

УДК 004.048

ББК 32.972

М97

Кэвин П. Мэрфи

М97 Вероятностное машинное обучение. Дополнительные темы. Том 1 / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2024. – 810 с.: ил.

ISBN 978-5-93700-120-7

Учебник повышенного типа для научных работников и аспирантов, специализирующихся в области машинного обучения и статистики и желающих глубже познакомиться с глубоким обучением, байесовским выводом, порождающими (генеративными) моделями и принятием решений в условиях неопределенности.

Дополняя ранее изданную книгу «Вероятностное машинное обучение. Введение», этот учебник более высокого уровня знакомит научных работников и аспирантов с деталями самых актуальных теорий и методов машинного обучения, включая глубокие порождающие модели, графовые модели, байесовский вывод, обучение с подкреплением и каузальность. В этом томе глубокое обучение излагается в контексте более широкого статистического контекста, а подходы к глубокому обучению унифицированы с подходами к вероятностному моделированию и выводу. Отдельные части книги написаны ведущими исследователями и специалистами в предметной области из таких организаций, как Google, DeepMind, Amazon, университет Пердью, Нью-Йоркский университет и Вашингтонский университет; в частности, по этой причине книга крайне важна для понимания животрепещущих проблем машинного обучения.

Рассматривается порождение многомерных выходов, например изображений, текста и графов. Обсуждаются методы проникновения в существо данных, основанные на моделях с латентными величинами. Уделено внимание обучению и тестированию при различных распределениях. Исследуется, как использовать вероятностные модели и вывод для каузального вывода и принятия решений.

На сайте книги имеется код на Python.

УДК 004.048

ББК 32.972

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 (анг.) 978-0-26204-843-9

ISBN (рус.) 978-5-93700-120-7

© 2023 Kevin P. Murphy

© Оформление, издание, перевод, ДМК Пресс, 2024

Оглавление

Предисловие от издательства	27
Предисловие.....	28
Соавторы	29
Прочие соавторы	30
Об обложке	30
Глава 1. Введение.....	31
ЧАСТЬ I. ОСНОВАНИЯ.....	35
Глава 2. Вероятность.....	37
2.1. Введение	37
2.1.1. Пространство вероятностей	37
2.1.2. Дискретные случайные величины	37
2.1.3. Непрерывные случайные величины	38
2.1.4. Аксиомы вероятностей.....	39
2.1.5. Условная вероятность	40
2.1.6. Формула Байеса.....	40
2.2. Некоторые распространенные распределения вероятностей.....	41
2.2.1. Дискретные распределения	41
2.2.1.1. Распределение Бернулли и биномиальное распределение	41
2.2.1.2. Категориальное и мультиномиальное распределения	42
2.2.1.3. Распределение Пуассона	42
2.2.1.4. Отрицательное биномиальное распределение	42
2.2.2. Непрерывные распределения на \mathbb{R}	43
2.2.2.1. Гауссово (нормальное) распределение	43
2.2.2.2. Полунормальное распределение	44
2.2.2.3. t-распределение Стьюдента	44
2.2.2.4. Распределение Коши	44
2.2.2.5. Распределение Лапласа	45
2.2.2.6. Субгауссово и супергауссово распределения	45
2.2.3. Непрерывные распределения на \mathbb{R}^+	46
2.2.3.1. Гамма-распределение.....	46
2.2.3.2. Экспоненциальное распределение.....	47
2.2.3.3. Распределение хи-квадрат.....	47
2.2.3.4. Обратное гамма-распределение.....	48
2.2.3.5. Распределение Парето.....	48
2.2.4. Непрерывные распределения на отрезке $[0, 1]$	50
2.2.4.1. Бета-распределение.....	50

2.2.5. Многомерные непрерывные распределения.....	50
2.2.5.1. Многомерное нормальное (гауссово) распределение.....	50
2.2.5.2. Многомерное распределение Стьюдента.....	50
2.2.5.3. Круговое нормальное (фон Физеса–Фишера) распределение ...	51
2.2.5.4. Матричное нормальное распределение (MN).....	51
2.2.5.5. Распределение Уишарта.....	52
2.2.5.6. Обратное распределение Уишарта.....	52
2.2.5.7. Распределение Дирихле.....	53
2.3. Гауссовы совместные распределения.....	55
2.3.1. Многомерное нормальное распределение.....	55
2.3.1.1. Определение.....	55
2.3.1.2. Гауссовы оболочки.....	56
2.3.1.3. Маргинальные и условные распределения для MVN.....	58
2.3.1.4. Информационная (каноническая) форма.....	58
2.3.1.5. Вывод: моментная форма.....	59
2.3.1.6. Вывод: информационная форма.....	61
2.3.2. Линейные гауссовы системы.....	62
2.3.2.1. Совместное распределение.....	62
2.3.2.2. Апостериорное распределение (формула Байеса для гауссовых распределений).....	63
2.3.2.3. Пример: объединение показаний датчиков с известным шумом измерений.....	64
2.3.3. Общий математический анализ гауссовых систем.....	65
2.3.3.1. Моментная и каноническая параметризация.....	65
2.3.3.2. Умножение и деление.....	66
2.3.3.3. Маргинализация.....	66
2.3.3.4. Обусловливание фактами.....	67
2.3.3.5. Преобразование линейно-гауссова условного распределения вероятностей в канонический потенциал.....	67
2.3.3.6. Пример: произведение гауссовых распределений.....	68
2.4. Экспоненциальное семейство.....	68
2.4.1. Определение.....	69
2.4.2. Примеры.....	70
2.4.2.1. Распределение Бернулли.....	70
2.4.2.2. Категориальное распределение.....	71
2.4.2.3. Одномерное гауссово распределение.....	72
2.4.2.4. Одномерное гауссово распределение с фиксированной дисперсией.....	72
2.4.2.5. Многомерное гауссово распределение.....	73
2.4.2.6. Примеры противоположного свойства.....	74
2.4.3. Логарифмическая функция разбиения является производящей функцией кумулянтов.....	74
2.4.3.1. Вывод среднего.....	75
2.4.3.2. Вывод дисперсии.....	75
2.4.3.3. Связь с информационной матрицей Фишера.....	76
2.4.4. Канонические (натуральные) и средние (моментные) параметры.....	76
2.4.5. Оценка максимального правдоподобия для экспоненциального семейства.....	77

2.4.6. Экспоненциальное дисперсионное семейство	78
2.4.7. Вывод максимальной энтропии экспоненциального семейства.....	78
2.5. Преобразования случайных величин	79
2.5.1. Обратимые преобразования (биекции)	80
2.5.2. Аппроксимация Монте-Карло	80
2.5.3. Интегральное преобразование вероятности	81
2.6. Марковские цепи	82
2.6.1. Параметризация	83
2.6.1.1. Марковские переходные ядра	83
2.6.1.2. Марковские матрицы переходов	83
2.6.1.3. Марковские модели высшего порядка	84
2.6.2. Приложение: языковое моделирование.....	85
2.6.3. Оценивание параметров	85
2.6.3.1. Оценка максимального правдоподобия	85
2.6.3.2. Проблема разреженных данных	86
2.6.3.3. Оценка апостериорного максимума	87
2.6.4. Стационарное распределение марковской цепи.....	87
2.6.4.1 Что такое стационарное распределение?.....	88
2.6.4.2. Вычисление стационарного распределения	89
2.6.4.3. Когда существует стационарное распределение?.....	89
2.6.4.4. Детальный баланс	91
2.7. Меры расхождения распределений вероятностей.....	92
2.7.1. f-расхождение.....	92
2.7.1.1. Расхождение КЛ.....	93
2.7.1.2. Альфа-расхождение	93
2.7.1.3. Расстояние Хеллингера	93
2.7.1.4. Расстояние хи-квадрат.....	94
2.7.2. Интегральные вероятностные метрики	94
2.7.3. Максимальное среднее расхождение (MCP).....	95
2.7.3.1. MCP как ИВМ	95
2.7.3.2. Вычисление MCP с помощью ядерного трюка	96
2.7.3.3. Вычисление за линейное время	96
2.7.3.4. Выбор подходящего ядра.....	97
2.7.4. Расстояние полной вариации	97
2.7.5. Оценка отношения плотностей с помощью бинарных классификаторов	99

Глава 3. Статистика..... 101

3.2. Байесовская статистика	101
3.2.1. Подбрасывание монеты	102
3.2.1.1. Правдоподобие.....	102
3.2.1.2. Априорное распределение	102
3.2.1.3. Апостериорное распределение	103
3.2.1.4. Апостериорная мода (оценка MAP)	104
3.2.1.5. Апостериорное среднее	105
3.2.1.6. Апостериорная дисперсия.....	105
3.2.1.7. Байесовские доверительные интервалы	106

3.2.1.8. Апостериорное предсказательное распределение	107
3.2.1.9. Предельное правдоподобие	108
3.2.2. Моделирование более сложных данных.....	109
3.2.3. Выбор априорного распределения	110
3.2.4. Вычислительные проблемы	111
3.2.5. Перестановочность и теорема де Финетти	111
3.3. Частотная статистика	112
3.3.1. Выборочные распределения	112
3.3.2. Бутстрэпная аппроксимация выборочного распределения	113
3.3.3. Асимптотическая нормальность выборочного распределения MLE	115
3.3.4. Информационная матрица Фишера	115
3.3.4.1. Определение.....	115
3.3.4.2. Эквивалентность информационной матрицы Фишера и гессiana отрицательного логарифмического правдоподобия	116
3.3.4.3. Пример: FIM для биномиального распределения	117
3.3.4.4. Пример: FIM для одномерного гауссова распределения	118
3.3.4.5. Пример: FIM для логистической регрессии	118
3.3.4.6. FIM для экспоненциального семейства.....	119
3.3.5. Противоречащие интуиции свойства частотной статистики.....	120
3.3.5.1. Доверительные интервалы.....	120
3.3.5.2. p-значения.....	121
3.3.5.3. Обсуждение	123
3.3.6. Почему не все исповедуют байесовский подход?.....	123
3.4. Сопряженные априорные распределения	125
3.4.1. Биномиальная модель	125
3.4.2. Мультиномиальная модель	125
3.4.3. Одномерная гауссова модель	126
3.4.3.1. Апостериорное μ при заданном σ^2	126
3.4.3.2. Апостериорное σ^2 при заданном μ	128
3.4.3.3. Апостериорное μ и σ^2 : сопряженное априорное распределение	130
3.4.3.4. Апостериорные μ и σ^2 : неинформативное априорное распределение	131
3.4.4. Многомерная гауссова модель	132
3.4.4.1. Апостериорное μ при заданной Σ	132
3.4.4.2. Апостериорная Σ при заданном μ	133
3.4.4.3. Апостериорные Σ и μ	134
3.4.5. Модель их экспоненциального семейства	139
3.4.5.1. Правдоподобие.....	139
3.4.5.2. Априорное распределение	139
3.4.5.3. Апостериорное распределение	139
3.4.5.4. Предельное правдоподобие	140
3.4.5.5. Апостериорное предсказательное распределение	140
3.4.5.6. Пример: распределение Бернулли.....	140
3.4.6. За пределами сопряженных пар	141
3.4.6.1. Смеси сопряженных априорных распределений	142

3.4.6.2. Робастные (с тяжелыми хвостами) априорные распределения	143
3.4.6.3. Априорные распределения для скалярных дисперсий	144
3.4.6.4. Априорные распределения для ковариационных матриц	144
3.5. Неинформативные априорные распределения	146
3.5.1. Априорные распределения с максимальной энтропией	146
3.5.2. Априорные распределения Джеффриса	147
3.5.2.1. Априорное распределение Джеффриса для биномиального распределения	148
3.5.2.2. Априорное распределение Джеффриса для мультиномиального распределения	149
3.5.2.3. Априорное распределение Джеффриса для среднего и дисперсии одномерного гауссова распределения	149
3.5.3. Инвариантные априорные распределения	150
3.5.3.1. Трансляционно-инвариантные априорные распределения	150
3.5.3.2. Масштабно-инвариантное априорное распределение	150
3.5.3.3. Обучение инвариантных априорных распределений	151
3.5.4. Референтные априорные распределения	151
3.6. Иерархические априорные распределения	152
3.6.1. Иерархическая биномиальная модель	153
3.6.1.1. Вывод апостериорного распределения	154
3.6.1.2. Пример: набор данных о крысах	154
3.6.2. Иерархическая гауссова модель	155
3.6.2.1. Пример: набор данных о восьми школах	156
3.6.2.2. Нецентрированная параметризация	157
3.6.3. Иерархические условные модели	158
3.7. Эмпирический байесовский анализ	159
3.7.1. Эмпирический байесовский анализ для иерархической биномиальной модели	159
3.7.2. Эмпирический байесовский анализ для иерархической гауссовой модели	160
3.7.3. Эмпирический байесовский анализ для марковской модели (n -граммное сглаживание)	161
3.7.4. Эмпирический байесовский анализ для несопряженных моделей	164
3.8. Выбор модели	164
3.8.1. Байесовский выбор модели	164
3.8.1.1. Пример: симметрична ли монета?	165
3.8.2. Байесовское усреднение моделей	166
3.8.3. Оценивание предельного правдоподобия	166
3.8.3.1. Аналитическое решение для сопряженных моделей	167
3.8.3.2. Оценка гармонического среднего	167
3.8.3.3. Другие методы Монте-Карло	167
3.8.3.4. Вариационный байесовский анализ	167
3.8.4. Связь между перекрестной проверкой и предельным правдоподобием	168
3.8.5. Условное предельное правдоподобие	169
3.8.6. Байесовская оценка с исключением по одному (LOO)	170

3.8.7. Информационные критерии	171
3.8.7.1. Минимальная длина описания (MDL).....	172
3.8.7.2. Байесовский информационный критерий (BIC).....	172
3.8.7.3. Информационный критерий Акаике	173
3.8.7.4. Широко применимый информационный критерий (WAIC) ...	173
3.9. Проверка модели.....	174
3.9.1. Проверки апостериорного предсказательного распределения.....	174
3.9.1.1. Пример: одномерное гауссово распределение	175
3.9.1.2. Пример: линейная регрессия	176
3.9.2. Байесовские p -значения.....	177
3.10. Проверка гипотез	178
3.10.1. Частотный подход	178
3.10.2. Байесовский подход.....	179
3.10.2.1. Подход на основе сравнения моделей	179
3.10.2.2. Несобственные априорные распределения приводят к проблемам с коэффициентами Байеса	179
3.10.2.3. Подход на основе оценивания параметров	180
3.10.2.4. Одновыборочный критерий доли (биномиальный критерий)	181
3.10.2.5. Двухвыборочный критерий относительных долей (критерий χ^2)	181
3.10.2.6. Одновыборочный критерий среднего (t-критерий).....	181
3.10.2.7. Парный выборочный критерий относительных средних (парный t-критерий)	182
3.10.2.8. Двухвыборочный критерий относительных средних (двухвыборочный t-критерий)	183
3.10.2.9. Проверка коэффициента корреляции	183
3.10.3. Распространенные статистические критерии соответствуют выводу в линейных моделях.....	184
3.10.3.1. Аппроксимация непараметрических критериев с применением преобразования рангов	185
3.10.3.2. Предсказанная величина на одной или двух группах (t-критерий)	185
3.10.3.3. Предсказанная величина с метрическими предикторами (корреляционный критерий)	186
3.10.3.4. Предсказанная величина с одним номинальным предиктором (односторонний ANOVA)	186
3.10.3.5. Предсказанная величина с несколькими номинальными предикторами (многосторонний ANOVA).....	188
3.10.3.6. Предсказанная по счетчикам величина с номинальными предикторами (критерий χ^2).....	188
3.10.3.7. Неметрические предсказанные величины.....	189
3.11. Отсутствие данных	189

Глава 4. Графовые модели..... 191

4.1. Введение	191
---------------------	-----

4.2. Ориентированные графовые модели (байесовские сети).....	191
4.2.1. Представление совместного распределения	191
4.2.2. Примеры	192
4.2.2.1. Марковские цепи	192
4.2.2.2. «Студенческая» сеть.....	193
4.2.2.3. Сигмоидные сети доверия	195
4.2.3. Гауссовы байесовские сети	196
4.2.4. Свойства условной независимости.....	197
4.2.4.1. Глобальные марковские свойства (d-разделение).....	197
4.2.4.2. Оправдание (парадокс Берксона)	200
4.2.4.3. Марковские одеяла	201
4.2.4.4. Другие марковские свойства.....	202
4.2.5. Генерирование (выборка).....	203
4.2.6. Вывод	203
4.2.6.1. Пример: вывод в студенческой сети.....	204
4.2.7. Обучение	204
4.2.7.1. Обучение на неполных данных.....	205
4.2.7.2. Пример: вычисление оценки MLE для СРТ	206
4.2.7.3. Пример: вычисление апостериорного распределения для СРТ.....	208
4.2.7.4. Обучение на неполных данных.....	208
4.2.7.5. Применение EM-алгоритма для аппроксимации СРТ в случае неполных данных	209
4.2.7.6. Использование СГС для аппроксимации СРТ в случае неполных данных	210
4.2.8. Блочная нотация	211
4.2.8.1. Пример: факторный анализ	212
4.2.8.2. Пример: наивный байесовский классификатор	213
4.2.8.3. Пример: ослабление наивного байесовского предположения	213
4.3. Неориентированные графовые модели (марковские случайные поля)	214
4.3.1. Представление совместного распределения	215
4.3.1.1. Теорема Хаммерсли–Клиффорда	215
4.3.1.2. Распределение Гиббса.....	216
4.3.2. Полностью видимые MRF (Айзинга, Поттса, Хопфилда и т. д.)	216
4.3.2.1. Модели Айзинга	216
4.3.2.2. Модели Поттса.....	219
4.3.2.3. Модели Поттса для предсказания структуры белков	220
4.3.2.4. Сети Хопфилда	221
4.3.3. MRF с латентными величинами (машины Больцмана и т. д.)	223
4.3.3.1. Обычные машины Больцмана	223
4.3.3.2. Ограниченные машины Больцмана (RBM)	223
4.3.3.3. Глубокие машины Больцмана	225
4.3.3.4. Глубокие сети доверия (DBN)	225
4.3.4. Модели максимальной энтропии	225
4.3.4.1. Логарифмически-линейные модели	226

4.3.4.2. Индукция признаков для модели правописания с максимальной энтропией	226
4.3.5. Гауссовы MRF	228
4.3.5.1. Стандартные GMRF	228
4.3.5.2. Нелинейные гауссовы MRF	229
4.3.6. Свойства условной независимости	230
4.3.6.1. Основные результаты	230
4.3.6.2. Неориентированная альтернатива d-разделению	231
4.3.7. Генерирование (выборка)	232
4.3.8. Вывод	232
4.3.9. Обучение	234
4.3.9.1. Обучение на неполных данных	234
4.3.9.2. Вычислительные проблемы	235
4.3.9.3. Оценка максимального псевдоправдоподобия	235
4.3.9.4. Обучение на неполных данных	237
4.4. Условные случайные поля (CRF)	238
4.4.1. Одномерные CRF	238
4.4.1.1. Выделение именных групп	239
4.4.1.2. Распознавание именованных сущностей	240
4.4.1.3. Грамматический разбор естественного языка	241
4.4.2. Двумерные CRF	242
4.4.2.1. Семантическая сегментация	242
4.4.2.2. Модели деформируемых частей	243
4.4.3. Оценивание параметров	245
4.4.3.1. Логарифмически-линейные потенциалы	245
4.4.3.2. Общий случай	246
4.4.4. Другие подходы к структурному предсказанию	246
4.5. Сравнение ориентированных и неориентированных ВГМ	247
4.5.1. Свойства УН	247
4.5.2. Преобразование между ориентированной и неориентированной моделями	248
4.5.2.1. Преобразование ОВГМ в НВГМ	248
4.5.2.2. Преобразование НВГМ в ОВГМ	249
4.5.3. Условные ориентированные и неориентированные ВГМ и проблема смещения метки	250
4.5.4. Комбинирование ориентированных и неориентированных графов	251
4.5.4.1. Цепные графы	251
4.5.4.2. Ациклические ориентированные смешанные графы	252
4.5.5. Сравнение ориентированных и неориентированных гауссовых ВГМ	253
4.5.5.1. Ковариационные графы	254
4.6. Расширения ВГМ	255
4.6.1. Фактор-графы	255
4.6.1.1. Двудольные фактор-графы	255
4.6.1.2. Фактор-графы Форни	257
4.6.2. Вероятностные схемы	258

4.6.3. Ориентированные реляционные ВГМ.....	259
4.6.4. Неориентированные реляционные ВГМ.....	262
4.6.4.1. Коллективная классификация.....	262
4.6.4.2. Марковские логические сети.....	262
4.6.5. Вероятностные модели с открытым универсумом	265
4.6.6. Программы как вероятностные модели.....	265
4.7. Структурные каузальные модели.....	266
4.7.1. Пример: причинная связь между образованием и богатством	267
4.7.2. Модели структурных уравнений	268
4.7.3. Оператор do и дополненные ОАГ.....	269
4.7.4. Контрфактические вопросы	270

Глава 5. Теория информации 273

5.1. Расхождение КЛ	273
5.1.1. Желательные свойства.....	274
5.1.2. Расхождение КЛ – единственная мера, обладающая желательными свойствами	276
5.1.2.1. Непрерывность расхождения КЛ	276
5.1.2.2. Неотрицательность расхождения КЛ.....	276
5.1.2.3. Расхождение КЛ инвариантно относительно перепараметризации	277
5.1.2.4. Монотонность для равномерных распределений	278
5.1.2.5. Цепное правило для расхождения КЛ	278
5.1.3. Размышления о расхождении КЛ	278
5.1.3.1. Единицы измерения расхождения КЛ.....	279
5.1.3.2. Асимметрия расхождения КЛ.....	279
5.1.3.3. Расхождение КЛ как ожидаемый вес свидетельства.....	280
5.1.4. Минимизация расхождения КЛ	280
5.1.4.1. Прямое и обратное расхождения КЛ	280
5.1.4.2. Моментная проекция (покрытие мод)	281
5.1.4.3. Информационная проекция (поиск мод).....	282
5.1.5. Свойства расхождения КЛ	283
5.1.5.1. Лемма о сжатии	283
5.1.5.2. Неравенство обработки данных для расхождения КЛ.....	284
5.1.6. Расхождение КЛ и оценка MLE.....	285
5.1.7. Расхождение КЛ и байесовский вывод	286
5.1.8. Расхождение КЛ и экспоненциальные семейства	287
5.1.8.1. Пример: расхождение КЛ между двумя гауссовыми распределениями	288
5.1.9. Аппроксимация расхождения КЛ информационной матрицей Фишера	288
5.1.10. Расхождение Брегмана	289
5.1.10.1. Расхождение КЛ – частный случай расхождения Брегмана	290
5.2. Энтропия	290
5.2.1. Определение.....	290
5.2.2. Дифференциальная энтропия для непрерывных случайных величин	291

5.2.3. Типичные множества.....	293
5.2.4. Перекрестная энтропия и перплексия	293
5.3. Взаимная информация.....	294
5.3.1. Определение.....	294
5.3.2. Интерпретация.....	294
5.3.3. Неравенство обработки данных.....	295
5.3.4. Достаточные статистики	296
5.3.5. Многомерная взаимная информация	297
5.3.5.1. Полная корреляция	297
5.3.5.2. Информация о взаимодействии (коинформация)	297
5.3.5.3. Синергия и избыточность	298
5.3.5.4. МВИ и каузальность	299
5.3.5.5. МВИ и энтропия	299
5.3.6. Вариационные границы взаимной информации.....	300
5.3.6.1. Верхняя граница	300
5.3.6.2. Нижняя граница БА.....	300
5.3.6.3. Нижняя оценка НУД.....	301
5.3.6.4. Нижняя оценка InfoNCE	301
5.3.7. Сети релевантности	302
5.4. Сжатие данных (кодирование источника)	303
5.4.1. Сжатие без потери информации.....	304
5.4.2. Сжатие с потерей информации и компромисс между скоростью и искажением	304
5.4.3. Кодирование с возвратом битов	307
5.5. Коды с исправлением ошибок (кодирование канала).....	308
5.6. Информационное бутылочное горлышко.....	309
5.6.1. Простое информационное бутылочное горлышко.....	310
5.6.2. Вариационное информационное бутылочное горлышко.....	311
5.6.3. Условное энтропийное бутылочное горлышко.....	313
Глава 6. Оптимизация.....	315
6.1. Введение	315
6.2. Автоматическое дифференцирование	315
6.2.1. Дифференцирование в функциональной форме	315
6.2.2. Дифференцирование цепочек, контуров и программ	320
6.2.2.1. Цепные композиции и правило дифференцирования сложной функции	320
6.2.2.2. От цепочек к контурам	322
6.2.2.3. От контуров к программам	325
6.3. Стохастическая оптимизация	326
6.3.1. Стохастический градиентный спуск.....	326
6.3.1.1. Выбор величины шага	326
6.3.1.2. Уменьшение дисперсии.....	327
6.3.1.3. Предобусловленный СГС	327
6.3.2. Применение СГС для оптимизации целевой функции в виде конечной суммы	328
6.3.3. Применение СГС для оптимизации параметров распределения	328

6.3.4. Оценка на основе функции вклада (REINFORCE)	329
6.3.4.1. Управляющие вариаты	330
6.3.4.2. Преобразование Рао–Блэкуэлла	330
6.3.5. Прием перепараметризации.....	330
6.3.5.1. Пример.....	331
6.3.5.2. Полная производная	332
6.3.5.3. Оценка приземления	332
6.3.6. Прием Gumbel-softmax	333
6.3.7. Стохастические графы вычислений.....	334
6.3.8. Сквозная оценка	334
6.4. Натуральный градиентный спуск.....	335
6.4.1. Определение натурального градиента.....	336
6.4.2. Интерпретации НГС.....	337
6.4.2.1. НГС как метод доверенной области.....	337
6.4.2.2. НГС как метод Гаусса–Ньютона	337
6.4.3. Преимущества НГС	337
6.4.4. Аппроксимация натурального градиента.....	339
6.4.5. Натуральные градиенты для экспоненциального семейства.....	340
6.4.5.1. Аналитическое вычисление для гауссова случая.....	341
6.4.5.2. Стохастическая аппроксимация в общем случае	342
6.4.5.3. Натуральный градиент функции энтропии	342
6.5. Алгоритмы ограниченной оптимизации.....	343
6.5.1. Общий алгоритм	343
6.5.2. Пример: логистическая регрессия	344
6.5.3. EM-алгоритм	346
6.5.3.1. Нижняя граница	346
6.5.3.2. E-шаг	347
6.5.3.3. M-шаг	348
6.5.4. Пример: применение EM-алгоритма к многомерному нормальному распределению с неполными данными.....	348
6.5.4.1. E-шаг	349
6.5.4.2. M-шаг	350
6.5.4.3. Инициализация.....	350
6.5.4.4. Пример.....	350
6.5.5. Пример: робастная линейная регрессия с использованием правдоподобия Стьюдента	351
6.5.6. Расширения EM	352
6.5.6.1. Вариационный EM-алгоритм.....	352
6.5.6.2. Жесткий EM-алгоритм.....	352
6.5.6.3. EM-алгоритм Монте-Карло	353
6.5.6.4. Обобщенный EM-алгоритм.....	353
6.5.6.5. ESM-алгоритм	354
6.5.6.6. Онлайнный EM-алгоритм.....	354
6.6. Байесовская оптимизация.....	355
6.6.1. Последовательная оптимизация на основе модели	355
6.6.2. Суррогатные функции	357
6.6.2.1. Гауссовские процессы	357
6.6.2.2. Байесовские нейронные сети.....	357
6.6.2.3. Другие модели	357

6.6.3. Функции сбора	358
6.6.3.1. Вероятность улучшения.....	358
6.6.3.2. Ожидаемое улучшение	358
6.6.3.3. Верхняя доверительная граница.....	359
6.6.3.4. Выборка Томпсона	359
6.6.3.5. Энтропийный поиск	359
6.6.3.6. Градиент знания.....	360
6.6.3.7. Оптимизация функции сбора.....	360
6.6.4. Прочие проблемы	361
6.6.4.1. Параллельные (пакетные) запросы	361
6.6.4.2. Условные параметры	361
6.6.4.3. Многоточностные суррогаты	362
6.6.4.4. Ограничения	362
6.7. Оптимизация без вычисления производных.....	362
6.7.1. Локальный поиск.....	363
6.7.1.1. Стохастический локальный поиск	363
6.7.1.2. Поиск с запретами.....	364
6.7.1.3. Случайный поиск	365
6.7.2. Имитация отжига	366
6.7.3. Эволюционные алгоритмы.....	366
6.7.4. Алгоритмы оценки распределения	369
6.7.5. Метод перекрестной энтропии	371
6.7.5.1. Дифференцируемый СЕМ.....	371
6.7.6. Эволюционные стратегии.....	372
6.7.6.1. Натуральные эволюционные стратегии	372
6.7.6.2. CMA-ES	372
6.8. Оптимальная транспортировка	372
6.8.1. Разминка: оптимальное паросочетание двух семейств точек	373
6.8.2. От оптимальных паросочетаний к формулировкам Канторовича и Монжа	374
6.8.2.1. Расщепление по массе	374
6.8.2.2. Формулировка Монжа и оптимальные отображения дифференциала	375
6.8.2.3. Формулировка Канторовича	376
6.8.2.4. Расстояния Вассерштейна	377
6.8.3. Решение задачи об оптимальной транспортировке	377
6.8.3.1. Двойственность и вогнутость стоимости.....	377
6.8.3.2. Двойственность Канторовича–Рубинштейна и потенциалы Липшица.....	378
6.8.3.3. Отображения Монжа как градиенты выпуклых функций: теорема Бренье.....	378
6.8.3.4. Решения в замкнутой форме для одномерных и гауссовых распределений	380
6.8.3.5. Точное вычисление с помощью решателей линейных программ	381
6.8.3.6. Обеспечение гладкости с помощью энтропийной регуляризации	382

6.9. Субмодулярная оптимизация	383
6.9.1. Интуитивные соображения, пример и подоплёка.....	384
6.9.1.1. Кофе, лимон, молоко и чай.....	384
6.9.2. Основные определения субмодулярности	386
6.9.3. Примеры субмодулярных функций.....	388
6.9.4. Субмодулярная оптимизация	391
6.9.4.1. Субмодулярная максимизация	392
6.9.4.2. Дискретные ограничения.....	393
6.9.4.3. Минимизация субмодулярной функции.....	395
6.9.5. Приложения субмодулярности в машинном обучении и ИИ	396
6.9.6. Эскизы, опорные множества, дистилляция и отбор признаков и подмножеств данных	396
6.9.6.1. Варианты проектирования алгоритма обобщения	398
6.9.7. Комбинаторные информационные функции	401
6.9.8. Кластеризация, разбиение данных и параллельное машинное обучение	402
6.9.9. Активное обучение и обучение с частичным привлечением учителя	403
6.9.10. Вероятностное моделирование.....	405
6.9.11. Структурные нормы и функции потерь	406
6.9.12. Заключительные замечания.....	407

ЧАСТЬ II. ВЫВОД..... 409

Глава 7. Алгоритмы вывода: общий обзор..... 411

7.1. Введение	411
7.2. Типичные схемы вывода	412
7.2.1. Глобальные латентные величины	412
7.2.2. Локальные латентные величины	413
7.2.3. Глобальные и локальные латентные величины	413
7.3. Точные алгоритмы вывода	414
7.4. Приближенные алгоритмы вывода.....	415
7.4.1. Аппроксимация MAP и свойственные ей проблемы	415
7.4.1.1. Оценка MAP не дает меры неопределенности.....	415
7.4.1.2. Оценка MAP часто не дает правильного представления об апостериорном распределении.....	416
7.4.1.3. Оценка MAP не инвариантна относительно перепараметризации	416
7.4.2. Сеточная аппроксимация	417
7.4.3. Аппроксимация Лапласа (квадратичная).....	418
7.4.4. Вариационный вывод	419
7.4.5. Метод Монте-Карло по схеме марковской цепи.....	420
7.4.6. Последовательный метод Монте-Карло	422
7.4.7. Сложные апостериорные распределения	423
7.5. Оценка приближенных алгоритмов вывода	423

Глава 8. Гауссова фильтрация и сглаживание	425
8.1. Введение	425
8.1.1. Цели вывода	425
8.1.2. Уравнения байесовской фильтрации.....	427
8.1.3. Уравнения байесовского сглаживания	428
8.1.4. Гауссов подход.....	429
8.2. Вывод для линейных гауссовых SSM	430
8.2.1. Примеры	430
8.2.1.1. Прослеживание и оценивание состояний	430
8.2.1.2. Онлайн-байесовская линейная регрессия (рекурсивный метод наименьших квадратов).....	431
8.2.1.3. Предсказание временных рядов	431
8.2.2. Фильтр Калмана	431
8.2.2.1. Шаг предсказания	432
8.2.2.2. Шаг обновления	432
8.2.2.3. Апостериорное предсказательное распределение	432
8.2.2.4. Вывод	433
8.2.2.5. Абстрактная формулировка	434
8.2.2.6. Численные проблемы	436
8.2.2.7. Версия с непрерывным временем	436
8.2.3. Сглаживатель Калмана	436
8.2.3.1. Алгоритм	437
8.2.3.2. Вывод	437
8.2.3.3. Двухфильтровое сглаживание.....	438
8.2.3.4. Временная и пространственная сложность	439
8.2.3.5. Прямая фильтрация – обратная выборка.....	439
8.2.4. Фильтрация и сглаживание в информационной форме	439
8.2.4.1. Фильтрация: алгоритм	439
8.2.4.2. Фильтрация: вывод.....	440
8.2.4.3. Сглаживание: алгоритм.....	441
8.2.4.4. Сглаживание: вывод	441
8.3. Вывод, основанный на локальной линеаризации	442
8.3.1. Разложение в ряд Тейлора.....	442
8.3.2. Обобщенный фильтр Калмана (ОКФ).....	444
8.3.2.1. Точность.....	444
8.3.2.2. ОКФ с итерациями	445
8.3.2.3. Пример: прослеживание точки, движущейся по спирали на двумерной плоскости	446
8.3.2.4. Пример: обучение нейронной сети	446
8.3.3. Обобщенный сглаживатель Калмана.....	446
8.4. Вывод, основанный на сигма-точечном преобразовании.....	447
8.4.1. Сигма-точечное преобразование	447
8.4.2. Сигма-точечный фильтр Калмана	449
8.4.3. Сигма-точечный сглаживатель Калмана.....	450
8.5. Другие варианты фильтра Калмана.....	450
8.5.1. Обобщенная гауссова фильтрация	450
8.5.1.1. Статистическая линейная регрессия	451

8.5.1.2. Аппроксимация моментов	452
8.5.1.3. Аппроксимация на основе линеаризации	452
8.5.1.4. Аппроксимация на основе гауссовой квадратуры	453
8.5.1.5. Аппроксимация на основе метода Монте-Карло	453
8.5.2. Гауссова фильтрация на основе условных моментов	453
8.5.3. Итерированные фильтры и сглаживатели	455
8.5.4. Ансамблевый фильтр Калмана	456
8.5.5. Робастные фильтры Калмана	458
8.5.6. Двойной ОФК	458
8.6. Фильтрация с предполагаемой плотностью	458
8.6.1. Связь с гауссовой фильтрацией	460
8.6.2. ADF для SLDS (гауссов суммирующий фильтр).....	461
8.6.3. ADF для онлайн-логистической регрессии.....	462
8.6.4. ADF для онлайн-глубоких нейронных сетей.....	466
8.7. Другие методы вывода для SSM	466
8.7.1. Сеточные аппроксимации	466
8.7.2. Распространение математического ожидания	468
8.7.3. Вариационный вывод	468
8.7.4. MCMC	468
8.7.5. Фильтрация частиц	469

Глава 9. Алгоритмы передачи сообщений 470

9.1. Введение	470
9.2. Распространение доверия по цепочкам.....	471
9.2.1. Скрытые марковские модели.....	471
9.2.1.1. Пример: НММ казино	471
9.2.1.2. Вывод апостериорного распределения	472
9.2.2. Алгоритм прямого хода.....	473
9.2.3. Алгоритм прямого-обратного хода	474
9.2.3.1. Рекурсия обратного хода	474
9.2.3.2. Пример.....	475
9.2.3.3. Двухчастные сглаженные маргинальные распределения.....	475
9.2.3.4. Численно устойчивая реализация	476
9.2.4. Прямая фильтрация – обратное сглаживание	477
9.2.5. Временная и пространственная сложность	478
9.2.6. Алгоритм Витерби	478
9.2.6.1. Прямой проход.....	479
9.2.6.2. Обратный проход.....	480
9.2.6.3. Пример.....	480
9.2.6.4. Временная и пространственная сложность	481
9.2.6.5. Список N лучших.....	481
9.2.7. Прямая фильтрация – обратная выборка	482
9.3. Распространение доверия на деревьях	482
9.3.1. Ориентированные и неориентированные деревья.....	482
9.3.2. Алгоритм «сумма–произведение».....	484
9.3.3. Алгоритм «максимум–произведение»	486
9.3.3.1. Связь между МММ и MAP.....	486

9.3.3.2. Связь между MPM и MAP	487
9.3.3.3. Связь между MPE и MAP	488
9.4. Петлевое распространение доверия	488
9.4.1. Петлевое распространение доверия в попарных неориентированных графах	489
9.4.2. Петлевое распространение доверия для фактор-графов	489
9.4.3. Гауссово распространение доверия	490
9.4.4. Сходимость	492
9.4.4.1. Когда LBP сходится?	492
9.4.4.2. Обеспечение сходимости LBP	492
9.4.4.3. Повышение скорости сходимости с помощью адаптивной схемы	494
9.4.5. Точность	495
9.4.6. Обобщенное распространение доверия	495
9.4.7. Выпуклое BP	495
9.4.8. Приложение: коды с исправлением ошибок	496
9.4.9. Приложение: распространение близости	497
9.4.10. Эмуляция BP с помощью графовых нейронных сетей	498
9.5. Алгоритм исключения переменной	500
9.5.1. Вывод алгоритма	500
9.5.2. Вычислительная сложность VE	501
9.5.3. Выбор хорошего порядка исключения	503
9.5.4. Вычислительная сложность точного вывода	504
9.5.5. Недостатки VE	505
9.6. Алгоритм дерева сочленений	506
9.7. Вывод как оптимизация	507
9.7.1. Вывод как обратное распространение	507
9.7.1.1. Пример: вывод в небольшой модели	508
9.7.2. «Шевеление, затем MAP»	508
9.7.2.1. Гауссов случай	509
9.7.2.2. Дискретный случай	510
Глава 10. Вариационный вывод	511
10.1. Введение	511
10.1.1. Вариационная целевая функция	511
10.1.1.1. Физическая интерпретация: минимизация вариационной свободной энергии	512
10.1.1.2. Статистическая интерпретация: максимизация нижней границы свидетельства (ELBO)	513
10.1.2. Форма вариационного апостериорного распределения	513
10.1.3. Оценивание параметров с помощью вариационного EM- алгоритма	515
10.1.3.1. Оценка MLE для моделей с латентными величинами	515
10.1.3.2. Эмпирический байесовский анализ полностью наблюдаемых моделей	516
10.1.4. Стохастический VI	517
10.1.5. Амортизированный VI	517
10.1.6. Полуамортизированный вывод	518

10.2. Градиентный VI	518
10.2.1. Перепараметризованный VI	519
10.2.1.1. Гауссово распределение с диагональной ковариационной матрицей (среднее поле)	520
10.2.1.2. Гауссово распределение с полной ковариационной матрицей	521
10.2.1.3. Гауссово распределение с ковариационной матрицей, равной сумме диагональной матрицы и матрицы низкого ранга	522
10.2.1.4. Другие вариационные апостериорные распределения	523
10.2.1.5. Пример: байесовский вывод параметров	523
10.2.1.6. Пример: оценка MLE для LVM	525
10.2.2. VI с автоматическим дифференцированием	525
10.2.2.1. Основная идея	526
10.2.2.2. Пример: ADVI для бета-биномиальной модели	526
10.2.2.3. Пример: ADVI для GMM	526
10.2.2.4. Более сложные апостериорные распределения	528
10.2.3. Вариационный вывод методом черного ящика	528
10.2.3.1. Оценивание градиента методом REINFORCE	528
10.2.3.2. Уменьшение дисперсии с помощью управляющих вариат.	528
10.3. VI методом покоординатного подъема	529
10.3.1. Вывод алгоритма CAVI	530
10.3.2. Пример: CAVI для модели Айзинга	531
10.3.3. Вариационный байесовский вывод	534
10.3.4. Пример: VB для одномерного гауссова распределения	534
10.3.4.1. Целевое распределение	535
10.3.4.2. Обновление $q(\mu \Psi_\mu)$	535
10.3.4.3. Обновление $q(\lambda \Psi_\lambda)$	535
10.3.4.4. Вычисление математических ожиданий	536
10.3.4.5. Иллюстрация	536
10.3.4.6. Нижняя граница	537
10.3.5. Вариационный байесовский EM-алгоритм	538
10.3.6. Пример: VBEM для GMM	539
10.3.6.1. Вариационное апостериорное распределение	539
10.3.6.2. Вывод $q(\theta)$ (вариационный M-шаг)	539
10.3.6.3. Вывод $q(z)$ (вариационный E-шаг)	541
10.3.6.4. Эффекты VBEM, индуцирующие автоматическую разреженность	542
10.3.6.5. Нижняя граница предельного правдоподобия	543
10.3.6.6. Выбор модели с помощью VBEM	545
10.3.7. Вариационная передача сообщений	545
10.3.8. Autoconj	546
10.4. Более точные вариационные апостериорные распределения	546
10.4.1. Структурное среднее поле	546
10.4.2. Иерархические (со вспомогательными величинами) апостериорные распределения	547
10.4.3. Апостериорные распределения в виде нормализующих потоков	547

10.4.4. Неявные апостериорные распределения	548
10.4.5. Комбинирование VI с MCMC-выводом.....	548
10.5. Более точные границы.....	548
10.5.1. Многовыборочная ELBO (IWAE-граница).....	548
10.5.1.1. Патологии оптимизации IWAE-границы	549
10.5.2. Термодинамическая вариационная целевая функция.....	550
10.5.3. Минимизация верхней границы свидетельства.....	550
10.6. Алгоритм пробуждения–засыпания.....	551
10.6.1. Фаза пробуждения	551
10.6.2. Фаза засыпания.....	552
10.6.3. Фаза сна наяву.....	553
10.6.4. Краткое описание алгоритма	554
10.7. Распространение математического ожидания	554
10.7.1. Алгоритм	555
10.7.2. Пример	556
10.7.3. EP как обобщение ADF	557
10.7.4. Вопросы оптимизации.....	557
10.7.5. Степенное EP и α -расхождение.....	558
10.7.6. Стохастическое EP.....	558
Глава 11. Методы Монте-Карло	560
11.1. Введение	560
11.2. Интегрирование методом Монте-Карло	560
11.2.1. Пример: <i>оценивание π методом Монте-Карло</i>	561
11.2.2. Точность интегрирования методом Монте-Карло	561
11.3. Генерирование случайных выборок из простых распределений.....	563
11.3.1. Выборка с помощью обратной cdf	563
11.3.2. Выборка из гауссова распределения (метод Бокса–Мюллера).....	564
11.4. Выборка с отклонением.....	565
11.4.1. Основная идея	565
11.4.2. Пример.....	566
11.4.3. Адаптивная выборка с отклонением	567
11.4.4. Выборка с отклонением в пространствах высокой размерности.....	567
11.5. Выборка по значимости	568
11.5.1. Прямая выборка по значимости	568
11.5.2. Самонормированная выборка по значимости	569
11.5.3. Выбор вспомогательного распределения	570
11.5.4. Выборка по значимости с отжигом	571
11.5.4.1. Оценивание нормировочных постоянных с использованием AIS	572
11.6. Управление дисперсией оценки Монте-Карло	572
11.6.1. Общие случайные числа	572
11.6.2. Преобразование Рао–Блэквелла	572
11.6.3. Управляющие вариаты	573
11.6.3.1. Пример.....	574

11.6.4. Антитетическая выборка.....	574
11.6.4.1. Пример.....	575
11.6.5. Метод квази-Монте-Карло	575

Глава 12. Метод Монте-Карло по схеме марковской цепи 577

12.1. Введение	577
12.2. Алгоритм Метрополиса–Гастингса.....	578
12.2.1. Основная идея	578
12.2.2. Почему алгоритм МГ работает	579
12.2.3. Вспомогательные распределения	581
12.2.3.1. Независимая выборка.....	581
12.2.3.2. Алгоритм случайного блуждания Метрополиса	581
12.2.3.3. Комбинирование вспомогательных распределений.....	582
12.2.3.4. МСМС с управлением от данных	583
12.2.3.5. Адаптивный МСМС	583
12.2.4. Инициализация	583
12.3. Выборка Гиббса	584
12.3.1. Основная идея	584
12.3.2. Выборка Гиббса – частный случай МГ	585
12.3.3. Пример: выборка Гиббса для моделей Айзинга.....	585
12.3.4. Пример: выборка Гиббса для моделей Поттса	587
12.3.5. Пример: выборка Гиббса для ГММ	587
12.3.5.1. Случай известных параметров.....	588
12.3.5.2. Случай неизвестных параметров.....	588
12.3.6. Метрополис внутри Гиббса.....	590
12.3.7. Блочная выборка Гиббса	590
12.3.8. Свернутая выборка Гиббса	591
12.4. МСМС со вспомогательной величиной	593
12.4.1. Выборка по уровням	593
12.4.2. Алгоритм Свендсена–Ванга	595
12.5. Гамильтонов метод Монте-Карло (НМС).....	597
12.5.1. Гамильтонова механика	597
12.5.2. Интегрирование уравнений Гамильтона	598
12.5.2.1. Метод Эйлера.....	598
12.5.2.2. Модифицированный метод Эйлера.....	599
12.5.2.3. Схема интегрирования с перешагиванием.....	599
12.5.2.4. Схемы интегрирования более высокого порядка	599
12.5.3. Алгоритм НМС.....	599
12.5.4. Настройка НМС	601
12.5.4.1. Выбор числа шагов с помощью алгоритма NUTS	601
12.5.4.2. Выбор размера шага.....	601
12.5.4.3. Выбор ковариационной матрицы (обратных масс)	601
12.5.5. НМС на римановом многообразии	602
12.5.6. Метод Монте-Карло по Ланжевену	602
12.5.7. Связь между СЕС и выборкой Ланжевена.....	603
12.5.8. Применение НМС к ограниченным параметрам	604
12.5.9. Ускорение НМС.....	605

12.6. Сходимость МСМС	605
12.6.1. Скорости перемешивания марковских цепей	606
12.6.2. Практическая диагностика сходимости	607
12.6.2.1. Трассировочные графики	608
12.6.2.2. Оценочное потенциальное уменьшение масштаба (EPSR)	610
12.6.3. Эффективный объем выборки	611
12.6.4. Улучшение скорости сходимости	613
12.6.5. Нецентрированные параметризации и воронка Нила	613
12.7. Стохастический градиентный МСМС	614
12.7.1. Динамика Ланжевена со стохастическим градиентом	615
12.7.2. Предобусловливание	615
12.7.3. Уменьшение дисперсии оценки градиента	616
12.7.4. SG-НМС	617
12.7.5. Недодемпфированная динамика Ланжевена	618
12.8. МСМС методом обратимого прыжка (межпространственный)	618
12.8.1. Основная идея	619
12.8.2. Пример	620
12.8.3. Обсуждение	622
12.9. Методы отжига	622
12.9.1. Имитация отжига	622
12.9.2. Параллельная закалка	625

Глава 13. Последовательный метод Монте-Карло 626

13.1. Введение	626
13.1.1. Постановка задачи	626
13.1.2. Фильтрация частиц для моделей пространства состояний	627
13.1.3. SMC-генераторы выборок для статического оценивания параметров	628
13.2. Фильтрация частиц	629
13.2.1. Выборка по значимости	629
13.2.2. Последовательная выборка по значимости	630
13.2.3. Последовательная выборка по значимости с перевыборкой	632
13.2.3.1. Бутстрэпный фильтр	633
13.2.3.2. Проблема вырождения пути	634
13.2.3.3. Оценка нормировочной постоянной	634
13.2.4. Методы перевыборки	635
13.2.4.1. Обратная CDF	635
13.2.4.2. Мультиномиальная перевыборка	636
13.2.4.3. Стратифицированная перевыборка	637
13.2.4.4. Систематическая перевыборка	637
13.2.4.5. Сравнение	637
13.2.5. Адаптивная перевыборка	637
13.3. Вспомогательные распределения	638
13.3.1. Локально-оптимальное вспомогательное распределение	639
13.3.2. Вспомогательные распределения, основанные на обобщенном и сигма-точечном фильтре Калмана	639
13.3.3. Вспомогательные распределения, основанные на аппроксимации Лапласа	640

13.3.3.1. Пример: нейронное декодирование	640
13.3.4. Вспомогательные распределения, основанные на SMC (вложенный SMC)	642
13.4. Фильтрация частиц с преобразованием Рао–Блэкуэлла (RBPF).....	642
13.4.1. Смесь фильтров Калмана	642
13.4.1.1. Улучшения	644
13.4.2. Пример: слежение за маневрирующим объектом.....	644
13.4.3. Пример: FastSLAM.....	646
13.5. Обобщения фильтра частиц	649
13.6. SMC-генераторы выборок	649
13.6.1. Составные части SMC-генератора выборок	650
13.6.2. Закалка правдоподобий (геометрическая траектория)	651
13.6.2.1. Пример: выборка из одномерного бимодального распределения	652
13.6.3. Закалка данных	653
13.6.3.1. Пример: IBIS для одномерного гауссова распределения	654
13.6.4. Выборка редких событий и экстремумы	655
13.6.5. SMC–ABC и вывод без использования правдоподобия	655
13.6.6. SMC ²	656
13.6.7. SMC с вариационной фильтрацией.....	656
13.6.8. Вариационный сглаживающий SMC.....	657
Предметный указатель	659

Предисловие

Я пишу более длинную [книгу], чем обычно, потому что для написания короткой мне не хватает времени.

– Блез Паскаль, перефразировано

Эта книга является продолжением [Mur22], но различные вопросы машинного обучения (МО) излагаются в ней более глубоко. Предыдущая книга была в основном посвящена методам обучения функций вида $f: \mathcal{X} \rightarrow \mathcal{Y}$, где f – некоторая нелинейная модель, например глубокая нейронная сеть, \mathcal{X} – множество возможных входов (как правило, $\mathcal{X} = \mathbb{R}^D$), а $\mathcal{Y} = \{1, \dots, C\}$ представляет множество меток для задач классификации или $\mathcal{Y} = \mathbb{R}$ для задач регрессии. Джуда Перл, хорошо известный исследователь ИИ, назвал такой вид МО «помпезной аппроксимацией кривых» (цитируется по работе [Har18]).

В этой книге мы расширим область МО на более трудные задачи. Например, мы рассмотрим обучение и тестирование при различных распределениях, порождение многомерных выходов, таких как изображения, текст и графы, когда пространством выходов является, например, $\mathcal{Y} = \mathbb{R}^{256 \times 256}$; мы обсудим обнаружение «внутренней сущности» данных на основе моделей латентных переменных, а также применение вероятностных моделей к каузальному выводу и принятие решений в условиях неопределенности.

Мы предполагаем, что читатель в какой-то мере знаком с МО и другими относящимися к делу математическими дисциплинами (например, теорией вероятностей, статистикой, линейной алгеброй, оптимизацией). Этот вводный материал можно найти в предшествующей книге [Mur22], равно как и еще в нескольких хороших книгах (например, [Lin+21b; DFO20]).

Написанный на Python код (в основном с применением JAX) для построения почти всех рисунков можно найти в Сети. В частности, если в подрисуночной подписи сказано «Построено программой `gauss_plot_2d.ipynb`», то следует искать соответствующий Jupyter-блокнот по адресу probml.github.io/notebooks#gauss_plot_2d.ipynb. Щелчок по ссылке на рисунок в pdf-версии книги откроет список таких блокнотов. Щелчок по ссылке на блокнот откроет рисунок в программе Google Colab, которая позволит легко воспроизвести рисунок самостоятельно и модифицировать исходный код, чтобы лучше понять методы. (Colab дает доступ к бесплатному GPU, что может быть полезно для некоторых вычислительно трудоемких демонстраций.)

Помимо кода, сайт probml.github.io/supp содержит дополнительные материалы, не включенные в книгу из-за недостатка места. Упражнения (с решениями) по темам, рассматриваемым в этой книге, см. в работе [Gut22].

Введение

Интеллект – это не способность к распознаванию образов и аппроксимации функций. Это способность к моделированию мира.

– Джош Тененбаум, NeurIPS 2021

Значительная часть современного машинного обучения посвящена задаче отображения входов на выходы (т. е. аппроксимации функций вида $f: \mathcal{X} \rightarrow \mathcal{Y}$), при решении которой часто используется «**глубокое обучение**» (см., например, [LBN15; Sch14; Sej20; BLN21]). Джуда Перл, хорошо известный исследователь ИИ, назвал это «помпезной аппроксимацией кривых» (цитируется по работе [Har18]). Это не совсем справедливо, т. к. в случае, когда \mathcal{X} и (или) \mathcal{Y} – пространства высокой размерности, например изображения, предложения языка, графы или последовательности решений либо действий, термин «аппроксимация кривой» лишь сбивает с толку, потому что интуитивные представления, работающие для одномерной прямой, зачастую не переносятся на многомерные пространства (см., например, [BPL21a]). Тем не менее цитированное высказывание намекает на то, чего, как многие полагают, не хватает нынешним попыткам «разрешить ИИ» с применением методов машинного обучения, а именно что все они уделяют чрезмерно много внимания предсказанию наблюдаемых паттернов, но недостаточно – «пониманию» истинной структуры, скрытой за этими паттернами.

Достижение «глубокого понимания» структуры, стоящей за наблюдаемыми данными, необходимо для прогресса науки вообще, равно как и для некоторых приложений, скажем, в здравоохранении (см., например, [DD22]), где выявление истинных причин или механизмов различных заболеваний – ключ к разработке лекарств. Кроме того, такое «глубокое понимание» необходимо для разработки *робастных* и *эффективных* систем. Под «робастными» мы понимаем методы, которые хорошо работают даже при наличии неожиданных изменений в распределении данных, к которым применяется система. Это важно во многих областях, в т. ч. в робототехнике (см., например, [Roy+21]). Под «эффективными» мы обычно понимаем статистически эффективные методы, способные быстро обучаться на небольших объемах данных (см. [Lu+21b]). Это важно, потому что в некоторых областях, например в здравоохранении и робототехнике, объем доступных данных ограничен, тогда как в других, например в лингвистике и машинном зрении, в них нет недостатка, т. к. их можно в любом количестве почерпнуть из интернета. Нас также интересуют вычислительно эффективные методы, хотя этот вопрос вторичен, поскольку вычислительные мощности про-

должают расти. (Мы отмечаем еще, что эта тенденция оказала решающее влияние на многие недавние достижения в области ИИ, см. [Sut19].)

Для разработки робастных и эффективных систем в этой книге предлагается подход на основе моделей, при котором мы пытаемся найти *экономные представления* истинного «**процесса, порождающего данные**» (DGP) по выборкам из одного или нескольких наборов данных (см. [Lak+17; Win+19; Sch20; Ben+21a; Cun22; MTS22]). На самом деле это напоминает научный метод, когда мы пытаемся объяснить наблюдения (точнее, их характерные черты), разрабатывая теории или модели. Один из способов формализовать этот процесс дает применение **байесовского вывода** к вероятностным моделям, описанное в работах [Jay03; Vox80; GS13]. Подробно алгоритмы вывода будут обсуждаться в части II книги¹. Но прежде, в части I, мы изложим подготовительные сведения, которые понадобятся в дальнейшем. (Читатели, уже знакомые с основами, могут пропустить эту часть.)

Располагая набором методов вывода (некоторые из них совсем простые, например вычисление оценки максимального правдоподобия каким-нибудь методом оптимизации, скажем стохастического градиентного спуска), мы можем обратиться к обсуждению различных видов моделей. Выбор модели зависит от задачи, объема имеющихся данных и метрик успеха. Мы дадим общий обзор четырех основных видов моделей: предсказания (например, классификация и регрессия), порождения (например, изображений или текста), обнаружения («осмысленной структуры» в данных) и управления (принятия оптимальных решений). Детали будут приведены ниже.

В части III обсуждаются модели предсказания. Они представляют собой условные распределения вида $p(y|x)$, где $x \in \mathcal{X}$ – некоторый вход (часто высокой размерности), а $y \in \mathcal{Y}$ – желаемый выход (часто низкой размерности). В этой части книги предполагается, что существует единственный правильный ответ, который мы хотим предсказать, но, возможно, недостоверно.

В части IV обсуждаются модели порождения. Они представляют собой условные распределения вида $p(x)$ или $p(x|c)$, где c – факультативные обуславливающие входы, а правильных выходов может быть несколько. Например, по заданному текстовому описанию c мы можем сгенерировать разные наборы изображений, «отвечающих» этому описанию. Вычисление таких моделей сложнее, чем моделей предсказания, потому что неясно, каким должен быть желаемый выход.

В части V обсуждаются модели с латентными величинами, т. е. модели с совместным распределением вида $p(z, x) = p(z)p(x|z)$, где z – скрытое состояние, а x – наблюдения, предположительно порожденные из z . Цель – вычислить $p(z|x)$, чтобы раскрыть какое-то (предположительно осмысленное или полезное) истинное состояние или паттерны в наблюдаемых данных. Мы также рассматриваем методы обнаружения паттернов, которым неявно обучились

¹ Отметим, что в сообществе глубокого обучения под термином «вывод» (inference) понимается применение функции к некоторым входам для вычисления выхода. Это никак не связано с байесовским выводом, перед которым стоит гораздо более трудная задача обращения функции, т. е. перехода от наблюдаемых выходов к возможным скрытым входам (причинам). Это, скорее, похоже на то, что в глубоком обучении называется «обучением» (training).

предсказательные модели вида $p(y|x)$, не полагаясь на явную порождающую модель данных.

Наконец, в части VI обсуждаются модели и алгоритмы, которые можно использовать для принятия решений в условиях неопределенности. Это естественно подводит нас к важной теме причинности, последней в этой книге.

Ввиду широкого охвата материала мы не можем вдаваться во все детали каждой темы. Однако мы старались изложить основные сведения. В некоторых случаях мы совершаем «глубокое погружение», доходя до уровня передовых исследований (по состоянию на 2022 год). Мы надеемся, что собрание всех этих тем в одном месте позволит вам провести связи между разнородными, на первый взгляд, областями, а значит, лучше понять предмет машинного обучения.

Часть I

Основания

Вероятность

2.1. ВВЕДЕНИЕ

В этом разделе мы формально определим, что будет пониматься под вероятностью, следуя изложению, принятому в работе [Cha21, глава 2]. Другие хорошие введения в эту тему см., например, в работах [GS97; BT08; Bet18; DFO20].

2.1.1. Пространство вероятностей

Определим **пространство вероятностей** как тройку $(\Omega, \mathcal{F}, \mathbb{P})$, где Ω – **выборочное пространство**¹, т. е. множество возможных исходов эксперимента, \mathcal{F} – **пространство событий**, т. е. множество всех возможных подмножеств Ω , а \mathbb{P} – **вероятностная мера**, т. е. отображение события $E \subseteq \Omega$ в число, принадлежащее отрезку $[0, 1]$ (т. е. $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$), которое удовлетворяет некоторым условиям непротиворечивости, обсуждаемым в разделе 2.1.4.

2.1.2. Дискретные случайные величины

Простейший случай – когда исходы эксперимента образуют счетное множество. Например, рассмотрим трехгранную игральную кость, грани которой помечены буквами A, B и C . (Мы выбрали три грани вместо шести для краткости.) Выборочное пространство $\Omega = \{A, B, C\}$ представляет все возможные исходы «эксперимента». Пространство событий – это все возможные подмножества выборочного пространства, т. е. $\mathcal{F} = \{\emptyset, \{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}\}$. Событие является элементом пространства событий. Например, событие $E_1 = \{A, B\}$ представляет исходы, когда на кости выпадает грань A или B , а событие $E_2 = \{C\}$ – исход, когда на кости выпадает грань C .

Определив пространство событий, мы должны определить вероятностную меру, т. е. способ вычислить «размер» или «вес» каждого множества в пространстве событий. В случае трехгранной кости предположим, что вероятности отдельных исходов (элементарных событий) заданы следующим образом: $\mathbb{P}[\{A\}] = 2/6$, $\mathbb{P}[\{B\}] = 1/6$, $\mathbb{P}[\{C\}] = 3/6$. Вероятности других событий можно вычислить путем сложения мер исходов, например $\mathbb{P}[\{A, B\}] = 2/6 + 1/6 = 1/2$. Мы формализуем эту идею в разделе 2.1.4.

¹ В отечественной литературе употребляется также термин «пространство элементарных событий». – *Прим. перев.*

Чтобы упростить обозначения, сопоставим каждому исходу в пространстве событий число. Это можно сделать, определив **случайную величину (св)**, т. е. функцию $X : \Omega \rightarrow \mathbb{R}$, которая отображает событие $\omega \in \Omega$ в число $X(\omega)$ на вещественной прямой. Например, можно определить случайную величину X для нашей трехгранной кости, положив $X(A) = 1, X(B) = 2, X(C) = 3$. В качестве другого примера рассмотрим эксперимент, в котором симметричная монета подбрасывается дважды. Выборочное пространство $\Omega = \{\omega_1 = (O, O), \omega_2 = (O, P), \omega_3 = (P, O), \omega_4 = (P, P)\}$, где O – орел, а P – решка. Пусть случайная величина X представляет число выпавших орлов. Тогда имеем $X(\omega_1) = 2, X(\omega_2) = 1, X(\omega_3) = 1$ и $X(\omega_4) = 0$.

Определим множество возможных значений случайной величины как ее **пространство состояний**, обозначаемое $X(\Omega) = \mathcal{X}$. Определим вероятность любого заданного состояния как

$$p_x(a) = \mathbb{P}[X = a] = \mathbb{P}[X^{-1}(a)], \quad (2.1)$$

где $X^{-1}(a) = \{\omega \in \Omega | X(\omega) = a\}$ – прообраз a . Здесь p_x называется **функцией вероятности** (probability mass function, или **pmf**) случайной величины X . В примере с двойным подбрасыванием симметричной монеты pmf равна $p_x(0) = \mathbb{P}[(P, P)] = 1/4, p_x(1) = \mathbb{P}[(P, O), (O, P)] = 2/4, p_x(2) = \mathbb{P}[(O, O)] = 1/4$. Функцию вероятности можно представить гистограммой или какой-нибудь параметрической функцией (см. раздел 2.2.1). Мы будем называть p_x **распределением вероятностей** случайной величины X . Индекс X будет часто опускаться, если из контекста понятно, о чем идет речь.

2.1.3. Непрерывные случайные величины

Можно также рассмотреть эксперименты с непрерывными исходами. В этом случае мы предполагаем, что выборочное пространство – подмножество множества вещественных чисел, $\Omega \in \mathbb{R}$, и определяем каждую непрерывную случайную величину как тождественную функцию $X(\omega) = \omega$.

Например, рассмотрим измерение продолжительности некоторого события (в секундах). Определим выборочное пространство как $\Omega = \{t : 0 \leq t \leq T_{\max}\}$. Поскольку это множество несчетное, невозможно перечислить все его подмножества, в отличие от дискретного случая. Вместо этого мы должны определить пространство событий в терминах **сигма-поля**, или **сигма-алгебры Бореля**. Говорят, что \mathcal{F} является σ -полем, если (1) $\emptyset \in \mathcal{F}$ и $\Omega \in \mathcal{F}$; (2) \mathcal{F} замкнуто относительно операции дополнения, т. е. если $E \in \mathcal{F}$, то $E^c \in \mathcal{F}$; и (3) \mathcal{F} замкнуто относительно операций счетного объединения и счетного пересечения, т. е. $\bigcup_{i=1}^{\infty} E_i \in \mathcal{F}$ и $\bigcap_{i=1}^{\infty} E_i \in \mathcal{F}$, при условии что $E_1, E_2, \dots \in \mathcal{F}$. Наконец, будем говорить, что \mathcal{B} является σ -полем Бореля, если это σ -поле, порожденное полузамкнутыми интервалами вида $(-\infty, b] = \{x : -1 < x \leq b\}$. Беря объединения, пересечения и дополнения таких интервалов, можно видеть, что \mathcal{B} содержит следующие множества:

$$(a, b), [a, b], (a, b], [a, b], \{b\}, -\infty \leq a \leq b \leq \infty. \quad (2.2)$$

В примере с продолжительностью событий можно дополнительно ограничить пространство событий только теми интервалами, для которых нижняя граница равна 0, а верхняя $\leq T_{\max}$.

Чтобы определить вероятностную меру, сопоставим каждому $x \in \Omega$ значение весовой функции $p_X(x) \geq 0$, называемой **функцией плотности вероятности** (англ. probability density function – **pdf**). Перечень распространенных pdf приведен в разделе 2.2.2. Тогда вероятность события $E = [a, b]$ можно записать в виде

$$\mathbb{P}([a, b]) = \int_E d\mathbb{P} = \int_a^b p(x)dx. \quad (2.3)$$

Можно также определить **функцию распределения** (англ. cumulative distribution function – **cdf**) случайной величины X следующим образом:

$$P_X(x) \triangleq \mathbb{P}[X \leq x] = \int_{-\infty}^x p_X(x')dx'. \quad (2.4)$$

Отсюда можно вычислить вероятность интервала по формуле

$$\mathbb{P}([a, b]) = p(a \leq X \leq b) = P_X(b) - P_X(a). \quad (2.5)$$

Термин «распределение вероятностей» может относиться к pdf p_X , или к cdf P_X , или даже к вероятностной мере \mathbb{P} .

Приведенные выше определения можно обобщить на многомерные пространства, $\Omega \subseteq \mathbb{R}^n$, и даже на более сложные выборочные пространства, например пространства функций.

2.1.4. Аксиомы вероятностей

Вероятностная мера, ассоциированная с пространством событий, должна удовлетворять **аксиомам вероятностей**, называемым также **аксиомами Колмогорова**¹:

- неотрицательность: $\mathbb{P}[E] \geq 0$ для любого $E \subseteq \Omega$;
- нормировка: $\mathbb{P}[\Omega] = 1$;
- аддитивность: для любой счетной последовательности попарно непересекающихся множеств $\{E_1, E_2, \dots\}$ имеет место равенство

$$\mathbb{P}[\cup_{i=1}^{\infty} E_i] = \sum_{i=1}^{\infty} \mathbb{P}[E_i]. \quad (2.6)$$

В конечном случае, когда имеется всего два непересекающихся множества, E_1 и E_2 , это равенство принимает вид:

$$\mathbb{P}[E_1 \cup E_2] = \mathbb{P}[E_1] + \mathbb{P}[E_2]. \quad (2.7)$$

Это соответствует вероятности события E_1 или E_2 в предположении, что эти события взаимно исключающие (непересекающиеся множества).

¹ Можно показать, что эти правила следуют из более общего набора предположений о рассуждениях в условиях неопределенности. Этот результат называется **теоремой Кокса** [Cox46; Cox61].

Из этих аксиом можно вывести **правило дополнения**:

$$\mathbb{P}[E^c] = 1 - \mathbb{P}[E], \quad (2.8)$$

где $E^c = \Omega \setminus E$ – дополнение E . (Это следует из того, что $\mathbb{P}[\Omega] = 1 = \mathbb{P}[E \cup E^c] = \mathbb{P}[E] + \mathbb{P}[E^c]$.) Можно еще показать, что $\mathbb{P}[E] \leq 1$ (доказательство от противного) и что $\mathbb{P}[\emptyset] = 0$ (это вытекает из первого следствия при $E = \Omega$).

Также можно доказать следующий результат, известный как **правило сложения**:

$$\mathbb{P}[E_1 \cup E_2] = \mathbb{P}[E_1] + \mathbb{P}[E_2] - \mathbb{P}[E_1 \cap E_2]. \quad (2.9)$$

Это верно для любой пары событий, даже пересекающихся.

2.1.5. Условная вероятность

Рассмотрим два события E_1 и E_2 . Если $\mathbb{P}[E_2] \neq 0$, то определим **условную вероятность** E_1 при условии E_2 как

$$\mathbb{P}[E_1|E_2] \triangleq \frac{\mathbb{P}[E_1 \cap E_2]}{\mathbb{P}[E_2]}. \quad (2.10)$$

Отсюда получаем **правило умножения**:

$$\mathbb{P}[E_1 \cap E_2] = \mathbb{P}[E_1|E_2]\mathbb{P}[E_2] = \mathbb{P}[E_2|E_1]\mathbb{P}[E_1]. \quad (2.11)$$

Условная вероятность измеряет, насколько вероятно событие E_1 , если известно, что событие E_2 уже произошло. Однако если события не связаны друг с другом, то вероятность не изменится. Формально говорят, что события E_1 и E_2 **независимы**, если

$$\mathbb{P}[E_1 \cap E_2] = \mathbb{P}[E_1]\mathbb{P}[E_2]. \quad (2.12)$$

Если $\mathbb{P}[E_1] > 0$ и $\mathbb{P}[E_2] > 0$, то это эквивалентно требованию $\mathbb{P}[E_1|E_2] = \mathbb{P}[E_1]$, или, что то же самое, $\mathbb{P}[E_2|E_1] = \mathbb{P}[E_2]$. Аналогично говорят, что E_1 и E_2 **условно независимы** при условии E_3 , если

$$\mathbb{P}[E_1 \cap E_2|E_3] = \mathbb{P}[E_1|E_3]\mathbb{P}[E_2|E_3]. \quad (2.13)$$

Из определения условной вероятности можно вывести **формулу полной вероятности**, которая утверждает следующее: если $\{A_1, \dots, A_n\}$ – разбиение выборочного пространства Ω , то для любого события $B \subseteq \Omega$ имеем

$$\mathbb{P}[B] = \sum_{i=1}^n \mathbb{P}[B|A_i]\mathbb{P}[A_i]. \quad (2.14)$$

2.1.6. Формула Байеса

Из определения условной вероятности можно вывести **формулу Байеса**, называемую также **теоремой Байеса**, которая утверждает, что для любых двух событий E_1 и E_2 , таких что $\mathbb{P}[E_1] > 0$ и $\mathbb{P}[E_2] > 0$, имеет место равенство

$$\mathbb{P}[E_1|E_2] = \frac{\mathbb{P}[E_2|E_1]\mathbb{P}[E_1]}{\mathbb{P}[E_2]}. \quad (2.15)$$

Для дискретной случайной величины X с K возможными состояниями формулу Байеса можно записать, воспользовавшись формулой полной вероятности:

$$p(X = k|E) = \frac{p(E|X = k)p(X = k)}{p(E)} = \frac{p(E|X = k)p(X = k)}{\sum_{k'=1}^K p(E|X = k')p(X = k')}. \quad (2.16)$$

Здесь $p(X = k)$ – **априорная вероятность**, $p(E|X = k)$ – **правдоподобие**, $p(X = k|E)$ – **апостериорная вероятность**, а $p(E)$ – нормировочная постоянная, называемая **маргинальным правдоподобием**.

Аналогично для непрерывной случайной величины X формулу Байеса можно записать в виде:

$$p(X = x|E) = \frac{p(E|X = x)p(X = x)}{p(E)} = \frac{p(E|X = x)p(X = x)}{\int p(E|X = x')p(X = x')dx'}. \quad (2.17)$$

2.2. НЕКОТОРЫЕ РАСПРОСТРАНЕННЫЕ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

Существует великое множество распределений вероятностей, используемых в различных моделях. Ниже мы кратко опишем наиболее употребительные. Другие распределения см. в дополнении к главе 2 (<https://github.com/probml/pml-book/blob/main/supp2.md>), а интерактивные визуализации – по адресу <https://ben18785.shinyapps.io/distribution-zoo/>.

2.2.1. Дискретные распределения

В этом разделе мы обсудим некоторые дискретные распределения, определенные на подмножествах неотрицательных целых чисел.

2.2.1.1. Распределение Бернулли и биномиальное распределение

Пусть $x \in \{0, 1, \dots, N\}$. **Биномиальное распределение** определяется следующим образом:

$$\text{Bin}(x|N, \mu) \triangleq \binom{N}{x} \mu^x (1 - \mu)^{N-x}, \quad (2.18)$$

где $\binom{N}{k} \triangleq \frac{N!}{(N-k)!k!}$ – число способов выбрать k предметов из N (оно называется **биномиальным коэффициентом** и произносится «С из N по k»).

Если $N = 1$, т. е. $x \in \{0, 1\}$, то биномиальное распределение сводится к **распределению Бернулли**

$$\text{Ber}(x|\mu) = \begin{cases} 1 - \mu & \text{если } x = 0 \\ \mu & \text{если } x = 1 \end{cases} \quad (2.19)$$

где $\mu = \mathbb{E}[x] = p(x = 1)$ – среднее.

2.2.1.2. Категориальное и мультиномиальное распределения

Если величина дискретная, $x \in \{1, \dots, K\}$, то можно использовать **категориальное распределение**:

$$\text{Cat}(x|\theta) \triangleq \prod_{k=1}^K \theta_k^{\mathbb{I}(x=k)}. \quad (2.20)$$

Альтернативно K -значную величину x можно представить унитарным двоичным вектором \mathbf{x} , что позволяет переписать формулу в виде:

$$\text{Cat}(\mathbf{x}|\theta) \triangleq \prod_{k=1}^K \theta_k^{x_k}. \quad (2.21)$$

Если k -й элемент \mathbf{x} показывает, сколько раз k наблюдалось в $N = \sum_{k=1}^K x_k$ испытаниях, то получаем **мультиномиальное распределение**:

$$\mathcal{M}(\mathbf{x}|N, \theta) \triangleq \binom{N}{x_1 \dots x_K} \prod_{k=1}^K \theta_k^{x_k}, \quad (2.22)$$

где **мультиномиальный коэффициент** определяется как

$$\binom{N}{k_1 \dots k_m} \triangleq \frac{N!}{k_1! \dots k_m!}. \quad (2.23)$$

2.2.1.3. Распределение Пуассона

Пусть $X \in \{0, 1, 2, \dots\}$. Говорят, что случайная величина имеет **распределение Пуассона** с параметром $\lambda > 0$, и записывают это в виде $X \sim \text{Poi}(\lambda)$, если ее функция вероятности имеет вид:

$$\text{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad (2.24)$$

где λ – среднее (и дисперсия) x .

2.2.1.4. Отрицательное биномиальное распределение

Пусть имеется урна с N шарами, из которых R красные и B синие. Допустим, что производится **выборка с возвращением** до тех пор, пока не будет извлечено $n \geq 1$ шаров. Пусть X – число синих шаров среди них. Можно показать, что $X \sim \text{Bin}(n, p)$, где $p = B/N$ – доля синих шаров; таким образом, X имеет биномиальное распределение, рассмотренное в разделе 2.2.1.1.

Теперь предположим, что извлечение красного шара рассматривается как «неудача», а извлечение синего – как «успех». Допустим, что извлечение шаров продолжается, пока не будет зарегистрировано r неудач. Пусть X – получившееся в результате число успехов (синих шаров); можно показать, что $X \sim \text{NegBinom}(r, p)$, т. е. имеет **отрицательное биномиальное распределение**, определяемое формулой

$$\text{NegBinom}(x|r, p) \triangleq \binom{x+r-1}{x} (1-p)^r p^x \quad (2.25)$$

для $x \in \{0, 1, 2, \dots\}$. (Если r вещественное, то заменяем $\binom{x+r-1}{x}$ на $\frac{\Gamma(x+r)}{x! \Gamma(r)}$, пользуясь тем фактом, что $(x-1)! = \Gamma(x)$.)

Это распределение имеет следующие моменты:

$$\mathbb{E}[x] = \frac{p r}{1-p}, \quad \mathbb{V}[x] = \frac{p r}{(1-p)^2}. \quad (2.26)$$

Это двухпараметрическое семейство дает бóльшую гибкость при моделировании, чем распределение Пуассона, поскольку позволяет представить среднее и дисперсию порознь. Это полезно, например, для моделирования «контагиозных» событий с положительно коррелированной встречаемостью, поскольку при этом дисперсия больше, чем в случае, когда события независимы. На самом деле распределение Пуассона – частный случай отрицательного биномиального распределения, т. к. можно показать, что $\text{Poi}(\lambda) = \lim_{r \rightarrow \infty} \text{NegBinom}(r, \lambda/(1+\lambda))$. Другой частный случай возникает при $r = 1$; получающееся распределение называется **геометрическим**.

2.2.2. Непрерывные распределения на \mathbb{R}

В этом разделе мы обсудим некоторые одномерные распределения, определенные на множестве вещественных чисел, $p(x)$ для $x \in \mathbb{R}$.

2.2.2.1. Гауссово (нормальное) распределение

Самым распространенным одномерным распределением является **гауссово распределение**, называемое также **нормальным**. (Обсуждение этих названий см. в [Mur22, раздел 2.6.4].) Функция плотности вероятности (pdf) для гауссова распределения имеет вид

$$\mathcal{N}(x|\mu, \sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad (2.27)$$

где $\sqrt{2\pi\sigma^2}$ – нормировочная постоянная, необходимая для того, чтобы интеграл плотности был равен 1. Параметр μ определяет среднее распределения, совпадающее с его модой. Параметр σ^2 определяет дисперсию. Иногда говорят о **точности** гауссова распределения, понимая под ней обратное среднее: $\lambda = 1/\sigma^2$. Высокая точность означает узкое распределение (с низкой дисперсией) с центром в μ .

Функция распределения (cdf) гауссова распределения имеет вид

$$\Phi(x; \mu, \sigma^2) \triangleq \int_{-\infty}^x \mathcal{N}(z|\mu, \sigma^2) dz. \quad (2.28)$$

Если $\mu = 0$ и $\sigma = 1$ (так называемое **стандартное нормальное распределение**), то мы пишем просто $\Phi(x)$.

2.2.2.2. Полунормальное распределение

В некоторых задачах нам нужно распределение на множестве неотрицательных вещественных чисел. Один из способов создать такое распределение – определить $Y = |X|$, где $X \sim \mathcal{N}(0, \sigma^2)$. Индуцированное распределение Y называется **полунормальным распределением** и имеет pdf

$$\mathcal{N}_+(y|\sigma) \triangleq 2\mathcal{N}(y|0, \sigma^2) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left(-\frac{y^2}{2\sigma^2}\right) \quad y \geq 0. \quad (2.29)$$

Можно считать, что это распределение $\mathcal{N}(0, \sigma^2)$, «сложенное» на себя.

2.2.2.3. t-распределение Стьюдента

Гауссово распределение чувствительно к выбросам, т. к. вероятность экспоненциально убывает с увеличением (квадрата) расстояния от центра. Более робастным является **t-распределение Стьюдента**, которое мы для краткости будем называть просто **распределением Стьюдента**. Его pdf имеет вид:

$$\mathcal{T}_\nu(x|\mu, \sigma^2) = \frac{1}{Z} \left[1 + \frac{1}{\nu} \left(\frac{x - \mu}{\sigma} \right)^2 \right]^{-\left(\frac{\nu+1}{2}\right)}, \quad (2.30)$$

$$Z = \frac{\sqrt{\nu\pi\sigma^2}\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+1}{2})} = \sqrt{\nu}\sigma B\left(\frac{1}{2}, \frac{\nu}{2}\right), \quad (2.31)$$

где μ – среднее, $\sigma > 0$ – масштабный параметр (не стандартное отклонение), а $\nu > 0$ называется **числом степеней свободы**, хотя правильнее было бы говорить о **степени нормальности** [Kru13], потому что при больших значениях ν распределение ведет себя как гауссово. Здесь $\Gamma(a)$ – **гамма-функция**, определяемая формулой

$$\Gamma(a) \triangleq \int_0^\infty x^{a-1} e^{-x} dx, \quad (2.32)$$

а $B(a, b)$ – **бета-функция**, определяемая формулой

$$B(a, b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \quad (2.33)$$

2.2.2.4. Распределение Коши

При $\nu = 1$ распределение Стьюдента называется распределением **Коши** или **Лоренца**. Его pdf определяется формулой

$$\mathcal{C}(x|\mu, \gamma) = \frac{1}{Z} \left[1 + \left(\frac{x - \mu}{\gamma} \right)^2 \right]^{-1}, \quad (2.34)$$

где $Z = \gamma\beta^{1/2}, 1/2) = \gamma\pi$. Это распределение примечательно наличием таких тяжелых хвостов, что интеграл, определяющий его среднее, расходится.

Половинное распределение Коши – это вариант распределения Коши (со средним 0), «сложенного» на себя, так что вся плотность вероятности сосредоточена в области положительных вещественных чисел. Оно имеет вид

$$C_+(x|\gamma) \triangleq \frac{2}{\pi\gamma} \left[1 + \left(\frac{x}{\gamma} \right)^2 \right]^{-1}. \quad (2.35)$$

2.2.2.5. Распределение Лапласа

Еще одно распределение с тяжелыми хвостами – **распределение Лапласа**, называемое также **двусторонним экспоненциальным распределением**. Его pdf имеет вид:

$$\text{Laplace}(x|\mu, b) \triangleq \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right). \quad (2.36)$$

Здесь μ – параметр сдвига, а $b > 0$ – масштабный параметр. См. график на рис. 2.1.

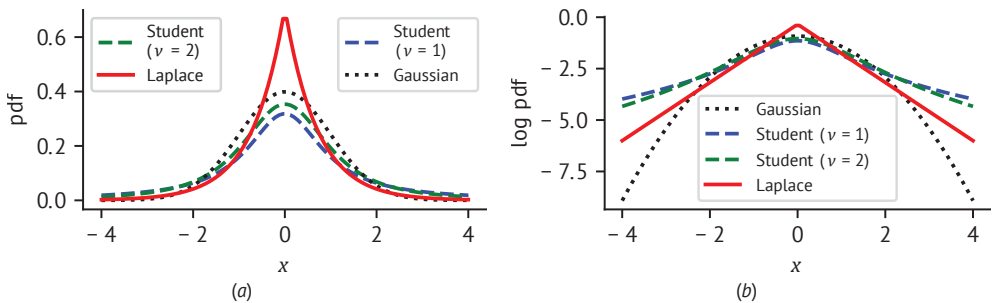


Рис. 2.1. (a) Функции плотности вероятности для $\mathcal{N}(0, 1)$, $\mathcal{T}_1(0, 1)$ и $\text{Laplace}(0, 1/\sqrt{2})$. И для гауссова распределения, и для распределения Лапласа среднее равно 0, а дисперсия – 1. Среднее и дисперсия распределения Стьюдента не определены при $\nu = 1$; (b) логарифм этих pdf. Отметим, что распределение Стьюдента не является логарифмически выпуклым при любом значении параметра, в отличие от распределения Лапласа. Тем не менее оба они унимодальны. Построено программой [student_laplace_pdf_plot.ipynb](#)

2.2.2.6. Субгауссово и супергауссово распределения

Существует два основных варианта гауссова распределения, известных как **супергауссово**, или лептокуртическое (от греческого слова «Lepto» – узкий), и **субгауссово**, или платикуртическое (от греческого слова «Platy» – широкий), распределения. Эти распределения отличаются **куртозисом**, т. е. мерой тяжести хвостов (насколько быстро уменьшается плотность при удалении от среднего). Точнее, куртозис определяется формулой

$$\text{kurt}(z) \triangleq \frac{\mu_4}{\sigma^4} = \frac{\mathbb{E}[(Z - \mu)^4]}{(\mathbb{E}[(Z - \mu)^2])^2}, \quad (2.37)$$

где σ – стандартное отклонение, а μ_4 – четвертый **центральный момент**. (Первым моментом является среднее, $\mu_1 = \mu$, вторым – дисперсия, $\mu_2 = \sigma^2$.) Для стандартного гауссова распределения куртозис равен 3, поэтому некоторые авторы определяют **избыточный куртозис** как куртозис минус 3.

У супергауссова распределения (например, распределения Лапласа) избыточный куртозис положителен, поэтому хвосты тяжелее, чем у гауссова. У субгауссова распределения, например равномерного, избыточный куртозис отрицателен, поэтому хвосты легче, чем у гауссова. См. иллюстрацию на рис. 2.2.

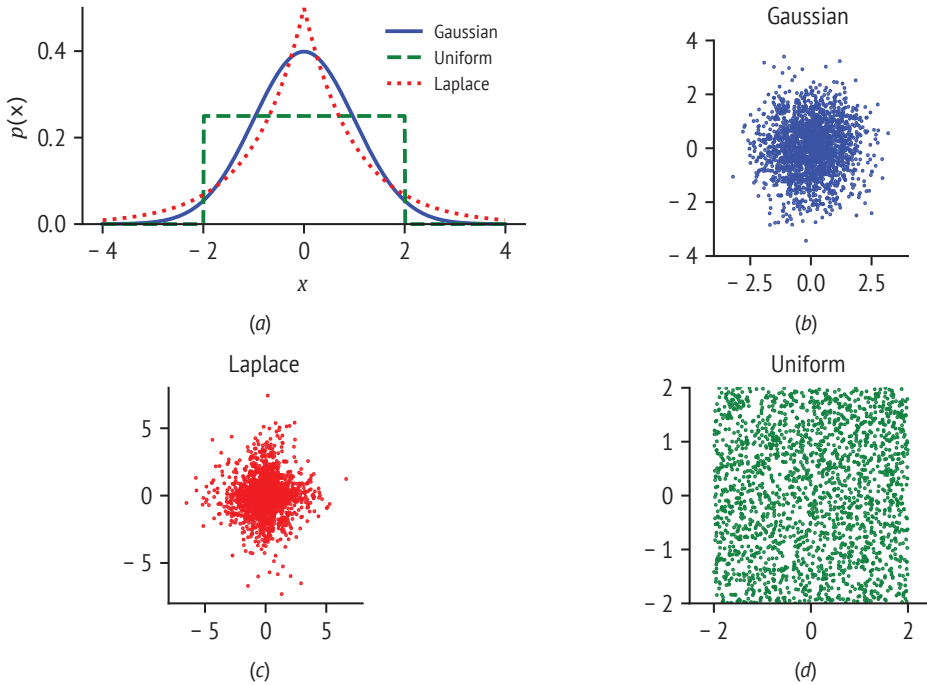


Рис. 2.2. Гауссово (синее), субгауссово (равномерное, зеленое) и супергауссово (Лапласа, красное) распределения в одномерном и двумерном случаях.

Построено программой `sub_super_gauss_plot.ipynb`

2.2.3. Непрерывные распределения на \mathbb{R}^+

В этом разделе мы обсудим некоторые одномерные распределения, определенные на множестве положительных вещественных чисел, $p(x)$ для $x \in \mathbb{R}^+$.

2.2.3.1. Гамма-распределение

Гамма-распределение – гибкое распределение случайных величин, принимающих положительные вещественные значения, $x > 0$. Оно определяется двумя параметрами: формы $a > 0$ и скорости $b > 0$:

$$\text{Ga}(x|\text{shape} = a, \text{rate} = b) \triangleq \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}. \quad (2.38)$$

Иногда это распределение параметризуют скоростью a и **масштабом** $s = 1/b$:

$$\text{Ga}(x|\text{shape} = a, \text{scale} = s) \triangleq \frac{1}{s^a \Gamma(a)} x^{a-1} e^{-x/s}. \quad (2.39)$$

См. иллюстрацию на рис. 2.3(а).

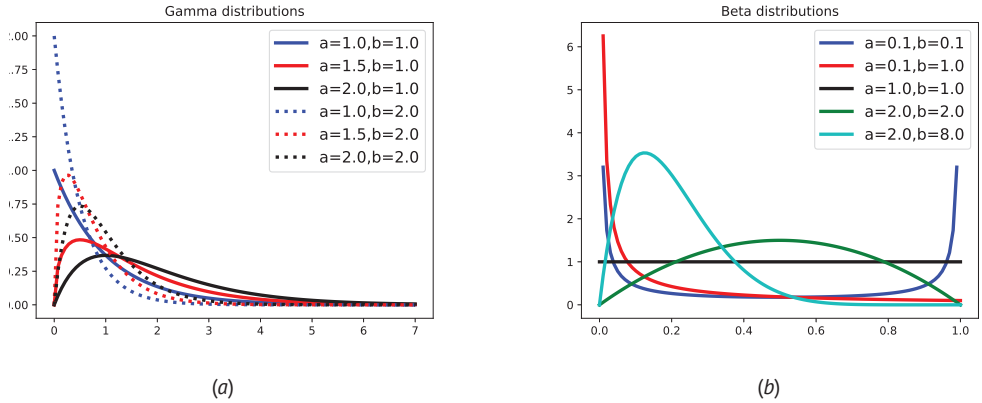


Рис. 2.3. (а) Некоторые гамма-распределения. Если $a \leq 1$, то мода расположена в точке 0, в противном случае мода отстоит от 0 на положительное расстояние. При увеличении скорости b уменьшается горизонтальный масштаб, т. е. всё сжимается влево и вверх. Построено программой `gamma_dist_plot.ipynb`; (б) некоторые бета-распределения. Если $a < 1$, мы имеем пик слева, а если $b < 1$, то пик справа. Если $a = b = 1$, то распределение равномерное. Если $a > 1$ и $b > 1$, то распределение унимодальное. Построено программой `beta_dist_plot.ipynb`

2.2.3.2. Экспоненциальное распределение

Экспоненциальное распределение – частный случай гамма-распределения, оно определяется формулой

$$\text{Expon}(x|\lambda) \triangleq \text{Ga}(x|\text{shape} = 1, \text{rate} = \lambda). \quad (2.40)$$

Это распределение определяет промежутки времени между событиями в пуассоновском процессе, т. е. таком, в котором события возникают непрерывно и независимо с постоянной средней скоростью λ .

2.2.3.3. Распределение хи-квадрат

Распределение хи-квадрат – частный случай гамма-распределения, оно определяется формулой

$$\chi_\nu^2(x) \triangleq \text{Ga}(x|\text{shape} = \frac{\nu}{2}, \text{rate} = \frac{1}{2}), \quad (2.41)$$

где ν называется числом степеней свободы. Это распределение суммы квадратов гауссовых случайных величин. Точнее, если $Z_i \sim \mathcal{N}(0, 1)$ и $S = \sum_{i=1}^{\nu} Z_i^2$, то $S \sim \chi_\nu^2$. Тогда если $X \sim \mathcal{N}(0, \sigma^2)$, то $X^2 \sim \sigma^2 \chi_1^2$. Так как $\mathbb{E}[\chi_1^2] = 1$ и $\mathbb{V}[\chi_1^2] = 2$, имеем

$$\mathbb{E}[X^2] = \sigma^2, \mathbb{V}[X^2] = 2\sigma^4. \quad (2.42)$$