

УДК 004.85  
ББК 32.971.3  
Г93

Гудфеллоу Я., Бенджио И., Курвилль А.  
Г93 Глубокое обучение / пер. с англ. А. А. Слинкина. – 2-е изд., испр. – М.: ДМК Пресс, 2018. – 652 с.: цв. ил.

**ISBN 978-5-97060-618-6**

Глубокое обучение — это вид машинного обучения, наделяющий компьютеры способностью учиться на опыте и понимать мир в терминах иерархии концепций. Книга содержит математические и концептуальные основы линейной алгебры, теории вероятностей и теории информации, численных расчетов и машинного обучения в том объеме, который необходим для понимания материала. Описываются приемы глубокого обучения, применяемые на практике, в том числе глубокие сети прямого распространения, регуляризация, алгоритмы оптимизации, сверточные сети, моделирование последовательностей и др. Рассматриваются такие приложения, как обработка естественных языков, распознавание речи, компьютерное зрение, онлайн-овые рекомендательные системы, биоинформатика и видеоигры.

Издание предназначено студентам вузов и аспирантам, а также опытным программистам, которые хотели бы применить глубокое обучение в составе своих продуктов или платформ.

УДК 004.85  
ББК 32.971.3

Права на издание книги на русском языке предоставлены агентством Александра Корженевского.

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.

Материал, изложенный в данной книге, многократно проверен. Но поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги.

ISBN 978-1-491-93799-0 (анг.)  
ISBN 978-5-97060-618-6 (рус.)

© 2017 Massachusetts Institute of Technology  
© Оформление, издание, перевод, ДМК Пресс, 2018

# Содержание

<b>Веб-сайт .....</b>	<b>14</b>
<b>Благодарности .....</b>	<b>15</b>
<b>Обозначения .....</b>	<b>18</b>
<b>Глава 1. Введение.....</b>	<b>21</b>
1.1. На кого ориентирована эта книга .....	29
1.2. Исторические тенденции в машинном обучении .....	29
1.2.1. Нейронные сети: разные названия и переменчивая фортуна .....	30
1.2.2. Увеличение размера набора данных .....	36
1.2.3. Увеличение размера моделей .....	36
1.2.4. Повышение точности и сложности и расширение круга задач .....	40
<b>Часть I. Основы прикладной математики и машинного обучения .....</b>	<b>43</b>
<b>Глава 2. Линейная алгебра .....</b>	<b>44</b>
2.1. Скаляры, векторы, матрицы и тензоры .....	44
2.2. Умножение матриц и векторов .....	46
2.3. Единичная и обратная матрица .....	47
2.4. Линейная зависимость и линейная оболочка .....	48
2.5. Нормы .....	50
2.6. Специальные виды матриц и векторов .....	51
2.7. Спектральное разложение матрицы .....	52
2.8. Сингулярное разложение .....	54
2.9. Псевдообратная матрица Мура–Пенроуза .....	55
2.10. Оператор следа .....	56
2.11. Определитель .....	56
2.12. Пример: метод главных компонент .....	57
<b>Глава 3. Теория вероятностей и теория информации .....</b>	<b>61</b>
3.1. Зачем нужна вероятность? .....	61
3.2. Случайные величины .....	63
3.3. Распределения вероятности .....	63
3.3.1. Дискретные случайные величины и функции вероятности .....	64
3.3.2. Непрерывные случайные величины и функции плотности вероятности .....	64
3.4. Маргинальное распределение вероятности .....	65
3.5. Условная вероятность .....	65
3.6. Цепное правило .....	66
3.7. Независимость и условная независимость .....	66
3.8. Математическое ожидание, дисперсия и ковариация .....	66
3.9. Часто встречающиеся распределения вероятности .....	68
3.9.1. Распределение Бернулли .....	68

3.9.2. Категориальное распределение.....	68
3.9.3. Нормальное распределение.....	69
3.9.4. Экспоненциальное распределение и распределение Лапласа .....	70
3.9.5. Распределение Дирака и эмпирическое распределение.....	71
3.9.6. Смеси распределений .....	71
3.10. Полезные свойства употребительных функций .....	73
3.11. Правило Байеса .....	74
3.12. Технические детали непрерывных величин .....	75
3.13. Теория информации.....	76
3.14. Структурные вероятностные модели.....	78
<b>Глава 4. Численные методы .....</b>	<b>82</b>
4.1. Переполнение и потеря значимости.....	82
4.2. Плохая обусловленность .....	83
4.3. Оптимизация градиентным методом .....	84
4.3.1. Не только градиент: матрицы Якоби и Гессе.....	86
4.4. Оптимизация с ограничениями .....	92
4.5. Пример: линейный метод наименьших квадратов .....	94
<b>Глава 5. Основы машинного обучения.....</b>	<b>96</b>
5.1. Алгоритмы обучения.....	97
5.1.1. Задача Т .....	97
5.1.2. Мера качества Р.....	100
5.1.3. Опыт Е .....	101
5.1.4. Пример: линейная регрессия .....	103
5.2. Емкость, переобучение и недообучение .....	105
5.2.1. Теорема об отсутствии бесплатных завтраков.....	110
5.2.2. Регуляризация .....	112
5.3. Гиперпараметры и контрольные наборы.....	114
5.3.1. Перекрестная проверка.....	115
5.4. Оценки, смещение и дисперсия.....	115
5.4.1. Точечное оценивание .....	116
5.4.2. Смещение.....	117
5.4.3. Дисперсия и стандартная ошибка.....	119
5.4.4. Поиск компромисса между смещением и дисперсией для минимизации среднеквадратической ошибки.....	121
5.4.5. Состоятельность.....	122
5.5. Оценка максимального правдоподобия .....	122
5.5.1. Условное логарифмическое правдоподобие и среднеквадратическая ошибка.....	123
5.5.2. Свойства максимального правдоподобия .....	125
5.6. Байесовская статистика .....	125
5.6.1. Оценка апостериорного максимума (MAP).....	128
5.7. Алгоритмы обучения с учителем.....	129
5.7.1. Вероятностное обучение с учителем.....	129

5.7.2. Метод опорных векторов.....	130
5.7.3. Другие простые алгоритмы обучения с учителем .....	132
5.8. Алгоритмы обучения без учителя .....	134
5.8.1. Метод главных компонент.....	135
5.8.2. Кластеризация методом $k$ средних .....	137
5.9. Стохастический градиентный спуск .....	138
5.10. Построение алгоритма машинного обучения .....	140
5.11. Проблемы, требующие глубокого обучения.....	141
5.11.1. Проклятие размерности .....	141
5.11.2. Регуляризация для достижения локального постоянства и гладкости.....	142
5.11.3. Обучение многообразий.....	145
<b>Часть II. Глубокие сети: современные подходы.....</b>	<b>149</b>
<b>Глава 6. Глубокие сети прямого распространения .....</b>	<b>150</b>
6.1. Пример: обучение XOR .....	152
6.2. Обучение градиентными методами .....	157
6.2.1. Функции стоимости.....	158
6.2.2. Выходные блоки .....	160
6.3. Скрытые блоки.....	169
6.3.1. Блоки линейной ректификации и их обобщения.....	170
6.3.2. Логистическая сигмоида и гиперболический тангенс .....	171
6.3.3. Другие скрытые блоки .....	172
6.4. Проектирование архитектуры .....	173
6.4.1. Свойства универсальной аппроксимации и глубина .....	174
6.4.2. Другие архитектурные подходы .....	177
6.5. Обратное распространение и другие алгоритмы дифференцирования.....	179
6.5.1. Графы вычислений .....	179
6.5.2. Правило дифференцирования сложной функции .....	181
6.5.3. Рекурсивное применение правила дифференцирования сложной функции для получения алгоритма обратного распространения.....	182
6.5.4. Вычисление обратного распространения в полносвязном МСП .....	185
6.5.5. Символьно-символьные производные .....	186
6.5.6. Общий алгоритм обратного распространения.....	188
6.5.7. Пример: применение обратного распространения к обучению МСП.....	191
6.5.8. Осложнения .....	192
6.5.9. Дифференцирование за пределами сообщества глубокого обучения .....	193
6.5.10. Производные высшего порядка .....	195
6.6. Исторические замечания .....	196
<b>Глава 7. Регуляризация в глубоком обучении .....</b>	<b>199</b>
7.1. Штрафы по норме параметров .....	200
7.1.1. Регуляризация параметров по норме $L^2$ .....	201
7.1.2. $L^1$ -регуляризация.....	204
7.2. Штраф по норме как оптимизация с ограничениями .....	206

7.3. Регуляризация и недоопределенные задачи.....	208
7.4. Пополнение набора данных.....	208
7.5. Робастность относительно шума.....	210
7.5.1. Привнесение шума в выходные метки.....	211
7.6. Обучение с частичным привлечением учителя.....	211
7.7. Многозадачное обучение.....	212
7.8. Ранняя остановка.....	213
7.9. Связывание и разделение параметров.....	219
7.9.1. Сверточные нейронные сети.....	220
7.10. Разреженные представления.....	220
7.11. Баггинг и другие ансамблевые методы.....	222
7.12. Прореживание.....	224
7.13. Состязательное обучение.....	232
7.14. Тангенциальное расстояние, алгоритм распространения по касательной и классификатор по касательной к многообразию.....	233
<b>Глава 8. Оптимизация в обучении глубоких моделей.....</b>	<b>237</b>
8.1. Чем обучение отличается от чистой оптимизации.....	237
8.1.1. Минимизация эмпирического риска.....	238
8.1.2. Суррогатные функции потерь и ранняя остановка.....	239
8.1.3. Пакетные и мини-пакетные алгоритмы.....	239
8.2. Проблемы оптимизации нейронных сетей.....	243
8.2.1. Плохая обусловленность.....	243
8.2.2. Локальные минимумы.....	245
8.2.3. Плато, седловые точки и другие плоские участки.....	246
8.2.4. Утесы и резко растущие градиенты.....	248
8.2.5. Долгосрочные зависимости.....	249
8.2.6. Неточные градиенты.....	250
8.2.7. Плохое соответствие между локальной и глобальной структурами.....	250
8.2.8. Теоретические пределы оптимизации.....	252
8.3. Основные алгоритмы.....	253
8.3.1. Стохастический градиентный спуск.....	253
8.3.2. Импульсный метод.....	255
8.3.3. Метод Нестерова.....	258
8.4. Стратегии инициализации параметров.....	258
8.5. Алгоритмы с адаптивной скоростью обучения.....	263
8.5.1. AdaGrad.....	264
8.5.2. RMSProp.....	264
8.5.3. Adam.....	265
8.5.4. Выбор правильного алгоритма оптимизации.....	266
8.6. Приближенные методы второго порядка.....	267
8.6.1. Метод Ньютона.....	267
8.6.2. Метод сопряженных градиентов.....	268
8.6.3. Алгоритм BFGS.....	271
8.7. Стратегии оптимизации и метаалгоритмы.....	272

8.7.1. Пакетная нормировка.....	272
8.7.2. Покоординатный спуск.....	275
8.7.3. Усреднение Поляка.....	276
8.7.4. Предобучение с учителем.....	276
8.7.5. Проектирование моделей с учетом простоты оптимизации.....	279
8.7.6. Методы продолжения и обучение по плану.....	279
<b>Глава 9. Сверточные сети.....</b>	<b>282</b>
9.1. Операция свертки.....	282
9.2. Мотивация.....	284
9.3. Пулинг.....	290
9.4. Свертка и пулинг как бесконечно сильное априорное распределение.....	293
9.5. Варианты базовой функции свертки.....	295
9.6. Структурированный выход.....	304
9.7. Типы данных.....	305
9.8. Эффективные алгоритмы свертки.....	306
9.9. Случайные признаки и признаки, обученные без учителя.....	307
9.10. Нейробиологические основания сверточных сетей.....	308
9.11. Сверточные сети и история глубокого обучения.....	314
<b>Глава 10. Моделирование последовательностей: рекуррентные и рекурсивные сети.....</b>	<b>316</b>
10.1. Развертка графа вычислений.....	317
10.2. Рекуррентные нейронные сети.....	320
10.2.1. Форсирование учителя и сети с рекурсией на выходе.....	323
10.2.2. Вычисление градиента в рекуррентной нейронной сети.....	325
10.2.3. Рекуррентные сети как ориентированные графические модели.....	327
10.2.4. Моделирование контекстно-обусловленных последовательностей с помощью РНС.....	330
10.3. Двухнаправленные РНС.....	332
10.4. Архитектуры кодировщик-декодер или последовательность в последовательность.....	333
10.5. Глубокие рекуррентные сети.....	336
10.6. Рекурсивные нейронные сети.....	337
10.7. Проблема долгосрочных зависимостей.....	339
10.8. Нейронные эхо-сети.....	341
10.9. Блоки с утечками и другие стратегии нескольких временных масштабов.....	343
10.9.1. Добавление прямых связей сквозь время.....	343
10.9.2. Блоки с утечкой и спектр разных временных масштабов.....	343
10.9.3. Удаление связей.....	344
10.10. Долгая краткосрочная память и другие вентильные РНС.....	344
10.10.1. Долгая краткосрочная память.....	345
10.10.2. Другие вентильные РНС.....	347
10.11. Оптимизация в контексте долгосрочных зависимостей.....	348
10.11.1. Отсечение градиентов.....	348

10.11.2. Регуляризация с целью подталкивания информационного потока .....	350
10.12. Явная память .....	351
<b>Глава 11. Практическая методология .....</b>	<b>355</b>
11.1. Показатели качества .....	356
11.2. Выбор базовой модели по умолчанию .....	358
11.3. Надо ли собирать дополнительные данные? .....	359
11.4. Выбор гиперпараметров .....	360
11.4.1. Ручная настройка гиперпараметров .....	360
11.4.2. Алгоритмы автоматической оптимизации гиперпараметров .....	363
11.4.3. Поиск на сетке .....	364
11.4.4. Случайный поиск .....	365
11.4.5. Оптимизация гиперпараметров на основе модели .....	366
11.5. Стратегии отладки .....	367
11.6. Пример: распознавание нескольких цифр .....	370
<b>Глава 12. Приложения .....</b>	<b>373</b>
12.1. Крупномасштабное глубокое обучение .....	373
12.1.1. Реализации на быстрых CPU .....	373
12.1.2. Реализации на GPU .....	374
12.1.3. Крупномасштабные распределенные реализации .....	376
12.1.4. Сжатие модели .....	376
12.1.5. Динамическая структура .....	377
12.1.6. Специализированные аппаратные реализации глубоких сетей .....	379
12.2. Компьютерное зрение .....	380
12.2.1. Предобработка .....	381
12.3. Распознавание речи .....	385
12.4. Обработка естественных языков .....	388
12.4.1. $n$ -граммы .....	388
12.4.2. Нейронные языковые модели .....	390
12.4.3. Многомерные выходы .....	391
12.4.4. Комбинирование нейронных языковых моделей с $n$ -граммами .....	397
12.4.5. Нейронный машинный перевод .....	397
12.4.6. Историческая справка .....	401
12.5. Другие приложения .....	402
12.5.1. Рекомендательные системы .....	402
12.5.2. Представление знаний, рассуждения и ответы на вопросы .....	405
<b>Часть III. Исследования по глубокому обучению .....</b>	<b>409</b>
<b>Глава 13. Линейные факторные модели .....</b>	<b>411</b>
13.1. Вероятностный РСА и факторный анализ .....	412
13.2. Анализ независимых компонент (ICA) .....	413
13.3. Анализ медленных признаков .....	415
13.4. Разреженное кодирование .....	417
13.5. Интерпретация РСА в терминах многообразий .....	419

<b>Глава 14. Автокодировщики</b>	<b>422</b>
14.1. Понижающие автокодировщики	423
14.2. Регуляризированные автокодировщики	423
14.2.1. Разреженные автокодировщики	424
14.2.2. Шумоподавляющие автокодировщики	426
14.2.3. Регуляризация посредством штрафования производных	427
14.3. Репрезентативная способность, размер слоя и глубина	427
14.4. Стохастические кодировщики и декодеры	428
14.5. Шумоподавляющие автокодировщики	429
14.5.1. Сопоставление рейтингов	430
14.6. Обучение многообразий с помощью автокодировщиков	433
14.7. Сжимающие автокодировщики	436
14.8. Предсказательная разреженная декомпозиция	440
14.9. Применения автокодировщиков	441
<b>Глава 15. Обучение представлений</b>	<b>443</b>
15.1. Жадное послойное предобучение без учителя	444
15.1.1. Когда и почему работает предобучение без учителя?	446
15.2. Перенос обучения и адаптация домена	451
15.3. Разделение каузальных факторов с частичным привлечением учителя	454
15.4. Распределенное представление	459
15.5. Экспоненциальный выигрыш от глубины	465
15.6. Ключ к выявлению истинных причин	466
<b>Глава 16. Структурные вероятностные модели в глубоком обучении</b>	<b>469</b>
16.1. Проблема бесструктурного моделирования	470
16.2. Применение графов для описания структуры модели	473
16.2.1. Ориентированные модели	473
16.2.2. Неориентированные модели	475
16.2.3. Статистическая сумма	477
16.2.4. Энергетические модели	478
16.2.5. Разделенность и d-разделенность	480
16.2.6. Преобразование между ориентированными и неориентированными графами	481
16.2.7. Факторные графы	486
16.3. Выборка из графических моделей	487
16.4. Преимущества структурного моделирования	488
16.5. Обучение и зависимости	489
16.6. Вывод и приближенный вывод	490
16.7. Подход глубокого обучения к структурным вероятностным моделям	491
16.7.1. Пример: ограниченная машина Больцмана	492
<b>Глава 17. Методы Монте-Карло</b>	<b>495</b>
17.1. Выборка и методы Монте-Карло	495
17.1.1. Зачем нужна выборка?	495



17.1.2. Основы выборки методом Монте-Карло .....	495
17.2. Выборка по значимости .....	497
17.3. Методы Монте-Карло по схеме марковской цепи .....	499
17.4. Выборка по Гиббсу.....	502
17.5. Проблема перемешивания разделенных мод .....	503
17.5.1. Применение темперирования для перемешивания мод .....	506
17.5.2. Глубина может помочь перемешиванию .....	506

## **Глава 18. Преодоление трудностей, связанных**

<b>со статической суммой.....</b>	<b>508</b>
18.1. Градиент логарифмического правдоподобия .....	508
18.2. Стохастическая максимизация правдоподобия и сопоставительное расхождение .....	510
18.3. Псевдоправдоподобие .....	517
18.4. Сопоставление рейтингов и сопоставление отношений .....	519
18.5. Шумоподавляющее сопоставление рейтингов.....	521
18.6. Шумосопоставительное оценивание .....	521
18.7. Оценивание статистической суммы.....	524
18.7.1. Выборка по значимости с отжигом .....	525
18.7.2. Мостиковая выборка .....	528

## **Глава 19. Приближенный вывод.....**

19.1. Вывод как оптимизация.....	530
19.2. ЕМ-алгоритм .....	532
19.3. МАР-вывод и разреженное кодирование .....	533
19.4. Вариационный вывод и обучение .....	535
19.4.1. Дискретные латентные переменные .....	536
19.4.2. Вариационное исчисление.....	541
19.4.3. Непрерывные латентные переменные.....	544
19.4.4. Взаимодействия между обучением и выводом .....	545
19.5. Обученный приближенный вывод .....	546
19.5.1. Бодрствование-сон.....	546
19.5.2. Другие формы обученного вывода .....	547

## **Глава 20. Глубокие порождающие модели.....**

20.1. Машины Больцмана.....	548
20.2. Ограниченные машины Больцмана.....	550
20.2.1. Условные распределения .....	550
20.2.2. Обучение ограниченных машин Больцмана.....	552
20.3. Глубокие сети доверия .....	553
20.4. Глубокие машины Больцмана .....	555
20.4.1. Интересные свойства.....	557
20.4.2. Вывод среднего поля в ГМБ .....	558
20.4.3. Обучение параметров ГМБ.....	560
20.4.4. Послойное предобучение .....	560

20.4.5. Совместное обучение глубоких машин Больцмана.....	563
20.5. Машины Больцмана для вещественных данных .....	566
20.5.1. ОМБ Гаусса–Бернулли .....	567
20.5.2. Неориентированные модели условной ковариации.....	568
20.6. Сверточные машины Больцмана.....	572
20.7. Машины Больцмана для структурных и последовательных выходов.....	573
20.8. Другие машины Больцмана.....	574
20.9. Обратное распространение через случайные операции .....	575
20.9.1. Обратное распространение через дискретные стохастические операции.....	577
20.10. Ориентированные порождающие сети .....	579
20.10.1. Сигмоидные сети доверия.....	580
20.10.2. Дифференцируемые генераторные сети .....	581
20.10.3. Вариационные автокодировщики .....	583
20.10.4. Порождающие состязательные сети .....	586
20.10.5. Порождающие сети с сопоставлением моментов.....	589
20.10.6. Сверточные порождающие сети.....	590
20.10.7. Авторегрессивные сети.....	591
20.10.8. Линейные авторегрессивные сети.....	591
20.10.9. Нейронные авторегрессивные сети.....	592
20.10.10. NADE.....	593
20.11. Выборка из автокодировщиков .....	595
20.11.1. Марковская цепь, ассоциированная с произвольным шумоподавляющим автокодировщиком .....	596
20.11.2. Фиксация и условная выборка.....	596
20.11.3. Возвратная процедура обучения.....	597
20.12. Порождающие стохастические сети.....	598
20.12.1. Дискриминантные GSN .....	599
20.13. Другие схемы порождения .....	599
20.14. Оценивание порождающих моделей .....	600
20.15. Заключение .....	603
<b>Список литературы .....</b>	<b>604</b>
<b>Предметный указатель .....</b>	<b>646</b>