

# Содержание

<b>Веб-сайт</b> .....	14
<b>Благодарности</b> .....	15
<b>Обозначения</b> .....	18
<b>Глава 1. Введение</b> .....	21
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> .....	43
<b>Глава 2. Линейная алгебра</b> .....	44
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> .....	61
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 вентиляльные РНС.....	344
10.10.1. Долгая краткосрочная память.....	345
10.10.2. Другие вентиляльные 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> .....	422
14.1. Понижающие автокодировщики .....	423
14.2. Регуляризованные 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> .....	443
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> .....	469
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> .....	495
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.....**

<b>19.1. Вывод как оптимизация.....</b>	<b>530</b>
19.2. EM-алгоритм .....	532
19.3. MAP-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.....**

<b>20.1. Машины Больцмана.....</b>	<b>548</b>
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>

# Веб-сайт

[www.deeplearning.book](http://www.deeplearning.book)

Книгу сопровождает указанный выше сайт, где представлены упражнения, слайды, исправления ошибок и другие материалы, полезные читателям и преподавателям.

# Благодарности

Эта книга не состоялась бы, если бы не помощь со стороны многих людей.

Мы благодарны тем, кто откликнулся на наше предложение о написании книги и помог спланировать ее содержание и структуру: Гийом Ален (Guillaume Alain), Кюнхюн Чо (Kyunghyun Cho), Чаглар Гюльчехре (Çaglar Gülçehre), Дэвид Крюгер (David Krueger), Гуго Ларошель (Hugo Larochelle), Разван Паскану (Razvan Pascanu) и Томас Рохе (Thomas Rohée).

Мы также благодарны всем, кто присылал отзывы о самой книге, иногда даже о нескольких главах: Мартен Абади (Martín Abadi), Гийом Ален, Йон Андруцопулос (Jon Androutsopoulos), Фред Берш (Fred Bertsch), Олекса Биланюк (Olexa Bilaniuk), Уфук Джан Бичиджи (Ufuk Can Biçici), Матко Босняк (Matko Bošnjak), Джон Буерсма (John Boersma), Грег Брокман (Greg Brockman), Александр де Бребиссон (Alexandre de Brébisson), Пьер Люк Каррье (Pierre Luc Carrier), Саратх Чандар (Sarath Chandar), Павел Чилински (Pawel Chilinski), Марк Дауст (Mark Daoust), Олег Дашевский, Лоран Дин (Laurent Dinh), Стефан Дрезейтль (Stephan Dreseitl), Джим Фан (Jim Fan), Мяо Фан (Miao Fan), Мейре Фортунато (Meire Fortunato), Фредерик Франсис (Frédéric Francis), Нандо де Фрейтас (Nando de Freitas), Чаглар Гюльчехре (Çaglar Gülçehre), Юрген Ван Гаэл (Jurgen Van Gael), Хавьер Алонсо Гарсиа (Javier Alonso García), Джонатан Хант (Jonathan Hunt), Гопи Джейярам (Gopi Jeeyaram), Чингиз Кабытаев (Chingiz Kabytayev), Лукаш Кайзер (Lukasz Kaiser), Варун Канаде (Varun Kanade), Азифулла Хан (Asifullah Khan), Акиель Хан (Akiel Khan), Джон Кинг (John King), Дидерик П. Кингма (Diederik P. Kingma), Ян Лекун (Yann LeCun), Рудольф Матей (Rudolf Mathey), Матиас Маттамала (Matías Mattamala), Абхинав Мауриа (Abhinav Maurya), Кэвин Мерфи (Kevin Murphy), Олег Мюрк (Oleg Mürk), Роман Новак (Roman Novak), Огастес К. Одена (Augustus Q. Odena), Симон Павлик (Simon Pavlik), Карл Пичотта (Carl Pichotta), Эдди Пирс (Eddie Pierce), Кари Пулли (Kari Pulli), Руссель Рахман (Roussel Rahman), Тапани Райко (Tapani Raiko), Анурга Ранджан (Anurag Ranjan), Йоханнес Ройт (Johannes Roith), Михаэла Роска (Mihaela Rosca), Халис Сак (Halis Sak), Сезар Салгадо (César Salgado), Григорий Сапунов, Ёсинори Сасаки (Yoshinori Sasaki), Майк Шустер (Mike Schuster), Джулиан Сербан (Julian Serban), Нир Шабат (Nir Shabat), Кен Ширрифф (Ken Shirriff), Андрэ Симпело (Andre Simpel), Дэвид Слейт (David Slate), Скотт Стэнли (Scott Stanley), Давид Суссилло (David Sussillo), Илья Суцкевер (Ilya Sutskever), Карлес Гелада Саец (Carles Gelada Sáez), Грэхэм Тейлор (Graham Taylor), Валентин Толмер (Valentin Tolmer), Массимильяно Томассоли (Massimiliano Tomassoli), Ан Тран (An Tran), Шубхенду Триведи (Shubhendu Trivedi), Алексей Умнов, Винсет Ванхоуке (Vincent Vanhoucke), Марко Висентини-Скарцанелла (Marco Visentini-Scanzanella), Матрин Вита (Martin Vita), Дэвид Уорд-Фарли (David Warde-Farley), Дастин Уэбб (Dustin Webb), Кэлвин Су (Kelvin Xu), Вэй Сюэ (Wei Xue), Ке Янг (Ke Yang), Ли Йо (Li Yao), Зигмунт Заяц (Zygmunt Zając) и Озан Чаглаян (Ozan Çaglayan).

Мы признательны и тем, кто делился своим мнением об отдельных главах.

- Обозначения: Чжан Юань Хан (Zhang Yuanhang).
- Глава 1 «Введение»: Юсуф Акгуль (Yusuf Akgul), Себастьян Братьерес (Sebastien Bratieres), Самира Эбрахими (Samira Ebrahimi), Чарли Горичаназ (Charlie

Gorichanaz), Брендан Лоудермилк (Brendan Loudermilk), Эрис Моррис (Eric Morris), Космин Пярвулеску (Cosmin Părvulescu) и Альфредо Солано (Alfredo Solano).

- Глава 2 «Линейная алгебра»: Амджад Алмахаири (Amjad Almahairi), Никола Баниц (Nikola Banic), Кэвин Беннетт (Kevin Bennett), Филипп Кастонге (Philippe Castonguay), Оскар Чанг (Oscar Chang), Эрик Фослер-Люссье (Eric Fosler-Lussier), Андрей Халявин, Сергей Орешков (Sergey Oreshkov), Иштван Петрас (István Petrás), Дэннис Прэнгл (Dennis Prangle), Томас Рохе (Thomas Rohée), Гитанджали Гулве Сехгал (Gitanjali Gulve Sehgal), Колби Толанд (Colby Toland), Алессандро Витале (Alessandro Vitale) и Боб Уэлланд (Bob Welland).
- Глава 3 «Теория вероятностей и теория информации»: Джон Филипп Андерсон (John Philip Anderson), Кай Арулкумаран (Kai Arulkumaran), Венсан Дюмуле (Vincent Dumoulin), Руй Фа (Rui Fa), Стефан Гоус (Stephan Gouws), Артем Оботуров, Антти Расмус (Antti Rasmus), Алексей Сурков и Фолькер Тресп (Volker Tresp).
- Глава 4 «Численные методы»: Тран Лам Аниан Фишер (Tran Lam AnIan Fischer) и Ху Ю Хуан (Hu Yuhuang).
- Глава 5 «Основы машинного обучения»: Дзмитри Бахданау (Dzmitry Bahdanau), Жюстен Доманге (Justin Domingue), Нихил Гарг (Nikhil Garg), Макото Оцука (Makoto Otsuka), Боб Пепин (Bob Pepin), Филип Попьен (Philip Popien), Бхарат Прабхакар (Bharat Prabhakar), Эммануэль Райнер (Emmanuel Rayner), Питер Шепард (Peter Shepard), Ки-Бонг Сонг (Kee-Bong Song), Чжен Сун (Zheng Sun) и Энди Ву (Andy Wu).
- Глава 6 «Глубокие сети прямого распространения»: Уриэль Бердуго (Uriel Berdugo), Фабрицио Боттарель (Fabrizio Bottarel), Элизабет Бэрл (Elizabeth Burl), Ишан Дуругкар (Ishan Durugkar), Джефф Хлыва (Jeff Hlywa), Ёнг Вук Ким (Jong Wook Kim), Давид Крюгер (David Krueger), Адития Кумар Прахарадж (Aditya Kumar Praharaj) и Стэн Сутла (Sten Sootla).
- Глава 7 «Регуляризация в глубоком обучении»: Мортен Колбэк (Morten Kolbæk), Читыдж Лауриа (Kshitij Lauria), Инкю Ли (Inkyu Lee), Сунил Мохан (Sunil Mohan), Хай Пхонг Пхан (Hai Phong Phan) и Джошуа Сэлисбэри (Joshua Salisbury).
- Глава 8 «Оптимизация обучения глубоких моделей»: Марсель Аккерман (Marcel Ackermann), Питер Армитейдж (Peter Armitage), Роуэл Атиенза (Rowel Atienza), Эндрю Брок (Andrew Brock), Теган Махарадж (Tegan Maharaj), Джеймс Мартенс (James Martens), Мостафа Натех (Mostafa Nategh), Кашиф Расул (Kashif Rasul), Клаус Штробль (Klaus Strobl) и Никола Тэрнер (Nicholas Turner).
- Глава 9 «Сверточные сети»: Мартен Аржовски (Martín Arjovsky), Евгений Бревдо (Eugene Brevdo), Константин Дивилов, Эрик Йенсен (Eric Jensen), Мехди Мирза (Mehdi Mirza), Алекс Пайно (Alex Raino), Марджори Сэйер (Marjorie Sayer), Райан Стаут (Ryan Stout) и Вентао Ву (Wentao Wu).
- Глава 10 «Моделирование последовательностей: рекуррентные и рекурсивные сети»: Гёкчен Ераслан (Gökçen Eraslan), Стивен Хиксон (Steven Hickson), Разван Паскану (Razvan Pascanu), Лоренхо фон Риттер (Lorenzo von Ritter), Руй Родригес (Rui Rodrigues), Дмитрий Сердюк, Донгуй Ши (Dongyu Shi) и Кай Ю Ян (Kaiyu Yang).

- Глава 11 «Практическая методология»: Даниэль Бекштейн (Daniel Beckstein).
- Глава 12 «Приложения»: Джордж Дал (George Dahl), Владимир Некрасов (Vladimir Nekrasov) и Рибана Рошер (Ribana Roscher).
- Глава 13 «Линейные факторные модели»: Джейнт Кушик (Jayanth Koushik).
- Глава 15 «Обучение представлений»: Кунал Гхош (Kunal Ghosh).
- Глава 16 «Структурные вероятностные модели в глубоком обучении»: Минь Ле (Minh Lê) и Антон Варфолом (Anton Varfolom).
- Глава 18 «Преодоление трудностей, связанных со статической суммой»: Сэм Боумен (Sam Bowman).
- Глава 19 «Приближенный вывод»: Ю Цзя Бао (Yujia Bao).
- Глава 20 «Глубокие порождающие модели»: Николас Чападос (Nicolas Charados), Даниэль Галвес (Daniel Galvez), Вэн Минь Ма (Wenming Ma), Фади Медхат (Fady Medhat), Шакри Мохамед (Shakir Mohamed) и Грегуар Монтавон (Grégoire Montavon).
- Библиография: Лукас Михельбахер (Lukas Michelbacher) и Лесли Н. Смит (Leslie N. Smith).

Мы также благодарим авторов, давших нам разрешение использовать в тексте изображения, рисунки и данные из их публикаций. Источники указываются в подрисуночных подписях.

Мы благодарны Лю Вану (Lu Wang), написавшему программу pdf2htmlEX, которой мы пользовались для подготовки варианта книги для веба, а также за помощь в улучшении качества получившегося HTML-кода.

Спасибо супруге Яна Даниэле Флори Гудфеллоу за терпеливую поддержку Яна на всем протяжении работы над книгой и за помощь в выверке текста.

Спасибо команде Google Brain за создание атмосферы интеллектуального общения, позволившей Яну уделять много времени работе над книгой и получать отзывы и наставления от коллег. Особенно хотим поблагодарить бывшего начальника Яна, Грегга Коррадо (Greg Corrado), и его нынешнего начальника, Сами Бенджио (Samy Bengio), за поддержку этого проекта. Наконец, мы благодарны Джеффри Хинтону (Geoffrey Hinton), который подбадривал нас, когда было трудно.

# Обозначения

В этом разделе приведен краткий перечень обозначений, используемых в книге. Большая часть соответствующего математического аппарата описана в главах 2–4.

## Числа и массивы

$a$	скаляр (целый или вещественный)
$\mathbf{a}$	вектор
$\mathbf{A}$	матрица
$\mathbf{A}$	тензор
$\mathbf{I}_n$	единичная матрица с $n$ строками и $n$ столбцами
$\mathbf{I}$	единичная матрица, размер которой определяется контекстом
$\mathbf{e}^{(i)}$	стандартный базисный вектор $[0, \dots, 0, 1, 0, \dots, 0]$ , содержащий 1 в $i$ -й позиции
$\text{diag}(\mathbf{a})$	квадратная диагональная матрица, на диагонали которой находятся элементы вектора $\mathbf{a}$
$a$	случайная скалярная величина
$\mathbf{a}$	случайный вектор
$\mathbf{A}$	случайная матрица

## Множества и графы

$\mathbb{A}$	множество
$\mathbb{R}$	множество вещественных чисел
$\{0, 1\}$	множество из двух элементов: 0 и 1
$\{0, 1, \dots, n\}$	множество целых чисел от 0 до $n$ включительно
$[a, b]$	замкнутый интервал вещественной прямой от $a$ до $b$ , включающий границы
$(a, b]$	интервал вещественной прямой, не включающий $a$ , но включающий $b$
$\mathbb{A} \setminus \mathbb{B}$	разность множеств, т. е. множество, содержащее все элементы $\mathbb{A}$ , не являющиеся элементами $\mathbb{B}$
$\mathcal{G}$	граф
$\text{Pa}_{\mathcal{G}}(x_i)$	родители $x_i$ в $\mathcal{G}$

## Индексирование

$a_i$	$i$ -й элемент вектора $\mathbf{a}$ , индексирование начинается с 1
$a_{-i}$	все элементы вектора $\mathbf{a}$ , кроме $i$ -го
$A_{i,j}$	элемент матрицы $\mathbf{A}$ в позиции $(i, j)$
$\mathbf{A}_{i,:}$	$i$ -я строка матрицы $\mathbf{A}$
$\mathbf{A}_{:,i}$	$i$ -й столбец матрицы $\mathbf{A}$
$A_{i,j,k}$	элемент трехмерного тензора $\mathbf{A}$ в позиции $(i, j, k)$
$\mathbf{A}_{:,:,i}$	двумерная срезка трехмерного тензора $\mathbf{A}$
$a_i$	$i$ -й элемент случайного вектора $\mathbf{a}$

## Операции линейной алгебры

$\mathbf{A}^T$	матрица, транспонированная к $\mathbf{A}$
$\mathbf{A}^+$	псевдообратная матрица Мура-Пенроуза

$A \odot B$  поэлементное произведение матриц  $A$  и  $B$  (произведение Адамара)  
 $\det(A)$  определитель  $A$

**Математический анализ**

$\frac{dy}{dx}$  производная  $y$  по  $x$   
 $\frac{\partial y}{\partial x}$  частная производная  $y$  по  $x$   
 $\nabla_x y$  градиент  $y$  по  $x$   
 $\nabla_x y$  матрица производных  $y$  относительно  $X$   
 $\nabla_x y$  тензор производных  $y$  относительно  $X$   
 $\frac{\partial f}{\partial x}$  матрица Якоби  $J \in \mathbb{R}^{m \times n}$  функции  $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$   
 $\nabla_x^2 f(x)$  гессиан функции  $f$  в точке  $x$   
 или  $H(f)(x)$   
 $\int f(x) dx$  определенный интеграл по всей области определения  $x$   
 $\int_S f(x) dx$  определенный интеграл по множеству  $S$

**Теория вероятностей и теория информации**

$a \perp b$  случайные величины  $a$  и  $b$  независимы  
 $a \perp b \mid c$  они условно независимы при условии  $c$   
 $P(a)$  распределение вероятности дискретной случайной величины  
 $p(a)$  распределение вероятности непрерывной случайной величины или величины, тип которой не задан  
 $a \sim P$  случайная величина  $a$  имеет распределение  $P$   
 $\mathbb{E}_{x \sim P} [f(x)]$  или  $\mathbb{E}f(x)$  математическое ожидание  $f(x)$  при заданном распределении  $P(x)$   
 $\text{Var}(f(x))$  дисперсия  $f(x)$  при заданном распределении  $P(x)$   
 $\text{Cov}(f(x), g(x))$  ковариация  $f(x)$  и  $g(x)$  при заданном распределении  $P(x)$   
 $H(x)$  энтропия Шеннона случайной величины  $x$   
 $D_{\text{KL}}(P \parallel Q)$  расхождение Кульбака–Лейблера между  $P$  и  $Q$   
 $N(x; \mu, \Sigma)$  нормальное распределение случайной величины  $x$  со средним  $\mu$  и ковариацией  $\Sigma$

**Функции**

$f: A \rightarrow \mathbb{B}$  функция  $f$  с областью определения  $A$  и областью значений  $\mathbb{B}$   
 $f \circ g$  композиция функций  $f$  и  $g$   
 $f(x; \theta)$  функция от  $x$ , параметризованная  $\theta$  (иногда, чтобы не утяжелять формулы, мы пишем  $f(x)$ , опуская аргумент  $\theta$ )  
 $\log x$  натуральный логарифм  $x$   
 $\sigma(x)$  логистическая сигмоида,  $1 / (1 + \exp(-x))$   
 $\xi(x)$  функция  $\log(1 + \exp(x))$   
 $\|x\|_p$  норма  $L^p$  вектора  $x$   
 $\|x\|$  норма  $L^2$  вектора  $x$   
 $x^+$  положительная часть  $x$ , т. е.  $\max(0, x)$   
 $\mathbf{1}_{\text{condition}}$  1, если условие *condition* истинно, иначе 0

Иногда мы применяем функцию  $f$  со скалярным аргументом к вектору, матрице или тензору:  $f(\mathbf{x})$ ,  $f(\mathbf{X})$  или  $f(\mathbf{X})$ . Это означает, что функция  $f$  применяется к каждому элементу массива. Например, запись  $\mathbf{C} = \sigma(\mathbf{X})$  означает, что  $C_{i,j,k} = \sigma(X_{i,j,k})$  для всех  $i, j, k$ .

### ***Наборы данных и распределения***

$P_{\text{data}}$	распределение, порождающее данные
$\hat{P}_{\text{data}}$	эмпирическое распределение, определенное обучающим набором
$\mathcal{X}$	обучающий набор примеров
$\mathbf{x}^{(i)}$	$i$ -й пример из входного набора данных
$y^{(i)}$ или $\mathbf{y}^{(i)}$	метка, ассоциированная с $\mathbf{x}^{(i)}$ при обучении с учителем
$\mathbf{X}$	матрица $m \times n$ , в строке $\mathbf{X}_{i,:}$ которой находится входной пример $\mathbf{x}^{(i)}$



Изобретатели давно мечтали создать думающую машину. Эти мечты восходят еще к Древней Греции. Персонажей мифов – Пигмалиона, Дедала, Гефеста – можно было бы назвать легендарными изобретателями, а их творения – Галатею, Талоса и Пандору – искусственной жизнью (Ovid and Martin, 2004; Sparkes, 1996; Tandy, 1997).

Впервые задумавшись о программируемых вычислительных машинах, человек задался вопросом, смогут ли они стать разумными, – за сотню с лишним лет до построения компьютера (Lovelace, 1842). Сегодня **искусственный интеллект (ИИ)** – бурно развивающаяся дисциплина, имеющая многочисленные приложения. Мы хотим иметь интеллектуальные программы, которые могли бы автоматизировать рутинный труд, понимали речь и изображения, ставили медицинские диагнозы и поддерживали научные исследования.

Когда наука об искусственном интеллекте только зарождалась, были быстро исследованы и решены некоторые задачи, трудные для человека, но относительно простые для компьютеров – описываемые с помощью списка формальных математических правил. Настоящим испытанием для искусственного интеллекта стали задачи, которые легко решаются человеком, но с трудом поддаются формализации, – задачи, которые мы решаем интуитивно, как бы автоматически: распознавание устной речи или лиц на картинке.

Эта книга посвящена решению таких интуитивных задач. Цель заключается в том, чтобы компьютер мог учиться на опыте и понимать мир в терминах иерархии понятий, каждое из которых определено через более простые понятия. Благодаря приобретению знаний опытным путем этот подход позволяет исключить этап формального описания человеком всех необходимых компьютеру знаний. Иерархическая организация дает компьютеру возможность учиться более сложным понятиям путем построения их из более простых. Граф, описывающий эту иерархию, будет глубоким – содержащим много уровней. Поэтому такой подход к ИИ называется **глубоким обучением**.

Ранние успехи ИИ в большинстве своем были достигнуты в относительно стерильной формальной среде, где от компьютера не требовались обширные знания о мире. Взять, к примеру, созданную IBM шахматную программу Deep Blue, которая в 1997 году обыграла чемпиона мира Гарри Каспарова (Hsu, 2002). Шахматы – это очень простой мир, состоящий всего из 64 клеток и 32 фигур, которые могут ходить лишь строго определенным образом. Разработка успешной стратегии игры в шахматы – огромное достижение, но трудность задачи – не в описании множества фигур и допустимых ходов на языке, понятном компьютеру. Для полного описания игры достаточно очень короткого списка формальных правил, который заранее составляется программистом.

Забавно, что абстрактные, формально поставленные задачи, требующие значительных умственных усилий от человека, для компьютера как раз наиболее просты. Компьютеры давно уже способны обыграть в шахматы сильнейших гроссмейстеров, но лишь в последние годы стали сопоставимы с человеком в части распознавания объектов или речи. В повседневной жизни человеку необходим гигантский объем знаний о мире. Знания эти субъективны и представлены на интуитивном уровне, поэтому выразить их формально затруднительно. Но чтобы вести себя «разумно», компьютерам нужны такие же знания. Одна из основных проблем искусственного интеллекта – как заложить эти неформальные знания в компьютер.

Авторы нескольких проектов в области ИИ пытались представить знания о мире с помощью формальных языков. Компьютер может автоматически рассуждать о предложениях на таком языке, применяя правила логического вывода. В основе таких подходов лежит **база знаний**. Ни один из этих проектов не привел к существенному успеху. Одним из самых известных был проект Сус (Lenat and Guha, 1989) – машина логического вывода и база утверждений на языке СусL. За ввод утверждений отвечал штат учителей-людей. Процесс оказывается крайне громоздким. Люди из всех сил пытаются придумать формальные правила, достаточно сложные для точного описания мира. Например, Сус не сумел понять рассказ о человеке по имени Фред, который бреется по утрам (Linde, 1992). Его машина вывода обнаружила в рассказе противоречие: он знал, что в людях нет электрических деталей, но, поскольку Фред держал электрическую бритву, система решила, что объект «Бреющийся Фред» содержит электрические детали. И задала вопрос: является ли Фред по-прежнему человеком, когда бреется.

Трудности, с которыми сталкиваются системы, опирающиеся на «зашифрованные в код» знания, наводят на мысль, что система с искусственным интеллектом должна уметь самостоятельно накапливать знания, отыскивая закономерности в исходных данных. Это умение называется **машинным обучением**. С появлением машинного обучения перед компьютерами открылась возможность подступиться к задачам, требующим знаний о реальном мире, и принимать решения, кажущиеся субъективными. Простой алгоритм машинного обучения – **логистическая регрессия** – может решить, следует ли рекомендовать кесарево сечение (Mor-Yosef et al., 1990). Другой простой алгоритм – **наивный байесовский классификатор** – умеет отделять нормальную электронную почту от спама.

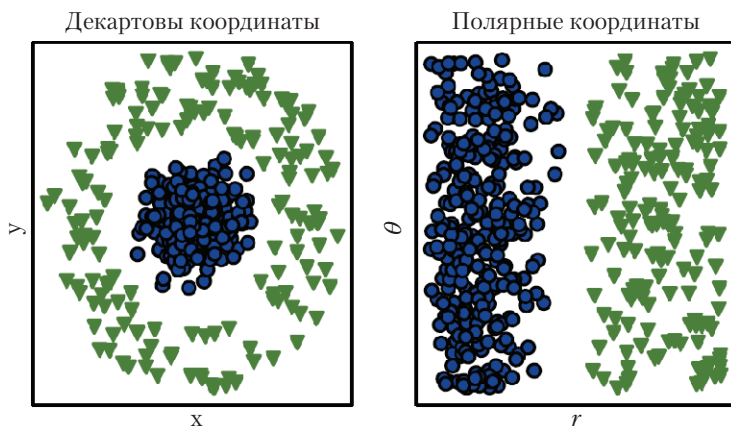
Качество этих простых алгоритмов сильно зависит от представления исходных данных. Так, система ИИ, выдающая рекомендации о показанности кесарева сечения, не осматривает пациента. Вместо этого врач сообщает системе относящуюся к делу информацию, например о наличии или отсутствии рубца на матке. Каждый отдельный элемент информации, включаемый в представление о пациенте, называется **признаком**. Алгоритм логистической регрессии анализирует, как признаки пациента коррелируют с различными результатами. Но он не может никаким образом повлиять на определение признаков. Если алгоритму предложить снимок МРТ, а не формализованные врачом сведения, то он не сможет выдать полезную рекомендацию. Отдельные пиксели снимка практически не коррелированы с осложнениями, которые могут возникнуть во время родов.

Эта зависимость от представления является общим явлением, проявляющимся как в информатике, так и в повседневной жизни. Если говорить об информатике, то такие операции, как поиск в коллекции данных, будут производиться многократно

быстрее, если коллекция структурирована и подходящим образом индексирована. Люди же легко выполняют арифметические операции с числами, записанными арабскими цифрами, но тратят куда больше времени, если используются римские цифры. Неудивительно, что выбор представления оказывает огромное влияние на качество и производительность алгоритмов машинного обучения. На рис. 1.1 приведен простой наглядный пример.

Многие задачи ИИ можно решить, если правильно подобрать признаки, а затем предъявить их алгоритму машинного обучения. Например, в задаче идентификации говорящего по звукам речи полезным признаком является речевой тракт. Он позволяет с большой точностью определить, кто говорит: мужчина, женщина или ребенок.

Но во многих задачах нелегко понять, какие признаки выделять. Допустим, к примеру, что мы пишем программу обнаружения автомобилей на фотографиях. Мы знаем, что у автомобилей есть колеса, поэтому могли бы считать присутствие колеса признаком. К сожалению, на уровне пикселей трудно описать, как выглядит колесо. Колесо имеет простую геометрическую форму, но распознавание его изображения может быть осложнено отбрасыванием теней, блеском солнца на металлических деталях автомобиля, наличием щитка, защищающего колесо от грязи, или объектов на переднем плане, частично загораживающих колесо, и т. д.



**Рис. 1.1** ❖ Пример различных представлений: предположим, что требуется разделить две категории данных, проведя прямую на диаграмме рассеяния. На левом рисунке данные представлены в декартовых координатах, и задача неразрешима. На правом рисунке те же данные представлены в полярных координатах и разделяются вертикальной прямой (рисунок подготовлен совместно с Дэвидом Уорд-Фарли)

Одно из решений этой проблемы – воспользоваться машинным обучением не только для того, чтобы найти отображение представления на результат, но и чтобы определить само представление. Такой подход называется **обучением представлений**. На представлениях, полученных в ходе обучения, часто удается добиться гораздо более высокого качества, чем на представлениях, созданных вручную. К тому же это позволяет системам ИИ быстро адаптироваться к новым задачам при минималь-

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

[e-Univers.ru](http://e-Univers.ru)