

УДК 004.4
ББК 32.972
Ш18

Шалев-Шварц Ш., Бен-Давид Ш.
Ш18 Идеи машинного обучения: от теории к алгоритмам / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2019. – 436 с.: ил.

ISBN 978-5-97060-673-5

Машинное обучение – один из самых быстро развивающихся разделов информатики с приложениями в самых разных областях. Цель этой книги – познакомить читателя с фундаментальными принципами машинного обучения и характерными для него алгоритмическими парадигмами. Книга содержит обширный свод основополагающих теоретических идей машинного обучения и математические выкладки, благодаря которым эти идеи становятся практическими алгоритмами. Вслед за изложением базовых основ дисциплины рассматривается широкий спектр тем, не нашедших достаточного отражения в предшествующих учебниках: вычислительная сложность обучения, понятия выпуклости и устойчивости, важные алгоритмы, включая стохастический градиентный спуск, нейронные сети и обучение структурированному выводу, а также совсем недавние теоретические концепции, например, РАС-байесовский подход и границы сжатия.

Издание ориентировано на студентов старших курсов, обучающихся информатике, техническим наукам, математике или статистике, а также может быть полезно исследователям, желающим углубить свои теоретические знания. Предполагается, что читатель знаком с основами теории вероятностей, линейной алгебры, математического анализа и теории алгоритмов.

УДК 004.4
ББК 32.972

Original English language edition published by Cambridge University Press is part of the University of Cambridge. Copyright © 2014 by Shai Shalev-Shwartz and Shai Ben-David. Russian-language edition copyright © 2019 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 978-1-107-05713-5 (анг.)
ISBN 978-5-97060-673-5 (рус.)

© 2014 Shai Shalev-Shwartz and Shai Ben-David
© Издание, перевод, оформление, ДМК Пресс, 2019

СОДЕРЖАНИЕ

Предисловие	15
Благодарности.....	16
Глава 1. Введение	17
1.1. Что такое обучение?.....	17
1.2. Когда необходимо машинное обучение?	19
1.3. Типы обучения	20
1.4. Связи с другими дисциплинами	22
1.5. Как читать эту книгу	23
1.5.1. Варианты построения курса на основе книги	24
1.6. Обозначения.....	25
ЧАСТЬ I. ОСНОВАНИЯ	28
Глава 2. Малый вперед	29
2.1. Формальная модель – схема статистического обучения.....	29
2.2. Минимизация эмпирического риска	31
2.2.1. Не все коту масленица – переобучение	31
2.3. Минимизация эмпирического риска с индуктивным смещением	32
2.3.1. Конечные классы гипотез.....	33
2.4. Упражнения	37
Глава 3. Формальная модель обучения	39
3.1. Вероятно почти корректное обучение	39
3.2. Более общая модель обучения	40
3.2.1. Отказ от предположения о реализуемости – агностическое РАС-обучение	41
3.2.2. Круг моделируемых проблем обучения	43
3.3. Резюме	45
3.4. Библиографические сведения.....	46
3.5. Упражнения	46
Глава 4. Обучаемость и равномерная сходимость	50
4.1. Равномерная сходимость – достаточное условие обучаемости	50
4.2. Конечные классы допускают агностическое РАС-обучение	51
4.3. Резюме	54
4.4. Библиографические сведения.....	54
4.5. Упражнения	54

Глава 5. Компромисс между смещением и сложностью	56
5.1. Теорема об отсутствии бесплатных завтраков	57
5.1.1. Теорема о бесплатных завтраках и априорное знание	59
5.2. Разложение ошибки	60
5.3. Резюме	61
5.4. Библиографические сведения	62
5.5. Упражнения	62
Глава 6. VC-размерность	63
6.1. Бесконечные классы могут быть обучаемыми	63
6.2. VC-размерность	64
6.3. Примеры	66
6.3.1. Ступенчатые функции	66
6.3.2. Интервалы	67
6.3.3. Осепараллельные прямоугольники	67
6.3.4. Конечные классы	68
6.3.5. VC-размерность и количество параметров	68
6.4. Фундаментальная теорема PAC-обучения	68
6.5. Доказательство теоремы 6.7	69
6.5.1. Лемма Зауэра и функция роста	70
6.5.2. Равномерная сходимость для классов небольшого эффективного размера	71
6.6. Резюме	74
6.7. Библиографические сведения	74
6.8. Упражнения	75
Глава 7. Неравномерная обучаемость	79
7.1. Неравномерная обучаемость	79
7.1.1. Характеристика неравномерной обучаемости	80
7.2. Структурная минимизация риска	81
7.3. Минимальная длина описания и бритва Оккама	85
7.3.1. Бритва Оккама	87
7.4. Другие концепции обучаемости – согласованность	88
7.5. Обсуждение различных понятий обучаемости	89
7.5.1. Еще раз о теореме об отсутствии бесплатных завтраков	92
7.6. Резюме	92
7.7. Библиографические сведения	93
7.8. Упражнения	93
Глава 8. Время обучения	96
8.1. Вычислительная сложность обучения	97
8.1.1. Формальное определение*	98
8.2. Реализация правила ERM	99
8.2.1. Конечные классы	100
8.2.2. Осепараллельные прямоугольники	100
8.2.3. Булевы конъюнкции	102
8.2.4. Обучение трехчленных ДНФ	103

8.3. Эффективно обучаемый, но не собственный алгоритм ERM	103
8.4. Трудность обучения*	104
8.5. Резюме	106
8.6. Библиографические сведения	106
8.7. Упражнения	106

ЧАСТЬ II. ОТ ТЕОРИИ К АЛГОРИТМАМ

Глава 9. Линейные предикторы

9.1. Полупространства	112
9.1.1. Линейное программирование для класса полупространств	113
9.1.2. Перцептрон для полупространств	114
9.1.3. VC-размерность класса полупространств	116
9.2. Линейная регрессия	117
9.2.1. Метод наименьших квадратов	118
9.2.2. Линейная регрессия для задач полиномиальной регрессии	119
9.3. Логистическая регрессия	120
9.4. Резюме	121
9.5. Библиографические сведения	122
9.6. Упражнения	122

Глава 10. Усиление

10.1. Слабая обучаемость	125
10.1.1. Эффективная реализация ERM для класса решающих пней	127
10.2. Алгоритм AdaBoost	128
10.3. Линейные комбинации базовых гипотез	131
10.3.1. VC-размерность $L(B, T)$	133
10.4. Применение AdaBoost для распознавания лиц	134
10.5. Резюме	135
10.6. Библиографические сведения	136
10.7. Упражнения	136

Глава 11. Выбор и контроль модели

11.1. Выбор модели с помощью SRM	139
11.2. Контроль	140
11.2.1. Зарезервированный набор	140
11.2.2. Контроль при выборе модели	141
11.2.3. Кривая выбора модели	142
11.2.4. k-групповая перекрестная проверка	143
11.2.5. Обучение–контроль–тестирование	144
11.3. Что делать, если обучить не удастся	144
11.4. Резюме	147
11.5. Упражнения	148

Глава 12. Выпуклые проблемы обучения

12.1. Выпуклость, липшицевость и гладкость	149
12.1.1. Выпуклость	149
12.1.2. Липшицевость	153

12.1.3. Гладкость.....	154
12.2. Выпуклые проблемы обучения.....	156
12.2.1. Обучаемость выпуклых проблем обучения.....	157
12.2.2. Выпуклые-липшицевы/гладкие-ограниченные проблемы обучения.....	158
12.3. Суррогатные функции потерь.....	159
12.4. Резюме	161
12.5. Библиографические сведения.....	161
12.6. Упражнения	161

Глава 13. Регуляризация и устойчивость

13.1. Минимизация регуляризированной потери.....	163
13.1.1. Гребневая регрессия.....	164
13.2. Устойчивые правила не подвержены переобучению	165
13.3. Регуляризация Тихонова как стабилизатор	166
13.3.1. Липшицева потеря	168
13.3.2. Гладкая неотрицательная потеря.....	169
13.4. Управление компромиссом между аппроксимацией и устойчивостью	170
13.5. Резюме	172
13.6. Библиографические сведения.....	172
13.7. Упражнения	173

Глава 14. Стохастический градиентный спуск

14.1. Градиентный спуск	177
14.1.1. Анализ метода ГС для выпуклых липшицевых функций.....	178
14.2. Субградиенты	180
14.2.1. Вычисление субградиентов.....	181
14.2.2. Субградиенты липшицевых функций	182
14.2.3. Субградиентный спуск.....	182
14.3. Стохастический градиентный спуск (СГС).....	183
14.3.1. Анализ СГС для выпуклых-липшицевых-ограниченных функций ...	183
14.4. Варианты	185
14.4.1. Добавление шага проецирования.....	185
14.4.2. Переменный размер шага	186
14.4.3. Другие способы усреднения	186
14.4.4. Строго выпуклые функции*	187
14.5. Обучение с помощью СГС.....	188
14.5.1. Применение СГС для минимизации риска	188
14.5.2. Анализ СГС для выпуклых-гладких проблем обучения.....	190
14.5.3. Применение СГС для минимизации регуляризированной потери.....	191
14.6. Резюме	192
14.7. Библиографические сведения	192
14.8. Упражнения	192

Глава 15. Метод опорных векторов

15.1. Зазор и SVM с жестким зазором.....	194
15.1.1. Однородный случай	197

15.1.2. Выборочная сложность правила Hard-SVM.....	197
15.2. SVM с мягким зазором и регуляризация по норме	198
15.2.1. Выборочная сложность Soft-SVM.....	200
15.2.2. Сравнение границ, основанных на зазоре и норме, с размерностью.....	201
15.2.3. Рамповая функция потерь*	201
15.3. Условия оптимальности и «опорные векторы»*	202
15.4. Двойственность*	203
15.5. Реализация Soft-SVM с помощью СГС	204
15.6. Резюме	205
15.7. Библиографические сведения	205
15.8. Упражнения	206
Глава 16. Ядерные методы	207
16.1. Погружение в пространство признаков	207
16.2. Ядерный трюк.....	209
16.2.1. Ядра как способ выразить априорное знание.....	213
16.2.2. Характеристика ядерных функций*	214
16.3. Реализация Soft-SVM с ядрами	215
16.4. Резюме	216
16.5. Библиографические сведения.....	217
16.6. Упражнения	217
Глава 17. Многоклассовая категоризация, ранжирование и сложные проблемы предсказания.....	219
17.1. Один против всех и все пары.....	220
17.2. Линейные многоклассовые предикторы	222
17.2.1. Как построить Ψ	222
17.2.2. Стоимостная классификация.....	224
17.2.3. ERM.....	224
17.2.4. Обобщенная кусочно-линейная потеря.....	225
17.2.5. SVM и СГС в многоклассовом случае	226
17.3. Предсказание структурированного выхода.....	228
17.4. Ранжирование.....	230
17.4.1. Линейные предикторы для ранжирования	232
17.5. Двудольное ранжирование и многомерные показатели качества	235
17.5.1. Линейные предикторы для двудольного ранжирования	237
17.6. Резюме.....	239
17.7. Библиографические сведения.....	239
17.8. Упражнения	240
Глава 18. Решающие деревья	242
18.1. Выборочная сложность	243
18.2. Алгоритмы на решающих деревьях.....	244
18.2.1. Реализации меры выигрыша.....	245
18.2.2. Редукция	246
18.2.3. Пороговые правила разбиения для вещественных признаков	247

18.3. Случайные леса	247
18.4. Резюме	248
18.5. Библиографические сведения	248
18.6. Упражнения	248

Глава 19. Ближайшие соседи

19.1. Метод k ближайших соседей	250
19.2. Анализ	251
19.2.1. Граница обобщаемости для правила 1-NN.....	252
19.2.2. Проклятие размерности	255
19.3. Эффективная реализация*	256
19.4. Резюме	256
19.5. Библиографические сведения	257
19.6. Упражнения	257

Глава 20. Нейронные сети

20.1. Нейронные сети прямого распространения	261
20.2. Обучение нейронных сетей.....	262
20.3. Выразительная способность нейронных сетей.....	263
20.3.1. Геометрическая интерпретация	265
20.4. Выборочная сложность нейронных сетей	266
20.5. Время обучения нейронных сетей.....	267
20.6. СГС и обратное распространение	268
20.7. Резюме.....	272
20.8. Библиографические сведения.....	272
20.9. Упражнения	273

ЧАСТЬ III. ДОПОЛНИТЕЛЬНЫЕ МОДЕЛИ ОБУЧЕНИЯ.....

Глава 21. Онлайновое обучение

21.1. Онлайновая классификация в реализуемом случае	277
21.1.1. Онлайновая обучаемость.....	279
21.2. Онлайновая классификация в нереализуемом случае	283
21.2.1. Алгоритм взвешенного большинства	284
21.3. Онлайновая выпуклая оптимизация.....	288
21.4. Алгоритм онлайнового перцептрона	290
21.5. Резюме	293
21.6. Библиографические сведения	293
21.7. Упражнения	294

Глава 22. Кластеризация

22.1. Алгоритмы кластеризации на основе связи	299
22.2. Метод k -средних и другие методы кластеризации на основе минимизации стоимости.....	300
22.2.1. Алгоритм k -средних.....	302
22.3. Спектральная кластеризация.....	303
22.3.1. Разрезание графа	304
22.3.2. Лапласиан графа и ослабленные разрезы графа	304

22.3.3. Ненормированная спектральная кластеризация	305
22.4. Метод информационного горлышка*	306
22.5. Общий взгляд на кластеризацию	307
22.6. Резюме	309
22.7. Библиографические сведения	309
22.8. Упражнения	309
Глава 23. Понижение размерности	312
23.1. Метод главных компонент (РСА)	313
23.1.1. Более эффективное решение для случая $d \gg m$	315
23.1.2. Реализация и демонстрация	315
23.2. Случайные проекции	317
23.3. Сжатое измерение сигнала	319
23.3.1. Доказательства*	321
23.4. РСА или сжатое измерение сигнала?	326
23.5. Резюме	327
23.6. Библиографические сведения	327
23.7. Упражнения	328
Глава 24. Порождающие модели	330
24.1. Оценка максимального правдоподобия	331
24.1.1. Оценка максимального правдоподобия для непрерывных случайных величин	332
24.1.2. Максимальное правдоподобие и минимизация эмпирического риска	333
24.1.3. Анализ обобщаемости	333
24.2. Наивная байесовская классификация	335
24.3. Линейный дискриминантный анализ	335
24.4. Скрытые переменные и ЕМ-алгоритм	336
24.4.1. ЕМ как алгоритм поочередной максимизации	338
24.4.2. ЕМ-алгоритм для смеси нормальных распределений (мягкий алгоритм k-средних)	340
24.5. Байесовское рассуждение	341
24.6. Резюме	343
24.7. Библиографические сведения	343
24.8. Упражнения	343
Глава 25. Отбор и порождение признаков	345
25.1. Отбор признаков	346
25.1.1. Фильтры	347
25.1.2. Подходы на основе жадного отбора	348
25.1.3. Нормы, индуцирующие разреженность	351
25.2. Манипулирование и нормировка признаков	353
25.2.1. Примеры преобразований признаков	355
25.3. Обучение признаков	356
25.3.1. Обучение словаря с помощью автокодировщиков	356
25.4. Резюме	358

25.5. Библиографические сведения	358
25.6. Упражнения	359

ЧАСТЬ IV. ДОПОЛНИТЕЛЬНЫЕ ГЛАВЫ

Глава 26. Радемахеровская сложность

26.1. Радемахеровская сложность	362
26.1.1. Исчисление Радемахера	367
26.2. Радемахеровская сложность линейных классов	369
26.3. Границы обобщаемости метода SVM	371
26.4. Границы обобщаемости для предикторов с малой нормой ℓ_1	373
26.5. Библиографические сведения	374

Глава 27. Числа покрытия

27.1. Покрытие	375
27.1.1. Свойства	375
27.2. От покрытия к радемахеровской сложности через сцепление	376
27.3. Библиографические сведения	378

Глава 28. Доказательство фундаментальной теоремы теории обучения

28.1. Верхняя граница для агностического случая	379
28.2. Нижняя граница для агностического случая	380
28.2.1. Доказательство того, что $m(\epsilon, \delta) \geq 0,5 \log(1/(4\delta))/\epsilon^2$	381
28.2.2. Доказательство того, что $m(\epsilon, 1/8) \geq 8d/\epsilon^2$	382
28.3. Верхняя граница для реализуемого случая	385
28.3.1. От ϵ -сетей к PAC-обучаемости	388

Глава 29. Многоклассовая обучаемость

29.1. Размерность Натараджана	389
29.2. Фундаментальная многоклассовая теорема	390
29.2.1. О доказательстве теоремы 29.3	390
29.3. Вычисление размерности Натараджана	391
29.3.1. Метод «один против всех»	391
29.3.2. Сведение многоклассовой категоризации к бинарной классификации в общем случае	392
29.3.3. Линейные многоклассовые предикторы	392
29.4. О хороших и плохих правилах ERM	394
29.5. Библиографические сведения	395
29.6. Упражнения	396

Глава 30. Границы сжатия

30.1. Границы сжатия	397
30.2. Примеры	399
30.2.1. Осепараллельные прямоугольники	399
30.2.2. Полупространства	399
30.2.3. Разделение полиномов	401

30.2.4. Разделение с зазором	401
30.3. Библиографические сведения.....	401
Глава 31. РАС-байесовский подход.....	402
31.1. РАС-байесовские границы.....	402
31.2. Библиографические сведения.....	404
31.3. Упражнения	405
Приложение А. Технические леммы	406
Приложение В. Концентрация меры.....	409
В.1. Неравенство Маркова.....	409
В.2. Неравенство Чебышева	410
В.3. Границы Чернова	411
В.4. Неравенство Хёфдинга	412
В.5. Неравенства Беннета и Бернштейна	413
В.5.1. Применение	414
В.6. Неравенство Слада.....	415
В.7. Концентрация случайных величин χ^2	415
Приложение С. Линейная алгебра	418
С.1. Основные определения.....	418
С.2. Собственные значения и собственные векторы	419
С.3. Положительно определенные матрицы.....	419
С.4. Сингулярное разложение	419
Литература	423
Предметный указатель	432