

Содержание

Предисловие	10
Глава 1. Установка и начало работы	13
Подготовка к установке	13
Установка R	13
Установка Python	14
Конфиденциальность	14
Установка Java	15
Установка H2O при помощи R (CRAN)	15
Установка H2O при помощи Python (pip).....	16
Наша первая задача машинного обучения.....	17
Обучение и предсказания в Python.....	21
Обучение и предсказания в R	23
Производительность и предсказания	25
Если вам не повезло	26
Веб-интерфейс Flow	27
Данные	28
Модели.....	29
Предсказания	31
Дополнительные сведения об интерфейсе Flow	32
Резюме.....	32
Глава 2. Импортирование и экспортирование данных	33
Требования к памяти	33
Подготовка данных.....	34
Загрузка данных в H2O.....	35
Загрузка файлов в формате CSV.....	35
Загрузка файлов в других форматах	37
Загрузка данных из R	38
Загрузка данных из Python	39
Операции с данными	40
«Ленивость», присвоение имен и удаление	40
Итоговые статистики	42
Операции со столбцами	42
Агрегирование строк.....	43
Индексация	44
Разделение данных в кластере H2O	46
Строки и столбцы	49
Выгрузка данных из H2O	52
Экспорт таблиц	52
Формат ROJO	53
Файлы моделей.....	54
Сохранение всех моделей	54
Резюме.....	55

Глава 3. Наборы данных	56
Набор данных об энергетической эффективности	56
Настройка и загрузка	57
Переменные	58
Разделение данных.....	59
Изучение данных.....	60
О наборе данных.....	64
Набор данных: рукописные цифры.....	64
Настройка и загрузка	65
Изучение данных.....	66
Как можно «помочь» модели	68
О наборе данных.....	70
Набор данных: результаты футбольных матчей	70
Корреляции	73
Пропущенные данные.....	76
Как обучать и тестировать?	77
Настройка и загрузка	77
Третий источник данных	78
Снова про пропущенные данные	80
Настройка и загрузка (снова).....	80
О наборе данных.....	83
Резюме.....	83
Глава 4. Общие параметры моделей	84
Поддерживаемые метрики	84
Метрики для регрессии	85
Метрики для классификации.....	85
Бинарная классификация	86
Основы	88
Объем выполняемой работы	89
Оценка и проверка	90
Ранняя остановка.....	90
Контрольные точки	92
Перекрестная проверка.....	94
Взвешивание наблюдений.....	95
Выборки и обобщающая способность.....	98
Регрессия.....	99
Контроль вывода результатов.....	100
Резюме.....	100
Глава 5. Случайный лес	101
Решающие деревья	101
Случайный лес	103
Параметры	103
Энергоэффективность зданий: случайный лес с настройками по умолчанию	105

Поиск по сетке	107
Полный перебор	108
Случайный поиск.....	110
Общая стратегия.....	112
Энергоэффективность зданий: настроенный случайный лес.....	113
MNIST: случайный лес с настройками по умолчанию	114
MNIST: настроенный случайный лес	116
Дополненные данные.....	119
Футбол: случайный лес с настройками по умолчанию.....	120
Футбол: настроенный случайный лес	122
Резюме.....	124
Глава 6. Градиентный бустинг.....	125
Бустинг	125
Хорошее, плохое... и непонятное	126
Параметры	127
Энергоэффективность зданий: градиентный бустинг с настройками по умолчанию	128
Энергоэффективность зданий: настроенный градиентный бустинг	130
MNIST: градиентный бустинг с настройками по умолчанию	133
MNIST: настроенный градиентный бустинг	134
Футбол: градиентный бустинг с настройками по умолчанию	137
Футбол: настроенный градиентный бустинг.....	138
Резюме.....	140
Глава 7. Линейные модели	141
Параметры GLM.....	141
Данные об энергоэффективности: GLM с настройками по умолчанию.....	145
Данные об энергоэффективности: настроенная GLM	147
MNIST: GLM с настройками по умолчанию	151
MNIST: настроенная GLM.....	153
Футбол: GLM с настройками по умолчанию	155
Футбол: настроенная GLM	156
Резюме.....	157
Глава 8. Глубокое обучение (нейронные сети).....	158
Что такое нейронные сети?.....	159
Количественные и категориальные переменные	160
Слои нейронной сети	161
Функции активации	163
Параметры	164
Регуляризация	164
Оценка качества	165
Энергоэффективность зданий: модель глубокого обучения с настройками по умолчанию	168
Энергоэффективность зданий: настроенная модель глубокого обучения.....	168

MNIST: модель глубокого обучения с настройками по умолчанию.....	174
MNIST: настроенная модель глубокого обучения	175
Футбол: модель глубокого обучения с настройками по умолчанию	179
Футбол: настроенная модель глубокого обучения	180
Резюме	185
Приложение: дополнительные параметры	185
Глава 9. Обучение на неразмеченных данных.....	187
Кластеризация методом k-средних.....	188
Автокодировщики	191
Вложенные автокодировщики.....	193
Метод главных компонент.....	194
GLRM.....	196
Пропущенные данные.....	196
GLRM.....	200
Избавляемся от R	200
Резюме	203
Глава 10. Все остальное.....	204
Документация.....	204
Установка актуальной версии	204
Сборка из исходных кодов	204
Запуск из командной строки	205
Кластеры.....	205
EC2	206
Другие облачные провайдеры	206
Hadoop	207
Spark / Sparkling Water	207
Наивный байесовский классификатор.....	207
Ансамбли.....	208
Стекинг: h2o.ensemble.....	208
Ансамбли для классификации.....	210
Резюме.....	210
Глава 11. Эпилог	211
Результаты для данных об энергоэффективности	211
Результаты для набора данных MNIST.....	213
Результаты для данных о футбольных матчах	214
Как далеко вы готовы зайти.....	216
Чем больше, тем лучше	217
Еще больше данных.....	218
Отбор сложных примеров.....	219
Автокодировщик	219
Сверточные сети	220
Ансамбли.....	221
Результаты.....	222
Резюме.....	223

Приложение 1. Deep Water	224
Установка.....	224
Сборка из исходных кодов	224
Amazon Machine Image	224
Образ Docker	224
Примеры данных	224
Обзор библиотеки Deep Water	224
Глубокое обучение в библиотеке H2O	225
Современные тенденции в глубоком обучении.....	225
Почему нужно использовать Deep Water	225
Начало работы: набор данных MNIST.....	226
Бекенды	227
CPU и GPU.....	227
Классификация изображений.....	229
Данные	229
Параметры изображений.....	229
Предварительно созданные архитектуры	229
Архитектуры, создаваемые пользователем.....	230
Предварительно обученные нейросети	230
Веб-интерфейс Flow	230
Поиск по сетке	233
Полный перебор	233
Случайный поиск.....	234
Контрольные точки	235
Ансамбли.....	237
Признаки скрытых слоев и меры сходства.....	238
Поддержка нескольких GPU.....	239
Развертывание моделей.....	240
МОJO.....	240
Prediction Service Builder	240
 Приложение 2. Ансамбли (стекинг моделей).....	 241
Вступление	241
Стекинг / Super Learner	241
Алгоритм.....	242
Вложенные ансамбли в библиотеке H2O.....	242
Пример	243
На языке R.....	243
На языке Python.....	245
Вопросы и ответы	247
Дополнительная информация	248
Список литературы.....	248
 Краткий предметный указатель.....	 249

Предисловие

Похоже, что машинное обучение наконец-то достигло стадии зрелости. У него было длинное детство, охватывающее период с 1950-х гг., когда были созданы первые программы (для игры в шашки), обучающиеся на основе опыта, а также первые нейронные сети. Исследователи в области искусственного интеллекта так часто обещали скорый прорыв в этой сфере, что мы давно перестали их слушать. Но, возможно, они были на верном пути все это время, и нужно было лишь увеличить вычислительные мощности еще на порядок, или же немного доработать алгоритмы, чтобы перейти к их полезному и продуктивному использованию.

В начале 90-х нейронные сети называли новым прорывом в области искусственного интеллекта. Я пытался применить их для игры в го, но они работали очень плохо даже по сравнению с результатами (также весьма посредственными), которые были получены инженерными методами на основе знания предметной области с использованием усеченных деревьев поиска. Также они плохо масштабировались. И когда 20 лет спустя я услышал о потрясающих успехах глубокого обучения в этой игре, я был смущен отличиями от нейросетей, от использования которых я тогда отказался. «Недостаточно много» – вот ответ; иногда нужно предоставить алгоритму больше вычислительных ресурсов (в данном случае на 5–6 порядков больше), чтобы он раскрыл свой потенциал.

H2O – это программное обеспечение для машинного обучения и анализа данных. Я начал использовать данную библиотеку из-за интереса к возможностям глубокого обучения (хотя в ней реализованы и другие алгоритмы) и сразу же был впечатлен. Ее особенности:

- является открытым (open source) продуктом, распространяется на условиях лицензии Apache;
- проста в использовании;
- масштабируется для работы с большими данными;
- хорошо документирована и имеет коммерческую поддержку;
- актуальной является уже третья версия, что говорит о зрелой архитектуре;
- поддержка различных операционных систем и языков программирования.

При разработке H2O целью было обеспечение эффективной работы при масштабировании на больших объемах данных.

В этой книге рассматривается практический аспект использования библиотеки H2O с минимумом математики и теоретических основ применяемых методов.

Разумеется, библиотека H2O не лишена недостатков, и вот некоторые из них. В ней нет поддержки вычислений на GPU, что, в частности, могло бы повысить эффективность работы алгоритмов глубокого обучения¹. Кластеры предназначены для работы с большими объемами данных, но для обработки относительно небольших выборок со сложной структурой вам может понадобиться один многоядерный сервер. Для кластеров не реализована поддержка высокой доступности. Библиотека H2O написана на языке Java и хорошо оптимизирована, но теоретически реализации на C++ могут работать быстрее. Не реализован алгоритм SVM (ма-

¹ В разработке находится проект DeepWater, который добавляет поддержку других библиотек для глубокого обучения, использующих GPU.

шины опорных векторов). Наконец, необходимость обеспечивать совместимость с несколькими платформами замедляет разработку.

Кто и зачем использует H2O?

Многие известные компании (www.h2o.ai/customers/) используют H2O для обработки своих данных, и сейчас их количество превышает 5000. В самой компании H2O.ai работает свыше 80 сотрудников, большая часть – разработчики.

Эта статистика может впечатлить вашего руководителя, но не вас самих как программистов. Для разработчиков на языках R или Python, которые уже используют необходимые им библиотеки машинного обучения, H2O предоставляет простой и эффективный способ работы с объемами данных, слишком большими для обработки на одном сервере. Алгоритмы, реализованные в H2O, работают значительно быстрее, по сравнению со SparkML (хотя в этой библиотеке реализовано больше методов, чем в H2O). Кроме того, продуманный выбор значений по умолчанию для параметров моделей позволяет писать компактный и понятный код. Одной из целей этой книги было научить вас настраивать модели, но, как мы увидим, иногда не получается улучшить результат, который дают настройки по умолчанию.

Для кого эта книга

Чтобы ограничить объем книги парой сотен страниц, я позволил себе некоторые вольности. Подразумевается, что вы знаете R или Python. Продвинутые возможности языков не используются, поэтому знаний других языков должно хватить для понимания примеров, но сами примеры приводятся только на R или на Python. Пользователи Python выиграют от знакомства с библиотекой pandas, и не в последнюю очередь благодаря тому, что она облегчает работу с данными.

Также я предполагаю наличие у вас определенной ментальной гибкости: чтобы не повторять каждый пример дважды, я предположил, что пользователи R смогут понять примеры на Python, и наоборот. Полезно ознакомиться с презентацией: <http://bit.ly/2gl4GRx> (в том числе пользователям R).

У вас должен быть определенный опыт выполнения операций с данными, хотя бы при помощи электронных таблиц или SQL. Также вы должны иметь общее представление о машинном обучении и искусственном интеллекте, об их применении в жизни. Возможно, вы читаете данную книгу, потому что хотите участвовать в этом процессе и желаете, чтобы происходящие изменения были этичными и благоприятными для всех. Если это так, я приветствую такой подход.

Также я рассчитываю, что вы немного знаете статистику. Ничего страшного знать не потребуется – название книги содержит слово «практическое» неспроста, и изложение теоретических основ алгоритмов машинного обучения сведено к минимуму, достаточному для настройки моделей, но недостаточному для самостоятельной реализации алгоритмов с нуля. Используйте Википедию или поисковые системы, чтобы узнать больше. Но вы точно должны знать, чем медиана отличается от среднего и от моды, а также что такое стандартное отклонение и нормальное распределение.

Более того, я надеюсь, что вы понимаете: статистика может вводить в заблуждение (хорошая иллюстрация: <https://xkcd.com/882/>), а модели машинного обучения могут переобучаться (<https://en.wikipedia.org/wiki/Overfitting>).

Также нужно уточнить, что подразумевается под «моим ПК», который часто упоминается в контексте времени работы алгоритмов. Это ноутбук средней мощности с 8 Гб ОЗУ и 4 ядрами (8 потоков). Этого достаточно для запуска примеров из книги; на самом деле должно хватить и 4 Гб ОЗУ. Тем не менее иногда я «жульничал» и запускал обработку на серверах облачных провайдеров. Это сделано исключительно из рациональных соображений: не хотелось ждать по 24 часа до получения результатов, прежде чем написать о них.

ПРИНЯТЫЕ ОБОЗНАЧЕНИЯ

В книге использованы следующие типографские обозначения:

- *курсивом* выделены имена и расширения файлов, а также параметры функций при их перечислении в списках;
- моноширинный шрифт использован для программного кода и его элементов (имен переменных и функций и др.), включенных в текст.



Предупреждения появляются в этом поле.



Советы появляются тут.



Примечания появляются в этом поле.

БЛАГОДАРНОСТИ

Прежде всего выражаю огромную благодарность техническим рецензентам: избитая фраза о том, что эта книга стала лучше благодаря вам, является правдой. Другая избитая фраза – все ошибки в книге остаются на моей совести. Итак, хочу сказать спасибо Кэтрин Джармул (Katharine Jarmul), Юлин Джуан (Yulin Zhuang), Хьюго Матиен (Hugo Mathien), Эрин ЛеДелл (Erin LeDell) и Тому Кралевичу (Tom Kraljevic); извините, если не принял каких-то из ваших предложений. Помимо Эрин и Тома, множество других людей из H2O.ai помогло мне своими ответами на вопросы, поэтому я благодарю Арно Кандела (Arno Candell), Томаса Никодима (Tomas Nykodym), Михала Курка (Michal Kurka), Навдипа Джилл (Navdeep Gill), СриСатиша Амбати (SriSatish Ambati), Лорен ДиПерна (Lauren DiPerna) и всех, кого я забыл упомянуть (извините!).

Спасибо Николь Таше (Nicole Tache) за то, что была редактором в течение первой половины процесса создания книги, и Дабби Хардин (Debbie Hardin) за то, что ее заменила (Николь решила, что единственный способ покинуть проект – это завести ребенка). Спасибо им обоим за их спокойствие, когда я так увлекался созданием моделей, что забывал про необходимость соблюдать сроки.

Спасибо моей семье за то, что отнеслись с пониманием к тому, сколько времени я вложил в эту книгу.

Наконец, спасибо всем остальным: людям, которые отвечают на вопросы на StackOverflow, размещают сообщения в блогах и создают видеокурсы, пишут книги, пополняют Википедию. Они помогли мне заполнить пробелы в моих знаниях. Не стесняйтесь сообщать мне об ошибках, найденных в книге, и о том, как ее можно улучшить.

Глава 1

Установка и начало работы

Вас очень обрадует простота установки библиотеки H2O. Сначала я покажу, как ее установить при помощи R, используя CRAN, а затем – как это сделать при помощи Python, используя pip¹.

После этого мы погрузимся в наш первый проект по машинному обучению, который включает загрузку данных, создание модели и предсказаний на ее основе, а также оценку наших успехов. К этому моменту вы сможете похвастаться, что вы в некотором роде являетесь экспертом по глубокому обучению.

После обзора того, как влияние случайности может ввести нас в заблуждение, в этой главе более детально будет рассмотрен веб-интерфейс Flow, который поставляется вместе с H2O.

ПОДГОТОВКА К УСТАНОВКЕ

В этой книге будут примеры и на языке R, и на языке Python, поэтому у вас должен быть установлен хотя бы один из них². Вам также будет нужна платформа Java. Если есть выбор, я рекомендую всегда использовать 64-битные версии, в том числе 64-версию ОС (на странице загрузки 64-битная версия часто обозначена как «x64», а 32-битная – как «x86»).

Вы можете спросить, имеет ли значение выбор R или Python. Нет, и вскоре будет объяснено, почему. Также не существует преимуществ по производительности в случае использования скриптов вместо более дружественных к пользователю графических сред, таких как Jupiter или RStudio.

Установка R

В ОС Linux менеджер пакетов вашего дистрибутива делает установку очень простой: `sudo apt-get install r-base` для Debian/Ubuntu/Mint/... или `sudo yum install R` для Red-Hat/Fedora/Centos/...

Пользователи Mac должны ознакомиться с <https://cran.r-project.org/bin/macosx/> и следовать данной инструкции.

¹ В главе 10 показаны некоторые альтернативные способы установки H2O (см. «Установка актуальной версии»), включая сборку из исходников. Вам это может понадобиться, если вы столкнулись с багом, который был исправлен только в последней версии для разработчиков, или если вы хотите начать самостоятельно вносить изменения в H2O.

² Здесь и далее под установкой языка подразумевается установка соответствующего интерпретатора. – *Прим. перев.*

При использовании ОС Windows скачайте и запустите установочный файл с <http://cran.rstudio.com/bin/windows/>, после чего следуйте инструкциям по установке. На странице **Выбор компонентов (Select Components)** предлагается установка 32- и 64-битных версий; я выбираю установку только 64-битной, но можно установить и обе.

Необязательным следующим шагом после установки R является установка RStudio; вы *можете* делать все необходимое для запуска H2O из командной строки, но RStudio значительно упрощает работу (особенно в Windows, где командная строка осталась на уровне 1995 года). Скачать и установить RStudio можно по адресу <https://www.rstudio.com/products/rstudio/download/>.


Установка Python

H2O одинаково хорошо работает и с Python 2.7, и с Python 3.5, примеры из книги также должны работать. Если вы используете более раннюю версию Python, возможно, потребуется выполнить ее обновление. Также будет нужен pip – менеджер пакетов для Python. Для Red-Hat/Fedora/Centos/... точный вид команды может быть различным, поэтому следуйте инструкциям по адресу https://packaging.python.org/en/latest/install_requirements_linux/.

В ОС Linux нужно выполнить команду `sudo apt-get install python-pip` для Debian/Ubuntu/Mint/... или `sudo apt-get install python3-pip` в случае использования Python3.

Для установки на Mac следуйте руководству «Использование Python на Macintosh» (<http://bit.ly/2gn4HF5>).

При использовании ОС Windows следуйте аналогичному руководству «Использование Python в Windows» (<http://bit.ly/1RCJ7VR>). Не забудьте выбрать 64-битную версию (если только вы не пользуетесь до сих пор 32-битной версией Windows, разумеется).

 Вы также можете захотеть испытать дистрибутив Anaconda (<https://www.continuum.io/downloads>), который содержит практически все пакеты, необходимые специалисту по анализу данных. Дополнительным преимуществом является возможность установки обычным пользователем, что будет полезным при отсутствии у вас root-привилегий. Доступны версии для Linux, Mac и Windows.

Конфиденциальность

H2O содержит код¹ для вызова сервиса Google Analytics при каждом запуске. Происходит анонимный сбор данных, поскольку отслеживается всего лишь используемая версия, но если это вас беспокоит или нарушает корпоративную политику безопасности, то создание пустого файла `.h2o_no_collect` в вашей домашней папке (`C:\Users\YourName\` в Windows) остановит данный процесс. Вы будете знать, что это сработало, увидев сообщение «Opted out of sending usage metrics.» в логах. Другой способ отключения сбора информации представлен в разделе «Запуск из командной строки» главы 10.

¹ <http://bit.ly/2f96Hyu> по состоянию на июнь 2016-го.

Установка Java

У вас должна быть установлена платформа Java, которую вы можете получить на странице загрузки <http://bit.ly/16mhlmY>. Выберите вариант JDK¹. Если вы думаете, что у вас уже есть Java JDK, но не уверены в этом, то можете двигаться дальше и установить H2O, а затем при необходимости вернуться к этому пункту и (пере)установить Java.

УСТАНОВКА H2O ПРИ ПОМОЩИ R (CRAN)

Если вы не используете R, можете перейти к разделу «Установка H2O при помощи Python (pip)».

Запустите R и введите команду `install.packages("h2o")`. Да, когда я говорил о простоте установки, я не шутил. Эта команда также установит все зависимости.

Если вы впервые используете CRAN², нужно будет выбрать используемое зеркало для скачивания (лучше всего в месте, ближайшем к вам, но можно выбрать и такое, куда бы вы хотели поехать и сделать селфи).

Если хотите установить H2O для всех пользователей на данном компьютере, запустите R с правами суперпользователя командой `sudo R`, затем введите команду `install.packages("h2o")`.

Давайте проверим, все ли работает, набрав команду `library(h2o)`. Если нет никаких тревожных сообщений, переходим к следующему шагу: `h2o.init()`. При должном везении вы увидите большое количество информации о запуске H2O, а также о вашем кластере, подобно представленной на рис. 1.1. В противном случае сообщение об ошибке укажет на отсутствующие зависимости или другие проблемы.

Версия H2O на CRAN может отставать на месяц или два от самой свежей и самой лучшей. Но не беспокойтесь об этом до тех пор, пока не столкнетесь с багом, исправленным в этой новейшей версии.

`h2o.init()` по умолчанию будет использовать только два ядра на вашем ПК и, возможно, четверть оперативной памяти³. Используйте команду `h2o.shutdown()`, которая делает именно то, о чем вы догадались. Затем запустите H2O снова, но с использованием всех доступных ядер: `h2o.init(nthreads = -1)`. Чтобы использовать все ядра и, скажем, 4 Гб ОЗУ: `h2o.init(nthreads = -1, max_mem_size = "4g")`.

¹ Варианты «Server JRE» или «JRE» могут работать с H2O, но я рекомендую всегда устанавливать JDK.

² CRAN является репозиторием пакетов для R. Дополнительную информацию можно найти на <https://cran.r-project.org/>.

³ См. <http://bit.ly/2gn5h6e>, чтобы узнать, как выполнить запрос и определить значения по умолчанию для вашей системы.

```

> h2o.init()

H2O is not running yet, starting it now...

Note: In case of errors look at the following log files:
      /tmp/Rtmp6btQEF/h2o_darren_started_from_r.out
      /tmp/Rtmp6btQEF/h2o_darren_started_from_r.err

java version "1.7.0_101"
OpenJDK Runtime Environment (IcedTea 2.6.6) (7u101-2.6.6-0ubuntu0.14.04.1)
OpenJDK 64-Bit Server VM (build 24.95-b01, mixed mode)

Starting H2O JVM and connecting: ..... Connection successful!

R is connected to the H2O cluster:
H2O cluster uptime:      3 seconds 535 milliseconds
H2O cluster version:    3.8.2.2
H2O cluster name:       H2O_started_from_R_darren_rge683
H2O cluster total nodes: 1
H2O cluster total memory: 1.71 GB
H2O cluster total cores: 8
H2O cluster allowed cores: 2
H2O cluster healthy:    TRUE
H2O Connection ip:      localhost
H2O Connection port:    54321
H2O Connection proxy:   NA
R Version:               R version 3.2.5 (2016-04-14)

Note: As started, H2O is limited to the CRAN default of 2 CPUs.
      Shut down and restart H2O as shown below to use all your CPUs.
      > h2o.shutdown()
      > h2o.init(nthreads = -1)

> |

```

Рис. 1.1 ❖ Запуск `h2o.init()` (в R)

УСТАНОВКА H2O ПРИ ПОМОЩИ PYTHON (PIP)

Если вы не используете Python, можете перейти к разделу «Наша первая задача машинного обучения».

Введите в командной строке `pip install -U h2o`. Все готово! Это невероятно просто.

Опция `-U` говорит о том, что нужно также обновить все зависимости. В Linux вам, вероятно, нужно будет обладать правами суперпользователя, поэтому используйте команду `sudo pip install -U h2o`. Или же можно выполнить установку для отдельного пользователя: `pip install -U -user h2o`.

Для проверки запустите Python, напечатайте `import h2o`, а затем, при отсутствии предостережений, `h2o.init()`. Будет выведена некоторая информация с таблицей в конце, где среди прочего будут указаны количество узлов (nodes), общее количество памяти и количество доступных ядер, как показано на рис. 1.2¹. При написании сообщения о баге не забудьте указать всю информацию из данной таблицы.

¹ Возможно, вы видите какие-то сообщения с предостережениями? Для данного скриншота они были отключены при помощи `import warnings; warnings.filterwarnings("ignore", category = DeprecationWarning)`, но обычно я просто игнорирую их.

```

In [3]: import h2o

In [4]: h2o.init()
Checking whether there is an H2O instance running at http://localhost:54321.... not found.
Attempting to start a local H2O server...
  Java Version: java version "1.7.0_111"; OpenJDK Runtime Environment (IcedTea 2.6.7) (7u111-
ild 24.111-b01, mixed mode)
  Starting server from /usr/local/h2o_jar/h2o.jar
  Ice root: /tmp/tmpsNCJ_v
  JVM stdout: /tmp/tmpsNCJ_v/h2o_darren_started_from_python.out
  JVM stderr: /tmp/tmpsNCJ_v/h2o_darren_started_from_python.err
  Server is running at http://127.0.0.1:54321
Connecting to H2O server at http://127.0.0.1:54321... successful.
-----
H2O cluster uptime:          02 secs
H2O cluster version:        3.10.0.7
H2O cluster version age:    5 days
H2O cluster name:           H2O_from_python_darren_a9x6hb
H2O cluster total nodes:    1
H2O cluster free memory:    1.710 Gb
H2O cluster total cores:    8
H2O cluster allowed cores:  8
H2O cluster status:         accepting new members, healthy
H2O connection url:         http://127.0.0.1:54321
H2O connection proxy:
Python version:              2.7.6 final
-----

```


Рис. 1.2 ❖ Запуск `h2o.init()` (в Python)

По умолчанию ваш экземпляр (instance) H2O может использовать все доступные ядра и (как правило) 25% оперативной памяти. Обычно это подходит для работы, но что, если вы хотите использовать ровно 4 Гб памяти и лишь два из восьми ядер? Сначала остановите H2O при помощи команды `h2o.shutdown()`, затем введите `h2o.init(nthreads=2, max_mem_size=4)`. Ниже приведена выдержка из таблицы, подтверждающая, что команда работает:

```

...
H2O cluster total free memory:
H2O cluster total cores:
H2O cluster allowed cores:
...

```

 Виртуальные окружения `virtualenv` не работают с H2O¹. Точнее, установка происходит, но запуск H2O невозможен. Если вы действительно хотите установить H2O таким способом, следуйте инструкциям по запуску H2O из командной строки в главе 10. В таком случае `h2o.init()` и все остальное будет работать.

НАША ПЕРВАЯ ЗАДАЧА МАШИННОГО ОБУЧЕНИЯ

Теперь, когда установлено все необходимое, давайте займемся делом. API для Python и R настолько подобны, что в этом примере мы будем рассматривать их параллельно. Если вы используете Python, изучите программный код 1.1, а если R – программный код 1.2. Они повторяют операции `import/library` и `h2o.init`, которые мы выполнили ранее. Не волнуйтесь, это не навредит.

¹ По крайней мере, для версии 3.10.x.x и более ранних.

Я подробно все объясню на следующих нескольких страницах, но хочу подчеркнуть, что это полный скрипт: он загружает данные, подготавливает их, создает модель – многослойную нейронную сеть (т. е. это глубокое обучение), которая сопоставима с наилучшей на данный момент для этого набора данных, и делает предсказания с ее помощью.

Набор данных Iris

Если вы раньше не слышали о наборе данных Iris, то это, должно быть, ваша первая книга по машинному обучению! Набор содержит 150 наблюдений для растений рода Ирис с результатами четырех измерений (длина и ширина чашелистиков и лепестков) и с указанием принадлежности к тому или иному виду. Представлены три вида, по 50 наблюдений для каждого из них.

Это очень популярный набор данных для машинного обучения, поскольку он достаточно маленький, чтобы обучение модели происходило быстро и его было удобно анализировать графически, и достаточно большой, чтобы быть интересным. Для него также не существует тривиального решения: ни одна из четырех переменных не позволяет идеально разделить классы.

Программный код 1.1 ❖ Глубокое обучение в Python с набором данных Iris

```
import h2o
h2o.init()

datasets = "https://raw.githubusercontent.com/DarrenCook/h2o/bk/datasets/"
data = h2o.import_file(datasets + "iris_wheader.csv") ❶
y = "class" ❷
x = data.names
x.remove(y)
train, test = data.split_frame([0.8]) ❸

m = h2o.estimators.deeplearning.H2ODeepLearningEstimator() ❹
m.train(x, y, train)
p = m.predict(test) ❺
```

Программный код 1.2 ❖ Глубокое обучение в R с набором данных Iris

```
library(h2o)
h2o.init(nthreads = -1)

datasets <- "https://raw.githubusercontent.com/DarrenCook/h2o/bk/datasets/"
data <- h2o.importFile(paste0(datasets, "iris_wheader.csv")) ❶
y <- "class" ❷
x <- setdiff(names(data), y)
parts <- h2o.splitFrame(data, 0.8) ❸
train <- parts[[1]]
test <- parts[[2]]

m <- h2o.deeplearning(x, y, train) ❹
p <- h2o.predict(m, test) ❺
```

Пункты ❶, ❷, ❸ отвечают за подготовку данных, ❹ – за обучение модели, ❺ – за использование модели.

❶ служит примером главной концепции, требующей понимания при использовании H2O: все данные находятся в кластере (сервере), а не на стороне клиента. *Даже если клиент и кластер являются одним и тем же компьютером.*

Следовательно, всякий раз, когда мы хотим обучить модель или сделать предсказание, мы должны разместить данные в кластере H2O; мы рассмотрим эту тему более детально в главе 2. Сейчас достаточно понимать, что эта строка создает таблицу с именем *iris_wheader.hex* в кластере. Эта команда распознает, что первая строка csv-файла содержит заголовки, которые автоматически используются в качестве имен столбцов. Она также понимает (проанализировав данные), что столбец *class* является категориальной переменной, а это значит, что мы будем решать задачу многоклассовой классификации, а не регрессии (см. врезку «Жаргон и соглашения»).

② определяет пару вспомогательных переменных: *y* будет именем целевой переменной, т. е. переменной, которую мы хотим научиться предсказывать, а *x* будет содержать имена переменных, на которых будет происходить обучение; в данном случае это означает все остальные переменные. Иными словами, мы будем пытаться использовать результаты четырех измерений (*sepal_len*, *sepal_wid*, *petal_len*, *petal_wid*) для предсказания, к какому виду принадлежит растение.

Жаргон и соглашения

Ваши данные представлены в виде *строк* (которые также называют *наблюдениями*) и *столбцов*. Они организованы в *таблицу* (*frame*, *data frame*)¹. Если вы знакомы с электронными таблицами или таблицами SQL, то таблицы в H2O – это примерно то же самое. Аналогом в R является объект класса *data.frame*, а в Python (библиотека *pandas*) – *DataFrame* (или словарь *dict* из списков *list* равной длины).

В H2O столбцы могут иметь следующие типы:

- *real* – число с плавающей точкой, то есть *numeric* в R, *float* в Python и *double* во многих других языках;
- *int* – целое число;
- *enum* – набор категорий или классов (*factor* в R или *categorical* в *pandas*);
- *time* – 64-битное целое число, миллисекунды от начала эры Unix (1 января 1970 г.). Может быть получено из разных форматов времени;
- *string* – текст. Практически все, что вы можете с ним сделать в H2O, – это конвертировать в тип *enum*; текстовые данные не могут непосредственно использоваться при построении моделей.

Решение об использовании типов *int* или *real* принимается системой H2O после анализа данных в столбце; вы можете лишь выбирать, использовать ли тип *enum* для числовых данных.

В случае *обучения с учителем* или *обучения на размеченных данных* (*supervised machine learning*) один из этих столбцов будет содержать то, что мы хотим научиться предсказывать. Такой столбец называют откликом, зависимой переменной, выходной переменной, правильным ответом и другими терминами. В этой книге я буду сохранять имя данного столбца в переменной *y*². В случае *обучения без учителя* или *обучения на неразмеченных данных* (*unsupervised machine learning*) переменная *y* не будет задана.

Некоторые из остальных или все остальные столбцы в таблице являются данными, на которых происходит обучение. Они называются по-разному: независимые пере-

¹ Здесь и далее термины *frame* и *data frame* переведены как «таблица», поскольку в русскоязычной литературе нет устоявшегося понятия «кадр данных» и речь всегда идет именно о таблицах. – *Прим. перев.*

² В API для R и Python она также называется *y*; в REST API ей соответствует *response_column*.

менные, признаки, входные данные, предикторы. Мы договоримся хранить список имен таких столбцов в переменной x ¹.

Еще несколько соглашений: все наши данные будут содержаться в переменной `data`; выборка, используемая для тренировки/обучения (`train`), будет находиться в переменной `train`; выборка для проверки (`validation`) – в переменной `valid`; выборка для тестирования (`test`) – в переменной `test`². И запомните, все эти переменные являются лишь указателями (`pointers`) для данных, которые на самом деле хранятся в кластере. В Python используется класс-обертка для указателя (`handle`), где также хранятся некоторые итоговые статистики; в R используется тот же принцип с реализацией в виде окружения (`environment`).

Я использовал короткие имена: это книга, и переносы в программном коде выглядят некрасиво; некоторые могут даже читать ее с телефона. Я рекомендую вам применять в своем коде осмысленные имена, то есть `premierLeagueScores2005_2015_train` вместо `train`. Когда ваш скрипт занимает тысячи строк и вы имеете дело с десятками наборов данных, это сэкономит ваше психическое здоровье.

❷ (разбивка на обучающую и тестовую выборки) является еще одним принципом, суть которого состоит в борьбе с переобучением (`overfitting`). В общих чертах: мы случайным образом отбираем 80% наших данных для обучения, а затем пытаемся использовать модель на оставшихся 20%, чтобы увидеть, насколько она хороша. В реально работающей системе эти 20% представляют садовников, которые приходят с новыми растениями и спрашивают, к какому виду они относятся.

Напомню, что в Python разбивка данных выглядит следующим образом. `split_frame()` является одной из функций (методов) класса `H2OFrame`. Указание `[0.8]` приводит к тому, что в первую часть попадают 80% данных, а остальные данные – во вторую:

```
train, test = data.split_frame([0.8])
```

В R функция `h2o.splitFrame()` принимает таблицу и возвращает список, элементы которого для удобства присвоены переменным `train` и `test`:

```
parts <- h2o.splitFrame(data, 0.8)
train <- parts[[1]]
test <- parts[[2]]
```

Разделение, которое происходит случайным образом для каждой строки, приводит к пропорции примерно 120/30 строк, но вы можете получить чуть больше строк в обучающей или тестовой выборке³.

Давайте подведем итоги. Как показано на рис. 1.3, клиент хранит лишь указатели для данных, находящихся в кластере H2O.

¹ Эта переменная во всех API называется x , но если вы когда-нибудь решите работать с REST API, то найдете там дополнение к x : список имен столбцов, которые *не нужно* использовать, с именем `ignored_columns`.

² Если вы незнакомы с этими концепциями, в главе 3 объясняются различия между проверочной и тестовой выборками, а также между проверкой и перекрестной проверкой (`cross-validation`).

³ Поэкспериментируйте с аргументом `seed` для получения точной разбивки. Например, мне повезло с `h2o.splitFrame(data, 0.8, seed = 99)`. В Python: `data.split_frame([0.8], seed = 99)`.

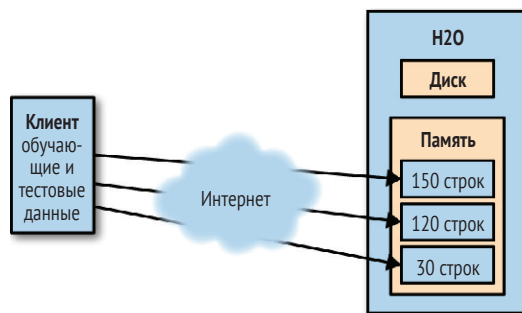


Рис. 1.3 ❖ Местоположение данных

Разумеется, наш «кластер» находится на локальном компьютере, то есть там же, где и клиент, в одной и той же оперативной памяти. Но вы должны думать о них как о находящихся на противоположных сторонах земного шара. Также представьте, что данные могут содержать миллиарды строк, слишком много, чтобы они могли поместиться в памяти клиента. Путем добавления компьютеров в кластер, если его общий объем памяти будет достаточно большим, можно обеспечить возможность загрузки и анализа таких данных клиентом, которым может выступать обычный ноутбук.

Обучение и предсказания в Python

Наконец, мы добрались до пункта ④ – машинного обучения. В Python процесс состоит из двух стадий:

1. Создадим объект для вашего алгоритма машинного обучения, задав при необходимости параметры для него:

```
m = h2o.estimators.deeplearning.H2ODeepLearningEstimator()
```

2. Запустим обучение, указав, какие данные использовать:

```
m.train(x, y, train)
```

Если вы предпочитаете стиль библиотеки `scikit-learn`, можете вместо этого использовать следующие команды:

```
from h2o.estimators.deeplearning import H2ODeepLearningEstimator
m = H2ODeepLearningEstimator()
m.train(x, y, train)
```

Поскольку мы не задаем никаких параметров конструктора, модель строится со значениями по умолчанию. Это значит (помимо прочего), что обучается нейросеть с двумя скрытыми слоями по 200 нейронов каждый в течение 20 эпох. В главе 8 будут даны определения понятий «нейроны» и «эпохи», но сейчас не будем об этом. Важно отметить, что со значениями по умолчанию обучение проходит очень быстро, оно занимает всего несколько секунд для этого набора данных.

Так же, как и в случае с данными, `m` является экземпляром класса-обертки указателя для модели, хранящейся в кластере H2O. Если вы выведете `m` на экран, то увидите много информации о том, как проходило обучение. Также можно использовать методы класса, чтобы получить только ту информацию, которая вас инте-

ресует, – например, `m.mse()` сообщит мне, что MSE (mean squared error – средне-квадратичная ошибка) равна 0.01096. Здесь есть элемент случайности, так что ваше значение, вероятно, будет немного отличаться¹.

`m.confusion_matrix(train)` возвращает *матрицу несоответствий* (confusion matrix), которая содержит информацию о правильно и неправильно определенных классах. Ниже представлены результаты для 120 наблюдений обучающей выборки:

Iris-setosa	Iris-versicolor	Iris-virginica	Error	Rate
42	0	0	0	0 / 42
0	37	1	0.0263158	1 / 38
0	1	39	0.025	1 / 40
42	38	40	0.0166667	2 / 120

В данном случае я вижу, что все 42 экземпляра вида *setosa* были определены безошибочно, но один экземпляр вида *versicolor* был определен как *virginica*, и один экземпляр *virginica* – как *versicolor*.

Последняя строка программного кода – **5** `p = m.predict(test)`. Она делает предсказания на основе модели и присваивает результат переменной `p`. Ниже приводятся несколько примеров этих предсказаний. В первом столбце указан предсказанный класс, в остальных – вероятности принадлежности к каждому из классов. Как вы можете видеть, модель уверена в своих ответах более чем на 99.5%:

predict	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	0.999016	0.000983921	1.90283E-019
Iris-setosa	0.998988	0.00101178	1.40209E-020
Iris-versicolor	5.22035E-005	0.997722	0.00222536
Iris-versicolor	0.000275126	0.995354	0.00437055

И здесь тоже таблица хранится в кластере H2O, поэтому вы видите лишь превью – первые 10 строк. Чтобы увидеть все 30 предсказаний, их нужно загрузить при помощи команды `p.as_data_frame()`. Если у вас не установлена библиотека `pandas`, вы увидите вложенный список:

```
[['predict', 'Iris-setosa', 'Iris-versicolor', 'Iris-virginica'],
 ['Iris-setosa', '0.9990160791818314', '9.83920818168421E-4',
 '1.9028267028039464E-19'], ['Iris-setosa', '0.9989882189908829', ...
 ..., ['Iris-virginica', '1.72617432126E-11', '1.0197263306598747E-4',
 '0.9998980273496721']]]
```

Вы можете анализировать данные в таком виде. Тем не менее API H2O для Python интегрирован с `pandas`. Если вы используете Python для работы с данными, то наверняка знакомы с этой библиотекой (а если нет, установите ее при помощи `pip install pandas`). При наличии установленной библиотеки `pandas` результат работы функции `p.as_data_frame()` будет таким:

¹ Я использовал `h2o.deeplearning(x, y, train, seed = 99, reproducible = TRUE)` для получения воспроизводимых значений.

```

      predict  Iris-setosa  Iris-versicolor  Iris-virginica
0  Iris-setosa  9.990161e-01      0.000984    1.902827e-19
1  Iris-setosa  9.989882e-01      0.001012    1.402089e-20
...

```

Что еще мы можем сделать? Этот вопрос детально рассматривается в главе 2. Мы можем получить долю правильных ответов при помощи `p["predict"] == test["class"]`.mean(). А команда `p["predict"].cbind(test["class"]).as_data_frame()` выведет результат в виде двух столбцов с предсказаниями и правильными метками классов:

```

      predict      class
0  Iris-setosa  Iris-setosa
1  Iris-setosa  Iris-setosa
...
11 Iris-versicolor Iris-versicolor
12 Iris-virginica Iris-versicolor
13 Iris-versicolor Iris-versicolor
14 Iris-virginica Iris-versicolor
15 Iris-versicolor Iris-versicolor
...

```

Обучение и предсказания в R

В R этап машинного обучения ❹ состоит из единственного вызова функции, в котором одновременно указываются параметры и используемые данные: `m <- h2o.deeplearning(x, y, train)` (на самом деле я запускал команду `m <- h2o.deeplearning(x, y, train, seed = 99, reproducible = TRUE)` для получения воспроизводимых результатов, но вам не стоит этого делать, потому что таким образом будет использоваться всего одно ядро, и это займет больше времени).

И вновь `m` является указателем для модели, хранящейся в кластере H2O. `h2o.mse(m)` сообщит мне, что MSE равна 0.01097. `h2o.confusionMatrix(m)` вернет матрицу несоответствий (по умолчанию – для обучающей выборки):

```

Confusion Matrix: vertical: actual; across: predicted
      setosa  versicolor  virginica  Error  Rate
setosa      42           0           0  0.0000 = 0 / 42
versicolor  0           37           1  0.0263 = 1 / 38
virginica   0           1           39  0.0250 = 1 / 40
Totals      42           38           40  0.0167 = 2 /120

```

Видим, что был получен идеальный результат для `setosa`, но по одному экземпляру вида `versicolor` и `versicolor` были определены ошибочно. Число в правой нижней ячейке соответствует общей доле ошибок – 1.67% (на обучающих данных).

Последняя строка программного кода – ❺ `p = h2o.predict(m, test)`. Она делает предсказания на основе модели. Как обычно, `p` является ссылкой на таблицу в кластере H2O. Если я выведу на экран `p`, то получу только первые шесть предсказаний. Чтобы увидеть их все, мне нужно загрузить данные. Будьте внимательны, работая с кластерами или с большими данными (Big Data™): вы должны сначала решить, как много данных вам нужно иметь на локальном компьютере, сколько времени уйдет на загрузку и справится ли вообще ваш компьютер с ними.

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru