

УДК 004.4
ББК 32.972
Х98

Хуттер Ф., Коттхофф Л., Ваншорен Х.

X98 Введение в автоматизированное машинное обучение (AutoML) / пер. с англ. В. С. Яценкова. – М.: ДМК Пресс, 2023. – 256 с.: ил.

ISBN 978-5-93700-196-2

Ошеломляющий успех коммерческих приложений машинного обучения (machine learning – ML) и быстрый рост этой отрасли создали высокий спрос на готовые методы ML, которые можно легко использовать без специальных знаний. Однако и сегодня успех практического применения в решающей степени зависит от экспертов – людей, которые вручную выбирают подходящие архитектуры и их гиперпараметры. Методы AutoML нацелены на устранение этого узкого места путем построения систем ML, способных к автоматической оптимизации и самонастройке независимо от типа входных данных. В этой книге впервые представлен всеобъемлющий обзор базовых методов автоматизированного машинного обучения (AutoML).

Издание послужит отправной точкой для изучения этой быстро развивающейся области; тем, кто уже использует AutoML в своей работе, книга пригодится в качестве справочника.

УДК 004.4
ББК 32.972



This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

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 978-1-80181-497-3 (англ.)

ISBN 978-5-93700-196-2 (рус.)

© Hutter F., Kotthoff L., Vanschoren J., 2019.
This book is an open access publication
© Перевод, оформление, издание,
ДМК Пресс, 2023

Содержание

| | |
|--|----|
| От издательства | 10 |
| Предисловие | 11 |
| Введение | 13 |
| ЧАСТЬ I. МЕТОДЫ AutoML | 17 |
| Глава 1. Оптимизация гиперпараметров | 18 |
| 1.1. Введение | 18 |
| 1.2. Постановка задачи | 20 |
| 1.2.1. Альтернативы оптимизации: ансамблирование и маргинализация | 21 |
| 1.2.2. Оптимизация по нескольким целям | 22 |
| 1.3. Оптимизация гиперпараметров методом черного ящика | 22 |
| 1.3.1. Оптимизация методом черного ящика без моделей | 22 |
| 1.3.2. Байесовская оптимизация..... | 24 |
| 1.3.2.1. Краткое введение в байесовскую оптимизацию | 25 |
| 1.3.2.2. Суррогатные модели | 26 |
| 1.3.2.3. Описание пространства конфигурации..... | 28 |
| 1.3.2.4. Ограниченная байесовская оптимизация | 29 |
| 1.4. Методы оптимизации с переменной точностью | 30 |
| 1.4.1. Прогнозирование на основе кривой обучения для ранней остановки | 31 |
| 1.4.2. Методы выбора алгоритма на основе приближений | 32 |
| 1.4.3. Адаптивный выбор точности..... | 35 |
| 1.5. Применение оптимизации гиперпараметров в AutoML | 36 |
| 1.6. Проблемы и перспективные направления исследований | 38 |
| 1.6.1. Бенчмарки и сопоставимость результатов | 38 |
| 1.6.2. Оптимизация на основе градиента | 40 |
| 1.6.3. Масштабируемость | 40 |
| 1.6.4. Переобучение и обобщение | 41 |
| 1.6.5. Построение конвейера произвольного размера..... | 42 |
| 1.7. Литература..... | 43 |

| | |
|---|-----|
| Глава 2. Метаобучение | 54 |
| 2.1. Введение | 54 |
| 2.2. Обучение на основе оценок моделей | 55 |
| 2.2.1. Независимые от задачи рекомендации | 56 |
| 2.2.2. Построение пространства конфигураций | 57 |
| 2.2.3. Перенос конфигурации | 58 |
| 2.2.3.1. Относительные ориентиры | 58 |
| 2.2.3.2. Суррогатные модели | 58 |
| 2.2.3.3. Многозадачное обучение с теплым стартом | 59 |
| 2.2.3.4. Другие методы | 60 |
| 2.2.4. Кривые обучения | 60 |
| 2.3. Обучение на основе свойств задачи | 61 |
| 2.3.1. Метапризнаки | 61 |
| 2.3.2. Обучение метапризнаков | 64 |
| 2.3.3. Оптимизация с теплым стартом на основе схожих задач | 64 |
| 2.3.4. Метамодели | 66 |
| 2.3.4.1. Ранжирование | 66 |
| 2.3.4.2. Прогнозирование производительности | 67 |
| 2.3.5. Синтез конвейера | 68 |
| 2.3.6. Настраивать или не настраивать? | 69 |
| 2.4. Обучение на основе предыдущих моделей | 69 |
| 2.4.1. Трансферное обучение | 69 |
| 2.4.2. Метаобучение в нейронных сетях | 70 |
| 2.4.3. Обучение на ограниченных данных | 71 |
| 2.4.4. За рамками обучения с учителем | 73 |
| 2.5. Заключение | 74 |
| 2.6. Литература | 75 |
| Глава 3. Поиск нейронной архитектуры | 85 |
| 3.1. Введение | 85 |
| 3.2. Пространство поиска | 87 |
| 3.3. Стратегия поиска | 90 |
| 3.4. Стратегия оценки производительности | 93 |
| 3.5. Перспективные направления | 96 |
| 3.6. Литература | 98 |
| ЧАСТЬ II. СИСТЕМЫ AutoML | 103 |
| Глава 4. Auto-WEKA: автоматический выбор модели и оптимизация гиперпараметров в WEKA | 104 |
| 4.1. Введение | 105 |
| 4.2. Предварительные условия | 106 |
| 4.2.1. Выбор модели | 106 |

| | |
|---|-----|
| 4.2.2. Оптимизация гиперпараметров | 107 |
| 4.3. Одновременный выбор алгоритмов и оптимизация гиперпараметров (CASH) | 108 |
| 4.3.1. Последовательный алгоритм конфигурации по модели (SMAC) | 109 |
| 4.4. Auto-WEKA | 110 |
| 4.5. Экспериментальная оценка | 112 |
| 4.5.1. Эталонные методы | 113 |
| 4.5.2. Результаты производительности, определенные перекрестной проверкой | 115 |
| 4.5.3. Результаты тестирования производительности | 115 |
| 4.6. Заключение | 117 |
| 4.6.1. Популярность Auto-WEKA в сообществе | 117 |
| 4.7. Литература..... | 118 |

Глава 5. Проект Hyperopt-sklearn

| | |
|--|-----|
| 5.1. Введение | 120 |
| 5.2. Оптимизация с помощью Hyperopt | 121 |
| 5.3. Выбор модели в scikit-learn как задача поиска | 123 |
| 5.4. Пример использования | 124 |
| 5.5. Эксперименты | 128 |
| 5.6. Текущее состояние и перспективные направления исследований..... | 130 |
| 5.7. Заключение..... | 133 |
| 5.8. Литература | 134 |

Глава 6. Auto-sklearn – эффективное и надежное автоматизированное машинное обучение

| | |
|---|-----|
| 6.1. Введение | 137 |
| 6.2. AutoML как задача CASH..... | 138 |
| 6.3. Новые методы повышения эффективности и надежности AutoML..... | 139 |
| 6.3.1. Поиск перспективных вариантов при помощи метаобучения | 140 |
| 6.3.2. Автоматизированное построение ансамбля моделей, оцененных во время оптимизации | 141 |
| 6.4. Практическая система автоматизированного машинного обучения..... | 142 |
| 6.5. Сравнение Auto-sklearn с Auto-WEKA и Hyperopt-sklearn | 146 |
| 6.6. Оценка предложенных улучшений AutoML..... | 148 |
| 6.7. Детальный анализ компонентов Auto-sklearn..... | 150 |
| 6.8. Обсуждение результатов и заключение | 151 |
| 6.8.1. Обсуждение результатов | 151 |
| 6.8.2. Практическое применение..... | 155 |
| 6.8.3. Расширения в PoSH Auto-sklearn..... | 155 |
| 6.8.4. Заключение и будущие исследования | 156 |
| 6.9. Литература | 157 |

| | |
|--|-----|
| Глава 7. На пути к автоматически настраиваемым глубоким нейронным сетям | 160 |
| 7.1. Введение | 160 |
| 7.2. Auto-Net 1.0 | 162 |
| 7.3. Auto-Net 2.0 | 164 |
| 7.4. Эксперименты | 170 |
| 7.4.1. Первичная оценка Auto-Net 1.0 и Auto-sklearn | 170 |
| 7.4.2. Результаты для наборов данных конкурса AutoML | 171 |
| 7.4.3. Сравнение AutoNet 1.0 и 2.0 | 173 |
| 7.5. Заключение | 174 |
| 7.6. Литература | 174 |

| | |
|---|-----|
| Глава 8. TPOT: инструмент оптимизации конвейеров на основе деревьев для автоматизации машинного обучения | 179 |
| 8.1. Введение | 180 |
| 8.2. Базовые принципы TPOT | 180 |
| 8.2.1. Конвейерные операторы машинного обучения | 181 |
| 8.2.2. Построение конвейеров на основе деревьев | 182 |
| 8.2.3. Оптимизация конвейеров на основе деревьев | 182 |
| 8.2.4. Эталонные данные | 183 |
| 8.3. Результаты | 183 |
| 8.4. Выводы и перспективные направления исследований | 187 |
| 8.5. Литература | 188 |

| | |
|--|-----|
| Глава 9. Проект Automatic Statistician | 190 |
| 9.1. Введение | 190 |
| 9.2. Базовые принципы Automatic Statistician | 192 |
| 9.2.1. Похожие исследования | 193 |
| 9.3. Automatic Statistician и данные временных рядов | 193 |
| 9.3.1. Грамматика операций над ядрами | 194 |
| 9.3.2. Процедура поиска и оценки | 195 |
| 9.3.3. Генерация описаний на естественном языке | 196 |
| 9.3.4. Сравнение с людьми | 198 |
| 9.4. Другие системы автоматической статистики | 198 |
| 9.4.1. Основные компоненты | 199 |
| 9.4.2. Проблемы и задачи | 200 |
| 9.4.2.1. Взаимодействие с пользователем | 200 |
| 9.4.2.2. Отсутствующие и беспорядочные данные | 200 |
| 9.4.2.3. Распределение ресурсов | 200 |
| 9.5. Заключение | 201 |
| 9.6. Литература | 201 |

| | |
|---|-----|
| ЧАСТЬ III. ПРОБЛЕМЫ AutoML | 205 |
| Глава 10. О чем говорят результаты конкурсов AutoML Challenge? | 206 |
| 10.1. Введение..... | 207 |
| 10.2. Формализация задачи и обзор условий..... | 210 |
| 10.2.1. Предметная область задачи..... | 210 |
| 10.2.2. Выбор полной модели..... | 211 |
| 10.2.3. Оптимизация гиперпараметров..... | 213 |
| 10.2.4. Стратегии поиска моделей..... | 214 |
| 10.3. Данные..... | 218 |
| 10.4. Протокол конкурса..... | 221 |
| 10.4.1. Бюджет времени и вычислительные ресурсы..... | 222 |
| 10.4.2. Метрики подсчета баллов..... | 222 |
| 10.4.3. Раунды и этапы в конкурсе 2015/2016..... | 225 |
| 10.4.4. Этапы конкурса 2018 года..... | 226 |
| 10.5. Результаты..... | 227 |
| 10.5.1. Оценки, полученные в конкурсе 2015/2016..... | 227 |
| 10.5.2. Результаты, полученные в конкурсе 2018 года..... | 230 |
| 10.5.3. Сложность наборов данных/задач..... | 230 |
| 10.5.4. Оптимизация гиперпараметров..... | 236 |
| 10.5.5. Метаобучение..... | 238 |
| 10.5.6. Методы, использованные в конкурсах..... | 239 |
| 10.6. Обсуждение..... | 245 |
| 10.7. Заключение..... | 246 |
| 10.8. Литература..... | 249 |
| Предметный указатель | 254 |