента стратегии	378
13.4. REINFORCE с базой.....	381
13.5. Методы исполнитель–критик	383
13.6. Метод градиента стратегии для непрерывных задач.....	385
13.7. Параметризация стратегии для непрерывных действий	388
13.8. Резюме	389
Библиографические и исторические замечания.....	390

Часть III. ЗАГЛЯНЕМ ПОГЛУБЖЕ..... 392

Глава 14. Психология 393

14.1. Предсказание и управление	394
14.2. Классическое обусловливание	395
14.2.1. Блокирующее обусловливание и обусловливание высшего порядка	397
14.2.2. Модель Рескорлы–Вагнера.....	399
14.2.3. TD-модель	401
14.2.4. Имитирование TD-модели.....	403
14.3. Инструментальное обусловливание	410
14.4. Отложенное подкрепление	415
14.5. Когнитивные карты	416
14.6. Привычное и целеустремленное поведение	418
14.7. Резюме.....	423
Библиографические и исторические замечания.....	425

Глава 15. Нейронауки	432
15.1. Основы нейронаук	433
15.2. Сигналы вознаграждения, сигналы подкрепления, ценности и ошибки предсказания	435
15.3. Гипотеза об ошибке предсказания вознаграждения	437
15.4. Дофамин	439
15.5. Экспериментальное подтверждение гипотезы об ошибке предсказания вознаграждения	443
15.6. Параллель между TD-ошибкой и дофамином	447
15.7. Нейронный исполнитель – критик	452
15.8. Правила обучения критика и исполнителя	456
15.9. Гедонистические нейроны	460
15.10. Коллективное обучение с подкреплением	462
15.11. Основанные на модели методы в мозге	466
15.12. Наркотическая зависимость	468
15.13. Резюме	469
Библиографические и исторические замечания	472
Глава 16. Примеры и приложения	481
16.1. TD-Gammon	481
16.2. Программы игры в шашки Сэмюэла	486
16.3. Стратегия выбора ставки в программе Watson	489
16.4. Оптимизация управления памятью	492
16.5. Игра в видеоигры на уровне человека	497
16.6. Мастерство игры в го	503
16.6.1. AlphaGo	506
16.6.2. AlphaGo Zero	509
16.7. Персонализированные веб-службы	513
16.8. Парение в восходящих потоках воздуха	516
Глава 17. Передовые рубежи	521
17.1. Общие функции ценности и вспомогательные задачи	521
17.2. Абстрагирование времени посредством опций	523
17.3. Наблюдения и состояния	526
17.4. Проектирование сигналов вознаграждения	532
17.5. Остающиеся вопросы	535
17.6. Экспериментальное подтверждение гипотезы об ошибке предсказания вознаграждения	539
Библиографические и исторические замечания	543
Предметный указатель	587