

УДК 004.4
ББК 32.972
Б91

Бурков А.

Б91 Инженерия машинного обучения / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2022. – 306 с.: ил.

ISBN 978-5-93700-125-2

Книга представляет собой подробный обзор передовых практик и паттернов проектирования в области прикладного машинного обучения. В отличие от многих учебников, уделяется внимание инженерным аспектам МО. Рассматриваются сбор, хранение и предобработка данных, конструирование признаков, а также тестирование и отладка моделей, развертывание и вывод из эксплуатации, сопровождение на этапе выполнения и в процессе эксплуатации. Главы книги можно изучать в любом порядке.

Издание будет полезно тем, кто собирается использовать машинное обучение в крупномасштабных проектах. Предполагается, что читатель знаком с основами МО и способен построить модель при наличии подходящим образом отформатированного набора данных.

Дизайн обложки разработан с использованием ресурса [freepik.com](https://www.freepik.com)

УДК 004.4
ББК 32.972

Copyright Title of English-language edition: "Machine Learning Engineering", ISBN 978-1-9995795-7-9 published by Andriy Burkov. Copyright © 2020 Andriy Burkov. Russian-language edition copyright © 2022 by DMK Press Publishing. All rights reserved.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN 978-1-9995795-7-9 (англ.)
ISBN 978-5-93700-125-2 (рус.)

© Andriy Burkov, 2020
© Перевод, оформление, издание,
ДМК Пресс, 2022

Содержание

От издательства	14
Вступительное слово	15
Предисловие	17
Глава 1. Введение	19
1.1. Обозначения и определения	19
1.1.1. Структуры данных	19
1.1.2. Заглавная сигма	21
1.2. Что такое машинное обучение	21
1.2.1. Обучение с учителем	22
1.2.2. Обучение без учителя	23
1.2.3. Обучение с частичным привлечением учителя	24
1.2.4. Обучение с подкреплением	24
1.3. Терминология машинного обучения	25
1.3.1. Данные, используемые прямо и косвенно	25
1.3.2. Первичные и аккуратные данные	26
1.3.3. Обучающие и зарезервированные наборы	27
1.3.4. Ориентир	28
1.3.5. Конвейер машинного обучения	28
1.3.6. Параметры и гиперпараметры	29
1.3.7. Классификация и регрессия	29
1.3.8. Обучение на основе модели и обучение на основе экземпляров	30
1.3.9. Поверхностное и глубокое обучение	30
1.3.10. Обучение и оценивание	30
1.4. Когда следует использовать машинное обучение	31
1.4.1. Когда задача слишком сложна для кодирования	31
1.4.2. Когда задача постоянно меняется	32
1.4.3. Когда речь идет о задаче восприятия	32
1.4.4. Когда это неизученное явление	32
1.4.5. Когда задача имеет простую целевую функцию	33
1.4.6. Когда это экономически выгодно	33
1.5. Когда не следует использовать машинное обучение	34
1.6. Что такое инженерия машинного обучения	34
1.7. Жизненный цикл проекта машинного обучения	36
1.8. Резюме	37
Глава 2. Прежде чем приступить к проекту	39
2.1. Определение приоритетов проекта машинного обучения	39
2.1.1. Последствия машинного обучения	39
2.1.2. Стоимость машинного обучения	40

2.2. Оценивание сложности проекта машинного обучения	41
2.2.1. Неизвестные	41
2.2.2. Упрощение задачи	42
2.2.3. Нелинейный прогресс	43
2.3. Определение цели проекта машинного обучения	43
2.3.1. Что модель может делать	43
2.3.2. Свойства успешной модели	44
2.4. Организация группы машинного обучения	45
2.4.1. Две традиции	45
2.4.2. Члены группы машинного обучения	46
2.5. Причины провалов проектов машинного обучения	47
2.5.1. Нехватка квалифицированных кадров	47
2.5.2. Отсутствие поддержки со стороны руководства	48
2.5.3. Отсутствующая инфраструктура данных	48
2.5.4. Трудности с разметкой данных	49
2.5.5. Разобщенные организации и отсутствие сотрудничества	49
2.5.6. Технически невыполнимые проекты	50
2.5.7. Нестыковка между техническими и коммерческими группами	50
2.6. Резюме	51

Глава 3. Сбор и подготовка данных

3.1. Вопросы к данным	53
3.1.1. Доступны ли данные?	54
3.1.2. Насколько велик объем данных?	54
3.1.3. Пригодны ли данные для использования?	56
3.1.4. Понятны ли данные?	58
3.1.5. Надежны ли данные?	58
3.2. Типичные проблемы с данными	60
3.2.1. Высокая стоимость	60
3.2.2. Плохое качество	62
3.2.3. Зашумленность	62
3.2.4. Смещение	63
Типы смещения	63
Как избежать смещения	67
3.2.5. Низкая предсказательная способность	69
3.2.6. Устаревшие примеры	70
3.2.7. Выбросы	70
3.2.8. Просачивание данных	71
3.3. Что считать хорошими данными	72
3.3.1. Хорошие данные информативны	72
3.3.2. Хорошие данные обладают хорошим покрытием	72
3.3.3. Хорошие данные отражают реальные входы	73
3.3.4. Хорошие данные несмещенные	73
3.3.5. Хорошие данные не являются результатом петли обратной связи	73
3.3.6. У хороших данных согласованные метки	74
3.3.7. Хорошие данные достаточно велики	74
3.3.8. Сводный перечень свойств хороших данных	74

3.4. Обработка данных о взаимодействии	75
3.5. Причины просачивания данных.....	75
3.5.1. Цель является функцией от признака	76
3.5.2. Признак скрывает цель.....	76
3.5.3. Признак из будущего.....	77
3.6. Разбиение данных.....	78
3.6.1. Просачивание во время разбиения.....	79
3.7. Обработка отсутствия атрибутов	80
3.7.1. Методы подстановки данных.....	80
3.7.2. Просачивание во время подстановки	82
3.8. Приращение данных.....	82
3.8.1. Приращение данных для изображений	82
3.8.2. Приращение данных для текста	84
3.9. Обработка несбалансированных данных.....	85
3.9.1. Выборка с избытком.....	86
3.9.2. Выборка с недостатком.....	87
3.9.3. Гибридные стратегии	87
3.10. Стратегии выборки данных.....	88
3.10.1. Простая случайная выборка.....	89
3.10.2. Систематическая выборка.....	90
3.10.3. Стратифицированная выборка.....	90
3.11. Хранение данных	90
3.11.1. Форматы данных	91
3.11.2. Уровни хранения данных	92
3.11.3. Версионирование данных	94
3.11.4. Документация и метаданные	96
3.11.5. Жизненный цикл данных.....	96
3.12. Дополнительные рекомендации по работе с данными	97
3.12.1. Воспроизводимость.....	97
3.12.2. Сначала данные, потом алгоритм.....	97
3.13. Резюме	98

Глава 4. Конструирование признаков

4.1. Зачем конструировать признаки.....	100
4.2. Как конструируются признаки.....	101
4.2.1. Конструирование признаков для текста	102
4.2.2. Почему мешок слов работает.....	105
4.2.3. Преобразование категориальных признаков в числа.....	105
4.2.4. Хеширование признаков	108
4.2.5. Тематическое моделирование	109
4.2.6. Признаки для временных рядов	112
4.2.7. Проявите свои творческие способности.....	114
4.3. Штабелирование признаков.....	115
4.3.1. Штабелирование векторов признаков	115
4.3.2. Штабелирование индивидуальных признаков	116
4.4. Свойства хороших признаков	117

4.4.1. Высокая предсказательная способность	117
4.4.2. Быстрое вычисление	117
4.4.3. Надежность	118
4.4.4. Некоррелированность	118
4.4.5. Другие свойства	118
4.5. Отбор признаков	119
4.5.1. Отрезание длинного хвоста	119
4.5.2. Boruta	120
4.5.3. L1-регуляризация	123
4.5.4. Зависящий от задачи отбор признаков	123
4.6. Синтезирование признаков	123
4.6.1. Дискретизация признаков	124
4.6.2. Синтез признаков из реляционных данных	125
4.6.3. Синтезирование признаков по данным	126
4.6.4. Синтезирование признаков по другим признакам	127
4.7. Обучение признаков на данных	128
4.7.1. Погружения слов	128
4.7.2. Погружения документов	130
4.7.3. Погружения всего, чего угодно	131
4.7.4. Выбор размерности погружения	132
4.8. Понижение размерности	132
4.8.1. Быстрое понижение размерности методом PCA	133
4.8.2. Понижение размерности с целью визуализации	133
4.9. Масштабирование признаков	133
4.9.1. Нормировка	134
4.9.2. Стандартизация	135
4.10. Просачивание данных при конструировании признаков	136
4.10.1. Возможные проблемы	136
4.10.2. Решение	136
4.11. Хранение и документирование признаков	136
4.11.1. Файл схемы	137
4.11.2. Хранилище признаков	138
4.12. Рекомендации по конструированию признаков	141
4.12.1. Генерируйте много простых признаков	141
4.12.2. Повторно используйте унаследованные системы	141
4.12.3. Используйте идентификаторы как признаки, когда это необходимо	142
4.12.4. ...но по возможности уменьшайте количество значений	142
4.12.5. Осторожнее со счетчиками	143
4.12.6. Отбирайте признаки, когда необходимо	143
4.12.7. Тщательно тестируйте код	144
4.12.8. Синхронизируйте код, модель и данные	144
4.12.9. Изолируйте код выделения признаков	144
4.12.10. Сериализуйте модель и экстрактор признаков совместно	145
4.12.11. Протоколируйте значения признаков	145
4.13. Резюме	145

Глава 5. Обучение модели с учителем (часть 1)	147
5.1. Прежде чем приступить к работе над моделью	148
5.1.1. Проверка согласованности со схемой	148
5.1.2. Определение достижимого уровня качества	148
5.1.3. Выбор метрики качества	149
5.1.4. Выбирайте правильный ориентир	149
5.1.5. Разбиение данных на три набора	151
5.1.6. Предварительные условия для обучения с учителем	152
5.2. Представление меток для машинного обучения	153
5.2.1. Многоклассовая классификация	153
5.2.2. Многозначная классификация	154
5.3. Выбор алгоритма обучения	154
5.3.1. Основные свойства алгоритма обучения	155
5.3.2. Выборочная проверка алгоритмов	156
5.4. Построение конвейера	158
5.5. Оценивание качества модели	159
5.5.1. Метрики качества для регрессии	159
5.5.2. Метрики качества для классификации	160
5.5.3. Метрики качества для ранжирования	165
5.6. Настройка гиперпараметров	168
5.6.1. Поиск на сетке	169
5.6.2. Случайный поиск	170
5.6.3. Поиск с измельчением	170
5.6.4. Другие методы	172
5.6.5. Перекрестная проверка	172
5.7. Обучение поверхностной модели	173
5.7.1. Стратегия обучения поверхностной модели	173
5.7.2. Сохранение и восстановление модели	174
5.8. Компромисс между смещением и дисперсией	175
5.8.1. Недообучение	175
5.8.2. Переобучение	176
5.8.3. Компромисс	177
5.9. Регуляризация	179
5.9.1. L1- и L2-регуляризации	179
5.9.2. Другие формы регуляризации	180
5.10. Резюме	180
Глава 6. Обучение модели с учителем (часть 2)	183
6.1. Стратегия обучения глубоких моделей	183
6.1.1. Стратегия обучения нейронной сети	184
6.1.2. Метрика качества и функция стоимости	184
6.1.3. Стратегии инициализации параметров	187
6.1.4. Алгоритмы оптимизации	187
6.1.5. Планы уменьшения скорости обучения	191
6.1.6. Регуляризация	192
6.1.7. Определение размера сети и настройка гиперпараметров	193

6.1.8. Работа с несколькими входами	195
6.1.9. Работа с несколькими выходами.....	196
6.1.10. Перенос обучения	197
6.2. Штабелирование моделей.....	199
6.2.1. Типы ансамблевого обучения	199
6.2.2. Алгоритм штабелирования моделей	200
6.2.3. Просачивание данных при штабелировании моделей.....	201
6.3. Борьба со сдвигом распределения.....	202
6.3.1. Обработка сдвига распределения	202
6.3.2. Состязательная проверка	202
6.4. Обработка несбалансированных наборов данных	203
6.4.1. Взвешивание классов	203
6.4.2. Ансамбль перераспределенных наборов данных.....	204
6.4.3. Другие методы	205
6.5. Калибровка модели.....	205
6.5.1. Хорошо откалиброванные модели.....	205
6.5.2. Методы калибровки	207
6.6. Поиск неполадок и анализ ошибок	208
6.6.1. Причины плохого поведения модели.....	208
6.6.2. Итеративное уточнение модели.....	209
6.6.3. Анализ ошибок.....	209
6.6.4. Анализ ошибок в комплексных системах	211
6.6.5. Использование расслоенных метрик.....	212
6.6.6. Исправление неправильных меток.....	212
6.6.7. Нахождение дополнительных примеров для пометки	213
6.6.8. Поиск неполадок при глубоком обучении	213
6.7. Рекомендации	215
6.7.1. Поставляйте хорошую модель.....	215
6.7.2. Доверяйте популярным реализациям с открытым исходным кодом	215
6.7.3. Оптимизируйте важную для бизнеса меру качества.....	216
6.7.4. При обновлении начинайте с нуля.....	216
6.7.5. Избегайте каскадов коррекций.....	217
6.7.6. Используйте каскадирование моделей с осторожностью	217
6.7.7. Пишите эффективный код, компилируйте и распараллеливайте	218
6.7.8. Тестируйте на старых и новых данных.....	219
6.7.9. Больше данных лучше, чем более умный алгоритм	220
6.7.10. Новые данные лучше более изощренных признаков.....	220
6.7.11. Радуйтесь крохотным достижениям	220
6.7.12. Обеспечьте воспроизводимость	220
6.8. Резюме	221
Глава 7. Оценивание модели	224
7.1. Офлайнное и онлайнное оценивания	225
7.2. A/B-тестирование	227
7.2.1. G-критерий	228
7.2.2. Z-критерий.....	231

7.2.3. Заключительные замечания и предупреждения.....	233
7.3. Многорукий бандит	233
7.4. Статистические границы качества модели	236
7.4.1. Статистический интервал для ошибки классификации	236
7.4.2. Бутстреп статистического интервала	237
7.4.3. Бутстреп интервала предсказания для регрессии	238
7.5. Оценивание адекватности тестового набора.....	239
7.5.1. Нейронное покрытие.....	239
7.5.2. Мутационное тестирование	240
7.6. Оценивание свойств модели	240
7.6.1. Робастность	241
7.6.2. Справедливость	241
7.7. Резюме	243

Глава 8. Развертывание модели 244

8.1. Статическое развертывание	245
8.2. Динамическое развертывание на устройстве пользователя.....	245
8.2.1. Развертывание параметров модели	246
8.2.2. Развертывание сериализованного объекта	246
8.2.3. Развертывание в браузере.....	246
8.2.4. Плюсы и минусы	246
8.3. Динамическое развертывание на сервере	247
8.3.1. Развертывание на виртуальной машине	247
8.3.2. Развертывание в контейнере.....	248
8.3.3. Бессерверное развертывание.....	250
8.3.4. Потокное развертывание модели.....	251
8.4. Стратегии развертывания	253
8.4.1. Разовое развертывание	253
8.4.2. Немое развертывание	254
8.4.3. Канареечное развертывание.....	254
8.4.4. Многорукие бандиты	255
8.5. Автоматизированное развертывание, версионирование и метаданные	255
8.5.1. Объекты, сопровождающие модель	255
8.5.2. Синхронизация версий.....	256
8.5.3. Метаданные версии модели.....	256
8.6. Рекомендации по развертыванию модели	257
8.6.1. Эффективность алгоритма	257
8.6.2. Развертывание глубоких моделей.....	260
8.6.3. Кеширование	260
8.6.4. Формат доставки модели и кода.....	261
8.6.5. Начинайте с простой модели.....	263
8.6.6. Тестируйте на посторонних	263
8.7. Резюме.....	264

Глава 9. Выполнение, мониторинг и сопровождение модели 266

9.1. Свойства среды выполнения модели.....	266
9.1.1. Безопасность и корректность	267

9.1.2. Простота развертывания	267
9.1.3. Гарантии правильности модели	268
9.1.4. Простота восстановления	268
9.1.5. Предотвращение расхождений между обучением и выполнением	268
9.1.6. Предотвращение скрытых петель обратной связи	269
9.2. Режимы выполнения модели	269
9.2.1. Выполнение в пакетном режиме.....	270
9.2.2. Обслуживание запроса со стороны человека	270
9.2.3. Обслуживание запроса со стороны машины.....	272
9.3. Выполнение модели на практике	273
9.3.1. Готовность к ошибкам.....	273
9.3.2. Отношение к ошибкам.....	274
9.3.3. Готовность к изменениям и отношение к ним	275
9.3.4. Готовность к особенностям человеческой природы и отношение к ним	277
Избегайте путаницы	277
Умерьте ожидания.....	277
Завоевывайте доверие	277
Не переутомляйте пользователя.....	278
Остерегайтесь фактора отторжения	278
9.4. Мониторинг модели	278
9.4.1. Что может пойти не так?.....	279
9.4.2. Что и как мониторить	280
9.4.3. Что протоколировать	282
9.4.4. Мониторинг неправомерного использования	283
9.5. Сопровождение модели	283
9.5.1. Когда обновлять	284
9.5.2. Как обновлять.....	285
9.6. Резюме	288
Глава 10. Заключение.....	290
10.1. Сухой остаток.....	290
10.2. Что еще почитать	294
10.3. Благодарности	295
Предметный указатель.....	297