

УДК 004.85
ББК 32.971.3
Л76

Лонца А.

Л76 Алгоритмы обучения с подкреплением на Python / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2020. – 286 с.: ил.

ISBN 978-5-97060-855-5

Эта книга поможет читателю овладеть алгоритмами обучения с подкреплением (ОП) и научиться реализовывать их при создании самообучающихся агентов.

В первой части рассматриваются различные элементы ОП, сфера его применения, инструменты, необходимые для работы в среде ОП. Вторая и третья части посвящены непосредственно алгоритмам. В числе прочего автор показывает, как сочетать Q-обучение с нейронными сетями для решения сложных задач, описывает методы градиента стратегии, TRPO и PPO, позволяющие повысить производительность и устойчивость, а также детерминированные алгоритмы DDPG и TD3. Читатель узнает о том, как работает техника подражательного обучения, познакомится с алгоритмами исследования на базе верхней доверительной границы (UCB и UCB1) и метаалгоритмом ESBAS.

Издание предназначено для тех, кто интересуется исследованиями в области искусственного интеллекта, применяет в работе глубокое обучение или хочет освоить обучение с подкреплением с нуля. Обязательное условие – владение языком Python на рабочем уровне.

УДК 004.85
ББК 32.971.3

First published in the English language under the title 'Reinforcement Learning Algorithms with Python. 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 978-1-78913-111-6 (англ.)
ISBN 978-5-97060-855-5 (рус.)

Copyright © Packt Publishing 2019
© Оформление, издание, перевод,
ДМК Пресс, 2020

Содержание

Об авторе	12
Предисловие	13
Часть I. АЛГОРИТМЫ И ОКРУЖАЮЩИЕ СРЕДЫ	18
Глава 1. Ландшафт обучения с подкреплением	19
Введение в ОП.....	20
Сравнение ОП и обучения с учителем	22
История ОП	23
Глубокое обучение	25
Элементы ОП	26
Стратегия	26
Функция ценности.....	28
Вознаграждение.....	29
Модель	30
Применение ОП	30
Игры.....	30
Робототехника и индустрия 4.0	31
Машинное обучение.....	32
Экономика и финансы	32
Здравоохранение	32
Интеллектуальные транспортные системы.....	33
Оптимизация энергопотребления и умные сети электроснабжения.....	33
Резюме	33
Вопросы	33
Для дальнейшего чтения.....	34
Глава 2. Реализация цикла ОП и OpenAI Gym	35
Настройка окружающей среды	36
Установка OpenAI Gym	36
Установка Roboschool	37
OpenAI Gym и цикл ОП.....	37
Разработка цикла ОП.....	38
Привыкаем к пространствам	41
Разработка моделей МО с помощью TensorFlow	42
Тензоры	43
Создание графа	45
Простой пример линейной регрессии	46

Введение в TensorBoard	49
Типы окружающих сред ОП	51
Зачем нужны различные среды?	51
Окружающие среды с открытым исходным кодом	52
Резюме	54
Вопросы	55
Для дальнейшего чтения	55

Глава 3. Решение задач методом динамического программирования

МППР	56
Стратегия	58
Доход	58
Функции ценности	59
Уравнение Беллмана	60
Классификация алгоритмов ОП	61
Безмодельные алгоритмы	62
Алгоритмы ОП, основанные на модели	63
Разнообразие алгоритмов	64
Динамическое программирование	64
Оценивание и улучшение стратегии	65
Итерация по стратегиям	66
Итерация по ценности	70
Резюме	72
Вопросы	73
Для дальнейшего чтения	73

Часть II. БЕЗМОДЕЛЬНЫЕ АЛГОРИТМЫ ОП

Глава 4. Применение Q-обучения и алгоритма SARSA

Обучение без модели	76
Порядок действий	76
Оценивание стратегии	77
Проблема исследования	77
TD-обучение	78
TD-обновление	79
Улучшение стратегии	79
Сравнение методов Монте-Карло и TD-методов	79
SARSA	80
Алгоритм	80
Применение SARSA к игре Taxi-v2	81
Q-обучение	86
Теория	86
Алгоритм	87
Применение Q-обучения к игре Taxi-v2	87
Сравнение SARSA и Q-обучения	89

Резюме.....	91
Вопросы.....	92
Глава 5. Глубокая Q-сеть.....	93
Глубокие нейронные сети и Q-обучение	93
Аппроксимация функций	94
Q-обучение с нейронными сетями	95
Неустойчивость глубокого Q-обучения	96
DQN.....	97
Решение.....	97
Алгоритм DQN	98
Архитектура модели.....	101
Применение DQN к игре Pong	102
Игры Atari	102
Предварительная обработка	103
Реализация DQN	105
Результаты.....	112
Вариации на тему DQN.....	113
Double DQN	114
Dueling DQN	117
n-шаговый DQN	118
Резюме.....	120
Вопросы.....	120
Для дальнейшего чтения.....	121

Глава 6. Стохастическая оптимизация и градиенты

стратегии.....	122
Методы градиента стратегии.....	122
Градиент стратегии	123
Теорема о градиенте стратегии.....	124
Вычисление градиента.....	125
Стратегия	126
Алгоритм ГС с единой стратегией.....	127
Устройство алгоритма REINFORCE.....	127
Реализация REINFORCE.....	129
Посадка космического корабля с помощью алгоритма REINFORCE	132
REINFORCE с базой	134
Реализация REINFORCE с базой.....	136
Обучение алгоритма исполнитель–критик.....	137
Как критик помогает обучаться исполнителю	137
n-шаговая модель AC.....	138
Реализация AC.....	139
Посадка космического корабля с помощью алгоритма AC	141
Дополнительные улучшения AC и полезные советы	142
Резюме.....	143
Вопросы.....	143
Для дальнейшего чтения.....	143

Глава 7. Реализация TRPO и PPO	144
Roboschool	144
Управление непрерывной системой.....	145
Метод естественного градиента стратегии	148
Интуитивное описание NPG	149
Немного математики	150
Осложнения в методе естественного градиента	152
Оптимизация стратегии в доверительной области	152
Алгоритм TRPO	153
Реализация алгоритма TRPO	156
Применение TRPO	160
Проксимальная оптимизация стратегии	163
Краткое описание	163
Алгоритм PPO	163
Реализация PPO	164
Применение PPO	166
Резюме.....	168
Вопросы.....	168
Для дальнейшего чтения.....	169
 Глава 8. Применения алгоритмов DDPG и TD3	170
Сочетание оптимизации градиента стратегии с Q-обучением	170
Детерминированный градиент стратегии.....	171
Алгоритм DDPG	174
Реализация DDPG	176
Применение DDPG к среде BipedalWalker-v2.....	180
Алгоритм TD3	182
Проблема смещения оценки в сторону завышения.....	182
Уменьшение дисперсии	184
Применение TD3 к среде BipedalWalker-v2.....	186
Резюме.....	187
Вопросы.....	188
Для дальнейшего чтения.....	188
 Часть III. ЗА ПРЕДЕЛАМИ БЕЗМОДЕЛЬНЫХ АЛГОРИТМОВ	189
 Глава 9. ОП на основе модели	190
Методы на основе модели.....	190
Общая картина обучения на основе модели	191
Достоинства и недостатки	195
Сочетание безмодельного и основанного на модели обучения	196
Полезная комбинация	196
Построение модели из изображений	198
Применение алгоритма ME-TRPO к задаче об обратном маятнике	199

Принцип работы ME-TRPO	200
Реализация ME-TRPO	200
Эксперименты в среде RoboSchool	204
Резюме	206
Вопросы	207
Для дальнейшего чтения	207
Глава 10. Подражательное обучение и алгоритм DAgger	208
Технические требования	208
Установка Flappy Bird	209
Подход на основе подражания	209
Пример: помощник водителя	210
Сравнение подражательного обучения и обучения с подкреплением	211
Роль эксперта в подражательном обучении	211
Структура IL	212
Игра Flappy Bird	214
Порядок взаимодействия с окружающей средой	215
Алгоритм агрегирования набора данных	216
Алгоритм DAgger	217
Реализация DAgger	217
Анализ результатов игры в Flappy Bird	221
Обратное обучение с подкреплением	222
Резюме	223
Вопросы	223
Для дальнейшего чтения	224
Глава 11. Оптимизация методом черного ящика	225
За рамками ОП	225
Краткий обзор ОП	226
Альтернатива	226
Основы эволюционных алгоритмов	227
Генетические алгоритмы	230
Эволюционные стратегии	230
Масштабируемые эволюционные стратегии	232
Основной принцип	233
Масштабируемая реализация	234
Применение масштабируемой ЭС к среде LunarLander	239
Резюме	241
Вопросы	241
Для дальнейшего чтения	242
Глава 12. Разработка алгоритма ESBAS	243
Исследование и использование	244
Задача о многоруком бандите	245
Подходы к исследованию	246
ϵ -жадная стратегия	246

Алгоритм UCB	247
Сложность исследования	248
Алгоритм ESBAS.....	249
Что такое выбор алгоритма	249
ESBAS изнутри	250
Реализация.....	252
Тестирование в среде Acrobot.....	255
Резюме.....	257
Вопросы.....	258
Для дальнейшего чтения.....	258

Глава 13. Практические подходы к решению проблем ОП..... 259

Рекомендуемые практики глубокого ОП	259
Выбор подходящего алгоритма	260
От простого к сложному.....	261
Проблемы глубокого ОП.....	263
Устойчивость и воспроизводимость результатов	263
Эффективность	264
Обобщаемость.....	265
Передовые методы	266
ОП без учителя.....	266
Перенос обучения.....	268
ОП в реальном мире.....	270
Лицом к лицу с реальным миром.....	270
Преодоление разрыва между имитационной моделью и реальным миром	271
Создание собственной окружающей среды.....	272
Будущее ОП и его влияние на общество	272
Резюме.....	273
Вопросы.....	274
Для дальнейшего чтения.....	274

Ответы на вопросы..... 275

Предметный указатель 281