

УДК 004.4
ББК 32.972
X12

Хапке Х., Нельсон К.

X12 Разработка конвейеров машинного обучения. Автоматизация жизненных циклов модели с помощью TensorFlow / пер. с англ. Н. Б. Желновой. – М.: ДМК Пресс, 2021. – 346 с.: ил.

ISBN 978-5-97060-886-9

Машинное обучение становится важным элементом почти во всех отраслях. В этой книге представлено четкое и понятное руководство по автоматизации развертывания, управления и повторного использования моделей машинного обучения. Шаг за шагом описывается конкретный пример проекта, на котором можно отработать основные навыки в этой сфере. Благодаря множеству примеров кода и ясным, лаконичным объяснениям вы сможете создать свой собственный конвейер машинного обучения и запустите его в кратчайшие сроки.

Книга поможет ученым и инженерам, специализирующимся в области машинного обучения и искусственного интеллекта, выйти за рамки работы с единичной моделью и успешно реализовать свои проекты в области науки о данных. Также издание будет полезно менеджерам проектов в области науки о данных, разработчикам программного обеспечения и инженерам DevOps, которые хотят, чтобы их организация ускорила свои проекты, использующие технологии машинного обучения и искусственного интеллекта.

Читателю понадобится знание основных концепций машинного обучения и хотя бы одного из фреймворков, используемых в машинном обучении (например, PyTorch, TensorFlow, Keras).

УДК 004.4
ББК 32.972

DMK Press Authorized Russian translation of the English edition of Building Machine Learning Pipelines

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 (англ.) 978-1492053194
ISBN (рус.) 978-5-97060-886-9

© 2020 Hannes Hapke and Catherine Nelson
© Оформление, издание, перевод, ДМК Пресс, 2021

Оглавление

Предисловие от издательства	13
Предисловие	14
Введение	17
Для кого предназначена эта книга.....	18
Почему мы используем TensorFlow и TensorFlow Extended.....	19
Обзор глав	19
Условные обозначения, используемые в этой книге	21
Использование примеров кода.....	22
Онлайн-обучение O'Reilly.....	22
Как с нами связаться	23
Благодарности	23
Глава 1. Введение	26
Почему и где используются конвейеры машинного обучения	26
Когда следует подумать о конвейерах машинного обучения?.....	28
Обзор этапов конвейера машинного обучения	28
Этап загрузки данных и управление версиями данных.....	29
Проверка данных.....	29
Предварительная обработка данных	30
Обучение и настройка модели	31
Анализ модели.....	31
Управление версиями модели.....	32
Развертывание модели	32
Петли обратной связи	32
Приватность данных	33
Оркестровка конвейера.....	33
Для чего нужна оркестровка конвейера	33
Направленные ациклические графы	34
Наш демонстрационный проект машинного обучения	35
Структура проекта	36
Наша модель машинного обучения	36
Цель демонстрационного проекта	37
Резюме.....	37

Глава 2. Введение в TensorFlow Extended	38
Что такое TFX?	39
Установка TFX	40
Обзор компонентов TFX	41
Что такое метаданные ML Metadata?	42
Альтернативы TFX	45
Знакомство с Apache Beam	46
Установка	46
Базовый конвейер	47
Запуск элементарного конвейера	50
Резюме	50
Глава 3. Загрузка данных	51
Концепции загрузки данных	51
Загрузка локальных файлов данных	53
Загрузка удаленных файлов данных	57
Загрузка данных напрямую из баз данных	58
Подготовка данных	60
Разбиение наборов данных	60
Связующие наборы данных	62
Управление версиями наборов данных	63
Стратегии загрузки данных	64
Структурированные данные	64
Текстовые данные для задач обработки естественного языка	64
Графические данные для задач компьютерного зрения	64
Резюме	65
Глава 4. Проверка данных	66
Для чего нужна проверка данных?	67
TFDV	68
Установка	69
Генерация статистических показателей для набора данных	69
Генерация схемы на основе данных	71
Распознавание ошибок в данных	72
Сравнение наборов данных	73
Обновление схемы	75
Отклонения и дрейф данных	76
Наборы данных с систематической ошибкой выборки	77
Получение среза данных в TFDV	78
Обработка больших наборов данных с помощью Google Cloud Platform	80
Интеграция TFDV в конвейер машинного обучения	83
Резюме	84

Глава 5. Предварительная обработка данных	85
Для чего нужна предварительная обработка данных	86
Предварительная обработка данных в контексте всего набора данных	86
Масштабирование шагов предварительной обработки	86
Как избежать отклонения при обучении и работе модели	86
Развертывание шагов предварительной обработки и модели машинного обучения как единого артефакта	88
Проверка результатов предварительной обработки в конвейере	88
Предварительная обработка данных с помощью TFT	89
Установка	90
Стратегии предварительной обработки	90
Лучшие практики	92
Функции TFT	93
Автономная работа TFT	95
Интеграция TFT в конвейер машинного обучения	97
Резюме	101
Глава 6. Обучение модели	102
Определение модели для нашего демонстрационного проекта	103
Компонент TFX Trainer	106
Функция <code>run_fn()</code>	106
Запуск компонента <code>Trainer</code>	110
Другие соображения относительно компонента <code>Trainer</code>	112
Использование TensorBoard в интерактивном конвейере	113
Стратегии распределения	115
Настройка модели	118
Стратегии настройки гиперпараметров	118
Настройка гиперпараметров в конвейерах TFX	119
Резюме	119
Глава 7. Анализ и проверка модели	120
Как проанализировать модель	121
Метрики классификации	121
Метрики регрессии	124
Анализ модели TensorFlow	125
Анализ одной модели в TFMA	126
Анализ нескольких моделей в TFMA	129
Анализ достоверности модели	130
Формирование срезов для прогнозов модели в TFMA	132
Проверка пороговых значений решений с использованием метрик справедливости	134
Проведение более детального анализа с помощью инструмента анализа альтернатив (What-If Tool)	136

Объяснение модели	140
Генерация объяснений с помощью WIT	142
Другие методы объяснения	143
Анализ и проверка модели в TFX	145
ResolverNode	145
Компонент Evaluator	146
Проверка при помощи компонента Evaluator	147
Компонент TFX Pusher	148
Резюме	148

Глава 8. Развертывание модели с помощью

TensorFlow Serving	149
Простой сервер моделей	150
Недостатки развертывания моделей с помощью API на основе Python	151
Отсутствие разделения кода	151
Отсутствие контроля версий модели	152
Неэффективный вывод модели	152
TensorFlow Serving	152
Обзор архитектуры TensorFlow	153
Экспорт моделей для TensorFlow Serving	153
Сигнатуры моделей	155
Методы сигнатуры	155
Проверка экспортированных моделей	157
Проверка модели	158
Тестирование модели	159
Установка TensorFlow Serving	160
Установка Docker	160
Установка на Ubuntu	160
Сборка TensorFlow Serving из исходного кода	161
Настройка сервера TensorFlow	161
Конфигурация при работе с одной моделью	162
Конфигурация при работе с несколькими моделями	164
REST или gRPC	166
REST	166
gRPC	166
Выполнение прогнозов на сервере моделей	167
Получение прогнозов модели с использованием REST	167
Работа с TensorFlow Serving через gRPC	169
А/В-тестирование модели с использованием TensorFlow Serving	172
Запрос метаданных модели с сервера моделей	173
REST-запросы метаданных модели	173
Запросы gRPC для метаданных модели	174

Пакетные запросы на вывод прогнозов модели	175
Настройка использования пакетного режима в прогнозировании	177
Другие функции оптимизации TensorFlow Serving	179
Альтернативы TensorFlow Serving	180
BentoML	180
Seldon	180
GraphPipe	181
Simple TensorFlow Serving	181
MLflow	181
Ray Serve	181
Развертывание моделей с использованием услуг поставщиков облачных решений	182
Сценарии использования	182
Пример развертывания с помощью облачных платформ Google	182
Развертывание модели с помощью конвейеров TFX	188
Резюме	189

Глава 9. Расширенные концепции развертывания моделей с помощью TensorFlow Serving

Разделение зон ответственности в процессе развертывания	190
Обзор рабочего процесса	191
Оптимизация загрузки удаленной модели	193
Оптимизация модели для развертываний	194
Квантование	194
Сокращение	195
Дистилляция	196
Использование TensorRT совместно с TensorFlow Serving	196
TFLite	197
Шаги по оптимизации моделей машинного обучения с помощью TFLite	197
Развертывание моделей TFLite с помощью TensorFlow Serving	199
Мониторинг экземпляров TensorFlow Serving	200
Установка Prometheus	200
Конфигурация TensorFlow Serving	202
Простое масштабирование с помощью TensorFlow Serving и Kubernetes	204
Дополнительная литература о Kubernetes и Kubeflow	205
Резюме	206

Глава 10. Расширенные концепции TensorFlow Extended

Расширенные концепции конвейеров машинного обучения	207
Одновременное обучение нескольких моделей	208

Экспорт моделей TFLite	209
Ограничения TFLite.....	210
Обучение модели с «теплым» запуском	212
Участие человека в конвейере машинного обучения	212
Настройка компонента Slack	214
Как использовать компонент Slack	214
Пользовательские компоненты TFX	215
Сценарии использования пользовательских компонентов	216
Создание пользовательского компонента с нуля.....	216
Повторное использование существующих компонентов	225
Резюме.....	228

Глава 11. Конвейеры, часть 1: Apache Beam и Apache Airflow.... 230

Какой инструмент оркестрации выбрать?.....	231
Apache Beam.....	231
Apache Airflow	231
Kubeflow Pipelines	231
Kubeflow Pipelines на платформе AI.....	232
Преобразование вашего интерактивного конвейера TFX в производственный конвейер.....	232
Преобразование элементарного интерактивного конвейера для Beam и Airflow	234
Введение в Apache Beam	235
Оркестрация конвейеров TFX с помощью Apache Beam	235
Введение в Apache Airflow.....	237
Установка и начальная настройка.....	237
Элементарный пример использования Airflow.....	239
Оркестрация конвейеров TFX с помощью Apache Airflow.....	242
Настройка конвейера	242
Запуск конвейера.....	244
Резюме.....	245

Глава 12. Конвейеры, часть 2:

Kubeflow Pipelines 246

Введение в Kubeflow Pipelines.....	247
Установка и начальная настройка.....	249
Доступ к установленному экземпляру Kubeflow Pipelines.....	251
Оркестрация конвейеров TFX с помощью Kubeflow Pipelines.....	252
Настройка конвейера	254
Запуск конвейера.....	258
Полезные функции Kubeflow Pipelines.....	264

Конвейеры, работающие на Google Cloud AI Platform	269
Настройка конвейера	269
Настройка конвейера TFX.....	273
Запуск и работа конвейера	276
Резюме.....	277
Глава 13. Петли обратной связи	279
Явная и неявная обратная связь.....	280
Маховик данных	281
Петли обратной связи в реальном мире	282
Конструктивные шаблоны для сбора отзывов	284
Пользователи предпринимают определенные действия в результате прогноза	284
Пользователи оценивают качество прогноза.....	285
Пользователи исправляют прогноз.....	285
Краудсорсинг аннотаций	286
Экспертные аннотации	287
Обратная связь автоматически предоставляется системой	287
Как отслеживать петли обратной связи	287
Отслеживание явной обратной связи	288
Отслеживание неявной обратной связи	289
Резюме.....	289
Глава 14. Приватность данных, используемых для машинного обучения	290
Введение в приватность данных	290
Почему мы заботимся о приватности данных?	291
Самый простой способ повысить приватность данных	291
Какие данные должны быть приватными?	292
Дифференцированная приватность.....	293
Локальная и глобальная дифференцированная приватность.....	294
Эпсилон-дельта и бюджет приватности	295
Дифференцированная приватность в машинном обучении	296
Введение в TensorFlow Privacy	296
Обучение с оптимизатором, использующим подход дифференцированной приватности	297
Расчет параметра ϵ	298
Введение в федеративное обучение.....	299
Федеративное обучение в TensorFlow.....	301
Зашифрованное машинное обучение.....	302
Зашифрованное обучение модели	303
Преобразование обученной модели для обслуживания зашифрованных прогнозов	304

Другие методы обеспечения приватности данных.....	305
Резюме.....	305
Глава 15. Будущее конвейеров машинного обучения и следующие шаги.....	307
Отслеживание экспериментов с моделью	307
Предложения в области управления релизами модели.....	308
Будущие возможности конвейеров.....	309
Использование TFX с другими фреймворками машинного обучения.....	310
Тестирование моделей машинного обучения	310
Системы непрерывной интеграции и развертывания для машинного обучения.....	311
Сообщество инженеров машинного обучения.....	311
Резюме.....	311
Приложение А. Введение в инфраструктуру машинного обучения.....	313
Приложение В. Настройка кластера Kubernetes в Google Cloud	326
Приложение С. Советы по работе с Kubeflow Pipelines	332
Предметный указатель	340