

*Моей спутнице жизни Фатиме,
а также Оливеру и Якубу.*

Павел

*Нико и Леунтье ван Рейн за то,
что научили меня тому, что важно в жизни.*

Ян

*Моим родителям, а также Мануэле,
Кике, Манель и Артуру.*

Карлос

*Аде, Элиасу, Кобе и Вирле за то,
что напомнили мне, как прекрасен мир.*

Хоакин

Содержание

От издательства	21
Предисловие	22
Часть I. ОСНОВНЫЕ КОНЦЕПЦИИ И АРХИТЕКТУРА	26
Глава 1. Введение	27
1.1. Структура книги.....	27
1.2. Основные концепции и архитектура (часть I).....	28
1.2.1. Основные понятия.....	28
Роль машинного обучения.....	28
Роль метаобучения.....	29
Определение метаобучения.....	29
Метаобучение или автоматизированное машинное обучение?.....	30
Происхождение термина «метаобучение».....	30
1.2.2. Основные типы задач.....	31
1.2.3. Базовая архитектура систем метаобучения и AutoML.....	32
1.2.4. Выбор алгоритма с использованием метаданных из предыдущих задач (главы 2,5).....	34
1.2.5. Оценка и сравнение различных систем (глава 3).....	34
1.2.6. Роль характеристик/метапризнаков набора данных (глава 4).....	35
1.2.7. Различные типы моделей метауровня (глава 5).....	36
1.2.8. Оптимизация гиперпараметров (глава 6).....	37
1.2.9. Автоматические методы формирования конвейера (глава 7).....	37
1.3. Передовые технологии и методы (часть II).....	38
1.3.1. Настройка пространств конфигураций и экспериментов (глава 8).....	38
1.3.2. Автоматические методы для ансамблей и потоков.....	39
Объединение базовых учеников в ансамбли (глава 9).....	39
Метаобучение ансамблевыми методами (глава 10).....	39
Рекомендации по выбору алгоритма для потоковых данных (глава 11) ...	39
1.3.3. Перенос метамodelей между задачами (глава 12).....	40
1.3.4. Метаобучение глубоких нейронных сетей (глава 13).....	41

1.3.5. Автоматизация обработки данных и проектирование сложных систем	42
Автоматизация науки о данных (глава 14).....	42
Автоматизация проектирования сложных систем (глава 15).....	43
1.4. Хранилища результатов экспериментов (часть III).....	44
1.4.1. Хранилища метаданных (глава 16).....	44
1.4.2. Обучение на метаданных в репозиториях (глава 17).....	45
1.4.3. Заключительные замечания (глава 18).....	45
1.5. Литература	46

Глава 2. Применение метаобучения к выбору алгоритма (рейтинг)

2.1. Введение	47
2.1.2. Структура этой главы	48
2.2. Различные типы рекомендаций	48
2.2.1. Лучший алгоритм в наборе	50
2.2.2. Подмножество лучших алгоритмов	50
Определение алгоритмов с сопоставимой производительностью.....	50
Объединение подмножеств	51
2.2.3. Линейное ранжирование.....	52
2.2.4. Квазилинейное (слабое) ранжирование.....	52
2.2.5. Неполный рейтинг.....	52
2.2.6. Поиск лучшего алгоритма в рамках заданного бюджета	53
2.3. Ранжирование моделей для выбора алгоритма	53
2.3.1. Создание метамодели в виде ранжированного списка	54
Получение оценок производительности	54
Объединение результатов производительности в единый рейтинг	56
Пример: нахождение среднего рейтинга	57
2.3.2. Использование метамодели ранжирования для прогнозов (стратегия top-n).....	57
Пример	59
2.3.3. Оценка рекомендуемых рейтингов	60
2.4. Использование комбинированного показателя точности и времени выполнения.....	60
2.5. Расширения и другие подходы	62
2.5.1. Использование метода ранжирования по среднему для рекомендации конвейеров	62
2.5.2. Ранжирование может понизить рейтинг алгоритмов	63
2.5.3. Подходы, основанные на многокритериальном анализе с DEA	64
2.5.4. Использование схожести наборов данных для определения соответствующих частей метаданных	64
2.5.5. Работа с неполным ранжированием.....	65
Агрегирование неполных рейтингов	65
2.6. Литература	66

Глава 3. Оценка рекомендаций систем метаобучения и AutoML	69
3.1. Введение	69
3.2. Методика оценки алгоритмов базового уровня	70
3.2.1. Ошибка обобщения	70
3.2.2. Стратегии оценки	71
3.2.3. Потеря и функция потерь	72
3.3. Нормализация производительности для алгоритмов базового уровня	72
Подстановка значений производительности по рангам	73
Масштабирование к интервалу 0–1	73
Преобразование значений в нормальное распределение	73
Преобразование в квантильные значения	74
Нормализация с учетом погрешности	74
3.4. Методика оценки метаобучения и систем AutoML	74
3.4.1. Однопроходная оценка с откладыванием	74
Цель систем метаобучения/AutoML	75
Выполнение внутренней оценки системами метаобучения/AutoML	75
Избегайте предвзятой оценки	76
3.4.2. Оценка на метауровне с перекрестной проверкой	76
Оценка на метауровне с поиском в таблице	77
3.5. Оценка рекомендаций путем измерения корреляции	77
Ранговая корреляция Спирмена	78
Взвешенная мера ранговой корреляции	79
3.6. Оценка влияния рекомендаций	79
3.6.1. Потери производительности и кривые потерь	80
3.6.2. Характеризация кривых потерь по AUC	81
3.6.3. Агрегирование кривых потерь после нескольких проходов CV	81
3.6.4. Статистические тесты при заданном бюджете времени	82
3.7. Некоторые полезные меры	83
3.7.1. Низкая точность	83
3.7.2. Нормализованный дисконтированный совокупный прирост	83
3.8. Литература	84
Глава 4. Характеристики набора данных (метапризнаки)	86
4.1. Введение	86
4.1.1. Что такое хорошие признаки набора данных?	87
4.1.2. Характеристики, зависящие от задач и данных	87
4.1.3. Характеристики алгоритмов	88
4.1.4. Разработка метапризнаков	88
4.2. Характеризация данных в задачах классификации	88
4.2.1. Простые, статистические и теоретико-информационные метапризнаки	89
Простые метапризнаки	89
Статистические метапризнаки	89
Теоретико-информационные метапризнаки	90

4.2.2. Метапризнаки на основе модели	91
4.2.3. Метапризнаки на основе производительности	91
Ориентиры	91
Относительные ориентиры	92
Ориентиры подвыборки и частичные кривые обучения.....	92
Вектор ориентиров производительности.....	92
4.2.4. Метапризнаки, основанные на концепции и сложности	93
Вариативность/неровность выходного пространства	93
Перекрытие отдельных признаков.....	94
Разделимость классов	94
Связь некоторых мер сложности с другими типами.....	94
4.3. Характеризация данных, используемая в задачах регрессии.....	95
4.3.1. Простые и статистические метапризнаки	95
Метапризнаки на основе корреляции.....	96
4.3.2. Меры на основе сложности задачи	96
4.3.3. Меры на основе сложности/модели	96
4.3.4. Меры гладкости.....	97
4.3.5. Меры нелинейности	97
4.4. Характеризация данных, используемых в задачах временных рядов	98
4.4.1. Общая статистика (описательная статистика).....	98
4.4.2. Характеристики в частотной области.....	98
4.4.3. Характеристики на основе автокорреляции	99
4.5. Характеризация данных, используемых в задачах кластеризации	99
4.5.1. Простые, статистические и теоретико-информационные метапризнаки.....	99
4.5.2. Метапризнаки на основе модели	100
4.5.3. Метапризнаки на основе производительности	100
4.5.4. Метаобучение или оптимизация на целевом наборе данных?	100
4.6. Получение новых признаков из базового набора	101
4.6.1. Генерация новых признаков путем агрегации	101
4.6.2. Генерация полного набора метапризнаков	101
4.6.3. Создание новых признаков с помощью PCA.....	102
4.6.4. Преобразование признаков путем отбора и проекции	102
4.6.5. Построение новых скрытых признаков с помощью матричного разложения	102
4.6.6. Создание новых признаков в виде встраиваний	103
4.7. Отбор метапризнаков	104
4.7.1. Статический отбор метапризнаков.....	104
4.7.2. Динамическая (итеративная) характеризация данных	105
4.8. Специфичные для алгоритма характеристики и проблемы представления	106
4.8.1. Характеристика данных, зависящая от алгоритма.....	106
Характеристика данных полезна для ранжирования пар алгоритмов ...	106
4.8.2. Проблемы представления.....	107
4.9. Установление сходства между наборами данных	107
4.9.1. Сходство на основе метапризнаков.....	107
4.9.2. Сходство, основанное на результатах работы алгоритмов	108

Косинусное подобие результатов производительности.....	108
Корреляционное подобие результатов производительности	109
4.10. Литература	109

Глава 5. Применение метаобучения к выбору алгоритма (продолжение).....

5.1. Введение.....	116
5.2. Использование регрессионных моделей в системах метаобучения.....	118
5.2.1. Эмпирические модели производительности	118
Использование метаданных из текущего набора данных	118
Подходы, использующие метаданные из других наборов данных.....	119
5.2.2. Нормализация производительности	120
5.2.3. Модели производительности.....	120
5.2.4. Деревья кластеризации.....	121
5.2.5. Преобразование прогнозов производительности в рейтинги	122
5.2.6. Прогнозирование производительности для каждого экземпляра	122
5.2.7. Преимущества и недостатки прогнозирования производительности	123
Преимущества.....	123
Недостатки	123
5.3. Использование классификации на метауровне для прогнозирования применимости.....	124
5.3.1. Алгоритмы классификации, используемые на метауровне.....	125
5.4. Методы, основанные на попарных сравнениях	125
5.4.1. Парные тесты, использующие ориентиры.....	126
5.4.2. Парный метод, основанный на частичных кривых обучения	126
Представление частичных кривых обучения.....	128
Проведение тестов на целевом наборе данных	128
Поиск наиболее похожих кривых обучения.....	128
Адаптация полученных кривых.....	129
Выполнение прогнозов для k ближайших наборов данных	129
Основные результаты	130
5.5. Парный метод для набора алгоритмов.....	130
Подробности приведены в следующих разделах.	130
Повтор сравнения для всех пар и создание частичного рейтинга	131
Определение лучшего алгоритма(ов).....	131
Оценка.....	132
Недостатки этого подхода.....	132
Использование частичного ранжирования для выполнения top-n алгоритмов	133
Расширение метода среднего ранжирования до частичного ранжирования.....	133
5.6. Итеративный подход к проведению парных тестов.....	133
Инициализация текущего лучшего алгоритма	134
Поиск лучшего парного теста.....	134

Вариант метода, учитывающий точность и время.....	135
Основные выводы	135
Связь с суррогатными моделями.....	136
5.7. Использование ART-деревьев и лесов	136
Построение набора парных моделей	136
Использование лесов ART для создания прогнозов.....	137
5.8. Активное тестирование.....	137
5.8.1. Активное тестирование, учитывающее точность и время выполнения	138
5.8.2. Активное тестирование с упором на аналогичные наборы данных.....	141
Сходство, основанное на полярности различий в производительности	141
5.8.3. Заключение	142
Определение наилучшего варианта для тестирования с помощью функций сбора	142
Использование метода AT для выбора и настройки рабочего процесса.....	142
Связь AT с сокращением пространств конфигураций.....	142
5.9. Непропозициональные подходы.....	143
5.10. Литература	143

Глава 6. Оптимизация гиперпараметров с помощью метаобучения.....

6.1. Введение.....	148
6.1.1. Обзор этой главы	150
6.2. Основные методы оптимизации гиперпараметров.....	151
6.2.1. Основные понятия.....	151
6.2.2. Основные методы оптимизации.....	152
Поиск по сетке	152
Случайный поиск.....	152
Расширение поиска с помощью гоночных методов	153
6.2.3. Эволюционные методы	154
6.2.4. Методы эвристического поиска	154
6.2.5. Гиперградиенты.....	155
6.2.6. Методы переменной точности	155
Последовательное деление пополам.....	156
Hyperband и расширения для последовательного деления пополам.....	157
6.3. Байесовская оптимизация.....	157
6.3.1. Последовательная оптимизация на основе модели.....	158
Функция сбора	159
Гауссовы процессы как суррогатные модели потерь.....	159
Случайные леса как суррогатные модели потерь.....	160
Примечание о предшествующих методах.....	160
6.3.2. Оценщик Парзена с древовидной структурой	160

6.4. Оптимизация гиперпараметров с помощью метаобучения.....	161
6.4.1. Горячий запуск: использование метазнаний при инициализации	161
Повторное использование лучшей конфигурации	162
Поиск глобально лучшей конфигурации	162
Ранжирование конфигураций	163
6.4.2. Использование метазнаний в байесовской оптимизации.....	164
Суррогатная совместная настройка (SCoT/MKL)	164
Гауссов процесс с многоядерным обучением (MKL-GP)	164
Многозадачная и переменная байесовская оптимизация.....	165
Ансамбль индивидуальных суррогатных моделей (SGPT)	165
Функция переноса-сбора (TAF)	166
Фокусировка внимания на высокоэффективных регионах с помощью QRF	167
6.4.3. Адаптивное сходство наборов данных	167
6.5. Заключительные замечания.....	167
Планирование эксперимента, исследование и использование	167
6.6. Заключение	168
6.7. Вопросы для обсуждения	168
6.8. Литература	169

Глава 7. Автоматизация проектирования конвейеров

7.1. Введение	173
7.1.1. Организация этой главы	174
7.1.2. Процесс KDD	175
7.2. Ограничение поиска при автоматизированном проектировании конвейера	176
7.2.1. Определение пространства альтернатив (декларативная предвзятость)	176
Роль онтологий	176
Что онтологии обычно не выражают	178
7.2.2. Различные способы добавления процедурной предвзятости.....	179
Использование эвристического ранжировщика	179
7.2.3. Контекстно-независимые грамматики	179
Пример	179
Индуктивный вывод CFG из примеров рабочих процессов	181
Ограничения CFG	182
7.3. Стратегии, используемые при создании конвейера	182
7.3.1. Операторы.....	182
7.3.2. Ручной выбор операторов	182
7.3.3. Ручное изменение существующих конвейеров.....	183
7.3.4. Использование планирования в разработке рабочего процесса.....	184
Абстрактные и конкретные операторы	184
Как работает планирование.....	185
Использование иерархического планирования	185
Инструмент оптимизации конвейера на основе дерева (TPOT).....	186

Общий помощник по автоматическому машинному обучению (GAMA)	187
Методы сокращения пространства поиска	187
Приоритизация поиска	188
Использование метазнаний в планировании	188
Методы ревизии рабочих процессов (конвейеров).....	188
7.4. Использование рейтингов успешных планов	189
Эффективность ранжирования конвейеров.....	190
Портфель успешных конвейеров	190
7.5. Литература.....	190

Часть II. ПЕРЕДОВЫЕ ТЕХНОЛОГИИ И МЕТОДЫ 195

Глава 8. Настройка пространств конфигураций и экспериментов 196

8.1. Введение	196
8.1.1. Структура этой главы	197
8.2. Типы пространств конфигураций	198
8.2.1. Пространства конфигураций, связанные с выбором алгоритма	198
8.2.2. Пространства конфигураций, связанные с оптимизацией гиперпараметров и CASH	198
Типы гиперпараметров	198
Непрерывные и дискретные пространства	199
Условные гиперпараметры и пространства	199
Выборка в непрерывных подпространствах	199
8.2.3. Пространства конфигураций, связанные с проектированием конвейера.....	200
8.3. Соответствие пространства конфигураций текущим задачам	201
8.3.1. Общие принципы построения пространств конфигураций	202
8.4. Значимость гиперпараметра и предельный вклад.....	203
8.4.1. Предельный вклад алгоритмов (конвейеров)	203
8.4.2. Определение значимости гиперпараметра для заданного набора данных.....	204
Прямой отбор.....	204
Абляционный анализ.....	204
Функциональный дисперсионный анализ	205
8.4.3. Определение значимости гиперпараметров для нескольких наборов данных	205
8.5. Сокращение пространства конфигураций.....	207
8.5.1. Сокращение портфелей алгоритмов/конфигураций	207
Выявление конкурентных алгоритмов	207
Пример	208
Использование алгоритма покрытия для выбора «неизбыточных» алгоритмов	209
Использование алгоритма покрытия с подобием на макроуровне	210

Использование алгоритма покрытия с подобием на микроуровне	210
8.5.2. Метод сокращения, основанный на комбинации мер	211
Подход, основанный на огибающей кривой	211
8.6. Пространства конфигураций в символическом обучении	212
8.6.1. Пространства версий	212
Управление предвзятостью предметно-ориентированного языка	213
Расширение предвзятости предметно-ориентированного языка	213
8.7. Какие наборы данных необходимы?	214
8.7.1. Использование существующих репозиториях наборов данных	214
8.7.2. Создание синтетических наборов данных	215
8.7.3. Создание вариантов существующих наборов данных	215
8.7.4. Сегментация большого набора данных или потока данных	216
8.7.5. Поиск наборов данных, обладающих различающей способностью	216
Использование характеристик наборов данных и 2D-следов	217
Использование корреляции рейтингов для характеристики разнообразия	218
8.8. Полные и неполные метаданные	218
8.8.1. Можно ли получить полные метаданные?	219
Слишком много ожидаемых экспериментов	219
Некоторые эксперименты могут привести к неудаче	219
Добавление новых наборов данных	220
Использование оценок вместо реальных значений	220
8.8.2. Необходимо ли иметь полные метаданные?	221
8.8.3. Имеет ли значение порядок тестов?	221
8.9. Использование стратегий многоруких бандитов для планирования экспериментов	221
8.9.1. Некоторые концепции и стратегии MAB	222
ϵ -жадная стратегия	222
ϵ -начальная стратегия	222
ϵ -убывающая стратегия	222
Метод сопоставления вероятностей (SoftMax)	223
Методы интервальной оценки и верхней доверительной границы	223
Стратегии ценообразования (POKER)	224
Контекстная задача многорукого бандита	224
8.10. Заключение	225
8.11. Литература	225

Глава 9. Объединение базовых учащихся в ансамбли

9.1. Введение	230
9.2. Бэггинг и бустинг	232
9.2.1. Бэггинг	232
9.2.2. Бустинг	233
9.3. Стекинг и каскадное обобщение	235
9.3.1. Стекинг	235
9.3.2. Каскадное обобщение	237
9.4. Каскадирование и делегирование	239

9.4.1. Каскадирование	239
9.4.2. Делегирование	241
9.5. Арбитраж	243
9.6. Деревья метарешений	246
9.7. Обсуждение	248
9.8. Литература	248
Глава 10. Метаобучение в ансамблевых методах	251
10.1. Введение	251
10.2. Основные характеристики ансамблевых систем	253
Хотим ли мы использовать существующий портфель решений?	253
Прогнозы для всего набора данных или для каждого примера?	253
Какой ансамблевый метод используется?	253
Модели генерируются с помощью одного или разных алгоритмов?	253
Метаданные извлекаются из текущих или прошлых наборов данных?	254
Какова задача обучения базового уровня?	254
10.3. Ансамблевые методы на основе выбора	254
10.4. Ансамблевое обучение (на наборе данных)	255
10.4.1. Метаобучение на этапах построения и сокращения	255
Генерация и сокращение	255
Повторное использование подходов на основе выбора для ансамблевого обучения	256
Выбор алгоритма ML на метауровне	257
Моделирование взаимозависимости моделей	257
Метапризнаки	258
Стратегия, используемая в Auto-sklearn	258
10.4.2. Метаобучение на этапе интеграции	258
Интеграция	258
Метод метаобучения	259
10.5. Динамический выбор моделей (для каждого экземпляра)	259
Повторное использование подходов на основе выбора для ансамблевого обучения	260
Структура слоев в системе ALMA	260
Моделирование взаимозависимости моделей	260
10.5.1. Метапризнаки	261
Использование признаков базового уровня и прогнозов в качестве метапризнаков	261
10.6. Генерация иерархических ансамблей	262
10.6.1. Иерархические ансамбли	262
10.6.2. Развитие иерархических ансамблей с эволюционными вычислениями	262
10.6.3. Метаобучение в методах иерархического ансамбля	263
10.7. Выводы и перспективные направления	264
10.8. Литература	264

Глава 11. Система рекомендации алгоритмов для потоковых данных	266
11.1. Введение	266
Формальное представление	268
11.1.1. Адаптация пакетных классификаторов к потоковым данным	269
11.1.2. Адаптация ансамблей к потоковым данным	270
11.1.3. Общая постановка задачи	270
11.2. Подходы на основе метапризнаков	271
11.2.1. Методы	272
11.2.2. Обучение метамоделей	273
11.2.3. Метапризнаки	274
11.2.4. Соображения относительно гиперпараметров	274
11.2.5. Метамодель	275
11.2.6. Оценка систем метаобучения для потоковых данных	276
11.2.7. Эталонные показатели	276
11.2.8. Промежуточный итог	277
11.3. Ансамблирование в области потоковых данных	278
11.3.1. Лучший классификатор на последнем интервале (Blast)	278
11.3.2. Коэффициенты затухания	279
11.3.3. Неоднородные ансамбли для случая дрейфа признаков	281
11.3.4. Соображения относительно выбора базовых классификаторов	281
11.3.5. Промежуточный итог	282
11.4. Повторяющиеся модели метауровня	283
11.4.1. Ансамбль, взвешенный по точности	283
11.4.2. Двухуровневая архитектура	284
11.5. Направления будущих исследований	285
11.6. Литература	286
Глава 12. Перенос знаний между задачами	289
12.1. Введение	289
12.2. Предыстория, терминология и обозначения	290
12.2.1. Когда применяется перенос обучения?	290
12.2.2. Различные типы переноса обучения	291
12.2.3. Что именно можно переносить?	293
12.3. Архитектуры, применяемые при переносе обучения	294
12.3.1. Перенос в нейронных сетях	294
12.3.2. Перенос обучения в ядерных методах	298
12.3.3. Перенос знаний в параметрических байесовских моделях	299
12.4. Теоретический базис «обучения обучению»	300
12.4.1. Сценарий «обучения обучению»	301
12.4.2. Границы ошибки обобщения для метаучеников	302
12.4.3. Другие теоретические исследования	303
Определение границ с использованием алгоритмической устойчивости	303
Границы в сценарии адаптации предметной области	304

12.4.4. Систематическая ошибка и дисперсия в метаобучении	304
Приложение А	305
12.5. Литература	307

Глава 13. Метаобучение и глубокие нейронные сети 311

13.1. Введение	311
13.2. Предыстория и обозначения	312
13.2.1. Метаабстракция для глубоких нейронных сетей	312
13.2.2. Общие процедуры обучения и оценки	313
N-классовое k-кратное обучение	314
13.2.3. Обзор остальной части этой главы	315
13.3. Метаобучение на основе метрик	316
Пример	317
13.3.1. Сиамские нейронные сети	319
13.3.2. Сопоставляющие сети	320
13.3.3. Графовые нейронные сети	322
13.3.4. Внимательные рекуррентные компараторы	323
Методы на основе метрик – краткий итог	324
13.4. Метаобучение на основе моделей	324
Пример	325
13.4.1. Нейронные сети с дополненной памятью	326
13.4.2. Метасети	328
13.4.3. Простой нейронный внимательный метаученик (SNAIL)	330
13.4.4. Условные нейронные процессы	332
Методы на основе моделей – краткие итоги	333
13.5. Метаобучение на основе оптимизации	333
Пример	334
13.5.1. Оптимизатор LSTM	334
13.5.2. Оптимизатор на основе обучения с подкреплением	336
13.5.3. Независимое от модели метаобучение (MAML)	336
13.5.4. Reptile	340
Методы на основе оптимизации – краткий итог	341
13.6. Обсуждение и перспективы исследований	342
13.6.1. Нерешенные проблемы	343
13.6.2. Перспективные направления исследований	343
13.7. Литература	344

Глава 14. Автоматизация науки о данных 348

14.1. Введение	348
14.2. Определение текущей проблемы/задачи	350
14.2.1. Понимание и описание проблемы	350
14.2.2. Создание дескрипторов задач	350
14.2.3. Определение типа задачи и целей	351
Роль целей обучения	351

Планирование целей обучения	352
14.3. Определение предметной области и знаний.....	352
Определение области путем сопоставления дескрипторов/ метапризнаков.....	353
Определение области путем классификации	353
Представление данных и целей	353
14.4. Получение данных	353
14.4.1. Выбрать существующие данные или спланировать получение новых?	354
14.4.2. Выявление данных и контекстных знаний, относящихся к предметной области.....	354
14.4.3. Получение данных и базовых знаний из разных источников.....	355
Получение данных из куба OLAP	355
14.5. Автоматизация предварительной обработки и преобразования данных	355
14.5.1. Преобразование/подготовка данных	357
Вывод типов данных	357
Некоторые способы подготовки данных	357
Система FOOFAN	357
Использование LLP в подготовке данных	358
Системы TDE и SYNTH.....	358
14.5.2. Выбор записей и сжатие модели	359
14.5.3. Автоматизация выбора метода предварительной обработки.....	360
14.5.4. Изменение степени детализации представления	361
Генерация агрегированных данных из куба OLAP.....	361
14.6. Автоматизация создания моделей и отчетов	362
14.6.1. Автоматизация создания и развертывания модели	362
14.6.2. Автоматическое создание отчетов	362
14.7. Литература.....	362

Глава 15. Автоматизация проектирования сложных систем 366

15.1. Введение	366
15.1.1. Обзор этой главы	368
15.2. Использование расширенного набора операторов	368
15.3. Изменение степени детализации путем введения новых понятий.....	369
Получение новых понятий из внешних источников	369
Автономное добавление новых понятий	370
15.3.1. Определение новых понятий путем кластеризации	370
15.3.2. Конструктивная индукция	370
15.3.3. Переформулировка теорий, состоящих из правил.....	370
Переформулировка теорий по специализации	370
Свертывание и развертывание	371
Поглощение	372
15.3.4. Введение новых понятий, выраженных в виде правил.....	372
15.3.5. Пропозиционализация	373
15.3.6. Автоматическое построение признаков в глубоких нейросетях	373

15.3.7. Повторное использование новых понятий для переопределения онтологий	374
15.4. Повторное использование новых понятий в продолжающемся обучении.....	374
Пример: использование приобретенных навыков для обучения более сложному поведению	374
15.5. Итеративное обучение	375
Пример: изучение определения сортировки вставками.....	377
15.6. Обучение решению взаимозависимых задач.....	378
15.7. Литература.....	379

Часть III. ОРГАНИЗАЦИЯ И ИСПОЛЬЗОВАНИЕ МЕТАДАННЫХ

381

Глава 16. Хранилища метаданных	382
16.1. Введение	382
16.2. Как устроен мир информации о машинном обучении	383
16.2.1. Потребность в качественных метаданных	383
16.2.2. Инструменты и инициативы	384
16.3. Что такое OpenML?.....	385
16.3.1. Наборы данных	385
16.3.2. Типы задач.....	386
16.3.3. Задачи.....	387
16.3.4. Потоки	388
16.3.5. Установки	388
16.3.6. Прогоны.....	389
16.3.7. Исследования и тестовые наборы.....	389
16.3.8. Интеграция OpenML в среды машинного обучения	391
Пример исследования с использованием существующих оценочных результатов.....	393
16.4. Литература	394

Глава 17. Обучение на метаданных в репозиториях

397

17.1. Введение	397
17.2. Анализ производительности алгоритмов на разных наборах данных	398
17.2.1. Сравнение различных алгоритмов	398
17.2.2. Влияние изменения некоторых настроек гиперпараметров	399
17.3. Анализ производительности алгоритмов на разных наборах данных	401
17.3.1. Эффект от использования разных классификаторов с гиперпараметрами по умолчанию	401
Оценка статистической значимости	402
17.3.2. Эффект оптимизации гиперпараметров.....	402
17.3.3. Выявление алгоритмов (рабочих процессов) со схожими прогнозами	405

17.4. Влияние определенных характеристик данных/рабочего процесса на производительность.....	407
17.4.1. Влияние выбора линейных и нелинейных моделей.....	407
17.4.2. Эффект от применения отбора признаков.....	408
17.4.3. Влияние конкретных настроек гиперпараметров	410
Настраиваемость алгоритмов.....	410
Настраиваемость гиперпараметров	411
Определение важности гиперпараметров в наборах данных с помощью ANOVA	411
17.5. Заключение.....	412
17.6. Литература.....	414

Глава 18. Заключительные соображения

18.1. Введение	416
18.2. Форма метазнания, используемая в различных подходах.....	417
18.2.1. Применение метазнаний в методах выбора алгоритма.....	418
Способы ранжирования, использующие априорные метаданные.....	418
Подходы, использующие динамические метаданные.....	418
18.2.2. Метазнания в подходах к оптимизации гиперпараметров	419
18.2.3. Использование метазнаний при разработке конвейеров	419
18.2.4. Метазнания в переносе обучения и в глубоких нейронных сетях... ..	420
18.3. Перспективные задачи и направления исследований.....	420
18.3.1. Разработка метапризнаков, связанных с характеристиками набора данных и производительностью	421
18.3.2. Дальнейшая интеграция подходов метаобучения и AutoML.....	421
18.3.3. Автоматизация подстройки к текущей задаче	421
Автоматизация получения метаданных.....	421
18.3.4. Автоматизация сокращения пространства конфигураций.....	422
Автоматизация сокращения алгоритмов базового уровня	422
Автоматизация сокращения пространства гиперпараметров.....	422
Автоматизация сокращения пространства рабочего процесса (конвейера).....	423
18.3.5. Автоматизация анализа потоков данных	423
18.3.6. Автоматизация настройки параметров нейронной сети.....	423
18.3.7. Автоматизация науки о данных	424
Постановка текущей проблемы/задачи	424
Выбор подходящего предметно-ориентированного метазнания	425
Получение данных	425
Изменение детализации представления	425
18.3.8. Автоматизация проектирования решений с более сложными структурами.....	426
18.3.9. Проектирование платформ метаобучения/AutoML.....	426
18.4. Заключение и обращение к читателям.....	426
18.5. Литература	426

Предметный указатель.....	427
----------------------------------	------------

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 поможете 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.

Предисловие

Первое издание этой книги вышло в свет в 2009 г., т. е. к моменту написания второго издания прошло более 10 лет. Поскольку за эти годы область машинного обучения существенно продвинулась вперед, мы решили подготовить второе издание. Мы постарались включить в него наиболее важные достижения, чтобы новая версия книги содержала актуальную информацию об этой области и была полезна исследователям, аспирантам и прикладным специалистам, работающим в этой области.

Каковы основные изменения? Прежде всего если вы просто сравните количество глав двух изданий, то заметите, что их стало вдвое больше. Примерно так же увеличилось количество страниц.

Отметим, что на момент написания первого издания термина *автоматизированное машинное обучение* (automated machine learning, AutoML) даже не существовало. Очевидно, что мы должны были описать это новаторское направление в новом издании, а также прояснить его связь с метаобучением. Кроме того, автоматизация методов проектирования цепочек операций – в настоящее время называемых *конвейерами* (pipeline) или *рабочими процессами* (workflow) – находилась в зачаточном состоянии. Разумеется, мы осознавали необходимость обновить существующий материал, чтобы не отставать от этого развития.

В последние годы исследования в области AutoML и метаобучения привлекают большое внимание не только исследователей, но и многих компаний, занимающихся искусственным интеллектом, включая, например, Google и IBM. Как можно использовать метаобучение для улучшения систем AutoML – это один из важнейших вопросов, на который в настоящее время пытаются ответить многие исследователи.

Эта книга нацелена в будущее. Как это обычно случается, чем лучше исследователи разбираются в какой-то области, тем больше перед ними возникает новых вопросов. Мы позаботились о том, чтобы включить некоторые из них в соответствующие главы.

Авторами первого издания были Павел Браздил, Кристоф Жиро-Каррье, Карлос Соарес и Рикардо Вилальта. Учитывая масштабные нововведения в этой области, мы решили укрепить команду, пригласив Хоакина Ваншорена и Яна ван Рейна присоединиться к проекту. К сожалению, Кристоф и Рикардо не смогли принять участие в работе над новым изданием. Тем не менее все авторы второго издания очень благодарны за их вклад в начало проекта.

Как устроена эта книга

Эта книга состоит из трех частей. В части I (главы 2–7) рассмотрены основные концепции и архитектура систем метаобучения и AutoML, а в части II (главы 8–15) обсуждаются различные расширения. Часть III (главы 16–18) описывает способы хранения и управления метаданными (например, хранилища метаданных) и заканчивается заключительными замечаниями.

Часть I. Основные понятия и архитектура

Глава 1 начинается с объяснения основных понятий, используемых в этой книге, таких как машинное обучение, метаобучение, автоматизированное машинное обучение и др. Затем она продолжается обзором базовой архитектуры системы метаобучения и служит введением к остальной части книги. Над этой главой работали все соавторы книги.

Глава 2 посвящена методам ранжирования на основе метаданных, поскольку их относительно легко реализовать, но это не умаляет их полезности в практических приложениях. Эта глава была написана П. Браздилом и Я. ван Рейном¹. Глава 3, написанная теми же авторами, посвящена теме оценки метаобучения и систем AutoML. В главе 4 обсуждаются различные показатели наборов данных, которые играют важную роль в качестве метапризнаков в системах метаобучения. Эта глава, как и следующая, также написана П. Браздилом и Я. ван Рейном. Главу 5 можно рассматривать как продолжение главы 2. В ней обсуждаются различные подходы к метаобучению, включая, например, попарные сравнения, которые применялись в прошлом. В главе 6 обсуждается оптимизация гиперпараметров. Она охватывает как базовые методы поиска, так и более продвинутые, применяемые в области AutoML. Эту главу написали три автора – П. Браздил, Я. ван Рейн и Х. Ваншорен. В главе 7 обсуждается вопрос автоматизации построения рабочих процессов или конвейеров, представляющих собой последовательности операций. Эта глава написана П. Браздилом, но в ней повторно использованы некоторые материалы из первого издания, подготовленного К. Жиро-Каррье.

Часть II. Передовые технологии и методы

Часть 2 (главы 8–15) продолжает темы части I, но охватывает различные расширения базовой методологии. Глава 8, написанная П. Браздилом и Я. ван Рейном, посвящена теме построения пространств конфигураций и планированию экспериментов. В двух последующих главах обсуждается конкретная тема ансамблей моделей. Глава 9, написанная Ш. Жиро-Каррье, дополняет материал этой книги. Она описывает различные способы организации набора алгоритмов базового уровня в ансамбли. Авторы второго издания не видели необходимости изменять эту главу, поэтому она сохранена в том виде, в каком появилась в первом издании.

Глава 10 продолжает тему ансамблей и показывает, как метаобучение можно использовать при построении ансамблей (ансамблевом обучении).

¹ Части глав 2 и 3 первого издания, написанные К. Соаресом и П. Браздилом, были повторно использованы и адаптированы для этой главы.

Эта глава была написана К. Соаресом и П. Браздилом. Последующие главы посвящены более конкретным темам. Глава 11, написанная Я. ван Рейном, описывает применение метаобучения для предоставления рекомендаций по выбору алгоритма потоковой обработки данных. Глава 12, написанная Р. Вилалтой и М. Месхи, посвящена переносу метамodelей и представляет собой вторую дополняющую главу этой книги. Это существенно обновленная версия аналогичной главы из первого издания, написанная Р. Вилалтой. Глава 13, написанная М. Хьюисманом, Я. ван Рейном и А. Плаатом, обсуждает метаобучение в глубоких нейронных сетях и представляет собой третью дополняющую главу этой книги. Глава 14 посвящена относительно новой теме автоматизации науки о данных. Эта глава была составлена П. Браздилом и содержит обзор различных идей и предложений его соавторов. Цель главы состоит в том, чтобы обсудить различные операции, обычно выполняемые в науке о данных, и рассмотреть вопрос о том, возможна ли здесь автоматизация и можно ли использовать в этом процессе метазнания. Цель главы 15, написанной П. Браздилом, также состоит в том, чтобы заглянуть в будущее и рассмотреть возможность автоматизации проектирования более сложных решений. В их число могут входить не только конвейеры операций, но и более сложные структуры управления (например, итерации) и автоматические изменения в базовом представлении.

Часть III. Организация и использование метаданных

Часть III охватывает некоторые практические вопросы и содержит последние три главы (16–18). В главе 16, написанной Х. Ваншореном и Я. ван Рейном, обсуждаются репозитории метаданных и, в частности, репозиторий, известный под названием OpenML. Этот репозиторий содержит данные о многих экспериментах по машинному обучению, проведенных в прошлом, и их соответствующие результаты. В главе 17, написанной Я. ван Рейном и Х. Ваншореном, показано, как можно изучать метаданные, чтобы получить более глубокое представление об исследованиях машинного обучения и метаобучения и, как следствие, сконструировать новые эффективные прикладные системы. Глава 18 завершает книгу краткими заключительными замечаниями о роли метазнания, а также представляет некоторые перспективные направления исследований. В основном глава была написана П. Браздилом, но содержит вклад других соавторов, в частности Я. ван Рейна и К. Соареса.

Благодарности

Мы выражаем благодарность всем тем, кто помог осуществить этот проект.

Мы признательны за грант 612.001.206 от Голландского исследовательского совета (NWO), направленный на финансирование издания этой книги.

Павел Браздил выражает благодарность Университету Порту, экономическому факультету, научно-исследовательскому институту INESC TEC и одному из его научных центров, а именно Лаборатории искусственного интеллекта и поддержки принятия решений (LIAAD), за их постоянную поддержку. Выполненная работа была частично поддержана национальными фунда-

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru