



БИБЛИОТЕКА
ПРОГРАММИСТА

Бретт Ланц



МАШИННОЕ ОБУЧЕНИЕ НА R

ЭКСПЕРТНЫЕ ТЕХНИКИ
ДЛЯ ПРОГНОСТИЧЕСКОГО АНАЛИЗА

Packt





**БИБЛИОТЕКА
ПРОГРАММИСТА**

Бретт Ланц

МАШИННОЕ ОБУЧЕНИЕ НА R

**ЭКСПЕРТНЫЕ ТЕХНИКИ
ДЛЯ ПРОГНОСТИЧЕСКОГО АНАЛИЗА**



**Санкт-Петербург · Москва · Екатеринбург · Воронеж
Нижний Новгород · Ростов-на-Дону · Самара · Минск**

2020

ББК 32.813
УДК 004.85
Л22

Ланц Бретт

Л22 Машинное обучение на R: экспертные техники для прогностического анализа. — СПб.: Питер, 2020. — 464 с.: ил. — (Серия «Библиотека программиста»).

ISBN 978-5-4461-1512-9

Язык R предлагает мощный набор методов машинного обучения, позволяющих быстро проводить нетривиальный анализ ваших данных.

Книга является руководством, которое поможет применять методы машинного обучения в решении ежедневных задач. Бретт Ланц научит всему необходимому для анализа данных, формирования прогнозов и визуализации данных.

Здесь вы найдете информацию о новых улучшенных библиотеках, советы об этических аспектах машинного обучения и проблемах предвзятости, а также познакомитесь с глубоким обучением.

16+ (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ББК 32.813
УДК 004.85

Права на издание получены по соглашению с Packt Publishing. Все права защищены. Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.

Информация, содержащаяся в данной книге, получена из источников, рассматриваемых издательством как надежные. Тем не менее, имея в виду возможные человеческие или технические ошибки, издательство не может гарантировать абсолютную точность и полноту приводимых сведений и не несет ответственности за возможные ошибки, связанные с использованием книги. Издательство не несет ответственности за доступность материалов, ссылки на которые вы можете найти в этой книге. На момент подготовки книги к изданию все ссылки на интернет-ресурсы были действующими.

ISBN 978-1788295864 англ.

© Packt Publishing 2019.

First published in the English language under the title «Machine Learning with R — Third Edition» — (9781788295864)

ISBN 978-5-4461-1512-9

© Перевод на русский язык ООО Издательство «Питер», 2020

© Издание на русском языке, оформление ООО Издательство «Питер», 2020

© Серия «Библиотека программиста», 2020

Краткое содержание

Об авторе	13
О научном редакторе.....	14
Предисловие.....	15
От издательства	19
Глава 1. Введение в машинное обучение	20
Глава 2. Управление данными и их интерпретация	51
Глава 3. Ленивое обучение: классификация с использованием метода ближайших соседей	88
Глава 4. Вероятностное обучение: классификация с использованием наивного байесовского классификатора	111
Глава 5. Разделяй и властвуй: классификация с использованием деревьев решений и правил	149
Глава 6. Прогнозирование числовых данных: регрессионные методы	197
Глава 7. Методы «черного ящика»: нейронные сети и метод опорных векторов.....	248
Глава 8. Обнаружение закономерностей: анализ потребительской корзины с помощью ассоциативных правил	294
Глава 9. Поиск групп данных: кластеризация методом k-средних	321
Глава 10. Оценка эффективности модели	349
Глава 11. Повышение эффективности модели.....	385
Глава 12. Специальные разделы машинного обучения.....	414

Оглавление

Об авторе.....	13
О научном редакторе.....	14
Предисловие.....	15
Для кого предназначена книга.....	15
О чем вы прочтете в издании	15
Что вам нужно для чтения книги	17
Загрузите файлы примеров кода.....	17
Цветные иллюстрации	17
Условные обозначения	18
От издательства	19
Глава 1. Введение в машинное обучение	20
Происхождение машинного обучения	21
Область применения машинного обучения и злоупотребление им	23
Успехи машинного обучения.....	24
Пределы возможностей машинного обучения	25
Этика машинного обучения.....	26
Как учатся машины.....	30
Хранение данных.....	32
Абстрагирование	32
Обобщение	35
Оценка	37
Машинное обучение на практике.....	38
Типы входных данных.....	39

Типы алгоритмов машинного обучения.....	41
Подбор алгоритмов по входным данным.....	44
Машинное обучение с использованием R.....	46
Установка R-пакетов.....	47
Загрузка и выгрузка R-пакетов.....	48
Установка RStudio.....	48
Резюме.....	50
Глава 2. Управление данными и их интерпретация.....	51
Структуры данных R.....	52
Векторы.....	52
Факторы.....	54
Списки.....	55
Фреймы данных.....	58
Матрицы и массивы.....	61
Управление данными в R.....	62
Сохранение, загрузка и удаление структур данных в R.....	63
Импорт и сохранение данных из CSV-файлов.....	64
Исследование данных и их интерпретация.....	65
Структуры данных.....	66
Числовые переменные.....	67
Категориальные переменные.....	79
Взаимосвязи между переменными.....	82
Резюме.....	86
Глава 3. Ленивое обучение: классификация с использованием метода ближайших соседей.....	88
Что такое классификация методом ближайших соседей.....	89
Алгоритм k-NN.....	89
Почему алгоритм k-NN называют ленивым.....	98
Пример: диагностика рака молочной железы с помощью алгоритма k-NN.....	99
Этап 1. Сбор данных.....	99
Этап 2. Исследование и подготовка данных.....	100

Шаг 3. Обучение модели на данных.....	104
Шаг 4. Оценка эффективности модели	106
Шаг 5. Повышение эффективности модели.....	108
Резюме	110

Глава 4. Вероятностное обучение: классификация с использованием наивного байесовского классификатора	111
Наивный байесовский классификатор.....	112
Основные понятия байесовских методов.....	112
Наивный байесовский алгоритм.....	119
Пример: фильтрация спама в мобильном телефоне с помощью наивного байесовского алгоритма	126
Шаг 1. Сбор данных.....	127
Шаг 2. Исследование и подготовка данных.....	128
Шаг 3. Обучение модели на данных.....	144
Шаг 4. Оценка эффективности модели	146
Шаг 5. Повышение эффективности модели.....	147
Резюме	148

Глава 5. Разделяй и властвуй: классификация с использованием деревьев решений и правил	149
Деревья решений	150
Разделяй и властвуй	152
Алгоритм дерева решений C5.0.....	156
Пример: распознавание рискованных банковских кредитов с помощью деревьев решений C5.0.....	161
Шаг 1. Сбор данных.....	161
Шаг 2. Исследование и подготовка данных.....	162
Шаг 3. Обучение модели на данных.....	165
Шаг 4. Оценка эффективности модели	169
Шаг 5. Повышение эффективности модели.....	170
Правила классификации	174
Отделяй и властвуй	175
Алгоритм 1R	178
Алгоритм RIPPER.....	181

Правила, построенные на основе деревьев решений	183
Когда деревья и правила становятся жадными	184
Пример: распознавание ядовитых грибов по алгоритму обучения на основе правил.....	187
Шаг 1. Сбор данных.....	187
Шаг 2. Исследование и подготовка данных.....	188
Шаг 3. Обучение модели на данных.....	189
Шаг 4. Оценка эффективности модели	192
Шаг 5. Повышение эффективности модели.....	192
Резюме	195
Глава 6. Прогнозирование числовых данных: регрессионные методы	197
Понятие регрессии.....	198
Простая линейная регрессия.....	200
Оценка методом наименьших квадратов	203
Корреляции	206
Множественная линейная регрессия.....	208
Пример: прогнозирование медицинских расходов с помощью линейной регрессии.....	213
Шаг 1. Сбор данных.....	213
Шаг 2. Исследование и подготовка данных.....	214
Шаг 3. Обучение модели на данных.....	220
Шаг 4. Определение эффективности модели	222
Шаг 5. Повышение эффективности модели.....	224
Регрессионные деревья и деревья моделей	231
Дополнение деревьев регрессией	232
Пример: оценка качества вина с помощью регрессионного дерева и дерева моделей	234
Шаг 1. Сбор данных.....	235
Шаг 2. Исследование и подготовка данных.....	236
Шаг 3. Обучение модели на данных.....	237
Шаг 4. Определение эффективности модели	241
Шаг 5. Повышение эффективности модели.....	243
Резюме	247

Глава 7. Методы «черного ящика»: нейронные сети и метод опорных векторов	248
Нейронные сети.....	249
От биологических нейронов — к искусственным	250
Функции активации.....	252
Топология сети	255
Обучение нейронной сети методом обратного распространения ошибки	259
Пример: моделирование прочности бетона с помощью нейронной сети	262
Шаг 1. Сбор данных.....	262
Шаг 2. Исследование и подготовка данных.....	263
Шаг 3. Обучение модели на данных.....	264
Шаг 4. Оценка эффективности модели	267
Шаг 5. Повышение эффективности модели	268
Метод опорных векторов	273
Классификация гиперплоскостями	274
Использование ядер в нелинейных пространствах.....	280
Пример: оптическое распознавание символов с помощью SVM	282
Шаг 1. Сбор данных.....	283
Шаг 2. Исследование и подготовка данных.....	284
Шаг 3. Обучение модели на данных.....	286
Шаг 4. Оценка эффективности модели	288
Шаг 5. Повышение эффективности модели.....	290
Резюме	293
Глава 8. Обнаружение закономерностей: анализ потребительской корзины с помощью ассоциативных правил	294
Ассоциативные правила.....	295
Алгоритм Apriori для поиска ассоциативных правил	296
Измерение интересности правила: поддержка и доверие	298
Построение набора правил по принципу Apriori	300

Пример: выявление часто покупаемых продуктов в соответствии с ассоциативными правилами	301
Шаг 1. Сбор данных.....	302
Шаг 2. Исследование и подготовка данных.....	303
Шаг 3. Обучение модели на данных.....	310
Шаг 4. Оценка эффективности модели	313
Шаг 5. Повышение эффективности модели.....	316
Резюме	320
Глава 9. Поиск групп данных: кластеризация методом k-средних	321
Что такое кластеризация	322
Кластеризация как задача машинного обучения	322
Алгоритм кластеризации методом k-средних.....	325
Сегментация рынка для подростков с использованием кластеризации методом k-средних.....	333
Шаг 1. Сбор данных.....	334
Шаг 2. Исследование и подготовка данных.....	335
Шаг 3. Обучение модели на данных.....	339
Шаг 4. Оценка эффективности модели	342
Шаг 5. Повышение эффективности модели.....	346
Резюме	347
Глава 10. Оценка эффективности модели	349
Измерение эффективности классификации.....	350
Прогнозы классификатора	350
Анализ матриц несоответствий	354
Использование матриц несоответствий для измерения эффективности	357
Не только точность: другие показатели эффективности	359
Визуализация компромиссов эффективности с помощью ROC-кривых.....	368
Оценка эффективности в будущем	374
Метод отложенных данных	375
Резюме	383

Глава 11. Повышение эффективности модели.....	385
Повышение эффективности готовых моделей.....	386
Автоматическая настройка параметров с помощью пакета caret	387
Повышение эффективности модели с помощью метаобучения.....	397
Понятие ансамблей.....	398
Бэггинг	400
Бустинг	402
Случайные леса	405
Резюме	413
Глава 12. Специальные разделы машинного обучения.....	414
Управление реальными данными и их подготовка	415
Очистка данных с помощью пакетов tidyverse.....	415
Чтение и запись данных во внешние файлы	419
Получение данных путем запросов к базам данных SQL	420
Работа с онлайн-данными и сервисами	425
Загрузка полного текста веб-страниц.....	426
Синтаксический анализ данных, полученных с веб-страниц.....	428
Работа со специфическими данными	435
Анализ данных в биоинформатике.....	436
Анализ и визуализация сетевых данных.....	436
Повышение эффективности R.....	441
Управление сверхбольшими наборами данных.....	442
Ускорение обучения благодаря параллельным вычислениям	445
Развертывание оптимизированных алгоритмов обучения	455
Вычисления на GPU	459
Резюме	462