

СОДЕРЖАНИЕ

От издательства	15
Предисловие	16
Часть I БАЗОВЫЕ КОМПОНЕНТЫ	19
1 Автоматизация решений и процессов в деятельности предприятия	20
1.1 Методика сценарного планирования	21
1.1.1. Стратегия: предприятие в целом	24
1.1.2 Тактика: подразделения, услуги и продукты	28
1.1.3 Реализация: клиенты, устройства, транзакции и интерфейсы	32
1.2 Возможности моделирования	35
1.3 Внедрение базовых моделей и AutoML	38
1.4 Краткие итоги главы	40
2 Прогнозные модели	41
2.1 Обзор с точки зрения системной инженерии	41
2.1.1 Семантические представления	42
2.1.2 Прогнозные модели	44
2.1.3 Генеративные модели	47
2.1.4 Модели управления	48
2.2 Метод максимального правдоподобия	50
2.2.1 Оценка правдоподобия	50
2.2.2 Оценка условного правдоподобия	52
2.2.3 Максимизация правдоподобия с использованием градиентного спуска	52
2.3 Модели с векторными входами	54
2.3.1 Линейный слой	54
2.3.2 Нелинейные слои	61
2.3.3 Остаточные блоки и обходные связи	65
2.3.4 Слои оценки распределения	66
2.3.5 Слои выборки	68
2.3.6 Слой встраивания с перекодировкой	69
2.3.7 Слои взаимодействия	73
2.3.8 Многоголовые и многобашенные архитектуры	74
2.4 Модели с последовательными входами	77
2.4.1 Проблемы моделирования последовательностей	77
2.4.2 Метод скользящего окна	78
2.4.3 Слой свертки	85

2.4.4	Рекуррентный слой.....	90
2.4.5	Слой долгой краткосрочной памяти.....	93
2.4.6	Механизм внимания	98
2.4.7	Слой трансформера	99
2.5	Модели с многомерными входными данными	107
2.5.1	Операция 2D-свертки.....	108
2.5.2	Слой двухмерной свертки	109
2.5.3	Слой двухмерной повышающей свертки.....	111
2.5.4	Глубокие двухмерные сверточные сети	112
2.5.5	Слой двухмерного трансформера	116
2.6	Модели обучения представлению	117
2.6.1	Функции потерь для изучения представлений с учителем.....	118
2.6.2	Автокодировщики	119
2.6.3	Представление элементов	124
2.7	Модели с графовыми входными данными.....	128
2.7.1	Задачи графового машинного обучения	128
2.7.2	Изучение представлений узлов	130
2.7.3	Графовые нейронные сети.....	136
2.8	Корректность модели	140
2.8.1	Несбалансированные данные	140
2.8.2	Данные наблюдений.....	143
2.9	Базовые модели.....	145
2.9.1	Стратегии предварительного обучения	146
2.9.2	Стратегии переноса обучения	147
2.9.3	Методы тонкой настройки	148
2.10	Краткие итоги главы.....	150
3	Генеративные модели	152
3.1	Регуляризация семантического пространства	153
3.2	Вариационный автокодировщик.....	154
3.2.1	Модели со скрытыми переменными и их оценка	155
3.2.2	Масштабируемая оценка модели с использованием ELBO	156
3.2.3	Предположения о нормальности.....	158
3.2.4	Сеть вариационного автокодировщика	159
3.2.5	Ограничения базового VAE	162
3.2.6	Условный вариационный автокодировщик.....	162
3.2.7	Иерархический вариационный автокодировщик	164
3.3	Вероятностные диффузионные модели шумоподавления.....	166
3.3.1	Прямой процесс	167
3.3.2	Обратный процесс	168
3.3.3	Обучение	169
3.3.4	Извлечение выборки	172
3.3.5	Условные диффузионные модели	172
3.4	Большие языковые модели.....	173
3.4.1	Языковое моделирование.....	174

3.4.2	Базовые языковые модели	174
3.4.3	Масштабирование архитектуры моделей	175
3.4.4	Свойства больших языковых моделей	182
3.4.5	Тонкая настройка с помощью инструкций	187
3.4.6	Цепные вызовы моделей	188
3.5	Краткие итоги главы	190
4	Управляющие модели	192
4.1	Основные методы принятия решений	192
4.2	Обучение на основе взаимодействия	196
4.3	Обучение с подкреплением и стохастический бандит	197
4.3.1	Жадные стратегии	198
4.3.2	Стратегия верхней доверительной границы	199
4.3.3	Выборка Томпсона	201
4.3.4	Нестационарные среды	203
4.4	Обучение с подкреплением: общий случай	205
4.4.1	Марковский процесс принятия решений	205
4.4.2	Стратегии и функции ценности	207
4.4.3	Оптимизация стратегии с помощью динамического программирования	208
4.4.4	Методы на основе ценности	210
4.4.5	Методы на основе стратегий	220
4.4.6	Комбинированные методы	224
4.5	Контрфактическая оценка стратегии	228
4.5.1	Выборка по важности	229
4.5.2	Выборка с отклонением действий	231
4.6	Краткие итоги главы	232
Часть II	АНАЛИЗ КЛИЕНТСКОЙ БАЗЫ	234
R1	Моделирование склонности	235
R1.1	Бизнес-задача	236
R1.1.1	Оценивание склонности	237
R1.1.2	Атрибуция событий	238
R1.2	Варианты решения	239
R1.3	Модели с агрегированными признаками	239
R1.4	Моделирование последовательности событий	242
R1.4.1	Оценивание склонности	242
R1.4.2	Атрибуция событий	243
R1.5	Прототип	245
R1.6	Пример	249
R1.7	Расширения и варианты	252
R1.7.1	Расширенные последовательные модели	252
R1.7.2	Сверточные модели	253
R1.7.3	Конструирование целевой метки	253

R1.7.4	Практическое применение.....	253
R1.8	Краткие итоги главы	254
R2	Изучение признаков клиента	255
R2.1	Бизнес-задача.....	256
R2.2	Варианты решения	258
R2.3	Обучение на последовательностях событий	258
R2.3.1	Изучение встраиваний товара.....	258
R2.3.2	Совмещение поведенческих и контентных признаков	259
R2.3.3	Изучение встраиваний клиентов.....	261
R2.3.4	Изучение встраиваний по данным из журналов.....	262
R2.4	Обучение на графах, тексте и изображениях	263
R2.5	Методы обучения с частичным привлечением учителя	265
R2.6	Методы на основе автокодировщика	267
R2.7	Прототип	267
R2.8	Пример	270
R2.9	Краткие итоги главы	272
R3	Динамическая персонализация.....	274
R3.1	Бизнес-задача.....	274
R3.2	Варианты решения	276
R3.3	Контекстно-свободные рекомендации	276
R3.4	Контекстные рекомендации	278
R3.4.1	UCB с теплым стартом	279
R3.4.2	Алгоритм LinUCB.....	280
R3.5	Оценка и бутстрэппинг.....	283
R3.6	Прототип	284
R3.7	Краткие итоги главы	287
R4	Следующее наилучшее действие.....	288
R4.1	Бизнес-задача.....	289
R4.1.1	Разработка целей и вознаграждений	290
R4.1.2	Разработка действий.....	293
R4.1.3	Моделирование и экспериментирование.....	294
R4.2	Варианты решения	295
R4.3	Расширенный подход к оцениванию	295
R4.4	Условные склонности	297
R4.5	Обучение с подкреплением	298
R4.6	Прототип	299
R4.7	Пример	305
R4.7.1	Бизнес-задача.....	306
R4.7.2	Архитектура решения	306
R4.7.3	Алгоритмы	308
R4.7.4	Схема действий, состояний и вознаграждений	308
R4.8	Краткие итоги главы	310

Часть III РАБОТА С КОНТЕНТОМ	311
R5 Визуальный поиск	312
R5.1 Бизнес-задача.....	313
R5.2 Варианты решения.....	315
R5.3 Поиск по стилю изображения.....	318
R5.3.1 Встраивание стилей.....	318
R5.3.2 Прототип.....	322
R5.4 Поиск в пользовательском семантическом пространстве.....	324
R5.4.1 Встраивания и атрибуты пользовательских изображений.....	325
R5.4.2 Прототип.....	328
R5.5 Обучение встраиванию без учителя.....	330
R5.6 Локализация и сегментация объектов.....	335
R5.6.1 Семантическая сегментация.....	336
R5.6.2 Прототип.....	338
R5.7 Краткие итоги главы.....	339
R6 Служба рекомендации товаров	342
R6.1 Бизнес-задача.....	343
R6.1.1 Общий обзор среды.....	343
R6.1.2 Варианты среды.....	345
R6.1.3 Метрики оценки и оптимизации.....	346
R6.2 Варианты решения.....	350
R6.2.1 Архитектура службы.....	350
R6.2.2 Архитектура модели.....	352
R6.3 Модели прогнозирования отклика.....	356
R6.3.1 Базовая реализация метода факторизации.....	356
R6.3.2 Нейронная совместная фильтрация.....	359
R6.3.3 Практический пример.....	360
R6.4 Модели прогнозирования взаимодействия.....	363
R6.5 Модели последовательностей.....	365
R6.5.1 Трансформер поведенческой последовательности.....	365
R6.5.2 Практический пример.....	367
R6.6 Графовые модели.....	368
R6.6.1 Практический пример: рекомендации с помощью Node2Vec.....	371
R6.6.2 Рекомендации с помощью GNN.....	373
R6.7 Варианты и дополнения.....	376
R6.8 Краткие итоги главы.....	377
R7 Управление знаниями	379
R7.1 Бизнес-задача.....	379
R7.2 Варианты решения.....	381
R7.3 Предварительная обработка данных.....	382
R7.3.1 Обнаружение атрибутов.....	384
R7.3.2 Извлечение атрибутов.....	385
R7.3.3 Гармонизация атрибутов.....	386

R7.4	Запрос структурированных данных.....	387
R7.5	Запрос неструктурированных данных.....	389
R7.5.1	Запрос с использованием одного промпта.....	389
R7.5.2	Запрос с использованием Map-Reduce	390
R7.5.3	Генерация ответа с дополненным поиском.....	391
R7.5.4	Диалоговый поиск	394
R7.5.5	Агенты	394
R7.6	Безопасность и конфиденциальность данных	397
R7.7	Оценка качества	398
R7.7.1	Предварительная обработка данных	398
R7.7.2	Запрос структурированных данных.....	399
R7.7.3	Запрос неструктурированных данных.....	400
R7.8	Краткие итоги главы.....	400
R8	Синтетический медиаконтент	402
R8.1	Бизнес-задача.....	402
R8.2	Варианты решения	404
R8.3	Модели синтеза изображений по языковому описанию.....	404
R8.3.1	Модель CLIP.....	406
R8.3.2	Прототип.....	408
R8.4	Генеративные модели преобразования текста в изображение	410
R8.4.1	Диффузионные модели шумоподавления для изображений.....	411
R8.4.2	Диффузионные модели со скрытым пространством.....	414
R8.5	Продвинутые механизмы обусловливания	421
R8.6	Краткие итоги главы	424
Часть IV	УПРАВЛЕНИЕ ДОХОДАМИ И ЗАПАСАМИ.....	426
R9	Прогнозирование спроса	427
R9.1	Бизнес-задача.....	427
R9.1.1	Операционная среда	427
R9.1.2	Модели спроса	429
R9.1.3	Задачи	430
R9.1.4	Приложения.....	432
R9.1.5	Метрики оценки	433
R9.2	Варианты решения	434
R9.3	Модели пространства состояний.....	435
R9.3.1	Простое экспоненциальное сглаживание.....	436
R9.3.2	Двойное экспоненциальное сглаживание	438
R9.3.3	Тройное экспоненциальное сглаживание	439
R9.3.4	Декомпозиция.....	439
R9.3.5	Вероятностный прогноз	440
R9.4	Регрессия временных рядов.....	441
R9.4.1	Вероятностный прогноз	443
R9.4.2	Область охвата модели	443
R9.4.3	Множественные горизонты прогнозирования.....	443

R9.4.4	Календарные признаки	444
R9.4.5	Лаговые признаки	445
R9.4.6	Признаки продукта	446
R9.4.7	Ценовые признаки	446
R9.4.8	Практический пример	447
R9.5	Модели последовательностей.....	453
R9.5.1	Модель DeerpAR	454
R9.5.2	Практический пример	457
R9.6	Составные модели	458
R9.6.1	Модель NeuralProphet	459
R9.6.2	Практический пример	462
R9.7	Иерархические модели	462
R9.7.1	Иерархический временной ряд	463
R9.7.2	Иерархическое прогнозирование с помощью согласования	464
R9.7.3	Иерархическое прогнозирование с помощью DeepVAR.....	465
R9.8	Методы подстановки в анализе спроса	467
R9.8.1	Устранение ограничений спроса.....	468
R9.8.2	Анализ сходства продуктов.....	470
R9.9	Дополнения и варианты	472
R9.9.1	Причинные эффекты	472
R9.9.2	Устранение последствий потрясений.....	472
R9.10	Краткое содержание главы.....	473
R10	Оптимизация цены и продвижения	475
R10.1	Бизнес-задача.....	476
R10.1.1	Процесс управления ценами	476
R10.1.2	Модель дохода.....	476
R10.1.3	Стратегический анализ	478
R10.1.4	Планирование и оценка	480
R10.1.5	Выполнение плана ценообразования	487
R10.1.6	Измерение результатов	488
R10.2	Варианты решения	491
R10.3	Дифференциация ценовой стратегии	493
R10.3.1	Дифференциация ценовой стратегии по продуктам	494
R10.3.2	Дифференциация ценовой стратегии по клиентам	498
R10.4	Моделирование реакции рынка	500
R10.4.1	Линейная модель.....	501
R10.4.2	Модель постоянной эластичности	503
R10.4.3	Моделирование перекрестных эффектов.....	505
R10.4.4	Модели реакции, зависящие от времени.....	507
R10.5	Оптимизация с использованием математического программирования.....	507
R10.5.1	Несколько продуктов	509
R10.5.2	Несколько временных интервалов	512
R10.5.3	Оптимизация в условиях неопределенности.....	512
R10.6	Оптимизация с использованием обучения с подкреплением	515

R10.6.1	Обоснование потребности в решателе	515
R10.6.2	Прототип.....	517
R10.7	Расширения и варианты	520
R10.7.1	Розничная торговля.....	520
R10.7.2	Потребительские услуги	521
R10.7.3	Производство потребительских товаров	521
R10.7.4	Промышленные товары и услуги	522
R10.8	Краткие итоги главы	523
R11	Динамическое ценообразование	525
R11.1	Бизнес-задача.....	526
R11.2	Варианты решения	528
R11.3	Ограниченное ценовое экспериментирование	528
R11.3.1	Разработка решения	528
R11.3.2	Прототип.....	530
R11.4	Непрерывное экспериментирование	533
R11.4.1	Разработка решения	533
R11.4.2	Прототип.....	537
R11.5.1	Варианты и расширения	538
R11.5.1	Байесовские модели спроса.....	538
R11.5.2	Несколько продуктов и ограниченные запасы.....	543
R11.6	Краткие итоги главы	543
R12	Оптимизация складских запасов.....	545
R12.1	Бизнес-задача.....	546
R12.1.1	Запасы в контексте производственных процессов	546
R12.1.2	Политика оптимизации запасов	547
R12.1.3	Процесс управления запасами	548
R12.1.4	Различные среды.....	549
R12.1.5	Показатели эффективности.....	553
R12.2	Варианты решения	555
R12.3	Обобщенное планирование	555
R12.4	Политики одноуровневого управления запасами.....	557
R12.4.1	Политика управления запасами	558
R12.4.2	Симулятор среды.....	560
R12.4.3	Сценарий 1: постоянный спрос, нулевое время выполнения заказа.....	562
R12.4.4	Сценарий 2: постоянные значения спроса и времени выполнения заказа	566
R12.4.5	Сценарий 3: стохастический спрос, постоянное время выполнения заказа	567
R12.4.6	Сценарий 4: стохастический спрос и стохастическое время выполнения заказа	570
R12.4.7	Упущенные продажи и отсутствие ограничений спроса	572
R12.5	Многоуровневые политики управления запасами.....	573
R12.5.1	Стохастические модели обслуживания.....	574
R12.5.2	Модели гарантированного обслуживания.....	579

R12.5.3	Управление запасами с помощью обучения с подкреплением ...	583
R12.6	Расширения и варианты.....	590
R12.6.1	Сезонные и скоропортящиеся продукты.....	591
R12.6.2	Несколько каналов продаж.....	592
R12.6.3	Несколько товаров и дифференциация политик.....	595
R12.6.4	Несколько товаров и скоординированное пополнение.....	596
R12.7	Краткое содержание главы.....	597

Часть V ПРОИЗВОДСТВЕННЫЕ ОПЕРАЦИИ И ИНТЕРНЕТ ВЕЩЕЙ.....599

R13	Обнаружение аномалий.....	600
R13.1	Бизнес-задача.....	600
R13.1.1	Мониторинг, оценка и обнаружение аномалий.....	601
R13.1.2	Прогнозное обслуживание.....	602
R13.2	Варианты решения.....	604
R13.3	Модели системы.....	605
R13.4	Мониторинг.....	608
R13.5	Оценка аномалий.....	610
R13.5.1	Основные модели нормальности.....	610
R13.5.2	Модели прогнозирования состояния.....	611
R13.5.3	Модели многообразия состояний.....	615
R13.5.4	Предварительная обработка показателей.....	617
R13.6	Обнаружение и классификация аномалий.....	619
R13.6.1	Пороговое значение.....	620
R13.6.2	Вероятность восстановления.....	620
R13.6.3	Обучение с подкреплением в обнаружении и классификации аномалий.....	622
R13.7	Прогноз остаточного срока службы.....	623
R13.7.1	Способ решения.....	623
R13.7.2	Прототип.....	625
R13.8	Краткое содержание главы.....	627
R14	Визуальный контроль качества.....	629
R14.1	Бизнес-задача.....	630
R14.1.1	Операционная среда.....	630
R14.1.2	Данные.....	631
R14.1.3	Цели.....	632
R14.2	Варианты решения.....	633
R14.3	Модели классификации дефектов, обучаемые с учителем.....	634
R14.4	Модели обнаружения аномалий.....	635
R14.4.1	Архитектура модели.....	636
R14.4.2	Структурное сходство.....	638
R14.4.3	Обнаружение аномалий с помощью переноса обучения.....	640
R14.5	Прототип.....	642
R14.6	Расширения и варианты.....	644
R14.7	Краткое содержание главы.....	646

А	Функции потерь	648
A.1	Функции потерь для регрессии	648
A.1.1	Среднеквадратическая ошибка.....	649
A.1.2	Корень среднеквадратической ошибки.....	651
A.1.3	Средняя абсолютная ошибка.....	651
A.1.4	Средняя абсолютная процентная ошибка.....	653
A.1.5	Потери Хьюбера.....	653
A.1.6	Квантильные потери.....	654
A.1.7	Потери Пуассона.....	656
A.2	Функции потерь для классификации	657
A.2.1	Бинарная перекрестная энтропия.....	658
A.2.2	Категориальная перекрестная энтропия.....	658
A.2.3	Расхождение Кульбака–Лейблера.....	659
A.3	Функции потерь для обучения представлению	662
A.3.1	Контрастные потери	662
A.3.2	Триплетные потери.....	663
A.3.3	Многоклассовые N -парные потери	665
A.3.4	Потери InfoNCE.....	666
A.3.5	Потери ArcFace.....	667
В	Метрики оценивания моделей	670
V.1	Метрики для регрессии.....	670
V.1.1	Средневзвешенная процентная ошибка.....	671
V.1.2	Взвешенная квантильная потеря.....	671
V.1.3	Резкость границ прогноза	672
V.2	Метрики качества классификации.....	672
V.2.1	Матрица несоответствий и связанные с ней метрики	673
V.2.2	Кривая ROC и AUC	675
V.2.3	Кривая точности–полноты	677
V.2.4	Оценка F1.....	678
V.3	Метрики для поиска	679
V.4	Метрики качества ранжирования	680
V.4.1	Коэффициент попаданий.....	681
V.4.2	Обобщенная средняя точность.....	681
V.4.3	Дисконтированная совокупная прибыль	682
V.5	Метрики качества генерации естественного языка	682
V.5.1	Точность и полнота совпадения.....	684
V.5.2	BLEU.....	684
V.5.3	ROUGE.....	686
V.5.4	BERTScore.....	686
V.5.5	G-Eval.....	688
	Библиография	690
	Предметный указатель	704

ОТ ИЗДАТЕЛЬСТВА

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте www.dmkpress.com, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу dmkpress@gmail.com. Сделав это, вы избавите других читателей от недопонимания и сможете нам улучшить последующие издания этой книги.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Благодарности

Здесь будут фамилии тех, кто помогал изданию этой книги, прислав в издательство найденные ошибки или ссылку на подозрительные материалы.

ПРЕДИСЛОВИЕ

Роль автоматизации и оптимизации на основе данных в операционной деятельности предприятия возростала на протяжении многих десятилетий, но за последние десять лет диапазон вариантов эффективного использования автоматизированных систем значительно расширился благодаря появлению методов глубокого обучения. Эти разработки сформировали разнообразную среду методов принятия решений и автоматизации, включая традиционные эконометрические модели и алгоритмы оптимизации, специализированные методы машинного обучения для компьютерного зрения и обработки естественного языка, которые, однако, часто могут быть адаптированы к другим областям, а также новые методы, такие как глубокое обучение с подкреплением, имеющие ограниченное применение в корпоративной практике. В этой книге будет показано, как предприятие может извлечь максимальную выгоду из сочетания традиционных методов моделирования, оптимизации и симуляции с методами глубокого обучения и обучения с подкреплением за счет интеллектуальной автоматизации широкого спектра операций предприятия, включая маркетинг, управление цепочками поставок и контроль производства. Назначение этой книги – представить инженерный базис, а также набор практических рецептов применения комбинаций этих методов в условиях реального предприятия.

ЦЕЛЕВАЯ АУДИТОРИЯ

Эта книга адресована специалистам по данным и менеджерам по аналитике и призвана сформировать у них навык системного подхода к решению задач принятия корпоративных решений и оптимизации с использованием методов глубокого обучения, обучения с подкреплением и вероятностного программирования. Работая над книгой, я задался целью предложить систематическую методику перевода различных вариантов деятельности предприятий в количественные, статистические и оптимизационные задачи и их декомпозиции на задачи машинного обучения. В начале книги дается обзор базовых компонентов системы. Подробные описания использования моделей и алгоритмов для конкретных случаев будут рассмотрены позже, но я не стремился обеспечить детальное изложение теории машинного обучения и ее математических основ. Ожидается, что читатель знаком с основными концепциями науки о данных и машинного обучения, а также имеет практический опыт статистического моделирования, включая как традиционные методы, так и методы глубокого обучения.

Книга также будет полезна специалистам по науке о данных и машинному обучению, имеющим опыт работы в биоинформатике, физике или других

областях, не связанных с типичной операционной деятельностью предприятия, и желающим узнать о специализированных методах моделирования для маркетинга, цепочек поставок и производственных приложений.

СТРУКТУРА КНИГИ И РЕКОМЕНДАЦИИ ПО ПРОЧТЕНИЮ

Я подхожу к задаче построения корпоративных ИИ-решений с точки зрения системного проектирования, рассматривая алгоритмы машинного обучения главным образом как готовые компоненты и уделяя особое внимание адаптации общих методов к конкретным сценариям использования на предприятии. Первые три главы посвящены описанию методики декомпозиции задач предприятия на задачи машинного обучения и оптимизации, а также рассмотрению основных категорий алгоритмов машинного обучения, необходимых для решения таких задач. Однако выбор и категоризация алгоритмов и методов несколько отличаются от канонической категоризации, представленной в большинстве учебников по машинному обучению, поскольку мы сосредоточимся исключительно на корпоративных приложениях. Затем читатель познакомится с рядом рецептов (R1–R14) для конкретных случаев применения в области маркетинга, цепочки поставок и производства. В большинстве рецептов я использую следующую трехуровневую структуру, чтобы охватить как теоретические, так и практические аспекты решений:

- *варианты решений*: в каждом рецепте определяется бизнес-задача и обсуждается несколько вариантов решения. Некоторые решения не требуют существенного изменения общих моделей или алгоритмов, и рецепт сфокусирован в основном на практических аспектах, таких как интеграция и эконометрические соображения. Другие решения требуют разработки специализированных алгоритмов, и они будут описаны со всеми математическими деталями;
- *прототипы*: в каждом рецепте показан процесс разработки одного или нескольких базовых прототипов, иллюстрирующих подход и основные свойства решения. Эти эталонные реализации обычно используют синтетические данные или симуляторы, чтобы избежать сложностей, связанных с наборами реальных данных. Обычно в рецепте описано, как работает прототип и как результаты доказывают жизнеспособность решения, но я старался не загромождать книгу подробностями низкоуровневой реализации; вместо этого в каждом рецепте приведены ссылки на соответствующие блокноты в репозитории исходного кода, чтобы читатель мог при необходимости глубже изучить реализацию;
- *подробные примеры*: для некоторых рецептов разработаны более полные эталонные реализации с применением крупных выборок данных, созданных на основе статистических параметров реальных наборов данных. Эти реализации помогают раскрыть проблемы, которые не проявляют себя в прототипах меньшего масштаба. Как и в случае с про-

тотипами, я не рассматриваю эти реализации на уровне исходного кода, а даю ссылки на репозиторий с полными блокнотами.

Книгу можно читать последовательно, чтобы систематически изучать идеи, заложенные в основу корпоративного ИИ, включая его основные компоненты и категории решений. Однако рецепты в основном независимы, и читатели, знакомые с основами глубокого обучения, могут бегло просмотреть первые три главы и далее изучать рецепты в любом порядке в соответствии со своими потребностями и приоритетами.

ПРИМЕРЫ ИСХОДНОГО КОДА И ДАННЫХ

Примеры реализации и прототипы, упомянутые в этой книге, а также несколько дополнительных моделей выпущены в виде проекта с открытым исходным кодом под названием TensorHouse. Этот проект доступен на <https://github.com/ikatsov/tensor-house>. Там создана специальная ветка `book-enterprise-ai-edition-2.1` с версией кода, совместимой с данной книгой. Я использую TensorFlow в качестве основной платформы для моделей глубокого обучения и несколько других платформ и библиотек для вспомогательных операций и специализированных функций.

Рецепты, представленные в этой книге, и примеры реализации дополняют друг друга. Рецепты предлагают всесторонний анализ бизнес-задач и вариантов их решения, но лишь краткое изложение фактических реализаций. Демонстрационные блокноты на GitHub содержат подробные пошаговые инструкции по реализации определенных решений, но не дублируют весь анализ и теоретические детали, представленные в книге.

ЧТО НОВОГО ВО ВТОРОМ ИЗДАНИИ?

Второе издание включает в себя два крупных обновления и множество мелких изменений. Во-первых, добавлен обширный новый материал по генеративному ИИ, охватывающий теоретические основы (глава 3), приложения для языкового моделирования (рецепт R7) и решения для генерации изображений (рецепт R8). Во-вторых, часть IV, посвященная управлению доходами и запасами, существенно переработана, и в нее добавлена отдельная глава, посвященная прогнозированию спроса (рецепт R9). Кроме того, существенно обновлена глава 2, посвященная прогнозным моделям и основам глубокого обучения; добавлены два приложения, предоставляющие исчерпывающий справочник по функциям потерь и показателям оценки, используемым в корпоративных приложениях; доработаны и улучшены все иллюстрации.

Часть I

БАЗОВЫЕ КОМПОНЕНТЫ

Первая часть этой книги посвящена методике преобразования задач автоматизации принятия производственных решений в задачи машинного обучения. Главу 1 мы начинаем с рассмотрения типичных уровней принятия решений в рамках деятельности предприятия и определения основных концепций, которые применимы к широкому спектру прикладных ситуаций. В главе 2 разрабатываем набор инструментов для изучения сопоставлений между объектами, их атрибутами и траекториями, который позволяет нам выводить скрытые свойства и прогнозировать будущие состояния объектов. В главе 3 мы дополняем этот набор инструментов методами создания сложных объектов, таких как изображения. Наконец, в главе 4 обсуждаем методы автоматизации принятия решений и управления объектами.

1

АВТОМАТИЗАЦИЯ РЕШЕНИЙ И ПРОЦЕССОВ В ДЕЯТЕЛЬНОСТИ ПРЕДПРИЯТИЯ

Корпоративный *искусственный интеллект* (ИИ) можно неофициально определить как набор методов улучшения операционной деятельности предприятия с использованием статистического обучения и вероятностных рассуждений. Этот набор очень обширен и включает в себя методы совершенствования стратегических решений на уровне всего предприятия. Среди них прогнозирование спроса, методы оптимизации решений на уровне отдельных бизнес-процессов, такие как таргетирование рекламных акций и управление запасами, а также методы, помогающие автоматизировать или оптимизировать отдельные транзакции, такие как модели обнаружения объектов и системы управления диалогами.

Хотя концепция корпоративного ИИ относительно нова, можно утверждать, что по своей сути она так же стара, как и сама концепция предприятия. Исторически сложилось так, что первой категорией задач, которые решали с использованием методов, основанных на данных, было принятие стратегических решений. Идея о том, что финансовое состояние и будущую траекторию экономического субъекта можно оценить с помощью агрегированных финансовых отчетов, была хорошо известна с древних времен. В этом смысле бухгалтеры были первыми специалистами по корпоративным данным. Собирая и агрегируя финансовые записи, бухгалтер создает краткое количественное представление предприятия в пространстве определенных показателей, таких как прибыль и доходы, и это пространство затем служит для оценки финансовых показателей предприятия, прогнозирования будущих результатов и составления отчетов, на которых основываются управленческие решения. Хотя может показаться, что связать бухгалтерский учет с ИИ – это слишком сложно, скоро вы убедитесь, что даже основные количественные методы, используемые в бухгалтерском учете, торговле акциями и инвести-

циях, можно органично и безопасно объединить с методами, которые обычно свойственны современному ИИ.

Массовый переход на второй уровень автоматизации принятия решений, то есть оптимизации тактических решений на уровне отдельных процессов внутри предприятия, произошел в основном в последней трети XX века благодаря появлению доступных и достаточно мощных компьютеров. Разработки того периода были сосредоточены преимущественно на решении задач численной и комбинаторной оптимизации в управлении цепочками поставок, транспортировке и производстве. Некоторые из этих приложений также включали статистический анализ корпоративных данных, но внедрение методов, основанных на данных, было ограничено низким уровнем цифровизации как бизнеса, так и потребителей.

Следующий уровень интеллектуальной автоматизации был достигнут в середине 2010-х годов, главным образом благодаря трем факторам. Началом послужила комплексная цифровизация корпоративной среды, за которой последовало массовое внедрение стратегий больших данных, вынудивших все корпоративные процессы непрерывно генерировать подробные, статистически анализируемые данные. Вторым фактором стала цифровизация потребительской среды, которая позволила персонализировать и оптимизировать продукты, услуги и их представления в реальном времени. Третьим (и последним) фактором стал революционный прогресс в статистических методах, особенно в глубоком обучении, позволивший понимать текстовые и визуальные данные с помощью программных систем. Это, в свою очередь, привело к появлению широкого спектра новых типов автоматизации, связанных с *обработкой естественного языка* (natural language processing, NLP) и использованием *компьютерного зрения* (computer vision, CV).

Эта книга в основном посвящена методам автоматизации мелкомасштабных решений и процессов в средах с большим объемом данных, что соответствует последнему уровню в приведенной выше классификации. Однако прежде чем мы углубимся в детали этих методов, необходимо сформировать понимание базовой методики, которая поможет нам правильно планировать и внедрять решения по автоматизации анализа данных и принятия решений внутри предприятия. Мы должны рассматривать проектирование отдельных компонентов с точки зрения системы, под правильным углом и в правильном контексте.

1.1 МЕТОДИКА СЦЕНАРНОГО ПЛАНИРОВАНИЯ

Принято считать, что корпоративные системы и механизмы создают преимущественно с учетом конкретных бизнес-целей, а степень успеха обычно измеряют с помощью *ключевых показателей эффективности* (key performance indicator, KPI). Например, можно разработать новый алгоритм рекомендации продуктов с целью повышения коэффициента конверсии сеанса с 3 до

3.5 %; можно создать систему оптимизации цен, которая увеличит прибыль на 200 млн долл., а на сборочной линии установить новую систему компьютерного зрения, чтобы снизить уровень брака на 20 %. Это одна из самых основополагающих и известных парадигм в мире бизнеса. Однако ее практическая реализация весьма сложна.

Одна из проблем заключается в том, что многие улучшения и действия влияют на несколько КРІ одновременно, и при положительном влиянии на одни показатели влияние на другие может быть отрицательным. Например, распродажи по сниженной цене могут увеличить товарооборот, но снизить прибыль, увеличение уровня складских запасов может улучшить доступность продукции и качество обслуживания клиентов, но увеличить затраты на хранение, а оптимальные для прибыли цены могут разрушить бизнес, которому необходимо увеличить свою долю на рынке, чтобы стать устойчивым. Анализ таких компромиссов и принятие решений по ним являются сложной задачей, а преобразование их в формальные задачи оптимизации – тем более.

Вторая проблема заключается в долгосрочном характере бизнес-операций, что иногда затрудняет определение и измерение одного ключевого показателя эффективности. Например, распродажи широко применяются розничными продавцами и производителями и, как известно, являются эффективным способом увеличения объема продаж, но прирост сбыта в период промоакций часто происходит за счет будущих продаж. Это особенно справедливо в отношении расходных материалов, таких как бумажные полотенца: рекламные акции часто побуждают потребителей покупать и накапливать большее количество продуктов, а затем ждать следующей распродажи. Это делает краткосрочные измерения роста продаж неточными или вводящими в заблуждение, в то время как долгосрочные измерения сложны и тоже не всегда достоверны, поскольку все другие части окружающей среды независимо и непредсказуемо меняются, искажая наблюдения.

Третья проблема – измеримость ключевых показателей эффективности. В предыдущем примере, касающемся стимулирования продаж, долгосрочные последствия являются одним, но не единственным фактором, который способен лишить измерения достоверности. Рекламные акции и изменения цен могут заставить потребителей переключаться с одного продукта или поставщика на другого, создавая эффекты перекрестных продуктов и перекрестных розничных продаж. Большинство продавцов знают о таких эффектах и хорошо понимают, что показатели роста продаж отдельных продуктов могут вводить в заблуждение, но более полные и точные измерения затруднены из-за более высоких требований к данным и сложности реализации. Во многих практических случаях имеет смысл разделить одну цель с множеством внутренних факторов на несколько показателей эффективности, которые можно отслеживать отдельно, и это возвращает нас к первой проблеме, связанной с множеством конфликтующих ключевых показателей эффективности.

Наконец, мы можем отнести к серьезным проблемам оценку решения. Разработка решения, целью которого является улучшение определенных показателей, обычно нуждается в инструментах оценки его производительности до внедрения в производство и сбора фактических результатов. Эта

проблема унаследовала все ранее упомянутые сложности, которые по сути принадлежали области описательной аналитики, и добавила к ним элементы прогнозирования и опережающего анализа.

Из сказанного следует, что для преобразования задач предприятия в задачи оптимизации и автоматизации решений обычно необходимо учитывать множество объектов, показателей, их корреляций и совместной динамики. Далее мы попытаемся объединить эти основные идеи в более формальную структуру, основанную на следующих понятиях.

Объект (entity, сущность): при автоматизации или оптимизации какого-либо аспекта предприятия мы обычно сосредоточиваемся на повышении производительности одного или нескольких объектов, которыми могут быть предприятие в целом, бизнес-подразделение, местоположение, продукт или клиент.

Сценарий (scenario): определяет действие или последовательность действий, которые мы потенциально можем выполнить для достижения определенных улучшений. Обычно мы можем выбирать между несколькими вариантами сценариев, включая универсальное бездействие.

Пространство полезности (utility space): объекты, с которыми мы работаем, могут быть описаны с использованием одной или нескольких метрик (KPI), а пространство, охватываемое измерениями метрики, принято называть пространством полезности. Например, в традиционной бизнес-аналитике обычно используются такие пространства полезности, как *Доход*Маржа* и *Риск*Вознаграждение*. Мы также будем называть пространства полезности с двумя или более измерениями *картами полезности (utility map)*, или *картами ценностей (utility value)*.

Траектория (trajectory): выполнение сценария заставляет объекты двигаться в пространстве метрик – в идеале в том направлении, которое выгодно бизнесу. Следовательно, сценарий оставляет след в пространстве полезности, который мы будем называть траекторией.

Граница Парето (Pareto frontier): если метрики, используемые для построения пространства полезности, находятся в обратной зависимости (например, стоимость и качество), мы обычно можем выбирать между несколькими сценариями, образующими различные компромиссы. Однако в такой ситуации невозможно улучшить все показатели одновременно, поэтому набор наилучших возможных компромиссов сформирует границу Парето, которую можно представить как поверхность в пространстве полезности. Однако мы можем сдвинуть границу Парето, если найдем возможность одновременного улучшения всех показателей пространства полезности за счет изменения некоторых базовых факторов.

Эти понятия проиллюстрированы на рис. 1.1, где организация движется в двухмерном пространстве полезности, и рассматриваются три альтернативных сценария воздействия, которые приводят к трем различным результатам, образующим границу.

На данный момент наша методика сценарного планирования довольно абстрактна, и не совсем ясно, как применить ее к практическим задачам и связать с методами *машинного обучения (machine learning, ML)*. Этот пробел будет постепенно закрыт в следующих разделах, где мы обсудим более кон-

кретные примеры для различных типов задач, а затем методы машинного обучения, необходимые для реализации этого подхода. Но цель данной главы заключается в том, чтобы рассмотреть различные примеры использования машинного обучения с точки зрения сценарного планирования, а более строгое описание теоретических основ будет представлено позже.

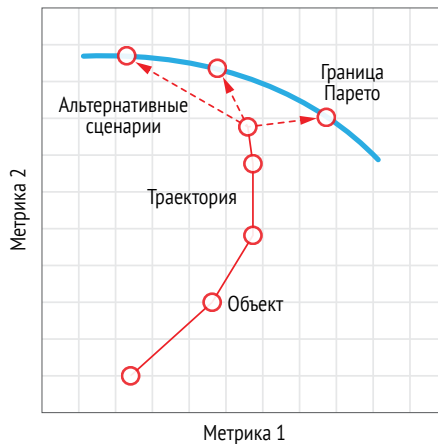


Рис. 1.1. Основные понятия методики сценарного планирования

1.1.1. Стратегия: предприятие в целом

Начнем с примеров стратегического анализа, в которых стоит задача изучить траекторию развития всей компании. Этот тип анализа обычно используется венчурными инвесторами для оценки стартапов и портфельными менеджерами для оценки публичных компаний. Вероятно, это максимально возможный уровень, на котором можно рассматривать применение методов автоматизации принятия решений.

Представьте, что нам нужно оценить компанию на ранней стадии развития, чтобы принять решение о том, стоит ли в нее инвестировать. Один из ключевых вопросов оценки заключается в том, насколько хорошо компания и ее продукт соответствуют рынку: выглядит ли компания готовой к росту или видны признаки сопротивления рынка и замедления развития (Hsu, 2019). В определенной степени на этот вопрос можно ответить, используя количественные методы, а результаты определяют дальнейшие действия как инвесторов, так и менеджмента компании.

Мы можем начать процедуру оценки с так называемого *анализа факторов роста* (growth accounting), который фокусируется на анализе компонентов дохода и эволюции этих компонентов с течением времени. В сценарном планировании мы определяем объекты как различные категории доходов, пространство полезности – как единственное измерение, выражаемое в долларах, а траектории – как эволюцию категорий доходов. Для анализа разложим выручку предприятия за период времени t следующим образом:

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru