

УДК 004.85
ББК 16.6
Б67

Бишоп К. М., Бишоп Х.

Б67 Глубокое обучение: принципы и концепции. Изд. 2-е с решениями / пер. с англ. В. И. Бахура. – М.: ДМК Пресс, 2025. – 850 с.: ил.

ISBN 978-5-93700-281-5

Эта книга предлагает исчерпывающее описание фундаментальных идей, лежащих в основе глубокого обучения. Она разбита на небольшие главы с последовательным изложением материала. Особое внимание уделяется практической ценности изучаемых методов в реальном мире. Сложные концепции рассмотрены в нескольких ракурсах, включая текстовые описания, диаграммы, математические формулы и программные псевдокоды. Второе издание дополнено решениями задач с пояснениями автора.

Книга адресована как новичкам в машинном обучении, так и опытным специалистам в этой области.

УДК 004.85
ББК 16.6

First published in English under the title.
An Introduction to Statistical Learning; with Deep Learning: Foundations and Concepts
by Christopher M. Bishop and Hugh Bishop.

This edition has been translated and published under licence from Springer Nature Switzerland AG.

Springer Nature Switzerland AG takes no responsibility and shall not be made liable for the accuracy of the translation.

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 978-3-031-45467-7 (англ.)

ISBN 978-5-93700-281-5 (рус.)

© Springer Nature Switzerland AG
2024

© Перевод, оформление, издание,
ДМК Пресс, 2025

Содержание

От издательства	15
Предисловие	16
Глава 1. Революция глубокого обучения	22
1.1 Влияние глубокого обучения.....	23
1.1.1 Медицинская диагностика.....	23
1.1.2 Структура белка	24
1.1.3 Синтез изображений.....	25
1.1.4 Большие языковые модели	26
1.2 Учебный пример.....	28
1.2.1 Синтетические данные.....	28
1.2.2 Линейные модели.....	30
1.2.3 Функция ошибки	30
1.2.4 Сложность модели	31
1.2.5 Регуляризация.....	35
1.2.6 Выбор модели.....	37
1.3 Краткая история машинного обучения	39
1.3.1 Однослойные сети	40
1.3.2 Обратное распространение	42
1.3.3 Глубокие сети	44
Глава 2. Вероятности	47
2.1 Правила вероятности.....	49
2.1.1 Пример медицинского обследования	49
2.1.2 Правила суммы и произведения.....	50
2.1.3 Теорема Байеса	53
2.1.4 Повторное медицинское обследование.....	53
2.1.5 Априорные и апостериорные вероятности.....	56
2.1.6 Независимые переменные.....	56
2.2 Плотность распределения вероятностей.....	56
2.2.1 Примеры распределений	58
2.2.2 Ожидания и ковариации	60
2.3 Гауссово распределение	61
2.3.1 Среднее значение и дисперсия	63
2.3.2 Функция правдоподобия.....	63
2.3.3 Ошибка максимального правдоподобия	65

2.3.4	Линейная регрессия	67
2.4	Преобразование плотностей	69
2.4.1	Многомерное распределение	72
2.5	Теория информации	73
2.5.1	Энтропия.....	73
2.5.2	Физическая перспектива.....	75
2.5.3	Дифференциальная энтропия.....	77
2.5.4	Максимальная энтропия	78
2.5.5	Дивергенция Кульбака–Лейблера.....	79
2.5.6	Условная энтропия	82
2.5.7	Взаимная информация.....	82
2.6	Байесовские вероятности	83
2.6.1	Параметры модели.....	84
2.6.2	Регуляризация.....	85
2.6.3	Байесовское машинное обучение	86
	Упражнения	87
Глава 3. Стандартные распределения		94
3.1	Дискретные переменные	95
3.1.1	Распределение Бернулли.....	95
3.1.2	Биноминальное распределение	96
3.1.3	Полиномиальное распределение	97
3.2	Многомерное гауссово распределение	99
3.2.1	Геометрия гауссова распределения	101
3.2.2	Моменты.....	104
3.2.3	Ограничения	105
3.2.4	Условное распределение	107
3.2.5	Маргинальное распределение.....	110
3.2.6	Теорема Байеса	113
3.2.7	Максимальное правдоподобие	115
3.2.8	Последовательная оценка	117
3.2.9	Гауссовы смеси.....	117
3.3	Периодические переменные	121
3.3.1	Распределение фон Мизеса	121
3.4	Семейство экспоненциальных распределений	127
3.4.1	Достаточная статистика	130
3.5	Непараметрические методы	131
3.5.1	Гистограммы.....	132
3.5.2	Ядерная оценка плотности	134
3.5.3	Методика ближайших соседей	137
	Упражнения	140
Глава 4. Однослойные сети: регрессия		147
4.1	Линейная регрессия.....	147
4.1.1	Базисные функции	148
4.1.2	Функция правдоподобия.....	150

4.1.3	Максимальное правдоподобие	151
4.1.4	Геометрия наименьших квадратов	153
4.1.5	Последовательное обучение	154
4.1.6	Регуляризованный метод наименьших квадратов	154
4.1.7	Множественные выходы	155
4.2	Теория принятия решений	157
4.3	Обратное отношение между смещением и дисперсией	161
	Упражнения	166
Глава 5. Однослойные сети: классификация		170
5.1	Дискриминантные функции	171
5.1.1	Два класса	171
5.1.2	Множественные классы	173
5.1.3	Кодирование «1 из K »	175
5.1.4	Наименьшие квадраты для классификации	175
5.2	Теория принятия решений	178
5.2.1	Коэффициент ошибок классификации	179
5.2.2	Ожидаемые потери	182
5.2.3	Опция отказа	183
5.2.4	Вывод и принятие решения	184
5.2.5	Точность классификатора	188
5.2.6	ROC-кривая	190
5.3	Генеративные классификаторы	193
5.3.1	Непрерывные входные данные	195
5.3.2	Решение методом максимального правдоподобия	197
5.3.3	Дискретные параметры	199
5.3.4	Экспоненциальное семейство	200
5.4	Дискриминационные классификаторы	201
5.4.1	Функции активации	201
5.4.2	Фиксированные базисные функции	202
5.4.3	Логистическая регрессия	203
5.4.4	Логистическая регрессия для нескольких классов	205
5.4.5	Пробит-регрессия	207
5.4.6	Канонические функции связей	209
	Упражнения	211
Глава 6. Глубокие нейронные сети		215
6.1	Ограничения фиксированных базисных функций	215
6.1.1	Проклятие размерности	216
6.1.2	Пространства большой размерности	219
6.1.3	Многообразие данных	221
6.1.4	Базисные функции на основе данных	223
6.2	Многослойные сети	224
6.2.1	Матрицы параметров	226
6.2.2	Универсальная аппроксимация	227
6.2.3	Функции активации скрытых элементов	228

6.2.4	Симметрии весового пространства	231
6.3	Глубокие сети	232
6.3.1	Иерархические представления	233
6.3.2	Распределенные представления	234
6.3.3	Обучение представлений	234
6.3.4	Трансферное обучение	236
6.3.5	Контрастивное обучение	238
6.3.6	Основные сетевые архитектуры	241
6.3.7	Тензоры	242
6.4	Функции ошибок	242
6.4.1	Регрессия	242
6.4.2	Бинарная классификация	244
6.4.3	Многоклассовая классификация	245
6.5	Сети смешанной плотности	246
6.5.1	Пример кинематики робота	247
6.5.2	Распределение условного смешивания	248
6.5.3	Градиентная оптимизация	251
6.5.4	Прогнозируемое распределение	252
	Упражнения	254

Глава 7. Градиентный спуск

7.1	Поверхности ошибок	260
7.1.1	Локальная квадратичная аппроксимация	261
7.2	Оптимизация методом градиентного спуска	264
7.2.1	Использование градиентной информации	264
7.2.2	Пакетный градиентный спуск	265
7.2.3	Стохастический градиентный спуск	265
7.2.4	Мини-батчи	266
7.2.5	Инициализация параметров	268
7.3	Сходимость	269
7.3.1	Импульс	271
7.3.2	График скорости обучения	274
7.3.3	RMSProp и Adam	274
7.4	Нормализация	277
7.4.1	Нормализация данных	277
7.4.2	Пакетная нормализация	278
7.4.3	Нормализация слоев	281
	Упражнения	282

Глава 8. Обратное распространение

8.1	Оценка градиентов	287
8.1.1	Однослойные сети	287
8.1.2	Общие сети с прямой передачей	288
8.1.3	Простой пример	291
8.1.4	Численное дифференцирование	292
8.1.5	Матрица Якоби	294

8.1.6	Матрица Гессе	296
8.2	Автоматическое дифференцирование	299
8.2.1	Прямой режим автоматического дифференцирования.....	301
8.2.2	Обратный режим автоматического дифференцирования.....	305
	Упражнения	306
Глава 9. Регуляризация.....		310
9.1	Индуктивное смещение	311
9.1.1	Обратные задачи	311
9.1.2	Теорема об отсутствии бесплатного обеда.....	312
9.1.3	Симметрия и инвариантность	314
9.1.4	Эквивариантность.....	317
9.2	Уменьшение весов.....	318
9.2.1	Последовательные регуляризаторы	320
9.2.2	Обобщенное уменьшение весов	323
9.3	Кривые обучения.....	324
9.3.1	Ранняя остановка	325
9.3.2	Двойной спуск.....	327
9.4	Совместное использование параметров	330
9.4.1	Мягкое разделение весов	331
9.5	Остаточные связи.....	333
9.6	Усреднение модели	337
9.6.1	Прореживание.....	340
	Упражнения	342
Глава 10. Сверточные сети.....		347
10.1	Компьютерное зрение	348
10.1.1	Данные изображений	349
10.2	Сверточные фильтры.....	350
10.2.1	Детекторы признаков	351
10.2.2	Эквивариантный перенос	352
10.2.3	Заполнение	355
10.2.4	Свертки со сдвигом	356
10.2.5	Многомерные свертки.....	356
10.2.6	Пулинг	358
10.2.7	Многослойные свертки.....	360
10.2.8	Примеры сетевых архитектур	361
10.3	Визуализация обученных CNN.....	364
10.3.1	Зрительная кора головного мозга.....	364
10.3.2	Визуализация обученных фильтров.....	366
10.3.3	Карты значимости.....	368
10.3.4	Состязательные атаки.....	369
10.3.5	Синтетические изображения.....	371
10.4	Определение объектов	372
10.4.1	Ограничительные рамки.....	373
10.4.2	Пересечение по объединению.....	374

10.4.3	Скользящие окна	375
10.4.4	Обнаружение в разных масштабах	377
10.4.5	Немаксимальное подавление	378
10.4.6	Быстрая региональная CNN	379
10.5	Сегментация изображений	380
10.5.1	Сверточная сегментация	380
10.5.2	Повышающая дискретизация	381
10.5.3	Полностью сверточные сети	383
10.5.4	Архитектура U-net	384
10.6	Перенос стиля	385
	Упражнения	387

Глава 11. Структурированные распределения

11.1	Модели графов	391
11.1.1	Ориентированные графы	391
11.1.2	Факторизация	392
11.1.3	Дискретные переменные	394
11.1.4	Гауссовы переменные	397
11.1.5	Бинарный классификатор	399
11.1.6	Параметры и наблюдения	400
11.1.7	Теорема Байеса	402
11.2	Условная независимость	403
11.2.1	Три примера графов	404
11.2.2	Объяснения	408
11.2.3	D-разделение	410
11.2.4	Наивный Байес	411
11.2.5	Генеративные модели	413
11.2.6	Покрытие Маркова	415
11.2.7	Графы в качестве фильтров	416
11.3	Модели последовательностей	417
11.3.1	Латентные переменные	420
	Упражнения	421

Глава 12. Трансформеры

12.1	Внимание	426
12.1.1	Обработка трансформеров	428
12.1.2	Коэффициенты внимания	430
12.1.3	Самовнимание	431
12.1.4	Сетевые параметры	432
12.1.5	Масштабируемое самовнимание	435
12.1.6	Многоголовое внимание	436
12.1.7	Слои трансформера	438
12.1.8	Вычислительная сложность	440
12.1.9	Позиционное кодирование	440
12.2	Естественный язык	444
12.2.1	Векторное представление слов	444

12.2.2	Лексическая обработка.....	446
12.2.3	Мультимножество слов.....	448
12.2.4	Модели авторегрессии.....	449
12.2.5	Рекуррентные нейронные сети.....	450
12.2.6	Обратное распространение во времени.....	452
12.3	Языковые модели трансформеров.....	453
12.3.1	Декодирующие трансформеры.....	454
12.3.2	Стратегии выборки.....	457
12.3.3	Кодирующие трансформеры.....	460
12.3.4	Трансформеры последовательности в последовательность.....	462
12.3.5	Большие языковые модели.....	464
12.4	Мультимодальные трансформеры.....	467
12.4.1	Визуальные трансформеры.....	468
12.4.2	Генеративные визуальные трансформеры.....	470
12.4.3	Аудиоданные.....	473
12.4.4	Преобразование текста в речь.....	474
12.4.5	Визуальные и языковые трансформеры.....	476
	Упражнения.....	478
Глава 13. Графовые нейронные сети.....		482
13.1	Машинное обучение на графах.....	483
13.1.1	Свойства графов.....	484
13.1.2	Матрица смежности.....	485
13.1.3	Эквивариантность перестановок.....	486
13.2	Нейронный обмен сообщениями.....	488
13.2.1	Сверточные фильтры.....	488
13.2.2	Графовые сверточные сети.....	490
13.2.3	Операторы агрегации.....	491
13.2.4	Операторы обновления.....	494
13.2.5	Классификация узлов.....	495
13.2.6	Классификация ребер.....	496
13.2.7	Классификация графов.....	496
13.3	Общие графовые сети.....	497
13.3.1	Графовые сети с вниманием.....	497
13.3.2	Встраивание ребер.....	498
13.3.3	Вложения графов.....	499
13.3.4	Чрезмерное сглаживание.....	500
13.3.5	Регуляризация.....	501
13.3.6	Геометрическое глубокое обучение.....	501
	Упражнения.....	502
Глава 14. Выборка.....		505
14.1	Основные алгоритмы выборки.....	505
14.1.1	Ожидаемые значения.....	505
14.1.2	Стандартные распределения.....	507
14.1.3	Выборка с отклонением.....	509

14.1.4	Адаптивная выборка с отклонением.....	511
14.1.5	Выборка по важности	513
14.1.6	Выборка и повторная выборка по значимости	515
14.2	Метод Монте-Карло с цепями Маркова.....	517
14.2.1	Алгоритм Метрополиса	517
14.2.2	Марковские цепи.....	519
14.2.3	Алгоритм Метрополиса–Гастингса	521
14.2.4	Выборка Гиббса.....	523
14.2.5	Выборка по предкам	527
14.3	Выборка Ланжевена	528
14.3.1	Модели на основе энергии.....	529
14.3.2	Максимизация правдоподобия	530
14.3.3	Динамика Ланжевена.....	532
	Упражнения	534

Глава 15. Дискретные латентные переменные

15.1	Кластеризация К-средних.....	538
15.1.1	Сегментация изображений	542
15.2	Гауссовы смеси.....	544
15.2.1	Функция правдоподобия.....	547
15.2.2	Максимальное правдоподобие	549
15.3	Алгоритм ожидания-максимизации.....	554
15.3.1	Гауссовы смеси.....	557
15.3.2	Сравнение с алгоритмом К-средних	559
15.3.3	Смеси распределений Бернулли	560
15.4	Нижняя граница доказательств	564
15.4.1	Новый взгляд на EM	566
15.4.2	Независимые и одинаково распределенные данные	568
15.4.3	Априорные параметры.....	568
15.4.4	Обобщенный EM.....	569
15.4.5	Последовательный EM.....	570
	Упражнения	571

Глава 16. Непрерывные латентные переменные

16.1	Анализ главных компонент.....	576
16.1.1	Определение максимальной дисперсии.....	577
16.1.2	Определение минимальной ошибки.....	579
16.1.3	Сжатие данных.....	582
16.1.4	Отбеливание данных	583
16.1.5	Данные высокой размерности	585
16.2	Вероятностные латентные переменные.....	586
16.2.1	Генеративная модель	587
16.2.2	Функция правдоподобия.....	588
16.2.3	Максимальное правдоподобие	590
16.2.4	Факторный анализ	594
16.2.5	Анализ независимых компонент.....	595

16.2.6	Фильтры Калмана.....	597
16.3	Нижняя граница доказательств	599
16.3.1	Максимизация ожидания.....	600
16.3.2	EM для PCA	603
16.3.3	EM для факторного анализа.....	604
16.4	Нелинейные модели латентных переменных	605
16.4.1	Нелинейные многообразия	606
16.4.2	Функция правдоподобия.....	608
16.4.3	Дискретные данные	609
16.4.4	Четыре метода генеративного моделирования.....	610
	Упражнения	612
Глава 17. Генеративные состязательные сети		617
17.1	Состязательное обучение.....	617
17.1.1	Функция потерь	619
17.1.2	Практическое обучение GAN.....	620
17.2	GAN для обработки изображений.....	623
17.2.1	CycleGAN.....	624
	Упражнения	628
Глава 18. Нормализующие потоки		631
18.1	Потоки сопряжения	633
18.2	Потоки авторегрессии	637
18.3	Непрерывные потоки	639
18.3.1	Нейронные дифференциальные уравнения	639
18.3.2	Обратное распространение нейронных ОДУ.....	640
18.3.3	Потоки нейронных ОДУ	642
	Упражнения	644
Глава 19. Автокодировщики		647
19.1	Детерминированные автокодировщики	647
19.1.1	Линейные автокодировщики	648
19.1.2	Глубокие автокодировщики.....	649
19.1.3	Разреженные автокодировщики.....	651
19.1.4	Шумоподавляющие автокодировщики.....	651
19.1.5	Маскированные автокодировщики.....	652
19.2	Вариационные автокодировщики.....	655
19.2.1	Амортизированный вывод	657
19.2.2	Метод перепараметризации.....	659
	Упражнения	663
Глава 20. Диффузионные модели		666
20.1	Прямой кодировщик.....	667
20.1.1	Диффузионное ядро.....	668
20.1.2	Условное распределение	669

20.2	Обратное декодирование	670
20.2.1	Обучение декодера	673
20.2.2	Нижняя граница доказательств	673
20.2.3	Переименование ELBO	675
20.2.4	Прогнозирование шума	677
20.2.5	Генерация новых выборок	679
20.3	Соответствие оценок	681
20.3.1	Оценка функции потерь	682
20.3.2	Модифицированная оценка потерь	682
20.3.3	Дисперсия шума	684
20.3.4	Стохастические дифференциальные уравнения	685
20.4	Управляемая диффузия	686
20.4.1	Наведение классификатора	687
20.4.2	Наведение без классификатора	688
	Упражнения	691
	Приложение А. Линейная алгебра	696
A.1	Матричные тождества	696
A.2	Следы и определители	697
A.3	Производные матрицы	698
A.4	Собственные векторы	700
	Приложение В. Вариационное исчисление	704
	Приложение С. Множители Лагранжа	707
	Решения упражнений	711
	Список литературы	820
	Предметный указатель	840

Предисловие

Глубокое обучение с использованием многослойных нейронных сетей, натренированных на больших массивах данных с целью решения сложных задач обработки информации, считается наиболее успешной парадигмой в области машинного обучения. За последнее десятилетие глубокое обучение произвело революцию во многих предметных областях, включая компьютерное зрение, распознавание речи и обработку естественного языка. Оно находит все большее применение в здравоохранении, промышленности, коммерции, финансовой сфере, науке и многих других отраслях. Совсем недавно было установлено, что масштабные нейронные сети, также известные как большие языковые модели с количеством обучаемых параметров порядка триллиона, демонстрируют первые признаки общих свойств искусственного интеллекта. Сегодня они выступают в качестве основного движущего фактора крупнейшего в истории технологического прорыва.

Основные задачи этой книги

Рост популярности глубокого обучения сопровождается стремительным увеличением количества и разнообразия научных публикаций в области машинного обучения наряду с ускорением инновационных процессов. Для новичков в этой области даже освоение базовых идей может показаться весьма сложной задачей, не говоря уж о переходе к передовым научным исследованиям. Исходя из этого, книга «Глубокое обучение: принципы и концепции» призвана обеспечить начинающим специалистам в области машинного обучения, а также тем, кто уже имеет опыт работы в этой сфере, глубокое понимание как фундаментальных идей, лежащих в основе глубокого обучения, так и ключевых концепций современных архитектур и методов глубокого обучения. Этот материал поможет читателю получить крепкую основу для будущей специализации. В связи с масштабностью и темпами изменений в этой области авторы намеренно отказались от попыток составить всеобъемлющий обзор всех последних исследований. Вместо этого значительная ценность книги заключается в представлении ключевых идей. Учитывая, что эта область, как можно ожидать, продолжит свое стремительное развитие, эти основы и концепции, по всей видимости, смогут выдержать испытание временем. Например, на момент написания книги большие языковые модели развивались

очень быстро, однако лежащая в их основе архитектура преобразования и механизм концентрации внимания оставались практически неизменными в течение последних пяти лет, в то время как множество основных принципов машинного обучения хорошо известны на протяжении десятилетий.

Ответственное применение технологий

Глубокое обучение представляет собой мощную технологию с обширными возможностями применения. Она обладает огромным потенциалом для создания новых ценностей и решения наиболее актуальных проблем современного общества. Однако эти же качества обуславливают возможность преднамеренного злоупотребления глубоким обучением или причинения неумышленного вреда. Мы не стали обсуждать этические и социальные аспекты использования глубокого обучения, поскольку эти темы настолько важны и сложны, что требуют более тщательного рассмотрения, чем это возможно в подобном техническом издании. Вместе с тем рассуждения на такие темы должны опираться на глубокие знания об основах технологии и ее работе, поэтому мы надеемся, что эта книга станет ценным вкладом в подобные серьезные исследования. При этом мы настоятельно рекомендуем читателям не оставлять без внимания более широкие последствия своей работы и изучать вопросы ответственного использования глубокого обучения и искусственного интеллекта наряду с изучением самой технологии.

Структура этой книги

Книга состоит из достаточно большого количества компактных глав, каждая из которых посвящена определенной теме. Книга построена по линейному принципу, поскольку каждая глава опирается только на материал, рассмотренный в предыдущих главах. Она хорошо подходит для преподавания курса машинного обучения в течение двух семестров для студентов или аспирантов, но в равной степени актуальна и для тех, кто активно занимается исследованиями или самообразованием.

Четкое понимание принципов машинного обучения может быть достигнуто только с помощью определенного уровня знаний математики. Говоря конкретно, в основе машинного обучения лежат три области математики: теория вероятности, линейная алгебра и многомерный математический анализ. Книга обеспечивает последовательное введение в необходимые концепции теории вероятности и включает приложение, в котором обобщены некоторые полезные практические методы линейной алгебры. Предполагается, что читатель уже знаком с основными концепциями многомерного анализа, однако в приложениях есть вводные сведения по вариационному исчислению и методу множителей Лагранжа. Основное внимание в книге тем

не менее уделено донесению до читателя четкого понимания изложенных концепций с акцентом на методы, имеющие реальную практическую ценность, а не на абстрактную теорию. Там, где это возможно, мы постарались представить более сложные концепции с нескольких взаимодополняющих точек зрения, включая текстовое описание, диаграммы и математические формулы. Кроме того, многие обсуждаемые в тексте ключевые алгоритмы кратко изложены в отдельных врезках. Они не связаны с вопросами эффективности вычислений, но служат дополнением к математическим выкладкам в тексте. Поэтому мы надеемся, что материал этой книги окажется доступным для читателей с разным уровнем подготовки.

В концептуальном плане эту книгу, скорее всего, стоит рассматривать как продолжение книги «Нейронные сети для распознавания образов» (*Neural Networks for Pattern Recognition*, Bishop, 1995b), в которой были впервые даны исчерпывающие сведения о нейронных сетях с позиций математической статистики. Ее также можно рассматривать как дополнение к книге «Распознавание образов и машинное обучение» (*Pattern Recognition and Machine Learning*, Bishop, 2006), в которой рассматривается более широкий спектр тем машинного обучения, несмотря на то что она была написана до начала революции глубокого обучения. Вместе с тем, чтобы обеспечить самодостаточность этой новой книги, соответствующий материал был перенесен из книги Bishop (2006) и переработан так, чтобы сосредоточиться именно на тех основополагающих идеях, которые важны при глубоком обучении. Это означает, что многие интересные темы в машинном обучении, рассмотренные в Bishop (2006), остаются интересными и сегодня, но в новой книге они не затрагиваются. Например, в Bishop (2006) довольно подробно рассматриваются байесовские методы, в то время как в этой книге они почти полностью исключены.

Справочная информация

Следуя принципу акцентирования внимания на ключевых идеях, мы не пытаемся представить всеобъемлющий обзор литературы, что в любом случае было бы невозможно с учетом масштабов и темпов изменений в этой сфере. Тем не менее мы приводим ссылки на некоторые важнейшие научные работы, а также обзорные статьи и другие источники для дополнительного чтения. Во многих случаях они также содержат важные детали практического применения, которые мы опускаем в тексте с целью не отвлекать читателя от обсуждения ключевых концепций.

На тему машинного обучения в целом и глубокого обучения в частности уже написано множество книг. Среди наиболее близких по уровню и стилю к этой книге можно отметить Bishop (2006), Goodfellow, Bengio and Courville (2016), Murphy (2022), Murphy (2023) и Prince (2023).

За последнее десятилетие специфика научной деятельности в области машинного обучения существенно изменилась: многие работы публикуются на

архивных сайтах до или даже вместо направления на экспертные конференции и в журналы. Самый популярный из таких сайтов – arXiv, что означает «архив», и доступен он по адресу <https://arXiv.org>.

Этот сайт позволяет обновлять статьи, что часто приводит к появлению нескольких версий в разные календарные годы, и это может привести к возникновению разночтений в отношении цитирования той или иной версии и того или иного года. Кроме того, на сайте имеется бесплатный доступ к PDF-файлу каждой статьи. Поэтому мы используем простой подход – ссылаться на статью в соответствии с годом ее первой загрузки, хотя рекомендуем читать самую последнюю версию.

Статьи на arXiv индексируются с использованием обозначения arXiv:YYMM.XXXXX, где YY и MM означают год и месяц первой загрузки соответственно. Последующие версии обозначаются добавлением номера версии N в форме arXiv:YYMM.XXXXXvN.

Упражнения

В конце каждой главы приводится подборка упражнений для закрепления ключевых идей, изложенных в тексте, или для их углубленного анализа и обобщения. Эти упражнения являются важной частью текста, и каждое из них оценивается по степени сложности: от (*), обозначающей простое упражнение на несколько минут, до (***) , обозначающей существенно более сложное упражнение. Настоятельно рекомендуется выполнять упражнения, поскольку активная работа с материалом значительно повышает эффективность обучения. Решения всех упражнений доступны в виде PDF-файла, который можно загрузить с веб-сайта книги.

Математические обозначения

Мы используем те же обозначения, что и в книге (Bishop 2006). Обзор математики в контексте машинного обучения представлен в книге (Deisenroth, Faisal and Ong 2020).

Векторы обозначаются строчными полужирными латинскими буквами, например \mathbf{x} , а матрицы обозначаются прописными полужирными латинскими буквами, например \mathbf{M} . Если не указано иное, предполагается, что все векторы являются векторами-столбцами. Надстрочный индекс T обозначает транспонирование матрицы или вектора, так что \mathbf{x}^T будет представлять собой строчный вектор. Обозначение (w_1, \dots, w_M) подразумевает вектор строки с M элементами, а соответствующий вектор столбца записывается как $\mathbf{w} = (w_1, \dots, w_M)^T$. Матрица тождественности $M \times M$ (также известная как единичная матрица) обозначается \mathbf{I}_M , и, если нет двусмысленности относительно ее размерности, она будет сокращаться до \mathbf{I} . Ее элементы I_{ij} равны 1, если $i = j$, и 0, если

$i \neq j$. Элементы единичной матрицы иногда обозначают как δ_{ij} . Обозначение $\mathbf{1}$ означает вектор-столбец, в котором все элементы имеют значение 1. Запись $\mathbf{a} \oplus \mathbf{b}$ обозначает конкатенацию векторов \mathbf{a} и \mathbf{b} , так что если $\mathbf{a} = (a_1, \dots, a_N)$ и $\mathbf{b} = (b_1, \dots, b_M)$, то $\mathbf{a} \oplus \mathbf{b} = (a_1, \dots, a_N, b_1, \dots, b_M)$. Запись $|x|$ обозначает модуль (положительную часть) скаляра x , также известного как абсолютное значение. Мы используем запись $\det \mathbf{A}$ для обозначения определителя матрицы \mathbf{A} .

Запись $x \sim p(x)$ означает, что x выбирается из распределения $p(x)$. Там, где возможны разночтения, мы будем использовать подстрочные индексы, как $p_x(\cdot)$, чтобы выразить, о какой плотности идет речь. Математическое ожидание функции $f(x, y)$ относительно случайной переменной x обозначается $\mathbb{E}_x[f(x, y)]$. В ситуациях, когда нет двусмысленности относительно того, по какой переменной производится усреднение, суффикс опускается, например $\mathbb{E}[x]$. Если распределение x зависит от другой переменной z , то соответствующее условное ожидание будет записано как $\mathbb{E}_x[f(x)|z]$. Аналогично дисперсия $f(x)$ обозначается $\text{var}[f(x)]$, а для векторных переменных ковариация записывается как $\text{cov}[\mathbf{x}, \mathbf{y}]$. Мы также будем использовать $\text{cov}[\mathbf{x}]$ как сокращенное обозначение для $\text{cov}[\mathbf{x}, \mathbf{x}]$.

Символ \forall означает «для всех», так что $\forall t \in \mathcal{M}$ обозначает все значения t в пределах множества \mathcal{M} . Мы используем \mathbb{R} для обозначения вещественных чисел. На графе множество соседних узлов i обозначается как $\mathcal{N}(i)$, которое не следует путать с гауссовым или нормальным распределением $\mathcal{N}(x|\mu, \sigma^2)$. Функционал обозначается $f[y]$, где $y(x)$ – это некоторая функция. Понятие функционала рассматривается в приложении В. Фигурные скобки $\{ \}$ обозначают множество. Запись $g(x) = \mathcal{O}(f(x))$ означает, что $|f(x)/g(x)|$ ограничено при $x \rightarrow \infty$. Например, если $g(x) = 3x^2 + 2$, то $g(x) = \mathcal{O}(x^2)$. Обозначение $\lfloor x \rfloor$ означает «нижнюю целую часть числа» x , т. е. наибольшее целое число, которое меньше или равно x .

Если имеется N независимых и одинаково распределенных (independent and identically distributed, i.i.d.) значений x_1, \dots, x_N D -мерного вектора $\mathbf{x} = (x_1, \dots, x_D)^T$, мы можем объединить эти результаты наблюдений в матрицу данных \mathbf{X} размерности $N \times D$, в которой n -я строка \mathbf{X} соответствует вектору строк \mathbf{x}_n^T . Так, n, i элемент \mathbf{X} соответствует i -му элементу n -го наблюдения \mathbf{x}_n и записывается как x_{ni} . Для одномерных переменных мы обозначаем такую матрицу через \mathbf{x} , т. е. вектор-столбец, n -й элемент которого равен x_n . Обратите внимание, что для обозначения \mathbf{x} (размерность N) используется другой шрифт, чтобы отличить его от x (размерность D).

Решения упражнений

Главы со 2 по 10 | Версия 1.0

Это версия 1.0 руководства по решениям для книги «Глубокое обучение: принципы и концепции» авторов М. Бишопа и Х. Бишопа (Springer, 2024). В нем представлены готовые решения для упражнений в главах 2–10. Полное руководство с решениями всех упражнений будет опубликовано после того, как оно будет получено от авторов книги. Бесплатная цифровая версия решений для покупателей первого издания доступна на веб-сайте книги по адресу www.dmkpress.com.

Отзывы о книге или связанных с ней материалах, включая данное руководство с решениями задач, пожалуйста, отправляйте авторам по адресу dmkpress@gmail.com.

От переводчика: по причине промежуточного характера этой редакции сборника решений упражнений в ряде фрагментов авторы еще не завершили разметку перекрестных упоминаний различных упражнений, глав книги и других материалов. Эти упоминания отмечены в переводе (как и в оригинале) двумя символами ?? (например, «...аргумент, аналогичный использованному в решении упражнения ??»).

Глава 2. Вероятности

2.1. Для начала определим $p(T = 1)$ путем изменения (2.20):

$$\begin{aligned} p(T = 1) &= p(T = 1|C = 0)p(C = 0) + p(T = 1|C = 1)p(C = 1) \\ &= \frac{3}{100} \times \frac{999}{1000} + \frac{90}{100} \times \frac{1}{1000} = \frac{3087}{100\,000} = 0,03087. \end{aligned} \quad (1)$$

Затем произведем оценку $p(C = 1|T = 1)$ путем модификации (2.22):

$$p(C = 1|T = 1) = \frac{p(T = 1|C = 1)p(C = 1)}{p(T = 1)} \quad (2)$$

$$= \frac{90}{100} \times \frac{1}{1000} \times \frac{100\,000}{3087} = \frac{90}{3087} \approx 0,029. \quad (3)$$

Таким образом, получается, что вероятность заболеть раком даже после положительного теста остается очень малой.

2.2. Необходимо учитывать, что каждое из чисел ряда 0, 1, 2, 3, 4, 5 и 6 выпадает только на одной из этих игральных костей, а это исключает выпадение ничьей при сравнении выпадений одной из костей против другой.

Сначала рассмотрим красную кость и отметим, что на ней имеется четыре копии числа 2 и две копии числа 6. В двух третях случаев при бросании красной кости выпадет 2, и в одной трети случаев выпадет 6. Следовательно, если бросать красную кость в паре с желтой (которая всегда выпадает на 3), желтая кость будет в среднем выигрывать в двух третях случаев, а в одной трети случаев будет проигрывать.

Теперь обратим внимание на синюю кость, на которой четыре раза встречается число 4 и два раза 0. Если бросать ее в паре с желтой костью, она будет выпадать с 4 в двух третях случаев и выигрывать в этом случае и 0 в одной трети случаев и в этом случае будет проигрывать.

Далее рассмотрим зеленую кость в сравнении с синей. Зеленая кость имеет три копии числа 1 и три копии 5. Чтобы вычислить вероятность выигрыша зеленой кости, для начала стоит отметить, что вероятность выпадения 5 на зеленой кости равна $1/2$, и в этом случае она обязательно выиграет у синей кости. Точно так же вероятность выпадения 1 на зеленой кости составляет $1/2$, и в этом случае вероятность ее выигрыша составляет $1/3$. Общая вероятность выигрыша зеленой кости определяется путем умножения вероятностей:

$$\left(\frac{1}{2} \times 1\right) + \left(\frac{1}{2} \times \frac{1}{3}\right) = \frac{2}{3}. \quad (4)$$

Наконец, рассмотрим вероятность выигрыша красной кости в паре с зеленой. Вероятность выпадения числа 6 на красной кости составляет $1/3$, и в этом случае она обязательно выигрывает. Точно так же вероятность выпадения числа 2 на красной кости составляет $2/3$, и в этом случае вероятность выигрыша этой кости составляет $1/2$. Общая вероятность выигрыша красной кости вновь получается путем умножения вероятностей:

$$\left(\frac{1}{3} \times 1\right) + \left(\frac{2}{3} \times \frac{1}{2}\right) = \frac{2}{3}. \quad (5)$$

Для получения дополнительной информации об этих кубиках можно обратиться по адресу microsoft.com/en-us/research/project/non-transitive-dice/.

2.3. Желаемое распределение можно выразить в форме с применением правил суммирования и произведения вероятностей:

$$p(\mathbf{y}) = \iint p(\mathbf{y}|\mathbf{u}, \mathbf{v}) p_{\mathbf{u}}(\mathbf{u}) p_{\mathbf{v}}(\mathbf{v}) d\mathbf{u} d\mathbf{v}. \quad (6)$$

Поскольку y является детерминированной функцией \mathbf{u} и \mathbf{v} , ее условное распределение задается дельта-функцией Дирака в виде

$$p(y|\mathbf{u}, \mathbf{v}) = \delta(y - \mathbf{u} - \mathbf{v}). \quad (7)$$

Подстановка (7) в (6) дает возможность выполнить интегрирование по \mathbf{v} и получить

$$p(y) = \int p_u(\mathbf{u}) p_v(y - \mathbf{u}) d\mathbf{u}, \quad (8)$$

как и требовалось.

2.4. Интегрируя равномерное распределение (2.33) по x , получим

$$\int_{-\infty}^{\infty} p(x) dx = \int_c^d \frac{1}{d-c} dx = \frac{d-c}{d-c} = 1. \quad (9)$$

И, следовательно, это распределение является нормализованным. Среднее значение этого распределения составляет

$$\mathbb{E}[x] = \int_a^b \frac{1}{b-a} x dx = \left[\frac{x^2}{2(b-a)} \right]_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.$$

Дисперсию можно определить, вычислив для начала

$$\mathbb{E}[x^2] = \int_a^b \frac{1}{b-a} x^2 dx = \left[\frac{x^3}{3(b-a)} \right]_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3},$$

и далее, используя (2.46), получить

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

2.5. Интегрируя экспоненциальное распределение (2.34) в пределах $0 \leq x \leq \infty$, получаем

$$\begin{aligned} \int_0^{\infty} p(x|\lambda) dx &= \int_0^{\infty} \lambda \exp(-\lambda x) dx \\ &= \int_0^{\infty} \exp(-y) dy \\ &= [-\exp(-y)]_0^{\infty} \\ &= 1, \end{aligned} \quad (10)$$

где используется замена переменных $y = \lambda x$. Таким образом, экспоненциальное распределение нормализовано. Точно так же, если проинтегрировать распределение Лапласа (2.35), получится

$$\begin{aligned}
\int_{-\infty}^{\infty} p(x|\mu, \lambda) dx &= \int_{-\infty}^{\infty} \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right) dx \\
&= \int_{-\infty}^{\mu} \frac{1}{2\gamma} \exp\left(\frac{x - \mu}{\gamma}\right) dx + \int_{\mu}^{\infty} \frac{1}{2\gamma} \exp\left(-\frac{x - \mu}{\gamma}\right) dx \\
&= \int_{-\infty}^0 \frac{1}{2} \exp(z) dz + \int_0^{\infty} \frac{1}{2} \exp(-z) dz \\
&= \left[\frac{1}{2} \exp(z)\right]_{-\infty}^0 + \left[-\frac{1}{2} \exp(-z)\right]_0^{\infty} \\
&= \frac{1}{2} + \frac{1}{2} = 1,
\end{aligned} \tag{11}$$

где в каждом из двух интегралов была произведена подстановка $z = (x - \mu)/\gamma$. Таким образом, получается, что распределение Лапласа также нормализовано.

2.6. Интегрируя эмпирическую плотность вероятности (2.37), получаем

$$\begin{aligned}
\int_{-\infty}^{\infty} p(x|\mathcal{D}) dx &= \frac{1}{N} \sum_{n=1}^N \int_{-\infty}^{\infty} \delta(x - x_n) dx \\
&= \frac{1}{N} \sum_{n=1}^N \int_{-\infty}^{\infty} \delta(y) dy \\
&= \frac{1}{N} \sum_{n=1}^N 1 = 1,
\end{aligned} \tag{12}$$

как и требовалось. Здесь производится подстановка $y = x - x_n$ в n -й интеграл суммирования, а затем применяется определение дельта-функции Дирака. Этот результат можно легко обобщить на многомерную переменную данных x .

2.7. Если подставить эмпирическое распределение (2.37) в определение математического ожидания по отношению к непрерывной плотности, заданной формулой (2.39), то получится

$$\begin{aligned}
\mathbb{E}[f] &= \int_{-\infty}^{\infty} p(x) f(x) dx \\
&\simeq \frac{1}{N} \sum_{n=1}^N \int_{-\infty}^{\infty} \delta(x - x_n) f(x) dx \\
&= \frac{1}{N} \sum_{n=1}^N \int_{-\infty}^{\infty} \delta(y_n) f(y_n + x_n) dy_n \\
&= \frac{1}{N} \sum_{n=1}^N f(x_n),
\end{aligned} \tag{13}$$

как и требовалось. Здесь применена замена переменной $y_n = x - x_n$ отдельно в каждом из интегралов, а также свойство функции Дирака $\delta(y_n)$, которая интегрируется до единицы и где ненулевой вклад вносит только $y_n = 0$.

2.8. При разложении квадрата получается

$$\begin{aligned}\mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] &= \mathbb{E}[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2] \\ &= \mathbb{E}[f(x)^2] - 2\mathbb{E}[f(x)]\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2 \\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2,\end{aligned}$$

как и требовалось.

2.9. Определение ковариации задается формулой (2.47) в виде

$$\text{cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y].$$

Используя (2.38) и тот факт, что $p(x, y) = p(x)p(y)$ при независимых x и y , получаем

$$\begin{aligned}\mathbb{E}[xy] &= \sum_x \sum_y p(x, y)xy \\ &= \sum_x p(x)x \sum_y p(y)y \\ &= \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

и, следовательно, $\text{cov}[x, y] = 0$. Случай, когда x и y являются непрерывными переменными, является аналогичным, при этом (2.38) заменяется на (2.39), а суммы заменяются интегралами.

2.10. Поскольку x и z независимы, их совместное распределение поддается факторизации $p(x, z) = p(x)p(z)$, и поэтому

$$\mathbb{E}[x + z] = \iint (x + z)p(x)p(z)dx dz \quad (14)$$

$$= \int xp(x)dx + \int zp(z)dz \quad (15)$$

$$= \mathbb{E}[x] + \mathbb{E}[z]. \quad (16)$$

Точно так же для дисперсий – сначала следует отметить, что

$$(x + z - \mathbb{E}[x + z])^2 = (x - \mathbb{E}[x])^2 + (z - \mathbb{E}[z])^2 + 2(x - \mathbb{E}[x])(z - \mathbb{E}[z]), \quad (17)$$

где последний член интегрируется по отношению к факторизованному распределению $p(x)p(z)$ до нуля. Таким образом,

$$\begin{aligned}\text{var}[x + z] &= \iint (x + z - \mathbb{E}[x + z])^2 p(x)p(z)dx dz \\ &= \int (x - \mathbb{E}[x])^2 p(x)dx + \int (z - \mathbb{E}[z])^2 p(z)dz \\ &= \text{var}(x) + \text{var}(z).\end{aligned} \quad (18)$$

В случае с дискретными переменными интегралы заменяются суммами, и вновь получаются те же результаты.

2.11. Используя определение математического ожидания (2.39), получаем

$$\begin{aligned}\mathbb{E}_y[\mathbb{E}_x[x|y]] &= \int p(y) \int p(x|y)x \, dx \, dy \\ &= \iint p(x, y)x \, dx \, dy \\ &= \int p(x)x \, dx = \mathbb{E}[x],\end{aligned}\quad (19)$$

где применено правило произведения вероятностей $p(x|y)p(y) = p(x, y)$. Теперь используем результат (2.46) для записи

$$\begin{aligned}\mathbb{E}_y[\text{var}_x[x|y]] + \text{var}_y[\mathbb{E}_x[x|y]] \\ = \mathbb{E}_y[\mathbb{E}_x[x^2|y]] - \mathbb{E}_y[\mathbb{E}_x[x|y]^2] + \mathbb{E}_y[\mathbb{E}_x[x|y]^2] - \mathbb{E}_y[\mathbb{E}_x[x|y]]^2.\end{aligned}\quad (20)$$

Теперь отметим, что второй и третий члены в правой части (20) аннулируются. Первый член в правой части (20) можно записать как

$$\begin{aligned}\mathbb{E}_y\mathbb{E}_x[x^2|y] &= \int p(y) \int p(x|y)x^2 \, dx \, dy \\ &= \iint p(x, y)x^2 \, dx \, dy \\ &= \int p(x)x^2 \, dx = \mathbb{E}[x^2].\end{aligned}\quad (21)$$

Таким же образом можно вновь использовать результат (2.46) для записи четвертого члена справа от (20) в виде $\mathbb{E}[x]^2$. Отсюда получаем

$$\mathbb{E}_y[\text{var}_x[x|y]] + \text{var}_y[\mathbb{E}_x[x|y]] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \text{var}[x],\quad (22)$$

что и требовалось.

2.12. Преобразование из декартовых координат в полярные определяется выражениями

$$x = r \cos \theta, \quad (23)$$

$$y = r \sin \theta, \quad (24)$$

и, таким образом, получаем $x^2 + y^2 = r^2$, где используется хорошо известное тригонометрическое тождество (3.127). Также видно, что матрица Якоби преобразования переменных равна

$$\begin{aligned}\frac{\partial(x, y)}{\partial(r, \theta)} &= \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} \\ &= \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} = r,\end{aligned}$$

где вновь используется (3.127). Таким образом, двойной интеграл в (2.125) принимает вид

$$I^2 = \int_0^{2\pi} \int_0^\infty \exp\left(-\frac{r^2}{2\sigma^2}\right) r dr d\theta \quad (25)$$

$$= 2\pi \int_0^\infty \exp\left(-\frac{u}{2\sigma^2}\right) \frac{1}{2} du \quad (26)$$

$$= \pi \left[\exp\left(-\frac{u}{2\sigma^2}\right) (-2\sigma^2) \right]_0^\infty \quad (27)$$

$$= 2\pi\sigma^2, \quad (28)$$

где используется замена переменных $r^2 = u$. Таким образом,

$$I = (2\pi\sigma^2)^{1/2}.$$

Наконец, с помощью преобразования $y = x - \mu$ интеграл гауссова распределения приобретает вид

$$\begin{aligned} \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx &= \frac{1}{(2\pi\sigma^2)^{1/2}} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2\sigma^2}\right) dy \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}} = 1, \end{aligned}$$

что и требовалось.

2.13. Из определения одномерного гауссова распределения (2.49) получаем

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} x dx. \quad (29)$$

Теперь выполним замену переменных $y = x - \mu$ и получим

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}y^2\right\} (y + \mu) dy. \quad (30)$$

Теперь стоит обратить внимание на то, что в множителе $(y + \mu)$ первый член в y соответствует нечетной подынтегральной функции, и по этой причине это интеграл должен исчезнуть (чтобы показать это в явном виде, запишем интеграл как сумму двух интегралов, один от $-\infty$ до 0, а другой от 0 до ∞ , а затем наглядно продемонстрируем, что эти два интеграла аннулируются). Во втором слагаемом μ является константой и выносится за пределы интеграла, оставляя нормализованное гауссово распределение, которое, в свою очередь, интегрируется до 1, и, таким образом, получается (2.52).

Чтобы получить (2.53), сначала подставляем выражение (2.49) для нормального распределения в результат нормализации (2.51) и затем перегруппировываем, что дает

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx = (2\pi\sigma^2)^{1/2}. \quad (31)$$

Теперь дифференцируем обе стороны (31) относительно σ^2 , а затем перегруппировываем, что дает

$$\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} (x-\mu)^2 dx = \sigma^2 \quad (32)$$

и это непосредственно доказывает, что

$$\mathbb{E}[(x-\mu)^2] = \text{var}[x] = \sigma^2. \quad (33)$$

Теперь разложим квадрат в левой части, что дает

$$\mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 = \sigma^2.$$

Используя (2.52), получаем (2.53), как и требовалось.

Наконец, (2.54) следует непосредственно из (2.52) и (2.53):

$$\mathbb{E}[x^2] - \mathbb{E}[x]^2 = (\mu^2 + \sigma^2) - \mu^2 = \sigma^2.$$

2.14. Для одномерного случая достаточно продифференцировать (2.49) по x , чтобы получить

$$\frac{d}{dx} \mathcal{N}(x|\mu, \sigma^2) = -\mathcal{N}(x|\mu, \sigma^2) \frac{x-\mu}{\sigma^2}.$$

Приравняв это выражение к нулю, получаем $x = \mu$.

2.15. Используем ℓ для обозначения $\ln p(\mathbf{X}|\mu, \sigma^2)$ из (2.56). По стандартным правилам дифференцирования получаем

$$\frac{d\ell}{d\mu} = \frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu).$$

Приравняв это к нулю и перенеся члены, содержащие μ , на другую сторону уравнения, получаем

$$\frac{1}{\sigma^2} \sum_{n=1}^N x_n = \frac{1}{\sigma^2} N\mu$$

и далее, умножив обе части на σ^2/N , получаем (2.57).

По аналогии получается

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{1}{2(\sigma^2)^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2\sigma^2},$$

и, приравнявая это к нулю, получаем

$$\frac{N}{2} \frac{1}{\sigma^2} = \frac{1}{2(\sigma^2)^2} \sum_{n=1}^N (x_n - \mu)^2.$$

Умножив обе части на $2(\sigma^2)^2/N$ и подставив μ_{ML} вместо μ , получаем (2.58).

2.16. Если $m = n$, то $x_n x_m = x_n^2$, и при использовании (2.53) получаем $\mathbb{E}[x_n^2] = \mu^2 + \sigma^2$, а если $n \neq m$, то две точки данных x_n и x_m независимы, и, следовательно, $\mathbb{E}[x_n x_m] = \mathbb{E}[x_n] \mathbb{E}[x_m] = \mu^2$, где использовано (2.52). Объединяя эти два результата, получаем (2.128).

Далее получаем

$$\mathbb{E}[\mu_{ML}] = \frac{1}{N} \sum_{n=1}^N \mathbb{E}[x_n] = \mu, \quad (34)$$

где использовано (2.52).

Наконец, рассмотрим $\mathbb{E}[\sigma_{ML}^2]$. Исходя из (2.57) и (2.58) с использованием (2.128), получаем

$$\begin{aligned} \mathbb{E}[\sigma_{ML}^2] &= \mathbb{E} \left[\frac{1}{N} \sum_{n=1}^N \left(x_n - \frac{1}{N} \sum_{m=1}^N x_m \right)^2 \right] \\ &= \frac{1}{N} \sum_{n=1}^N \mathbb{E} \left[x_n^2 - \frac{2}{N} x_n \sum_{m=1}^N x_m + \frac{1}{N^2} \sum_{m=1}^N \sum_{l=1}^N x_m x_l \right] \\ &= \left\{ \mu^2 + \sigma^2 - 2 \left(\mu^2 + \frac{1}{N} \sigma^2 \right) + \mu^2 + \frac{1}{N} \sigma^2 \right\} \\ &= \left(\frac{N-1}{N} \right) \sigma^2, \end{aligned} \quad (35)$$

что и требовалось.

2.17. Из определения (2.61) и с использованием (2.52) и (2.53) получаем

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2] &= \mathbb{E} \left[\frac{1}{N} \sum_{n=1}^N (x_n - \mu)^2 \right] \\ &= \frac{1}{N} \sum_{n=1}^N \mathbb{E}[x_n^2 - 2x_n \mu + \mu^2] \\ &= \frac{1}{N} \sum_{n=1}^N (\mu^2 + \sigma^2 - 2\mu\mu + \mu^2) \\ &= \sigma^2, \end{aligned} \quad (36)$$

как и требовалось.

2.18. Дифференцируя (2.66) по σ^2 , получаем

$$\frac{\partial}{\partial \sigma^2} \ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma^2) = \frac{1}{2\sigma^4} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 - \frac{N}{2} \frac{1}{\sigma^2}. \quad (37)$$

Приравнявая производную к нулю и производя перегруппировку, получаем

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N \{y(x_n, \mathbf{w}_{\text{ML}}) - t_n\}^2, \quad (38)$$

как и требовалось.

2.19. Если сделать предположение, что функция $y = f(x)$ является *строго* монотонной, что необходимо для исключения возможности появления пиков бесконечной плотности в $p(y)$, то тогда будет гарантировано существование обратной функции $x = f^{-1}(y)$. Затем можно использовать (2.71) для записи

$$p(y) = q(f^{-1}(y)) \left| \frac{df^{-1}}{dy} \right|. \quad (39)$$

Поскольку единственным ограничением f является ее монотонность, она может произвольно распределять вероятность по x над y . Это показано на рис. 2.12 на стр. 71 как часть решения ???. Исходя из (39), непосредственно получается, что

$$|f'(x)| = \frac{q(x)}{p(f(x))}.$$

2.20. Матрица Якоби для преобразования из $(x_1; x_2)$ в $(y_1; y_2)$ определяется следующим образом:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix}. \quad (40)$$

Для конкретного преобразования, определяемого в (2.78) и (2.79), получается

$$\frac{\partial y_1}{\partial x_1} = 1 + 5 \operatorname{sech}^2(x_1), \quad (41)$$

$$\frac{\partial y_1}{\partial x_2} = 0, \quad (42)$$

$$\frac{\partial y_2}{\partial x_1} = x_1^2, \quad (43)$$

$$\frac{\partial y_2}{\partial x_2} = 1 + 5 \operatorname{sech}^2(x_2). \quad (44)$$

2.21. Из описания в вводной части раздела 2.5 следует, что

$$h(p^2) = h(p) + h(p) = 2h(p).$$

Затем допустим, что для всех $k \leq K$, $h(p^k) = kh(p)$. Для $k = K + 1$ выполняется

$$h(p^{K+1}) = h(p^K p) = h(p^K) + h(p) = Kh(p) + h(p) = (K + 1)h(p).$$

Кроме того,

$$h(p^{n/m}) = nh(p^{1/m}) = \frac{n}{m} mh(p^{1/m}) = \frac{n}{m} h(p^{m/m}) = \frac{n}{m} h(p),$$

и, следовательно, в соответствии с принципом непрерывности получаем, что $h(p^x) = xh(p)$ для любого действительного числа x .

Теперь рассмотрим положительные действительные числа p и q , а также действительное число x , такие, что $p = q^x$. Из приведенного выше рассуждения следует, что

$$\frac{h(p)}{\ln(p)} = \frac{h(q^x)}{\ln(q^x)} = \frac{xh(q)}{x \ln(q)} = \frac{h(q)}{\ln(q)}$$

и, следовательно, $h(p) \propto \ln(p)$.

2.22. Поскольку необходима максимизация энтропии (2.86) при условии, что вероятности в сумме равны единице, получаем

$$\sum_i p(x_i) = 1. \quad (45)$$

Для обеспечения этого ограничения введем множитель Лагранжа λ и, таким образом, получим максимизацию

$$-\sum_i p(x_i) \ln p(x_i) + \lambda \left(\sum_i p(x_i) - 1 \right). \quad (46)$$

Устанавливая производную по $p(x_i)$ равной нулю, получаем

$$-\ln p(x_i) - 1 + \lambda = 0. \quad (47)$$

Решая уравнение для $p(x_i)$, получаем

$$p(x_i) = \exp(-1 + \lambda). \quad (48)$$

Поскольку правая часть не зависит от i , это указывает на то, что все вероятности равны. Из (45) следует, что $p(x_i) = 1/M$. Подставляя этот результат в (2.86), получаем, что значение энтропии при ее максимуме равно $\ln M$.

2.23. Энтропия дискретной переменной x с M состояниями может быть записана в виде

$$H(x) = -\sum_{i=1}^M p(x_i) \ln p(x_i) = \sum_{i=1}^M p(x_i) \ln \frac{1}{p(x_i)}. \quad (49)$$

Функция $\ln(x)$ является вогнутой \curvearrowright , поэтому можно применить неравенство Йенсена в виде (2.102), но с обратным неравенством, так что

$$H(x) \leq \ln \left(\sum_{i=1}^M p(x_i) \ln \frac{1}{p(x_i)} \right) = \ln M. \quad (50)$$

2.24. Получить требуемую функциональную производную можно путем простого анализа. Если же требуется более формальный подход, можно использовать методы, описанные в приложении В. Рассмотрим сначала функционал

$$I[p(x)] = \int p(x) f(x) dx.$$

После небольшой модификации $p(x) \rightarrow p(x) + \epsilon \eta(x)$ получаем

$$I[p(x) + \epsilon \eta(x)] = \int p(x) f(x) dx + \epsilon \int \eta(x) f(x) dx,$$

и, таким образом, из (В.3) следует, что функциональная производная задается формулой

$$\frac{\delta I}{\delta p(x)} = f(x).$$

Аналогичным образом если определить

$$J[p(x)] = \int p(x) \ln p(x) dx,$$

то при небольшом изменении $p(x) \rightarrow p(x) + \epsilon \eta(x)$ получаем

$$\begin{aligned} J[p(x) + \epsilon \eta(x)] &= \int p(x) \ln p(x) dx \\ &+ \epsilon \left\{ \int \eta(x) \ln p(x) dx + \int p(x) \frac{1}{p(x)} \eta(x) dx \right\} + O(\epsilon^2) \end{aligned}$$

и, следовательно,

$$\frac{\delta J}{\delta p(x)} = p(x) + 1.$$

Используя эти два результата, получаем следующий результат для функциональной производной:

$$-\ln p(x) - 1 + \lambda_1 + \lambda_2 x + \lambda_3(x - \mu)^2.$$

После перегруппировки получаем (2.97).

Для устранения множителей Лагранжа подставляем (2.97) в каждое из трех ограничений (2.93), (2.94) и (2.95) по очереди. Решение проще всего получить путем сравнения со стандартной формой гауссовой функции и обращая внимание на то, что результаты

$$\lambda_1 = 1 - \frac{1}{2} \ln(2\pi\sigma^2), \quad (51)$$

$$\lambda_2 = 0, \quad (52)$$

$$\lambda_3 = \frac{1}{2\sigma^2} \quad (53)$$

действительно удовлетворяют трем ограничениям.

Здесь следует отметить, что в вопросе есть опечатка (прим. авторов), и правильно он должен читаться так: «Используйте вариационное исчисление для того, чтобы показать, что стационарная точка функционала, показанного непосредственно перед (1.108), задается в (1.108)».

Многомерную версию этого вывода см. в упражнении 3.8.

2.25. Подставляя правую часть (2.98) в аргумент логарифма правой части (2.91), получаем

$$\begin{aligned} H[x] &= -\int p(x) \ln p(x) dx \\ &= -\int p(x) \left(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x - \mu)^2}{2\sigma^2} \right) dx \\ &= \frac{1}{2} \left(\ln(2\pi\sigma^2) + \frac{1}{\sigma^2} \int p(x) (x - \mu)^2 dx \right) \\ &= \frac{1}{2} (\ln(2\pi\sigma^2) + 1), \end{aligned}$$

где на последнем шаге используется (2.95).

2.26. Расхождение Кульбака–Лейблера имеет вид

$$KL(p||q) = -\int p(x) \ln q(x) dx + \int p(x) \ln p(x) dx.$$

Подставляя функцию гауссова распределения вместо $q(\mathbf{x})$, получаем

$$\begin{aligned} \text{KL}(p\|q) &= -\int p(\mathbf{x}) \left\{ -\frac{1}{2} \ln|\Sigma| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} d\mathbf{x} + \text{const} \\ &= \frac{1}{2} \left\{ \ln|\Sigma| + \text{Tr}(\Sigma^{-1} \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]) \right\} + \text{const} \\ &= \frac{1}{2} \left\{ \ln|\Sigma| + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} - 2\boldsymbol{\mu}^T \Sigma^{-1} \mathbb{E}[\mathbf{x}] + \text{Tr}(\Sigma^{-1} \mathbb{E}[\mathbf{x}\mathbf{x}^T]) \right\} + \text{const}. \end{aligned} \quad (54)$$

Дифференцируя это по $\boldsymbol{\mu}$ с использованием результатов из приложения А и приравнявая результат к нулю, получаем

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]. \quad (55)$$

Точно так же, дифференцируя (54) по Σ^{-1} и вновь используя результаты из приложения А, а также (55) и (2.48), получаем

$$\Sigma = \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T = \text{cov}[\mathbf{x}]. \quad (56)$$

2.27. Из (2.100) получаем

$$\text{KL}(p\|q) = -\int p(x) \ln q(x) dx + \int p(x) \ln p(x) dx. \quad (57)$$

Используя (2.49) и (2.51)–(2.53), можно переписать первый интеграл в правой части (57) как

$$\begin{aligned} -\int p(x) \ln q(x) dx &= \int \mathcal{N}(x|\mu, \sigma^2) \frac{1}{2} \left(\ln(2\pi s^2) + \frac{(x-m)^2}{s^2} \right) dx \\ &= \frac{1}{2} \left(\ln(2\pi s^2) + \frac{1}{s^2} \int \mathcal{N}(x|\mu, \sigma^2) (x^2 - 2xm + m^2) dx \right) \\ &= \frac{1}{2} \left(\ln(2\pi s^2) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2} \right). \end{aligned} \quad (58)$$

Второй интеграл в правой части (57) получаем из (2.91) как отрицательную дифференциальную энтропию гауссовой функции. Таким образом, из (57), (58) и (2.99) получаем

$$\begin{aligned} \text{KL}(p\|q) &= \frac{1}{2} \left(\ln(2\pi s^2) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2} - 1 - \ln(2\pi\sigma^2) \right) \\ &= \frac{1}{2} \left(\ln\left(\frac{s^2}{\sigma^2}\right) + \frac{\sigma^2 + \mu^2 - 2\mu m + m^2}{s^2} - 1 \right). \end{aligned}$$