

Оглавление

Предисловие	13
Вступление	15
Благодарности	17
О книге	18
Структура книги	18
Как читать эту книгу	19
Целевая аудитория	20
Формат кода, загрузки и требования к ПО	20
От издательства	21
Об авторах	22
Часть I. Последовательность действий при машинном обучении.	23
Глава 1. Что такое машинное обучение?	24
1.1. Как обучаются машины	25
1.2. Принятие решений на основе данных	30
1.2.1. Традиционные подходы	32
1.2.2. Подход с машинным обучением	36
1.2.3. Пять преимуществ машинного обучения.	42
1.2.4. Сложности	43

1.3. Рабочий процесс: от данных до внедрения	44
1.3.1. Сбор и подготовка данных	45
1.3.2. Обучение модели на данных	46
1.3.3. Оценка производительности модели	49
1.3.4. Оптимизация производительности модели	50
1.4. Усовершенствованные способы повышения эффективности	51
1.4.1. Предварительная обработка данных и проектирование признаков	51
1.4.2. Непрерывное совершенствование моделей	54
1.4.3. Масштабирование моделей	54
1.5. Заключение	55
1.6. Терминология	56
Глава 2. Реальные данные	57
2.1. Первый этап: сбор данных	59
2.1.1. Определяем набор входных признаков	62
2.1.2. Наблюдаемое значение целевой переменной	64
2.1.3. Достаточный объем обучающих данных	65
2.1.4. Репрезентативность обучающей выборки	68
2.2. Подготовка данных к моделированию	69
2.2.1. Категориальные признаки	70
2.2.2. Отсутствующие данные	73
2.2.3. Основы проектирования признаков	76
2.2.4. Нормализация данных	78
2.3. Визуализация данных	80
2.3.1. Мозаичные диаграммы	81
2.3.2. Диаграммы размаха	83
2.3.3. Графики плотности	86
2.3.4. Диаграммы рассеяния	88

2.4. Заключение	89
2.5. Терминология	90
Глава 3. Моделирование и прогнозирование	91
3.1. Основы моделирования с машинным обучением	92
3.1.1. Поиск связи между входными данными и целевой переменной	93
3.1.2. Зачем нужна хорошая модель	95
3.1.3. Типы методов моделирования	97
3.1.4. Обучение с учителем и без	100
3.2. Классификация: распределение по классам	101
3.2.1. Построение классификатора и получение предсказаний	103
3.2.2. Классификация сложных нелинейных данных	108
3.2.3. Классификация в случае множества классов	111
3.3. Регрессия: предсказание численных значений	113
3.3.1. Построение регрессора и генерация прогнозов	115
3.3.2. Регрессия для сложных нелинейных данных	119
3.4. Заключение	121
3.5. Терминология	122
Глава 4. Оценка и оптимизация модели	123
4.1. Оценка прогностической точности на новых данных	125
4.1.1. Проблема: переобучение и чрезмерно оптимистическая оценка модели	125
4.1.2. Решение: скользящий контроль	129
4.1.3. На что следует обращать внимание при перекрестной проверке	134
4.2. Оценка моделей классификации	135
4.2.1. Точность для отдельных классов и таблица сопряженности	138

4.2.2. Компромиссы при оценке точности и ROC-кривые	140
4.2.3. Многоклассовая классификация	144
4.3. Оценка моделей регрессии	147
4.3.1. Простые показатели эффективности регрессионных моделей	148
4.3.2. Исследование остатков	150
4.4. Оптимизация модели путем подбора параметров	152
4.4.1. Параметры настройки ML-алгоритмов	152
4.4.2. Сеточный поиск	154
4.5. Заключение	158
4.6. Терминология	159

Глава 5. Основы проектирования признаков 161

5.1. Мотивация: в чем польза проектирования признаков?	162
5.1.1. Что такое проектирование признаков?	162
5.1.2. Пять причин проектирования признаков	163
5.1.3. Проектирование признаков и знание предметной области	165
5.2. Основные этапы проектирования признаков	166
5.2.1. Пример: рекомендация события	167
5.2.2. Обработка даты и времени	170
5.2.3. Извлечение признаков из обычного текста	172
5.3. Выбор признаков	174
5.3.1. Прямой отбор и обратное исключение	178
5.3.2. Отбор признаков для исследования данных	180
5.3.3. Практический пример отбора признаков	181
5.4. Заключение	184
5.5. Терминология	186

Часть II. Практическое применение 187

Глава 6. Пример: чаевые для таксистов 188

6.1. Данные: сведения о чаевых и плате за проезд 189

6.1.1. Визуализация данных 190

6.1.2. Формулировка задачи и подготовка данных 194

6.2. Моделирование 197

6.2.1. Базовая линейная модель 197

6.2.2. Нелинейный классификатор 199

6.2.3. Добавление категориальных признаков 202

6.2.4. Добавление временных признаков 203

6.2.5. Аналитическая оценка модели 205

6.3. Заключение 206

6.4. Терминология 207

Глава 7. Усовершенствованное проектирование признаков 208

7.1. Более сложные текстовые признаки 209

7.1.1. Модель «мешок слов» 209

7.1.2. Тематическое моделирование 213

7.1.3. Расширение содержимого 217

7.2. Признаки, извлекаемые из изображений 219

7.2.1. Простые признаки 220

7.2.2. Извлечение объектов и форм 222

7.3. Признаки временных рядов 228

7.3.1. Типы временных рядов 228

7.3.2. Предсказания на основе временных рядов 231

7.3.3. Признаки классических временных рядов 232

7.3.4. Проектирование признаков для потоков событий 238

7.4. Заключение	239
7.5. Терминология	241
Глава 8. Пример обработки естественного языка	243
8.1. Изучение данных и сценарии их применения	244
8.1.1. Первый взгляд на набор данных	245
8.1.2. Анализ набора данных.	246
8.1.3. Так какой же будет наша задача?	247
8.2. Генерация базовых NLP-признаков и построение первого варианта модели	252
8.2.1. Признаки из «мешка слов»	253
8.2.2. Модель на базе наивного байесовского классификатора.	255
8.2.3. Нормализация признаков, полученных из «мешка слов», алгоритмом tf-idf.	260
8.2.4. Оптимизация параметров модели.	262
8.3. Усовершенствованные алгоритмы и тонкости процесса внедрения	267
8.3.1. Word2vec-признаки	268
8.3.2. Модель на базе алгоритма «случайный лес»	270
8.4. Заключение	273
8.5. Терминология	274
Глава 9. Масштабирование процесса машинного обучения	275
9.1. Перед началом масштабирования	276
9.1.1. Определяем важные аспекты.	277
9.1.2. Прореживание обучающей выборки вместо масштабирования?	280
9.1.3. Масштабируемые системы управления данными.	282
9.2. Масштабирование конвейера ML-моделирования	285
9.2.1. Масштабирование обучающих алгоритмов	286

9.3. Масштабирование предсказаний	291
9.3.1. Масштабирование объема предсказаний	292
9.3.2. Масштабирование скорости предсказаний	293
9.4. Заключение	296
9.5. Терминология	298
Глава 10. Пример с цифровой рекламой	300
10.1. Показ рекламы	302
10.2. Данные, связанные с цифровой рекламой	303
10.3. Проектирование признаков и стратегия моделирования	304
10.4. Размер и форма данных	306
10.5. Сингулярное разложение	309
10.6. Оценка и оптимизация ресурсов	312
10.7. Моделирование	314
10.8. Метод k-ближайших соседей	315
10.9. «Случайные леса»	318
10.10. Другие практические моменты	319
10.11. Заключение	321
10.12. Терминология	322
10.13. Подводим итоги	323
Приложение. Популярные алгоритмы машинного обучения . . .	326

Предисловие

В последние годы машинное обучение (ML — machine learning) превратилось в большой бизнес — фирмы используют его, чтобы заработать денег, прикладные исследования бурно развиваются как в индустриальной, так и в академической среде, а любопытные разработчики везде ищут возможность поднять свой уровень владения темой. Но возникший *спрос* намного превышает скорость *появления* хороших методик для изучения применяемых на практике техник. Наша книга призвана удовлетворить данный спрос.

Прикладное машинное обучение совмещает в себе равные доли математических принципов и полученных эмпирическим путем приемов, — другими словами, это настоящее искусство. Слишком сильная концентрация только на одном из этих аспектов в ущерб другому — проигрышная стратегия. Тут важен баланс.

Долгое время самым лучшим и единственным способом постижения машинного обучения было получение ученой степени в одной из областей, которые (по большей части независимо друг от друга) развивали статистические методы и техники оптимизации. Основной упор эти программы делали на ключевые алгоритмы, в том числе на их теоретические свойства и ограничения, а также на характерные особенности относящихся к данной сфере задач. Впрочем, параллельно не менее ценные знания накапливались неофициальным образом — в процессе неформального общения на конференциях, обмена информацией и сценариями обработки данных между коллегами из исследовательских лабораторий. Именно эти знания, по большому счету, и позволили установить, какие алгоритмы больше всего подходят в каждой ситуации, как обрабатывать данные на каждом этапе и как связать между собой различные этапы рабочего процесса.

Сейчас мы живем в эпоху открытого исходного кода, с готовыми к использованию высококачественными алгоритмами, доступными на сайте