ОБРАЗОВАТЕЛЬНАЯ МАНГА

МАШИННОЕ ОБУЧЕНИЕ

СОТРУДНИКУ ГОРОДСКОЙ АДМИНИСТРАЦИИ КИЁХАРА КАДЗУМА ПОРУЧЕНО ЗАДАНИЕ, КОТОРОЕ БЕЗ МАШИННОГО ОБУЧЕНИЯ НЕ ВЫПОЛНИТЬ. ПОД РУКОВОДСТВОМ СВОЕЙ ДАВНЕЙ ЗНАКОМОЙ МИЯНО САЯКА ОН ОСВАИВАЕТ ПРЕМУДРОСТИ РАБОТЫ С ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ – ОТ САМЫХ АЗОВ ДО ГЛУБОКОГО ОБУЧЕНИЯ.

ВМЕСТЕ С ГЕРОЯМИ МАНГИ ЧИТАТЕЛИ УЗНАЮТ О ТОМ, ЧТО ТАКОЕ РЕГРЕССИЯ И КАК ПРОВОДИТЬ КЛАССИФИКАЦИЮ, ОЗНАКОМЯТСЯ С ПРИНЦИПАМИ ОЦЕНКИ ТЕСТОВЫХ ДАННЫХ И ОСОБЕННОСТЯМИ РАБОТЫ НЕЙРОННЫХ СЕТЕЙ. В ЗАКЛЮЧИТЕЛЬНОЙ ЧАСТИ ИЗЛАГАЮТСЯ МЕТОДЫ ОБУЧЕНИЯ БЕЗ УЧИТЕЛЯ.

МАНГА ПРЕДНАЗНАЧЕНА ДЛЯ ТЕХ, КТО НАЧИНАЕТ ЗНАКОМСТВО С МАШИННЫМ ОБУЧЕНИЕМ И ОСВОИЛ МАТЕМАТИКУ НА УРОВНЕ ПЕРВЫХ КУРСОВ УНИВЕРСИТЕТА.

ISBN 978-5-97060-830-2

Интернет-магазин:
www.dmkpress.com

Оптовая продажа:
КТК «Галактика»
books@alians-kniga.ru

ДМК ИЗДАТЕЛЬСТВО
www.дмк.рф

ЗАНИМАТЕЛЬНАЯ МАНГА

МАШИННОЕ ОБУЧЕНИЕ

Араки Масахиро
Ватари Макана

ЗАНИМАТЕЛЬНАЯ МАНГА
МАШИННОЕ ОБУЧЕНИЕ

Араки Масахиро
Ватари Макана
Office SAWA, Ltd.

Ohmsha
ДМК ИЗДАТЕЛЬСТВО

Занимательное машинное обучение

Манга

マンガでわかる

機械学習

荒木　雅弘／著
渡　まかな／作画
ウェルテ／制作

Ohmsha

ОБРАЗОВАТЕЛЬНАЯ МАНГА

ЗАНИМАТЕЛЬНОЕ МАШИННОЕ ОБУЧЕНИЕ

Араки Масахиро

Художник Ватари Макана

Перевод
А. С. Слащевой

ДМК ПРЕСС
ИЗДАТЕЛЬСТВО

Москва
ДМК Пресс, 2020

УДК 004.4
ББК 32.972
А79

Араки М., Ватари М.

А79 Занимательная манга. Машинное обучение: манга / Араки Масахиро (автор), Ватари Макана (худ.); пер. с яп. А. С. Слащевой. — М.: ДМК Пресс, 2020. — 214 с. : ил. — (Серия «Образовательная манга»). — Доп. тит. л. яп.

ISBN 978-5-97060-830-2

Сотруднику городской администрации Киёхара Кадзума поручено задание, которое без машинного обучения не выполнить. Под руководством своей давней знакомой Мияно Саяка он осваивает премудрости работы с искусственным интеллектом – от самых азов до глубокого обучения.

Вместе с героями манги читатели узнают о том, что такое регрессия и как проводить классификацию, ознакомятся с принципами оценки тестовых данных и особенностями работы нейронных сетей. В заключительной части излагаются методы обучения без учителя.

Издание предназначено для тех, кто начинает знакомство с машинным обучением и освоил математику на уровне первых курсов университета.

УДК 004.4
ББК 32.972

Manga de Wakaru Kikai Gakushu (Manga Guide: Machine Learning)
By Araki Masaxiro (Author), Illustration by Vatari Makana
Published by Ohmsha, Ltd.
Russian language edition copyright © 2020 by DMK Press

Все права защищены. Никакая часть этого издания не может быть воспроизведена в любой форме или любыми средствами, электронными или механическими, включая фотографирование, ксерокопирование или иные средства копирования или сохранения информации, без письменного разрешения издательства.

ISBN 978-4-274-22244-3 (яп.)
ISBN 978-5-97060-830-2 (рус.)

Copyright © 2018 by and Office sawa, Ltd.
© Издание, перевод, ДМК Пресс, 2019

ПРЕДИСЛОВИЕ

В этой книге я представил несколько репрезентативных методов машинного обучения и попытался по возможности просто изложить их суть. Ее предполагаемая аудитория – те, кто только начинает знакомство с машинным обучением и уже владеет математикой на уровне первых курсов университета. Но если вы не дружите с математикой, то можете ознакомиться с разъяснениями в конце каждой главы и примерно понять, какие задачи решаются с помощью этих методов.

Особенность данной книги в том, что в начале каждой главы ставится задача, а затем постепенно объясняются методы машинного обучения, необходимые для ее решения. В таблице ниже перечислены задачи и методы, которые будут рассматриваться в каждой главе.

Глава	Задача	Метод
1	Прогноз количества участников мероприятия	Линейная регрессия
2	Определение вероятности заболевания диабетом	Логистическая регрессия, решающее дерево
3	Оценка результатов обучения	Метод проверки на резервированных данных, перекрестная проверка
4	Сортировка винограда	Сверточная нейронная сеть
5	Определение вероятности заболевания диабетом (повтор)	Ансамблевые методы
6	Рекомендация события	Кластерный анализ, матричное разложение

В каждой главе будет предложено лишь введение в тот или иной метод. Если вы хотите применить его на практике для решения какой-либо задачи, я советую обратиться к учебникам, которые указаны в списке рекомендованной литературы в конце книги.

Наконец, я благодарю всех сотрудников издательства Ohmsha за возможность написать эту книгу. Я также благодарен г-же Ватари Макана и всем сотрудникам Уильтэ, которые превратили мою рукопись в веселую мангу.

Июль 2018 года,
Араки Масахиро

СОДЕРЖАНИЕ

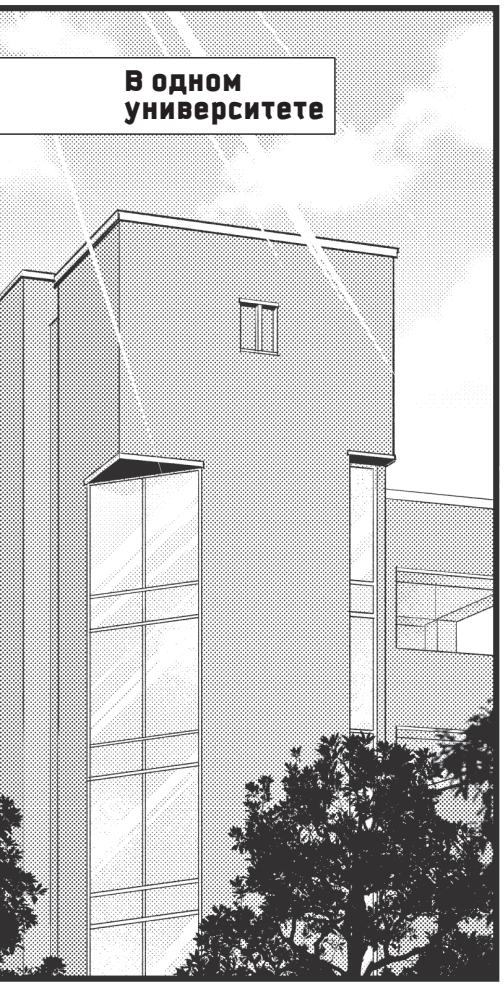
Глава 3
ОЦЕНКА РЕЗУЛЬТАТОВ 77

Глава 4
ГЛУБОКОЕ ОБУЧЕНИЕ 97

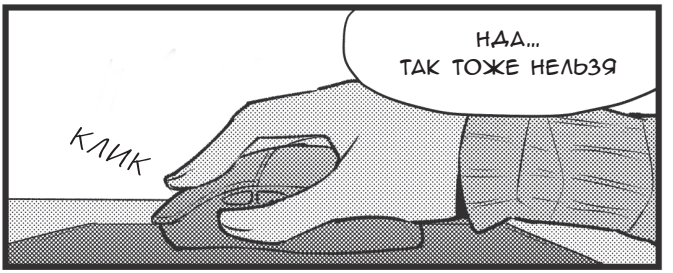
Глава 5
АНСАМБЛЕВЫЕ МЕТОДЫ 139

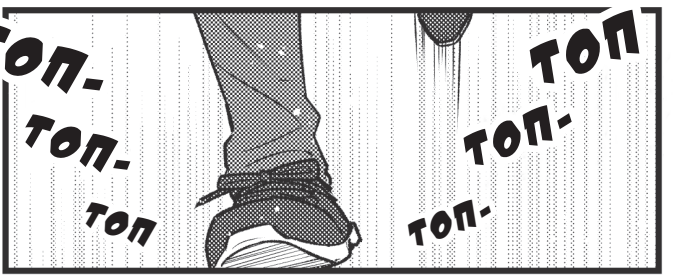
ПРОЛОГ

ПОГОВОРИМ О МАШИННОМ ОБУЧЕНИИ


В одном университете


топ-
топ-
топ
топ-
топ
топ-


ндд...
так тоже нельзя
клик

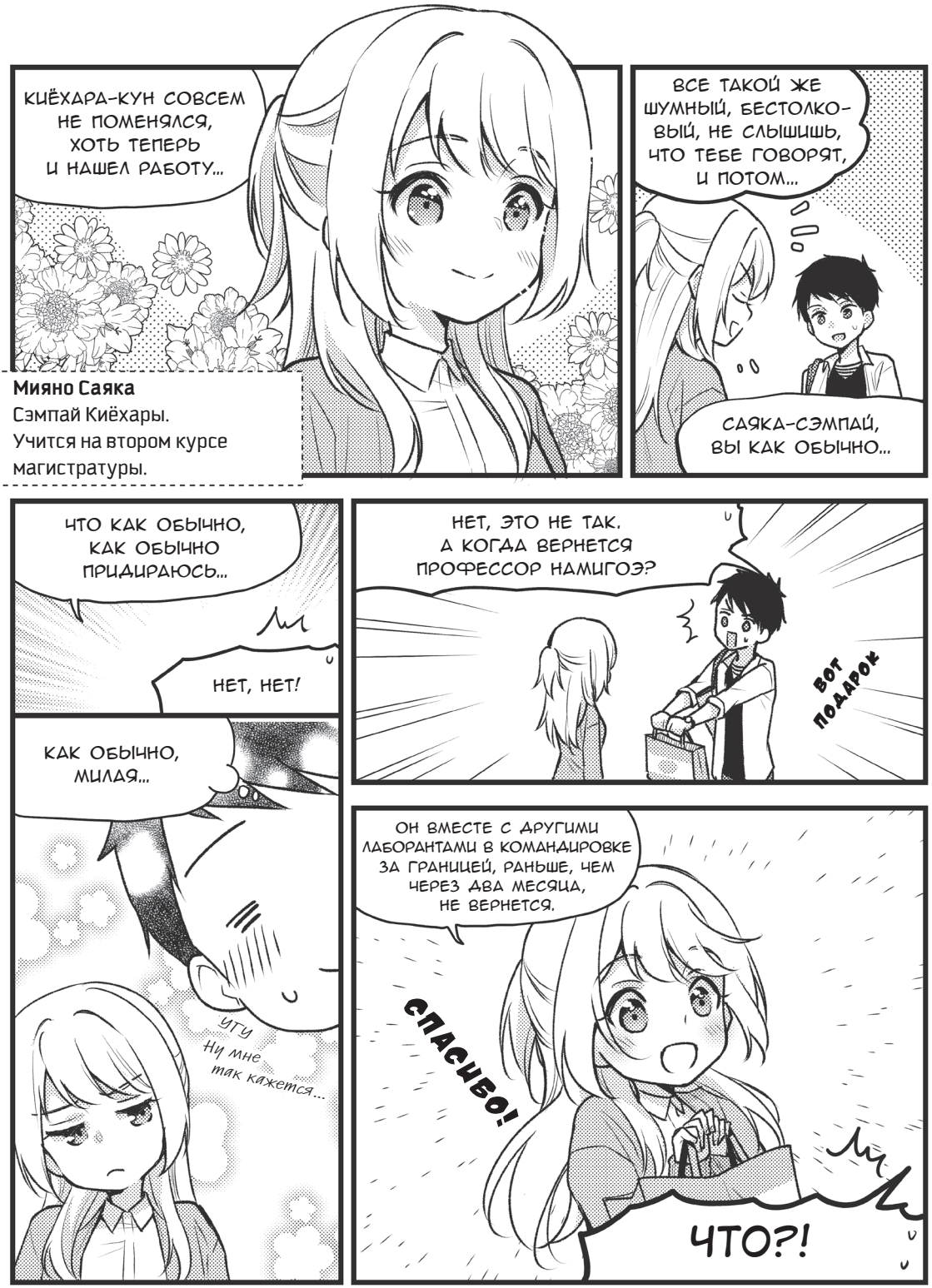

топ-
топ-
топ
топ
топ-
топ-


хм... неплохо бы найти и это,
но на компромисс идти я не хочу.


топ-
топ-
топ
профессор намигоэ?
бух!

Киёхара Кадзума
Вот уже год как работает в городской администрации.
Изучал машинное обучение в университете на факультете компьютерных технологий, но так ничему толком не выучился.

Сэмпай – название старших по курсу/классу учеников или студентов. Антоним – слово «кохай», которое обозначает младшего по курсу. – *Прим. перев.*



Мияно Саяка
Сэмпай Киёхары.
Учится на втором курсе магистратуры.


БЛИН, МНЕ ЭТО НИКАК НЕ ПОДХОДИТ!
ЧТО-ТО СЛУЧИЛОСЬ?
ТРЯСЕТСЯ
В УЖАСЕ
Я ХОТЕЛ ПОПРОСИТЬ ПРОФЕССОРА НАМИГОЭ РАССКАЗАТЬ МНЕ ПРО МАШИННОЕ ОБУЧЕНИЕ...
МАШИННОЕ ОБУЧЕНИЕ?
КИЁХАРА, ТЫ ЖЕ РАБОТАЕШЬ У СЕБЯ В МЕСТНОЙ АДМИНИСТРАЦИИ. ТЫ ЖЕ ГОВОРИЛ: "ХОЧУ СПОКОЙНО РАБОТАТЬ ДОМА И ЖИТЬ В СВОЕ УДОВОЛЬСТВИЕ". ЗАЧЕМ ТЕБЕ МАШИННОЕ ОБУЧЕНИЕ?
Киёхара-кун в университете
О-ХО-ХО...
Что? Работу искать? Хочу не слишком напряженно работать в хорошем месте
У ТЕБЯ ЧТО-ТО СТРЯСЛОСЬ? МОЖЕТ, МНЕ РАССКАЖЕШЬ?
.....


У НАС ЕСТЬ КОНСУЛЬТАНТ, КОТОРЫЙ ПОРУЧИЛ МНЕ ПОРАБОТАТЬ С ЕГО ИИ-ПРОГРАММОЙ, ПРОГНОЗИРУЮЩЕЙ КОЛИЧЕСТВО ГОСТЕЙ НА ПУБЛИЧНЫХ МЕРОПРИЯТИЯХ НАШЕЙ АДМИНИСТРАЦИИ.
ЭТОТ КОНСУЛЬТАНТ –
К МОМЕНТУ НАСТУПЛЕНИЯ СИНГУЛЯРНОСТИ ПОЯВИТСЯ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, МОЗГ, СРАВНИМЫЙ С ЧЕЛОВЕЧЕСКИМ ИЛИ ДАЖЕ ПРЕВОСХОДЯЩИЙ ЕГО, КОТОРЫЙ СДЕЛАЕТ ВСЮ ЧЕЛОВЕЧЕСКУЮ РАБОТУ НЕНУЖНОЙ!
ВООДУШЕВЛЕННО

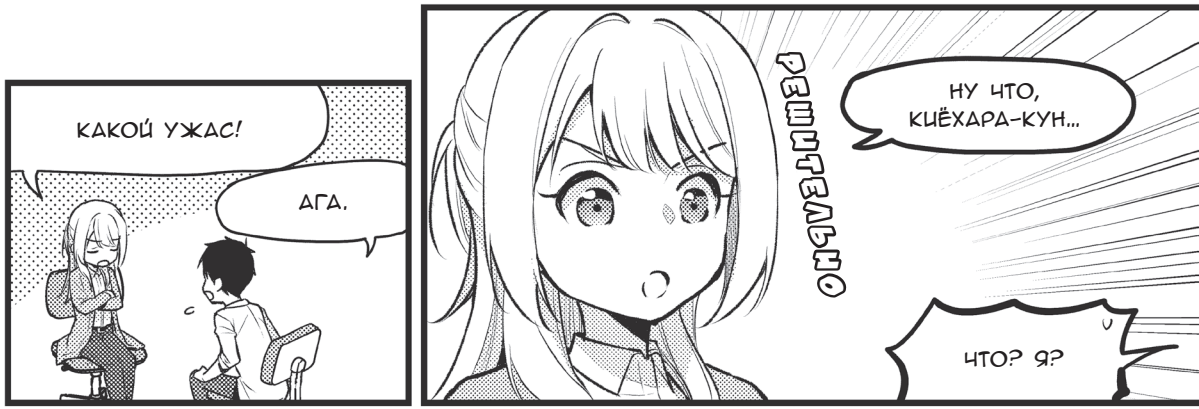

...ПРИМЕРНО ТАКОЙ ТИП... Я ПОДСЧИТАЛ КОЛИЧЕСТВО ГОСТЕЙ ЧЕРЕЗ ИИ-ПРОГРАММУ, КОТОРУЮ ОН РАЗРАБОТАЛ...
УФ!
СУДЯ ПО ПРОГНОЗАМ ИИ, КОЛИЧЕСТВО ГОСТЕЙ УМЕНЬШИТСЯ!
СПАСИБО!

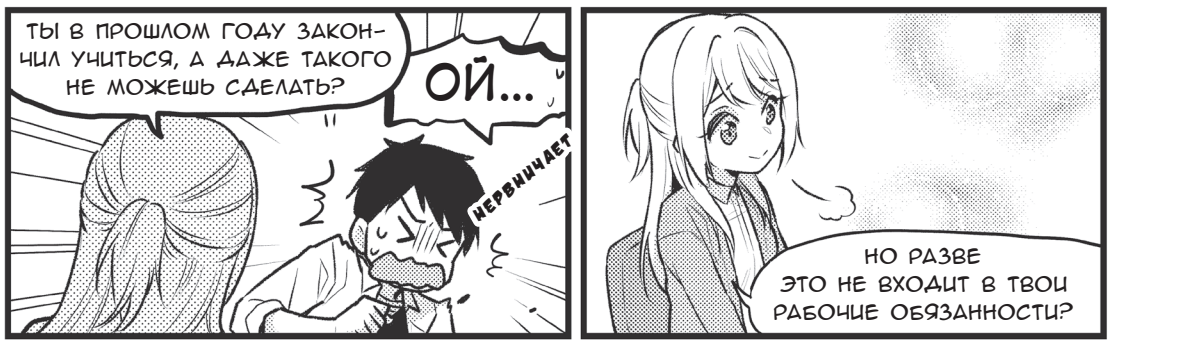

Количество гостей
2 года назад
1 год назад
Этот год (прогноз)
ЦИФРЫ БЫЛИ ПОДОЗРИТЕЛЬНЫМИ, И КОГДА Я ИХ ПРОВЕРИЛ...
ОЙ!
...ТО УВИДЕЛ, ЧТО КОЛИЧЕСТВО ГОСТЕЙ ДВА ПОСЛЕДНИХ ГОДА СНИЖАЕТСЯ ПО ПРЯМОЙ!


Я ДОЛОЖИЛ ОБ ЭТОМ ОТВЕТСТВЕННОМУ ЗА РЕКЛАМУ, НО ТОТ НЕ ОБРАТИЛ ВНИМАНИЯ...

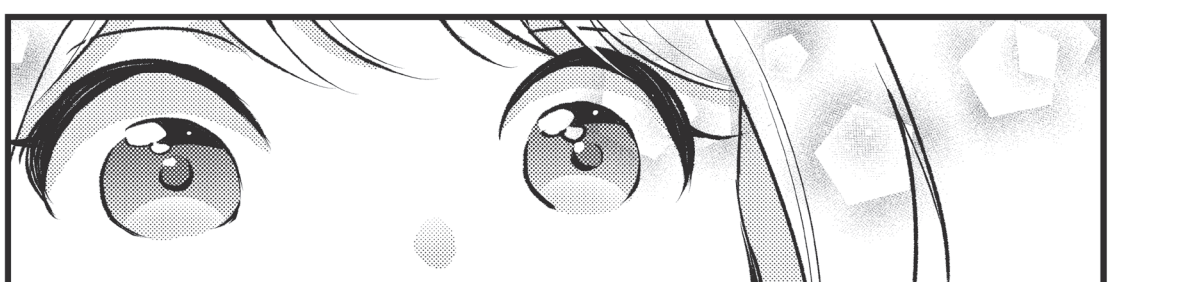

ПОЭТОМУ Я ПОДУМАЛ, ЧТО ЕСЛИ БЫ СМОГ, ИСПОЛЬЗУЯ ДАННЫЕ ЗА ДЕСЯТЬ ЛЕТ, СДЕЛАТЬ ПРОГНОЗ ПРИ ПОМОЩИ МАШИННОГО ОБУЧЕНИЯ, ТО У МЕНЯ ПОЛУЧИЛОСЬ БЫ ЕГО УБЕДИТЬ...
МОЖЕТ, НА ЭТО КОЛИЧЕСТВО ОСАДКОВ В СЕЗОН ДОЖДЕЙ ВЛИЯЕТ?..
НО Я НИЧЕГО УМЕЮ...

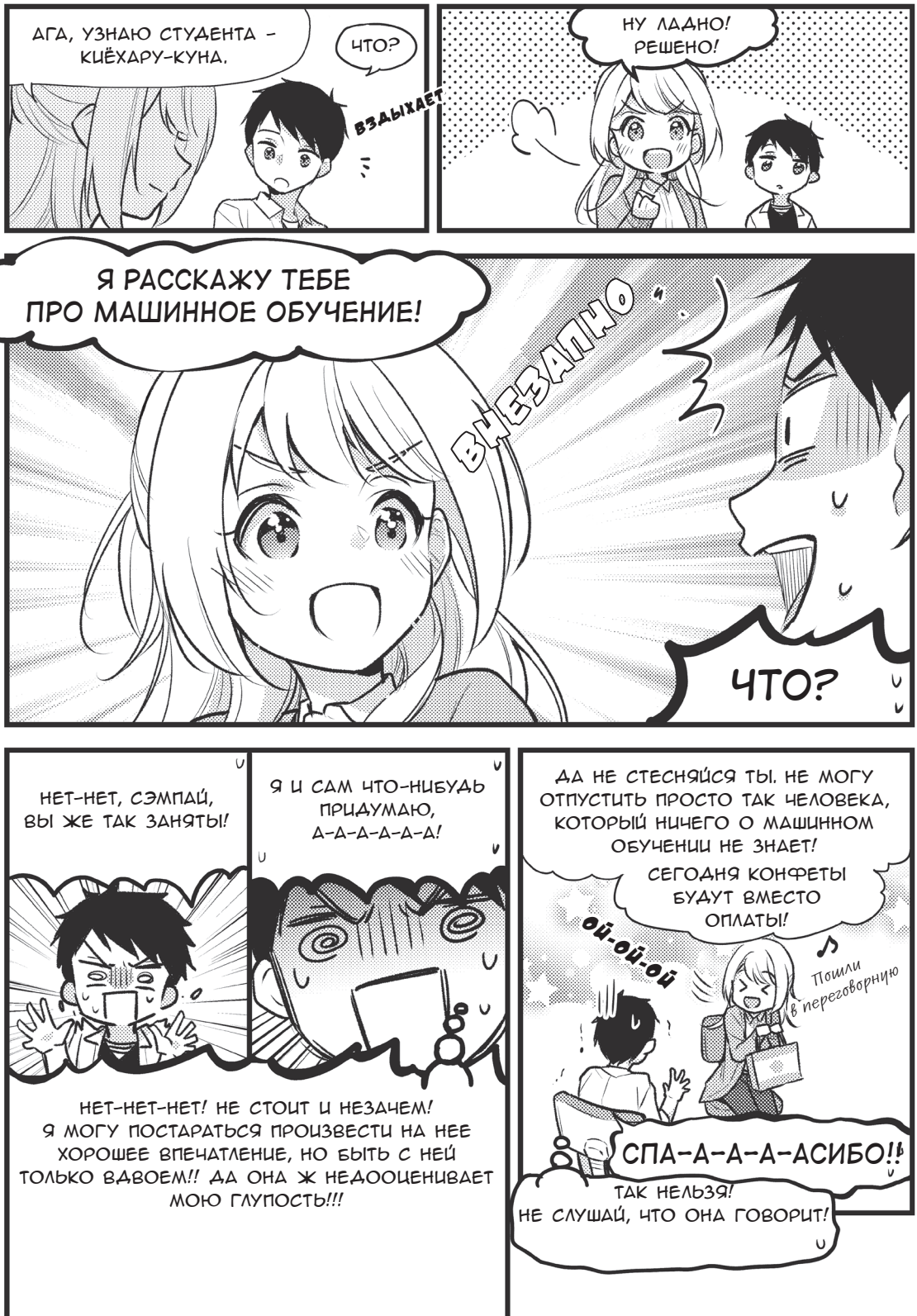

ПОЭТОМУ Я ПРИШЕЛ К ПРОФЕССОРУ НАМИГОЭ, ЧТОБЫ ОН РАССКАЗАЛ МНЕ О МАШИННОМ ОБУЧЕНИИ...

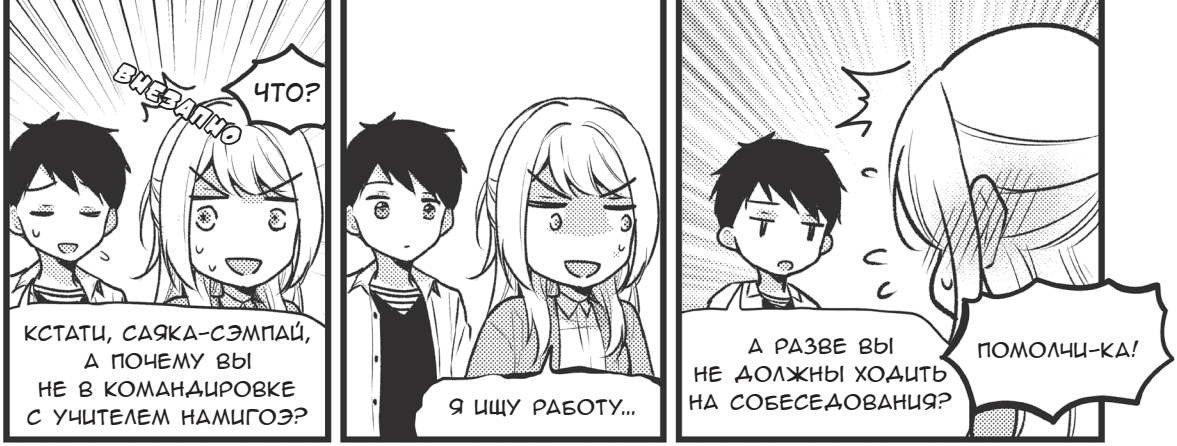

КАКОЙ УЖАС!
АГА.
РАЗНЕЗЬНО
НУ ЧТО, КИЁХАРА-КУН...
ЧТО? Я?

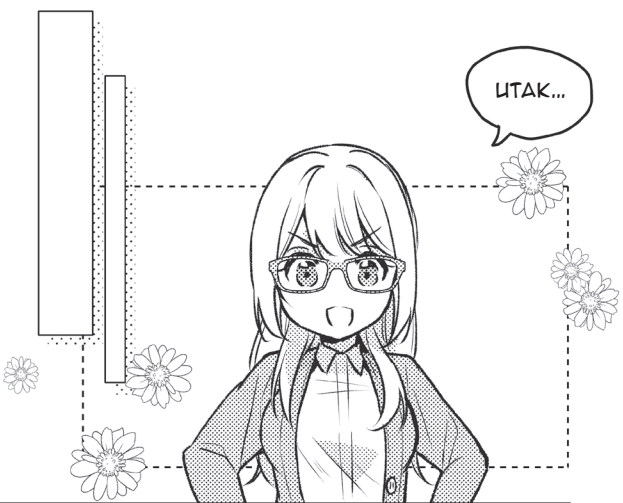

ТЫ В ПРОШЛОМ ГОДУ ЗАКОНЧИЛ УЧИТЬСЯ, А ДАЖЕ ТАКОГО НЕ МОЖЕШЬ СДЕЛАТЬ?
ОЙ...
НЕРВНИЧАЕТ
НО РАЗВЕ ЭТО НЕ ВХОДИТ В ТВОИ РАБОЧИЕ ОБЯЗАННОСТИ?


НУ... НЕ СОВСЕМ... ЕСЛИ БЫ ЭТОТ КОНСУЛЬТАНТ НЕ НАЧАЛ ВСЕХ ДЕРГАТЬ, ВСЕ БЫ ОБОШЛОСЬ...
НУ А КОГДА ПРИШЛОСЬ ВЗЯТЬСЯ ЗА ДЕЛО, ТО Я СРАЗУ ЖЕ ПОМЧАЛСЯ СЮДА...

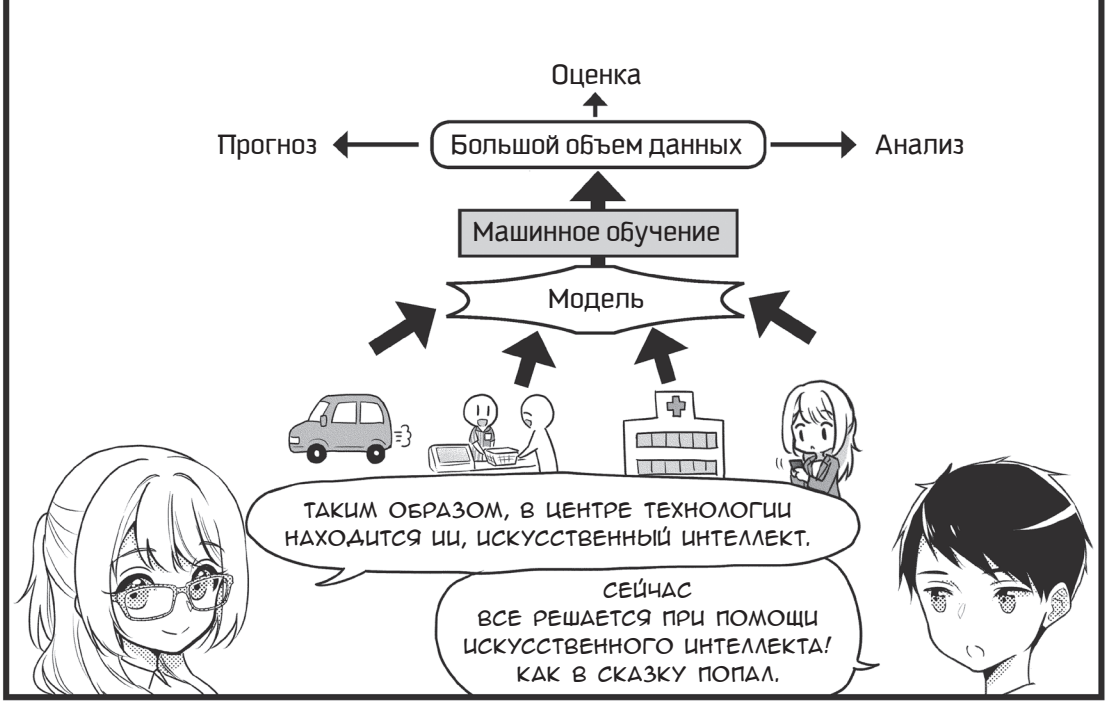



АГА, УЗНАЮ СТУДЕНТА – КИЁХАРУ-КУНА.
ЧТО?
ВЗДЫХАЕТ
НУ ЛАДНО! РЕШЕНО!
Я РАССКАЖУ ТЕБЕ ПРО МАШИННОЕ ОБУЧЕНИЕ!
ВНЕЗАПНО
ЧТО?
НЕТ-НЕТ, СЭМПАЙ, ВЫ ЖЕ ТАК ЗАНЯТЫ!
Я И САМ ЧТО-НИБУДЬ ПРИДУМАЮ, А-А-А-А-А-А!
НЕТ-НЕТ-НЕТ! НЕ СТОИТ И НЕЗАЧЕМ! Я МОГУ ПОСТАРАТЬСЯ ПРОИЗВЕСТИ НА НЕЕ ХОРОШЕЕ ВПЕЧАТЛЕНИЕ, НО БЫТЬ С НЕЙ ТОЛЬКО ВДВОЕМ!! ДА ОНА Ж НЕДООЦЕНИВАЕТ МОЮ ГЛУПОСТЬ!!!
ДА НЕ СТЕСНЯЙСЯ ТЫ. НЕ МОГУ ОТПУСТИТЬ ПРОСТО ТАК ЧЕЛОВЕКА, КОТОРЫЙ НИЧЕГО О МАШИННОМ ОБУЧЕНИИ НЕ ЗНАЕТ!
СЕГОДНЯ КОНФЕТЫ БУДУТ ВМЕСТО ОПЛАТЫ!
ОЙ-ОЙ-ОЙ
Пошли в переговорную
СПА-А-А-А-АСИБО!!
ТАК НЕЛЬЗЯ! НЕ СЛУШАЙ, ЧТО ОНА ГОВОРИТ!

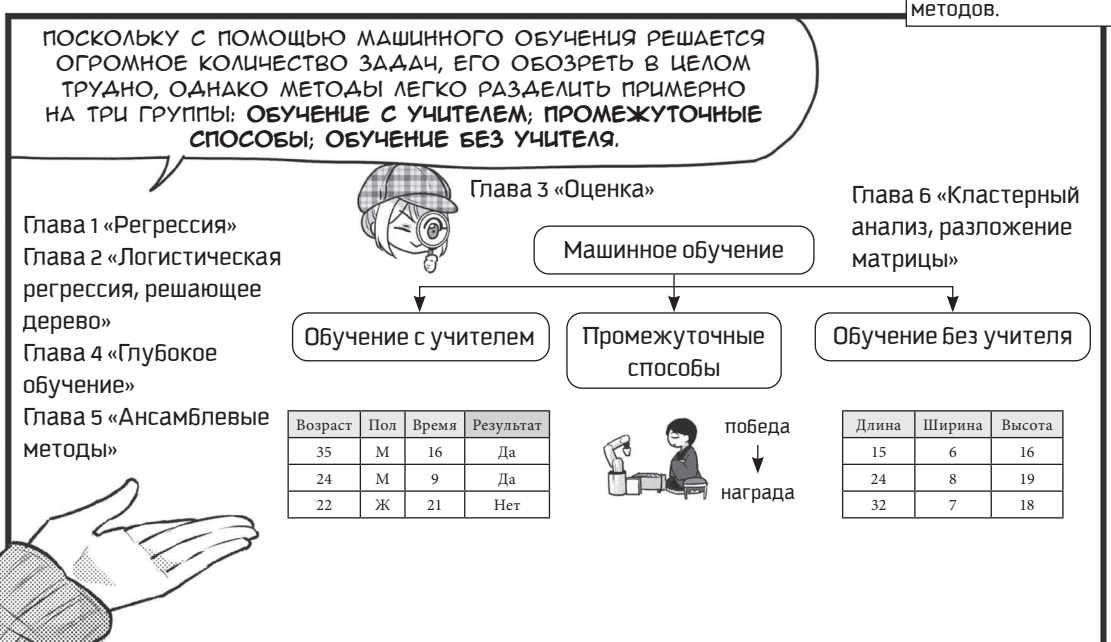




ПОСТРОЕНИЕ НА ОСНОВАНИИ БОЛЬШОГО ОБЪЕМА ДАННЫХ МОДЕЛИ, КОТОРАЯ МОЖЕТ ОЦЕНИВАТЬ И ДЕЛАТЬ ПРОГНОЗЫ.

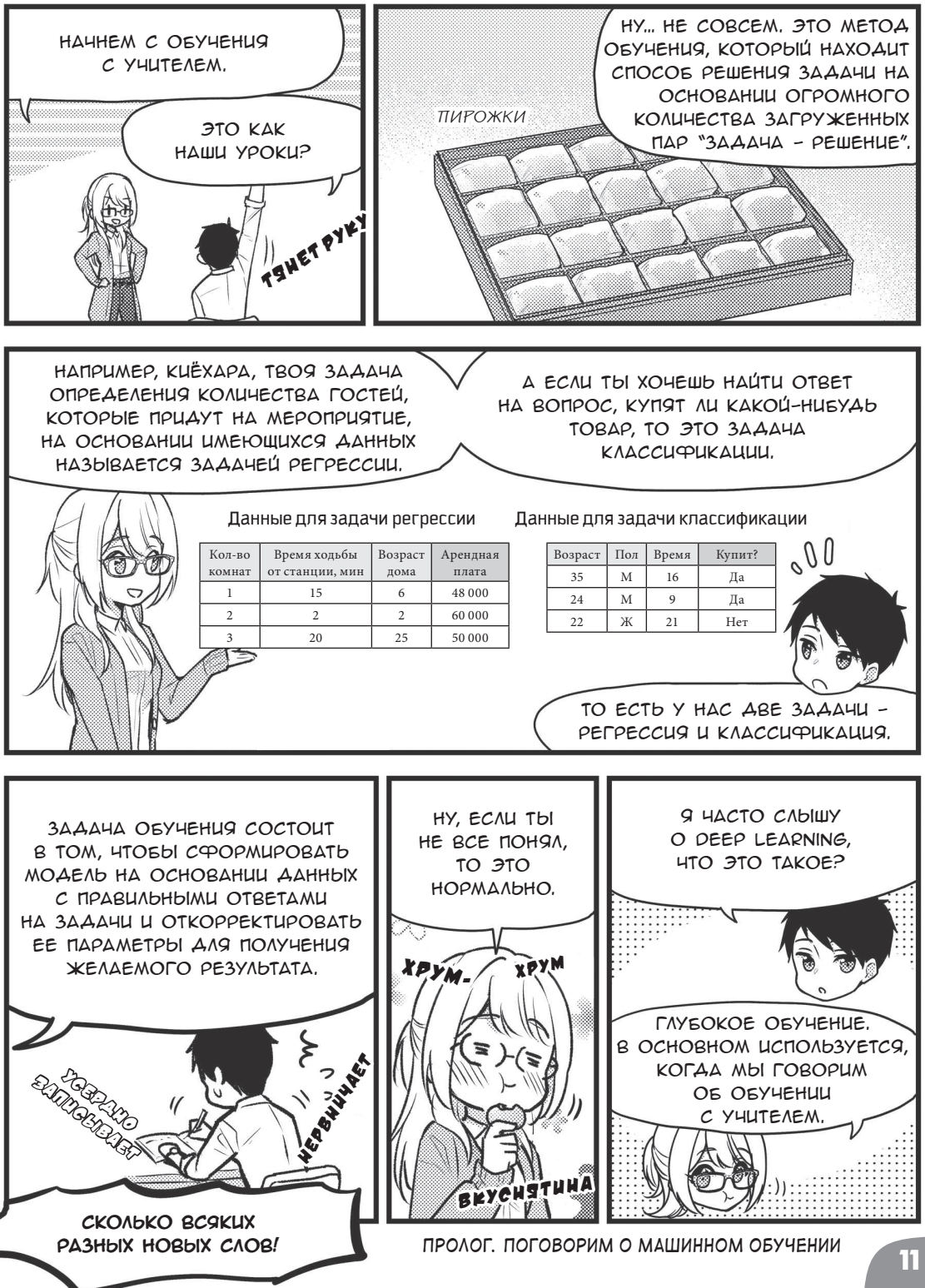


Data mining – метод обнаружения скрытых паттернов в огромном объеме данных при помощи статистики и математических методов.



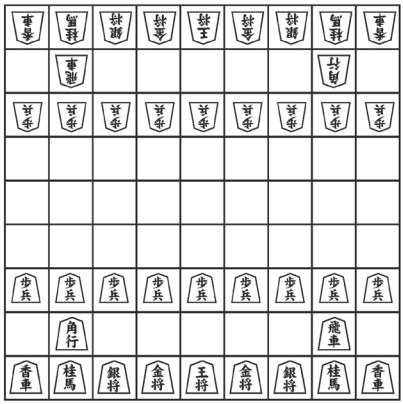
Возраст	Пол	Время	Результат
35	М	16	Да
24	М	9	Да
22	Ж	21	Нет

Длина	Ширина	Высота
15	6	16
24	8	19
32	7	18

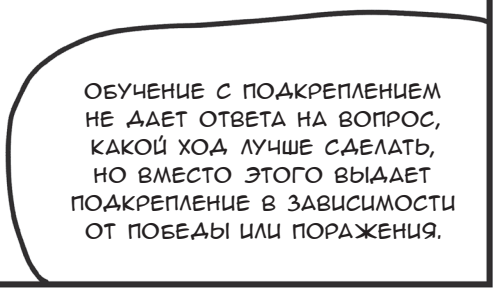


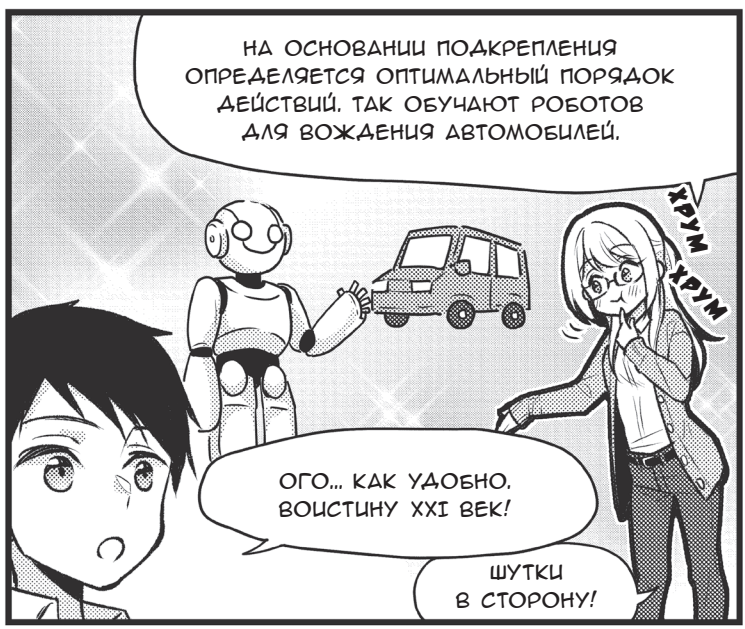
Кол-во комнат	Время ходьбы от станции, мин	Возраст дома	Арендная плата
1	15	6	48 000
2	2	2	60 000
3	20	25	50 000

Возраст	Пол	Время	Купит?
35	М	16	Да
24	М	9	Да
22	Ж	21	Нет



После каждого хода неизвестно, какой следующий ход лучше сделать*





* Приведенная таблица – расстановка для игры в сёги, японские шахматы. – *Прим. перев.*

ЧТО ЖЕ КАСАЕТСЯ ОБУЧЕНИЯ БЕЗ УЧИТЕЛЯ, ТО ЖЕЛАЕМОГО ОТВЕТА НА ВОПРОСЫ СРЕДИ ДАННЫХ НЕТ.

Обучение без учителя

Неразмеченные данные

Длина	Ширина	Высота
15	6	16
24	8	19
32	7	18

А КАК ТОГДА ПРОХОДИТ ОБУЧЕНИЕ?

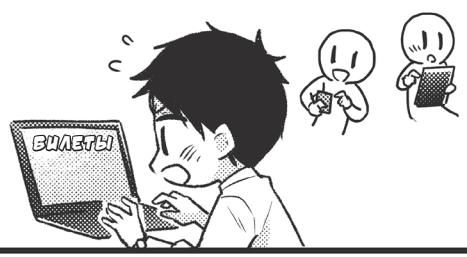
НУ, КАК ОБЫЧНО. ЦЕЛЬ СИСТЕМ БЕЗ УЧИТЕЛЯ – ЭТО ОБНАРУЖИТЬ В БОЛЬШОМ ОБЪЕМЕ ДАННЫХ ЗНАНИЯ, КОТОРЫЕ МОГУТ ПРИГОДИТЬСЯ ЧЕЛОВЕКУ.

ОНИ ИСПОЛЬЗУЮТСЯ, НАПРИМЕР, ДЛЯ РЕКОМЕНДАЦИИ ТОВАРОВ ПРИ ПОКУПКЕ В ИНТЕРНЕТ-МАГАЗИНАХ ИЛИ ПРИ ПОИСКЕ СТРАННОСТЕЙ В ИНФОРМАЦИИ О ДЕЙСТВИЯХ МЕХАНИЗМОВ.

История покупок

ID	#1	#2	#3	#4
115		1		
124		1		1
232				1

Рекомендуемые артисты



ПОКА ЧТО, НАДЕЮСЬ, ТЫ ВСЕ ПОНЯЛ ОТНОСИТЕЛЬНО МАШИННОГО ОБУЧЕНИЯ?

ХРУМ

ХРУМ

ДА...

ЭТО УЖЕ ВОСЬМОЙ.

В кабинете у Саяка (1)

Саяка и старшеклассница Ай

Давно не виделись, Ай-тян. С тех пор как мы были у дедушки, да?

Возможно, тогда собирались все его внуки. Чем ты сегодня занималась?

Рассказывала младшему товарищу про машинное обучение. Он хоть и учился у нас, но, боюсь, очень мало что понял… Ай-тян, ты же хочешь учиться в математическом классе, может, тоже зайдешь послушать?

Машинное обучение? ИИ? Это, наверное, очень сложно!

Там используются математические модели, но это все делается компьютерами. А основы этих моделей может понять и старшеклассник.

Я писала программы для астрономических расчетов. Математика – мой любимый предмет, но пойму ли я?

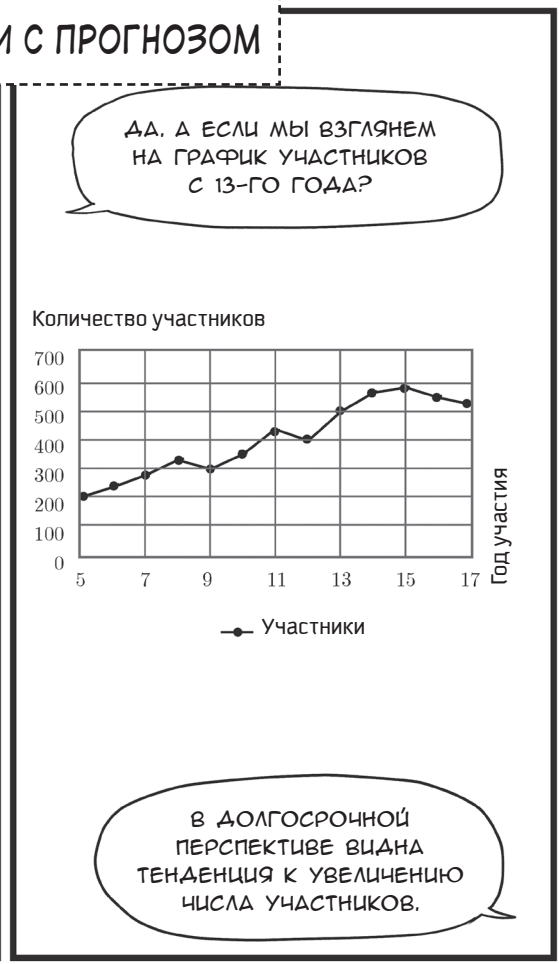
Конечно, поймешь. Мы начнем с регрессии. Приходи послушать!

Хорошо, если что-то будет трудно, я буду задавать вопросы!

ГЛАВА 1

ЧТО ТАКОЕ РЕГРЕССИЯ

1.1. СЛОЖНОСТИ С ПРОГНОЗОМ



ИМЕННО ТАК! В ЗАВИСИМОСТИ ОТ ТОГО, КАКОЙ ПЕРИОД МЫ ВЗЯЛИ, ТЕНДЕНЦИИ АБСОЛЮТНО ПРОТИВОПОЛОЖНЫЕ.

И ЭТО ТОТ СПОСОБ, КОТОРЫЙ ИСПОЛЬЗОВАЛ КОНСУЛЬТАНТ?

ДА. ДАЖЕ ЕСЛИ ТЕНДЕНЦИЯ И ПРАВИЛЬНАЯ, ГРАФИК НЕОБЯЗАТЕЛЬНО БУДЕТ ПОХОДИТЬ НА ПРЯМУЮ.

ВДОХНОВЛЕННО

1.2. ОПРЕДЕЛЯЕМ ЗАВИСИМЫЕ И НЕЗАВИСИМЫЕ ПЕРЕМЕННЫЕ

ЧТО Ж, КИЁХАРА, КАК МЫ ТЕПЕРЬ БУДЕМ ПРЕДСКАЗЫВАТЬ КОЛИЧЕСТВО ГОСТЕЙ?

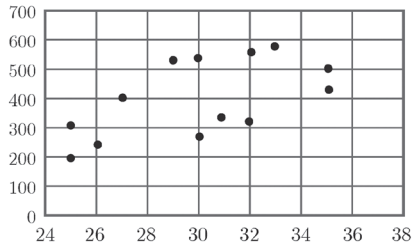
СТАРАТЕЛЬНО ЗАПИСЫВАЕТ

НУ... Я ДУМАЮ, ВЛИЯНИЕ ОКАЗЫВАЮТ ПОГОДА, ТЕМПЕРАТУРА, В ОСОБЕННОСТИ КОЛИЧЕСТВО ОСАДКОВ В СЕЗОН ДОЖДЕЙ...

А ТЕПЕРЬ ПОПРОБУЕМ ВЫБРАТЬ МЕТОД ПРЕДСКАЗАНИЯ В ЗАВИСИМОСТИ ОТ ФАКТОРОВ. ТЕ РЕЗУЛЬТАТЫ, КОТОРЫЕ МЫ ХОТИМ СПРОГНОЗИРОВАТЬ, НАЗЫВАЮТСЯ **ЗАВИСИМЫМИ ПЕРЕМЕННЫМИ**, А ФАКТОРЫ, ВЛИЯЮЩИЕ НА РЕЗУЛЬТАТ, – **НЕЗАВИСИМЫМИ ПЕРЕМЕННЫМИ**.

ДА!

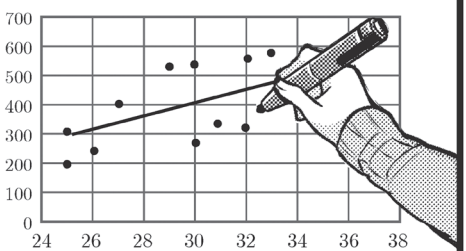
ПУСТЬ У НАС БУДЕТ ОДНА НЕЗАВИСИМАЯ ПЕРЕМЕННАЯ. РАССМОТРИМ ВЛИЯНИЕ ТЕМПЕРАТУРЫ В ДЕНЬ МЕРОПРИЯТИЯ НА КОЛИЧЕСТВО ГОСТЕЙ.



НА ЭТОМ ГРАФИКЕ ГОРИЗОНТАЛЬНАЯ ОСЬ – НЕЗАВИСИМАЯ ПЕРЕМЕННАЯ (ПОГОДА), А ВЕРТИКАЛЬНАЯ – КОЛИЧЕСТВО УЧАСТНИКОВ (ЗАВИСИМАЯ ПЕРЕМЕННАЯ). ОТМЕТИМ ТОЧКИ, РАВНЫЕ КОЛИЧЕСТВУ УЧАСТНИКОВ, И ПРОВЕДЕМ ПО НИМ ПРЯМУЮ.

ЭТО НЕВОЗМОЖНО?

КАК-ТО ТАК. ПО ВОЗМОЖНОСТИ ПРОВОДИМ ЕЕ НЕДАЛЕКО ОТ ВСЕХ ТОЧЕК. КОГДА МЫ ПРОВОДИМ ЛИНИЮ, ОНА ВЫГЛЯДИТ ТАК:



СКРИП СКРИП

Итак, обозначим наклон графика w_1, сдвиг графика относительно вертикальной оси – w_0, температуру за x, а количество участников – за y

$$y = w_1 x + w_0$$

– и получим эту формулу.

Если температура растет, то и количество участников растет, а это непохоже на правду.

Ага. Видимо, независимых переменных несколько.

Несколько? Это ужасно?

Сложность в том, что среди независимых переменных есть такие, которые оказывают влияние на зависимые переменные, а есть и такие, влияние которых невелико.

Влияние велико

Погода

Влияние мало

Влажность

Чтобы просто смоделировать такие условия, необходим способ, который будет упорядочивать зависимые переменные в зависимости от учитываемого **веса** независимых переменных.

Это **линейная регрессия**.

Температура × вес — учитывание

Влажность × вес — цена

→ Зависимая переменная

Что такое линейная регрессия?

Ага! Вот и регрессия!

Если независимых переменных две, то обозначим их как x_1 и x_2, их вес как w_1 и w_2 соответственно, а их взвешенная сумма вместе с постоянной составляющей w_0 записывается...

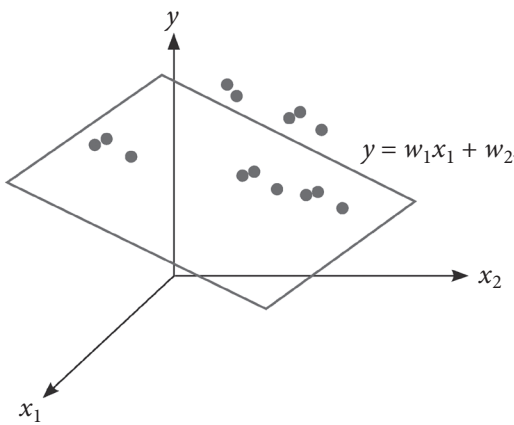
$$y = w_1x_1 + w_2x_2 + w_0$$

...вот так.

Очень похоже на недавнюю формулу.

Скрип скрип

Мы же строили двумерный график?



Можно построить трехмерный график, как на картинке, и спрогнозировать количество участников. Получится такая вот поверхность.

Фух!

И с каждой новой независимой переменной будет увеличиваться количество измерений?

Все так... Обобщаем: если в данных существует d независимых переменных, то точки на графике будут располагаться в $(d+1)$-мерном пространстве, и поэтому нам надо решить задачу нахождения d-мерной гиперплоскости.

Что?
То есть если независимых переменных будет 10, то нам нужно будет найти десятимерную гиперплоскость?
А это вообще возможно?

Да! Придется попросить помощи у математики!

1.3. НАХОДИМ ФУНКЦИЮ ЛИНЕЙНОЙ РЕГРЕССИИ

Шаг 1

Пусть у нас есть d независимых переменных. Тогда обозначим d-мерный столбец-вектор $\boldsymbol{x}$. В соответствии с этим вес, отличный от постоянной w_0, также будет обозначен d-мерным вектором-столбцом $\boldsymbol{w}$, и уравнение гиперплоскости примет такой вид:

$$y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0, \tag{1.1}$$

где T – обозначение транспонирования.

Шаг 2

До этого мы обозначили за y величину взвешенной суммы независимых переменных, а теперь обозначим так величину зависимых переменных в данных для обучения. Взвешенную сумму независимых переменных мы обозначим как $\hat{c}(\boldsymbol{x})$. Значок над буквой означает, что мы не можем гарантировать правильность полученных данных. Таким образом уравнение (1.1) примет вид (1.2):

$$\hat{c}(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0. \tag{1.2}$$

Тот факт, что функция линейной регрессии не слишком отличается от использованных данных, означает, что величина линейной функции $\hat{c}(\boldsymbol{x})$, куда входят независимые переменные $\boldsymbol{x}$, также мало отличается от величины зависимой переменной y. Цель в том, чтобы сделать эту разницу как можно меньше. Однако если эта разница проявляется в наборе данных, то случаи, где величина зависимой переменной выше величины линейной функции, накладываются на те, где величина зависимой переменной ниже величины линейной функции, и они компенсируют друг друга.

Шаг 3

Для определения «отклонения» линейной функции от имеющихся данных возводим в квадрат разницу между линейной функцией $\hat{c}(\boldsymbol{x})$ и зависимой переменной y; т.е. находим **квадрат ошибки**. Уменьшение квадрата ошибки путем корректировки веса линейной функции называется обучением по **методу наименьших квадратов**.

Таким образом, добавив к $\boldsymbol{x}$ в уравнении (1.2) 0-мерность и определив его величину как равную 1, а также добавив w_0 к 0-мерному $\boldsymbol{w}$, получим, что функция регрессии будет записываться как внутреннее произведение $(d + 1)$-мерного вектора (1.3).

$$\hat{c}(x) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}. \tag{1.3}$$

Шаг 4

Оценим коэффициенты w этого уравнения, используя обучающие данные $\{(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_n, y_n)\}$. Постараемся максимально возможно уменьшить разницу между величинами зависимой переменной y и функции $\hat{c}(\boldsymbol{x})$, рассчитанной по уравнению 1.3. Ошибка определяется значением коэффициентов $\boldsymbol{w}$ в уравнении (1.3) обозначив ее $E(\boldsymbol{w})$, получим следующее уравнение:

$$E(\boldsymbol{w}) = \sum_{i=1}^{n} (y - \hat{c}(\boldsymbol{x}))^2, \tag{1.4}$$

$$= \sum_{i=1}^{n} (y - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)^2. \tag{1.5}$$

Шаг 5

Чтобы избавиться от трудоемких вычислений суммы, представим независимые переменные матрицами, а зависимые – векторами. Обозначим матрицу, имеющую n позиций по вертикали, которая получилась путем транспонирования независимой переменной $\boldsymbol{x}$ d-мерного вектора-столбца как $\boldsymbol{X}$, y – вектор-столбец величин независимой переменной $\boldsymbol{y}$, а $\boldsymbol{w}$ – вектор-столбец коэффициентов.

В итоге отклонение примет следующий вид:

$$E(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}). \tag{1.6}$$

Чтобы минимизировать отклонение, нужно найти такие величины коэффициентов $\boldsymbol{w}$, при которых производная функции ошибки равняется 0, то есть:

$$\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = 0, \tag{1.7}$$

$$\boldsymbol{w} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \tag{1.8}$$

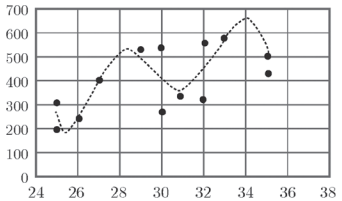
где $\boldsymbol{A}^{-1}$ – матрица, обратная матрице $\boldsymbol{A}$.

Таким образом, мы можем аналитически найти по обучающим данным веса $\boldsymbol{w}$ при помощи минимизации суммы квадратов ошибки. При подстановке $\boldsymbol{w}$ в уравнение (1.3) получится функция линейной регрессии $\hat{c}(\boldsymbol{x})$.

1.4. РЕГУЛЯРИЗАЦИЯ РЕЗУЛЬТАТА

ЕСЛИ КОЭФФИЦИЕНТЫ ЛИНЕЙНЫЕ, ТО ВЕСА МОГУТ БЫТЬ РАССЧИТАНЫ ТАКИМ ЖЕ ОБРАЗОМ, ДАЖЕ ЕСЛИ ЭТО УРАВНЕНИЯ ВЫСОКОГО ПОРЯДКА. ТАК МОЖНО ПРОВОДИТЬ ОБУЧЕНИЕ ПРИ ПОМОЩИ САМЫХ СЛОЖНЫХ УРАВНЕНИЙ РЕГРЕССИИ.

ЕСЛИ Я ЗНАЮ ЧИСЛЕННЫЙ ВЕС, ТО ОН МОЖЕТ БЫТЬ ПРОСТО ПОДСТАВЛЕН В ЭТУ ФОРМУЛУ?



НЕ СОВСЕМ ТАК. В ЭТОЙ ФОРМУЛЕ ДАЖЕ ПРИ НЕБОЛЬШОМ ИЗМЕНЕНИИ ДАННЫХ НА ВХОДЕ РЕЗУЛЬТАТЫ НА ВЫХОДЕ МОГУТ СИЛЬНО ОТЛИЧАТЬСЯ, И НЕЛЬЗЯ БУДЕТ ПОЛУЧИТЬ ХОРОШИЙ РЕЗУЛЬТАТ ПРИ ИСПОЛЬЗОВАНИИ ДАННЫХ, ОТЛИЧНЫХ ОТ ОБУЧАЮЩИХ.

СЛОЖНОВАТО. А КАК СДЕЛАТЬ ХОРОШИЙ ПРОГНОЗ?

ЕСЛИ НЕБОЛЬШОЕ ИЗМЕНЕНИЕ КАКОГО-ЛИБО ПАРАМЕТРА НА ВХОДЕ ДАЕТ БОЛЬШОЕ ИЗМЕНЕНИЕ НА ВЫХОДЕ, ЭТО ОЗНАЧАЕТ, ЧТО КОЭФФИЦИЕНТ УРАВНЕНИЯ СЛИШКОМ ВЕЛИК, ПОЭТОМУ ЕГО НЕОБХОДИМО УМЕНЬШИТЬ.

ТАК БЫСТРО?

НО, ВПРОЧЕМ, ЕСТЬ И ДРУГАЯ ТОЧКА ЗРЕНИЯ НА ВЕС, В СЛУЧАЯХ, КОГДА ПРАВИЛЬНОСТЬ ПРОГНОЗА ВАЖНЕЕ, ЧЕМ ОБЪЯСНИМОСТЬ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ.

ЧПОК

ПИРОЖОК РАЗЛОМЫ ПОПОЛАМ

ОБЪЯСНИМОСТЬ ВАЖНЕЕ ТОЧНОСТИ?

С НАБИТЫМ РТОМ

ДОПУСТИМ, ВМЕСТО ТОГО ЧТОБЫ ПРОГНОЗИРОВАТЬ, КАКИЕ ФАКТОРЫ ПОВЛИЯЮТ НА КАЧЕСТВО ТОВАРА, НАДО НАЙТИ ТЕ, КОТОРЫЕ ВЛИЯЮТ НА НЕГО БОЛЬШЕ ВСЕГО.

ТАК МОЖНО НАЙТИ ТЕ ХАРАКТЕРИСТИКИ, КОТОРЫЕ ДЕЛАЮТ ТОВАР БРАКОВАННЫМ.

Пирожок
Составляющие:
– мука
– начинка
– сахар

→ Что влияет больше?

Если говорить конкретнее, можно подставить в качестве веса переменных в формулу линейной регрессии 0 и посмотреть, изменится ли количество измерений.

Иными словами, следует найти способ, чтобы коэффициент w в уравнении линейной регрессии уменьшился, если величина его большая, или же стал равен нулю.

И этот способ называется **регуляризацией**.

Он нужен, чтобы коэффициент не был слишком большим...

Метод регуляризации, при котором мы уменьшаем большие коэффициенты, называется **Ридж-регрессией**,

а чтобы увеличить количество величин, равных нулю, используется **Лассо-регрессия**.

Регуляризация осуществляется добавлением дополнительного члена к уравнению ошибки.

Начнем с объяснения Ридж-регрессии.

Мы добавляем дополнительный член, квадрат параметра w.

Чтобы уменьшить это...

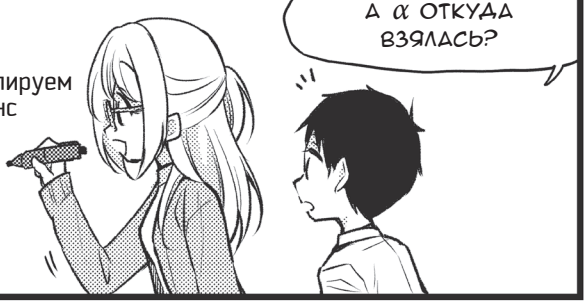
...надо увеличить это

$$E(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \alpha \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$$

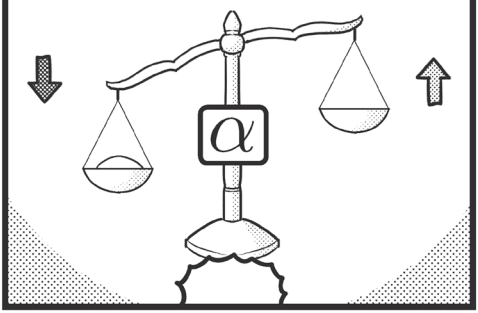
Регулируем баланс

Если величина веса слишком маленькая...

...то она далека от правильной



α – это вес дополнительного члена регуляризации. Если параметр большой, то эффект регуляризации становится важнее эффективности, если маленький, то эффективность становится важнее.



Используя ридж-регрессию, мы находим величину w, когда градиент по w функции ошибки равен 0, как и в случае нахождения этого параметра методом наименьших квадратов.

$$\boldsymbol{w} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \alpha\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

$\boldsymbol{I}$ – единичная матрица.

Кстати, а почему ридж-регрессия так называется?

А теперь поговорим о регрессии "лассо". Это регуляризация, при которой w становится абсолютной величиной.

Ridge означает гребень горы, и единичная матрица на него похожа*.

* Есть другие версии.

Хм...

Как я и говорила, ридж-регрессия – это регуляризация для уменьшения величины параметра.

Ridge – это квадрат w, лассо – это абсолютная величина w.

А что такое лассо?

Лассо – это петля для ловли кого-нибудь.

Вроде это слово используют, когда говорят о ковбоях.

Представь, что во множество параметров кидают лассо и выбирают самые маленькие из них.

Изначально это аббревиатура фразы **least absolute shrinkage and selection operator**.

Надо же.

Уравнение оценки ошибок регрессии "лассо" выглядит...

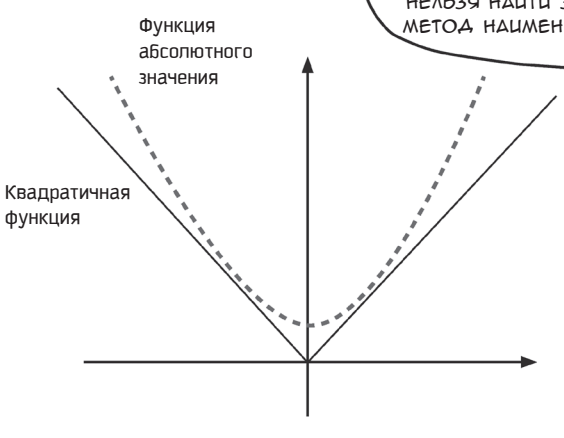
$$E(\boldsymbol{w}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \alpha \sum_{i=1}^{n} |w_j|$$

...так:

Поскольку w_0 – постоянное слагаемое в уравнении, его величина не повлияет на величину уравнения регрессии, и его обычно не регуляризируют. Здесь если вес дополнительного члена

увеличивается, растет число величин с весом, равным 0.

А как можно объяснить регрессию "лассо"?



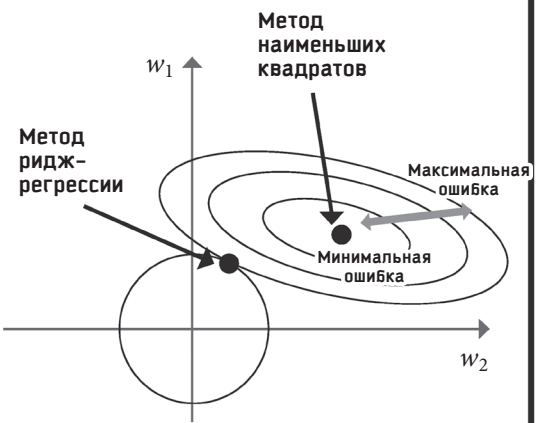
Поскольку туда входит абсолютная величина, недифференцируемая в точке начала координат, нельзя найти значение аналитически, используя метод наименьших квадратов, поэтому верхний предел/максимум дополнительного члена регуляризации ограничивается дифференцируемой квадратичной функцией. Был предложен метод, чтобы многократно обновлять ее параметры с целью уменьшения ошибки.

С ПОМОЩЬЮ МЕТОДА "ЛАССО" МОЖНО ПРОРЕДИТЬ НЕЗАВИСИМЫЕ ПЕРЕМЕННЫЕ С ВЕСАМИ, НЕ РАВНЫМИ НУЛЮ, И НАЙТИ ТЕ, КОТОРЫЕ ОКАЗЫВАЮТ ВЛИЯНИЕ.

УДОБНО.

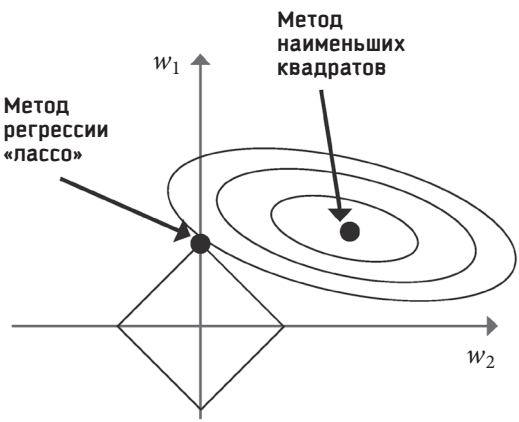
А ТЕПЕРЬ Я ОБЪЯСНЮ, ЧЕМ РИДЖ-РЕГРЕССИЯ ОТЛИЧАЕТСЯ ОТ РЕГРЕССИИ "ЛАССО".

ПОЖАЛУЙСТА!



КАК ПОКАЗАНО НА РИСУНКЕ, ПРИ ИСПОЛЬЗОВАНИИ РИДЖ-РЕГРЕССИИ ОГРАНИЧЕНИЕ ДИАПАЗОНА ПАРАМЕТРОВ ОКРУЖНОСТЬЮ (ОБЩИЙ СЛУЧАЙ d-МЕРНОЙ ГИПЕРСФЕРЫ) НЕ ПОЗВОЛЯЕТ КАЖДОМУ ВЕСУ ПРИНИМАТЬ БОЛЬШОЕ ЗНАЧЕНИЕ. В ОБЩЕМ СЛУЧАЕ ТОЧКА КАСАНИЯ ИЗОЛИНИИ ФУНКЦИИ ОШИБКИ ЯВЛЯЕТСЯ ТОЧКОЙ НА ОКРУЖНОСТИ, КОТОРАЯ ЯВЛЯЕТСЯ ЗНАЧЕНИЕМ ВЕСА ПАРАМЕТРА.

ПОЭТОМУ ВЕЛИЧИНА ПАРАМЕТРА УМЕНЬШАЕТСЯ.



А В СЛУЧАЕ РЕГРЕССИИ "ЛАССО", ПРИ УСЛОВИИ ЧТО ОПРЕДЕЛЕНА СУММА ПАРАМЕТРОВ, ДИАПАЗОН ПАРАМЕТРОВ ОГРАНИЧЕН ОБЛАСТЬЮ (РОМБОМ), УГЛЫ КОТОРОГО ЛЕЖАТ НА КАЖДОЙ ОСИ, КАК ПОКАЗАНО НА РИСУНКЕ.

И ОДИН ИЗ УГЛОВ РОМБА КАСАЕТСЯ ИЗОЛИНИИ ФУНКЦИИ ОШИБКИ.

КАЖЕТСЯ, В УГЛАХ БОЛЬШИНСТВО ПАРАМЕТРОВ СТАНОВЯТСЯ РАВНЫМИ 0.

ЭТО ВЛИЯНИЕ РЕГУЛЯРИЗАЦИИ РЕГРЕССИИ "ЛАССО".

Для начала загрузим библиотеку. В scikit-learn подготовлено несколько наборов данных, выберем методы из пакета datasets. Для регрессии это линейная регрессия, ридж-регрессия и регрессия «лассо».

```
from sklearn.datasets import load_boston
from sklearn.linear_model import LinearRegression, Ridge, Lasso
```

В качестве данных для анализа мы возьмем 13 параметров из стандартной выборки boston dataset, куда входят уровень преступности, количество комнат, географическое положение и прочие данные, связанные с недвижимостью.

Атрибут data экземпляра boston, который создан при помощи приведенного ниже кода, является матрицей, в которой признаковое описание объекта располагается в виде столбцов (для 13-мерного признака будет 506 векторов-строк), а атрибут target будет введен в качестве вектора-столбца – цены каждого свойства.

Можно показать детали данных boston с помощью атрибута descr функцией print (boston.DESCR).

```
boston = load_boston()
X = boston.data
y = boston.target
```

Код в scikit-learn удобен для обучения. А теперь используем учебный набор данных.

```
lr1 = LinearRegression()
```

В этом экземпляре можно вызвать метод fit, который выполняет обучение с набором признаковых описаний $\boldsymbol{X}$ и точной информацией $\boldsymbol{y}$ в качестве аргументов.

```
lr1.fi t(X, y)
```

Когда получено уравнение линейной регрессии, можно узнать прогнозируемое значение, которое в качестве аргумента имеет 13-мерный вектор $\boldsymbol{x}$, с помощью метода predict, который выведет прогнозируемое значение.

А теперь попробуем провести регуляризацию. Для начала используем сумму квадратов и коэффициенты формулы линейной регрессии, которые мы только что разобрали.

```
print("Linear Regression")
for f, w in zip(boston.feature_names, lr1.coef_) :
    print("{0:7s}: {1:6.2f}". format(f, w))
print("coef = {0:4.2f}".format(sum(lr1.coef_**2)))
```

```
Linear Regression
CRIM   :  -0.11
ZN     :   0.05
INDUS  :   0.02
CHAS   :   2.69
NOX    : -17.80
RM     :   3.80
AGE    :   0.00
DIS    :  -1.48
```

```
RAD    :   0.31
TAX    :  -0.01
PTRATIO:  -0.95
B      :   0.01
LSTAT  :  -0.53
coef = 341.86
```

Попробуем также провести ридж-регрессию. Поскольку в данные входят X и y, лучше всего начать с постройки экземпляра, по которому можно провести обучение.
Если есть параметр, который нужно указать, он задается в качестве аргумента экземпляра в формате «имя параметра = значение». Вес α дополнительного параметра регуляризации примем равным 10,0.

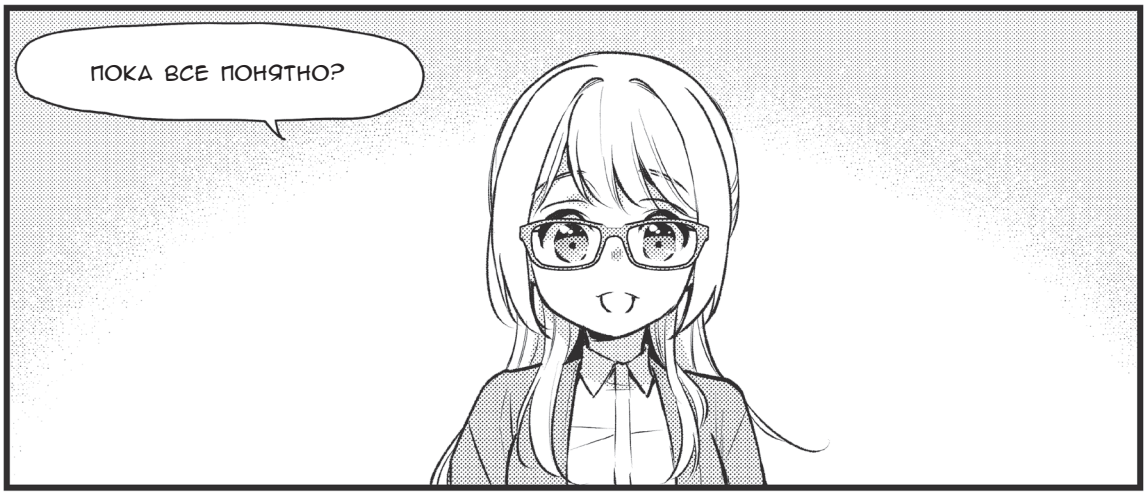
```
lr2 = Ridge(alpha=10.0)
lr2.fi t(X, y)
print("Ridge")
for f, w in zip(boston.feature_names, lr2.coef_) :
    print("{0:7s}: {1:6.2f}". format(f, w))
print("coef = {0:4.2f}".format(sum(lr2.coef_**2)))
```

```
Ridge
CRIM   :  -0.10
ZN     :   0.05
INDUS  :  -0.04
CHAS   :   1.95
NOX    :  -2.37
RM     :   3.70
AGE    :  -0.01
DIS    :  -1.25
RAD    :   0.28
TAX    :  -0.01
PTRATIO:  -0.80
B      :   0.01
LSTAT  :  -0.56
coef = 25.73
```

Мы видим, что сумма квадратов коэффициентов абсолютно мала. А теперь применим регрессию «лассо». Вес α дополнительного параметра регуляризации примем равным 2,0 и заметим, что несколько коэффициентов равны 0.

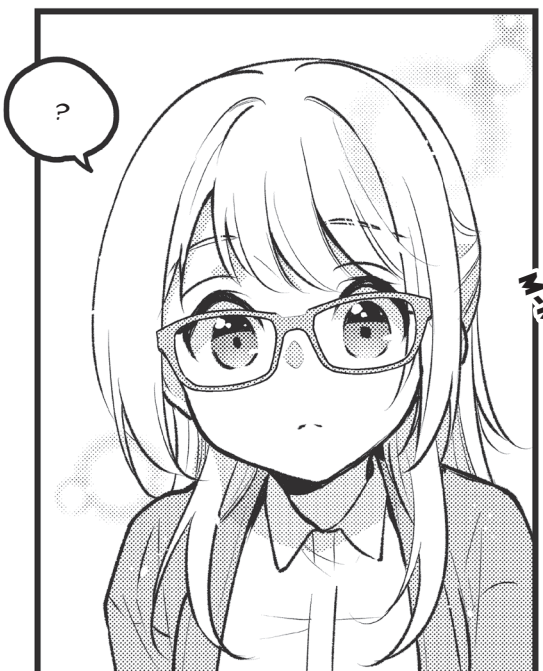
```
lr3 = Lasso(alpha=2.0)
lr3.fi t(X, y)
print("Lasso")
for f, w in zip(boston.feature_names, lr3.coef_) :
    print("{0:7s}: {1:6.2f}". format(f, w))
print("coef = {0:4.2f}".format(sum(lr3.coef_**2)))
```

```
Lasso
CRIM   :  -0.02
ZN     :   0.04
INDUS  :  -0.00
CHAS   :   0.00
NOX    :  -0.00
RM     :   0.00
AGE    :   0.04
DIS    :  -0.07
RAD    :   0.17
TAX    :  -0.01
PTRATIO:  -0.56
B      :   0.01
LSTAT  :  -0.82
coef = 1.02
```


ПОКА ВСЕ ПОНЯТНО?

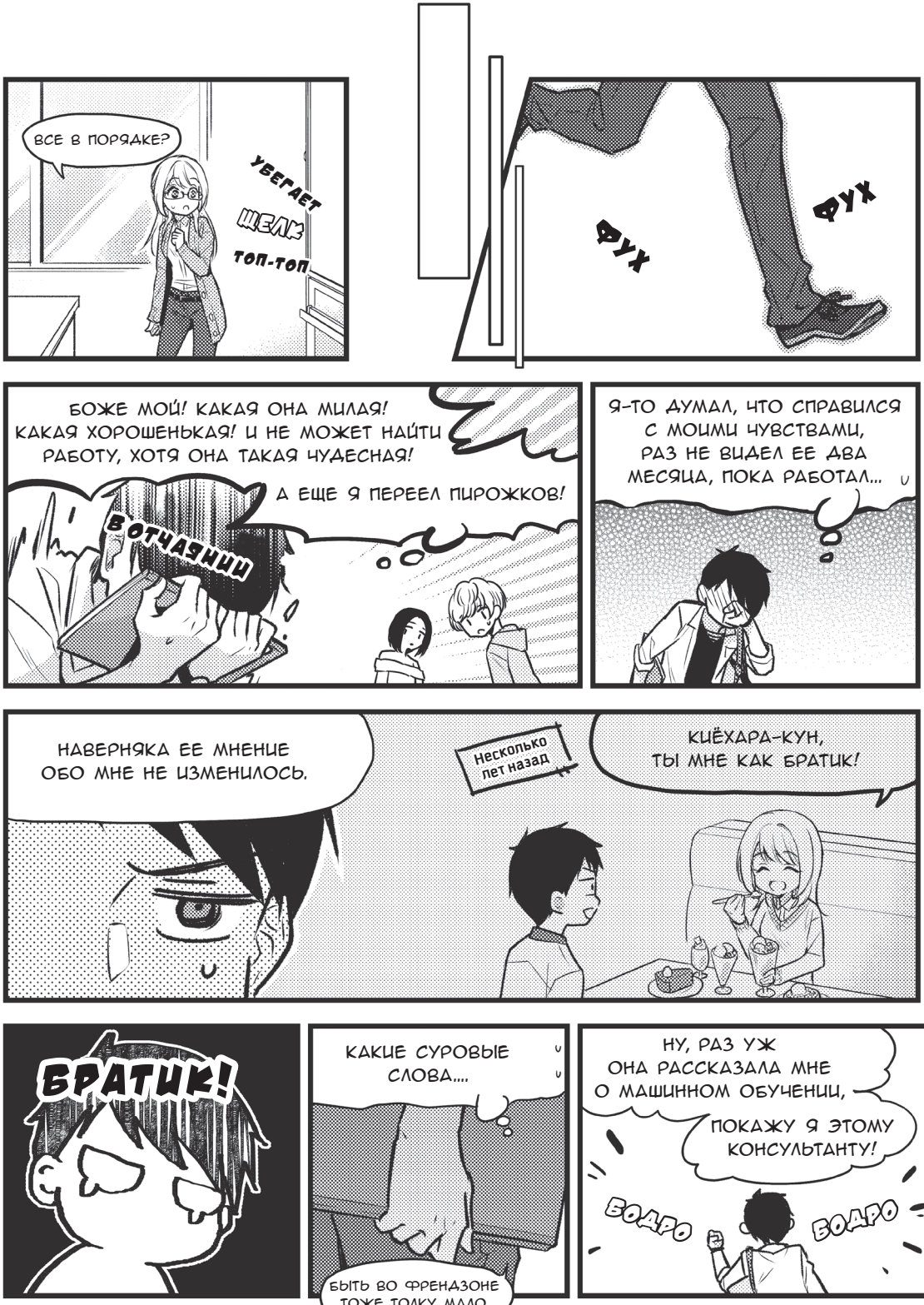

Н-НУ...
ДРОЖИТ
ОСТАЛОСЬ ПОПРОБОВАТЬ НА ПРАКТИКЕ...


НЕТ! ОНА МНЕ ТАК ДОВЕРЯЕТ!
Я НЕ МОГУ БОЛЬШЕ ТАК СИДЕТЬ ВДВОЕМ!


?

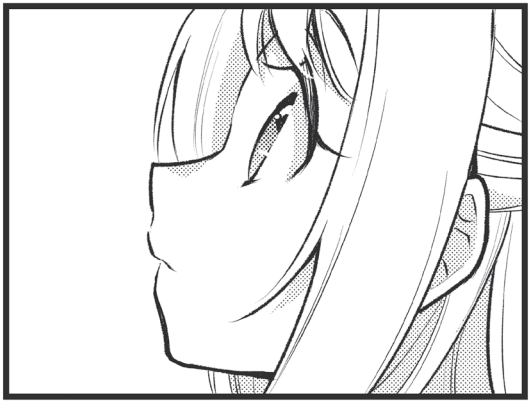

КАКАЯ
М-М-М-М-МИЛЛИЯ

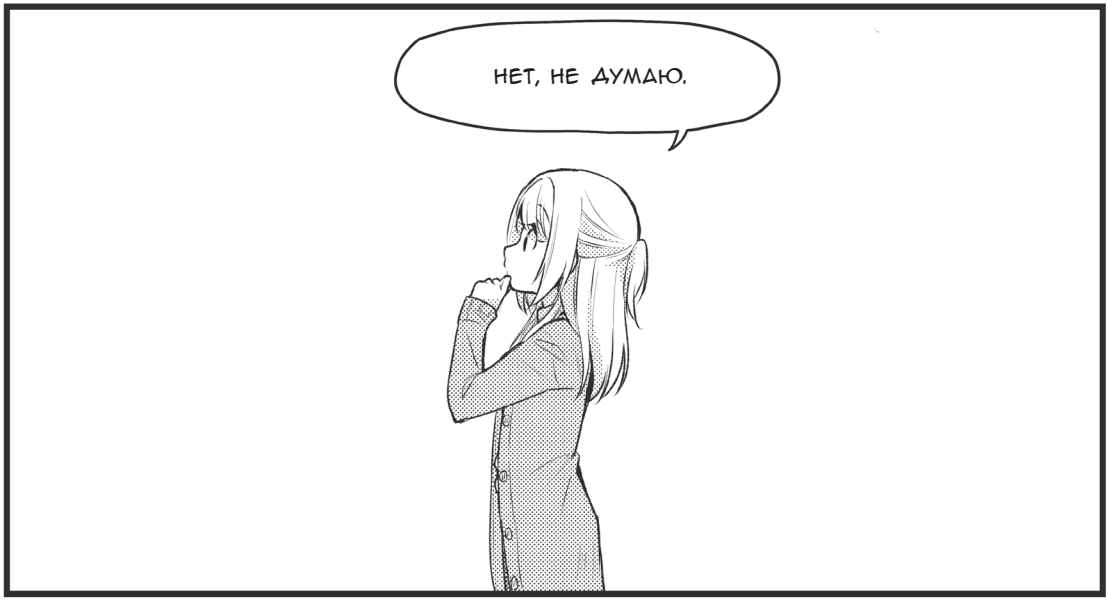

Я САМ СПРАВЛЮСЬ!
СПАСИБО БОЛЬШОЕ!
ТОП
ТОП
Д?


ВСЕ В ПОРЯДКЕ?
УБЕГАЕТ
ЩЕЛК
ТОП-ТОП
ФУХ
ФУХ
БОЖЕ МОЙ! КАКАЯ ОНА МИЛАЯ! КАКАЯ ХОРОШЕНЬКАЯ! И НЕ МОЖЕТ НАЙТИ РАБОТУ, ХОТЯ ОНА ТАКАЯ ЧУДЕСНАЯ!
А ЕЩЕ Я ПЕРЕЕЛ ПИРОЖКОВ!
В ОТЧАЯНИИ
Я-ТО ДУМАЛ, ЧТО СПРАВИЛСЯ С МОИМИ ЧУВСТВАМИ, РАЗ НЕ ВИДЕЛ ЕЕ ДВА МЕСЯЦА, ПОКА РАБОТАЛ...
НАВЕРНЯКА ЕЕ МНЕНИЕ ОБО МНЕ НЕ ИЗМЕНИЛОСЬ.
Несколько лет назад
КЮЁХАРА-КУН, ТЫ МНЕ КАК БРАТИК!
БРАТИК!
КАКИЕ СУРОВЫЕ СЛОВА....
БЫТЬ ВО ФРЕНДЗОНЕ ТОЖЕ ТОЛКУ МАЛО.
НУ, РАЗ УЖ ОНА РАССКАЗАЛА МНЕ О МАШИННОМ ОБУЧЕНИИ,
ПОКАЖУ Я ЭТОМУ КОНСУЛЬТАНТУ!
БОДРО
БОДРО


Не думаю, что Киёхара-кун вообще чему-то научился, хотя учился четыре года.
Но он быстро все схватывает, может, ему надо было побольше мотивации.
Ой!
Хлоп!
Пирожки кончились


Хотя он и учиться не хотел, и работать, а все-таки пришел сюда в выходной, чтобы поучиться.
Может быть, что-нибудь изменилось?




Нет, не думаю.

В кабинете у Саяка (2)

Математическое повторение (1)

Вот об этом мы говорили с Киёхара-куном. Ай-тян, ты все поняла?

Вы много говорили о векторах и матрицах. Вектор – это последовательность чисел, заключенная в скобки. Есть двухмерный вектор (a, b), трехмерный (a, b, c). А что такое d-мерный вектор?

Если d больше четырех, то мы не можем представить себе пространство, это трудно. Но необязательно представлять пространство, можно просто представить очень много чисел, выстроенных в ряд.

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}$$

А что такое вектор-столбец – ряд чисел, выстроенных в ряд?

Нет, не совсем так. Просто когда есть несколько признаков, их обычно представляют вертикально. В машинном обучении часто складывают матрицы и векторы, и матрица обычно слева, а состав матрицы удобно выражать как произведение матриц.

А в школе матрицы не проходят.

Можно сказать, что матрицы – это числа, записанные в виде прямоугольника.

Напомним, что столбцы идут сверху вниз, а строки слева направо.

Ага!

Матрица, которая имеет 2 столбца и 2 строки, называется матрицей 2 на 2 (2×2). Сумма матриц – это матрица, элементы которой равны сумме соответствующих элементов слагаемых матриц, но умножение матриц – очень трудная операция.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1\times5+2\times7 & 1\times6+2\times8 \\ 3\times5+4\times7 & 3\times6+4\times8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$$

Значение n-й строки и m-го столбца матрицы, которая является результатом умножения, определяется так: извлекаются n-я строка из первой матрицы и m-й столбец из второй и перемножаются, начиная с первого числа, а затем складываются.

А если число строк первой матрицы не совпадает с числом столбцов второй, то матрицы просто перемножить нельзя.

Да. Все так. При умножении матрицы (i×j) на матрицу (j×k) результатом будет матрица (i×k), если говорить совсем просто.

$$\begin{pmatrix} \square & \square & \square \\ \square & \square & \square \end{pmatrix} \cdot \begin{pmatrix} \square & \square \\ \square & \square \\ \square & \square \end{pmatrix} = \begin{pmatrix} \square & \square \\ \square & \square \end{pmatrix}$$

i строк, j столбцов j строк, k столбцов i строк, k столбцов

Теперь рассмотрим транспонированные и обратные матрицы. Транспонированная матрица $\boldsymbol{X}$ обозначается как $\boldsymbol{X}^T$ и получается при замене столбцов на строки и обратно.

Как мы транспонировали вектор w в формуле 1.1?

Векторы лучше представлять как особый случай. Например, d-мерный столбец-вектор может рассматриваться как матрица из d строк и 1 столбца.

А-а... Вот как. То есть если $\boldsymbol{w}$ – матрица из d строк и 1 столбца, то $\boldsymbol{w}^{\mathrm{T}}$ – матрица из 1 строки и d столбцов. Произведение $(\boldsymbol{w}^{\mathrm{T}}x)$ $\boldsymbol{x}$ матрицы из d строк и 1 столбца на транспонированую $\boldsymbol{w}^{\mathrm{T}}$ дает матрицу из одной строки и одного столбца. Но это разве не обычное число?

Именно. Обычное число называется скаляром. $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ – скаляр, и w_0 – тоже скаляр. Их сумма y – тоже скаляр.

Угу.

А теперь посмотрим на более сложное уравнение 1.8. Матрицу, обратную матрице $\boldsymbol{A}$, обозначим как $\boldsymbol{A}^{-1}$. Ай-тян, какое число обратно 5?

Обратное число – это результат деления единицы на него. Обратное число для 5 – 1/5.

Да. В основном обратная матрица работает так же. В мире матриц число 1 называется единичной матрицей. Единичная матрица с одинаковым количеством столбцов и строк обозначается *I*. Внутри нее по диагонали расположены единицы, а остальные величины – нули.

А почему ее величина соответствует единице?

Попробуем перемножить ее с другой матрицей. Ничего не изменилось?

Ничего. Результат произведения такой же.

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

Например, матрица $\boldsymbol{A}^{-1}$, обратная матрице $\boldsymbol{A}$, в которой два столбца и две строки, будет вычисляться по представленной ниже формуле.
Обычно вычисления матриц, обратной матрице с d столбцами и d строками, доверяют компьютеру.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -c \\ -c & a \end{pmatrix}$$

Если матрицу $\boldsymbol{A}$ перемножить на обратную ей левую матрицу $\boldsymbol{A}^{-1}$, то получится единичная матрица $\boldsymbol{I}$.

Есть еще непонятные места?

Вот этот знак – Σ.

Это греческая буква сигма, используется для записи суммы. С использованием сигмы формула $\sum_{i=1}^{d} w_i x_i$ записывается просто так.

И еще, как дифференцировать функцию, заданную вектором.

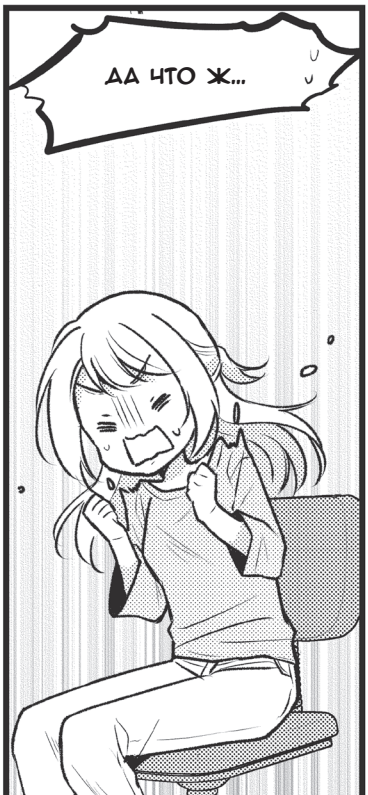
Ну… это… Проще говоря, можно представить вектор как обычную переменную и дифференцировать.

Вот и все! А теперь надо искать вес по формуле 1.8.

ГЛАВА 2

КАК ДЕЛАТЬ КЛАССИФИКАЦИЮ?


САЯКА-СЭМПАЙ ЗДЕСЬ?
БУ-
УХ!

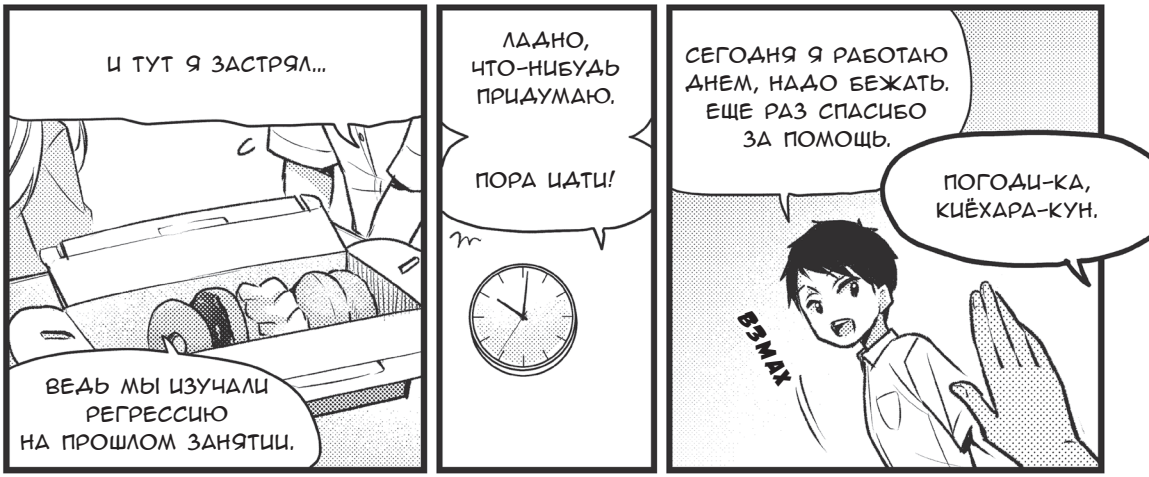

ДА ЧТО Ж...


...ТАКОЕ! В ЛАБОРАТОРИЮ НАДО ВХОДИТЬ
ТИШЕ!
ИЗВИНИТЕ...


ЧТО ОПЯТЬ СЛУЧИЛОСЬ?
ОПЯТЬ КОНСУЛЬТАНТ?


О-О-О!
Сладости!
ДА! СПАСИБО, БЛАГОДАРЯ ТЕБЕ ВСЕ ПРОШЛО КАК ПО МАСЛУ!
ВРУЧАЕТ
БЛАГОДАРЯ ЛИНЕЙНОЙ РЕГРЕССИИ Я СМОГ СПРОГНОЗИРОВАТЬ, ЧТО ГОСТЕЙ БУДЕТ БОЛЬШЕ.
УРД!
ПОСЛЕ ОБРАБОТКИ ДАННЫХ ТАК, КАК ВЫ РАССКАЗАЛИ, Я ПОГОВОРИЛ С ОТВЕТСТВЕННЫМИ ЛИЦАМИ И УБЕДИЛ ИХ ЗАКАЗАТЬ ПОБОЛЬШЕ НАПИТКОВ.
И МЕРОПРИЯТИЕ ПРОШЛО ПО ПЛАНУ И ИМЕЛО УСПЕХ!

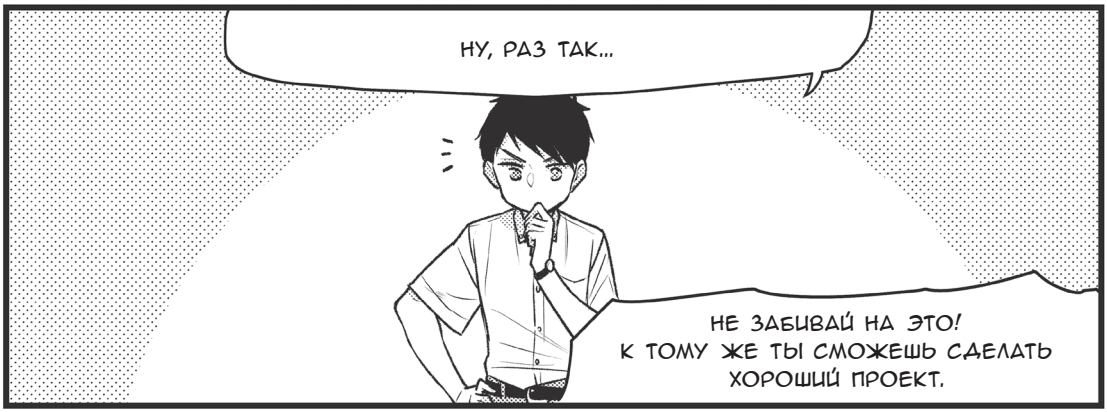

ХОРОШО, ЧТО ТЫ НЕ ДАЛ ОБВЕСТИ СЕБЯ ВОКРУГ ПАЛЬЦА.
ХРУМ
ХРУМ
ДА! СПАСИБО ОГРОМНОЕ, СЭМПАЙ!
И ТЕПЕРЬ ТЫ, КИЁХАРА-КУН, МОЖЕШЬ ПОСВЯТИТЬ СЕБЯ СВОИМ РАБОЧИМ ОБЯЗАННОСТЯМ!
НУ...
Я ЧТО-ТО НЕ ТО СКАЗАЛА?
ДА НЕТ...
ТЕПЕРЬ ВСЕ В МЭРИИ ДУМАЮТ ОБО МНЕ ИНАЧЕ...
А ОН, ВИДИМО, РАЗБИРАЕТСЯ В МАШИННОМ ОБУЧЕНИИ!
ЧТО?!
ОБО МНЕ ВДРУГ ЗАГОВОРИЛИ, И ТЕПЕРЬ ВСЕ СПРАШИВАЮТ У МЕНЯ СОВЕТА ПРО МАШИННОЕ ОБУЧЕНИЕ.
НАПРИМЕР?
В МОЕМ ОТДЕЛЕ ЗДРАВООХРАНЕНИЯ И БЛАГОСОСТОЯНИЯ ХОТЯТ УМЕНЬШИТЬ РАСХОДЫ НА МЕСТНУЮ МЕДИЦИНУ И В ТО ЖЕ ВРЕМЯ УВЕЛИЧИТЬ ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ ЛЮДЕЙ.
ВЕБ-ТЕСТ
Тест на диабет
ДА
НЕТ
Ответьте на вопросы - и узнаете
Начать
У НАС УВЕЛИЧИЛОСЬ КОЛИЧЕСТВО БОЛЬНЫХ ДИАБЕТОМ, И МЕНЯ ПОПРОСИЛИ ОПРЕДЕЛИТЬ, ИСХОДЯ ИЗ ДИАГНОСТИЧЕСКИХ ДАННЫХ, КТО НАХОДИТСЯ В ГРУППЕ РИСКА.
ДАННЫЕ ВСЕ УЖЕ СОБРАНЫ.
ТАК... ДАННЫХ ПОРЯДОЧНО, НО ЕСТЬ И ТЕ, У КОГО НЕ ВСЕ МЕДОСМОТРЫ ПРОЙДЕНЫ, ИЛИ ЖЕ НЕ ВСЕ ВРАЧИ... РЕГРЕССИЮ НЕ ПОСТРОИТЬ, НУЖНА КЛАССИФИКАЦИЯ ПО КАТЕГОРИЯМ.

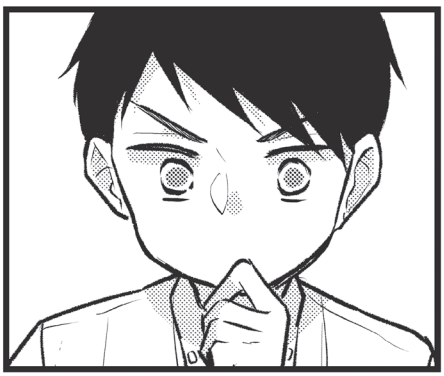

И ТУТ Я ЗАСТРЯЛ...
ВЕДЬ МЫ ИЗУЧАЛИ РЕГРЕССИЮ НА ПРОШЛОМ ЗАНЯТИИ.
ЛАДНО, ЧТО-НИБУДЬ ПРИДУМАЮ.
ПОРА ИДТИ!
СЕГОДНЯ Я РАБОТАЮ ДНЕМ, НАДО БЕЖАТЬ. ЕЩЕ РАЗ СПАСИБО ЗА ПОМОЩЬ.
ПОГОДИ-КА, КИЁХАРА-КУН.
ВЗМАХ


ХОЧЕШЬ, Я ПОМОГУ ТЕБЕ РАЗОБРАТЬСЯ С КЛАССИФИКАЦИЕЙ?
УДИВЛЕННО
ЧТО?


НЕТ-НЕТ, Я ЗАШЕЛ ТЕБЯ ОТБЛАГОДАРИТЬ!
НЕТ-НЕТ-НЕТ! ХВАТИТ С МЕНЯ БЕЗОТВЕТНОЙ ЛЮБВИ! Я БОЛЬШЕ ЭТОГО НЕ ВЫНЕСУ, ОНА МНЕ ТАК НРАВИТСЯ!


КИЁХАРА-КУН, ТЫ ЖЕ ХОЧЕШЬ, ЧТОБЫ ВСЕ ЖИТЕЛИ БЫЛИ ЗДОРОВЫМИ?
ЧТО?

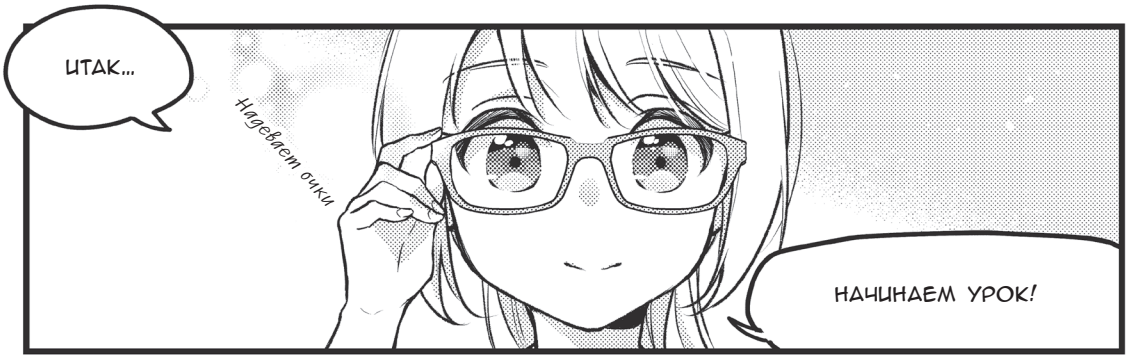

НУ, РАЗ ТАК...
НЕ ЗАБИВАЙ НА ЭТО!
К ТОМУ ЖЕ ТЫ СМОЖЕШЬ СДЕЛАТЬ ХОРОШИЙ ПРОЕКТ.




Д-ДА, ХОЧУ...
С ЧЕГО БЫ ТАКОЕ ДАВЛЕНИЕ...
ДАВЛЕНИЕ
КИЁХАРА-КУН, ТЫ БУДЕШЬ МОИМ УЧЕНИКОМ! Я НЕ УСПОКОЮСЬ, ЕСЛИ НЕ ПОЗАБОЧУСЬ О ТЕБЕ!


У ТЕБЯ ЕСТЬ ВРЕМЯ В СУББОТУ?
ДА, ЕСТЬ...
ВСТРЕТИМСЯ В СЕМЕЙНОМ РЕСТОРАНЧИКЕ В ДВА ЧАСА!
ХОРОШО... НЕТ, ВЕЛИКОЛЕПНО!
Denisan
IN
Denisan
Суббота


А МЫ ВЕДЬ ТУТ УЖЕ БЫЛИ...
ТЫ МНЕ КАК БРАТИК!
КИЁХАРА-КУН, ТЫ БУДЕШЬ МОИМ УЧЕНИКОМ!
ЗА НЕСКОЛЬКО ЛЕТ ПРОДВИНУЛСЯ ОТ БРАТИКА ДО УЧЕНИКА, НАЗЫВАЕТСЯ.
ПРОГРЕСС ЛИ ЭТО? РЕГРЕСС? НЕ ЗНАЮ!!!
В ОТЧАЯНИИ
КИЁХАРА-КУН!
ПОТИШЕ! ТЫ ЖЕ В РЕСТОРАНЕ!
САЯКА-СЭНСЭЙ!
ВЗВОЛНОВАННО
КАКАЯ ОНА МИЛАЯ!



2.1. Приводим данные в порядок

Определимся, какие данные нужны для пациентов в группе риска заболевания диабетом.

Есть данные за 10 лет, но они не совсем полные.

Пол	Возраст	ИМТ	Уровень глюкозы	Давление	Диабет
Ж	65	22	180	135	Нет
М	60	28	200	140	Да
М	75	21		120	Нет
Ж	72	25		140	Нет
М	65	26	210		Да
М	80	19	175	112	Нет

Недостающие значения

2.2. ОПРЕДЕЛЯЕМ КЛАСС ДАННЫХ

ПОЗИТИВНЫЙ ИЛИ НЕГАТИВНЫЙ.

ХВАЛЯТ ИЛИ РУГАЮТ ПРОДУКТ.

ПОЗИТИВНЫЙ

НЕГАТИВНЫЙ

САМАЯ ПРОСТАЯ ИЗ ЗАДАЧ КЛАССИФИКАЦИИ – ЭТО **БИНАРНАЯ КЛАССИФИКАЦИЯ**.

БИНАРНАЯ КЛАССИФИКАЦИЯ – ЭТО РАЗДЕЛЕНИЕ НА ДВЕ ЧАСТИ?

ИМЕННО. БОЛЕН ЧЕЛОВЕК ИЛИ НЕТ, ИДЕТ ПИСЬМО В СПАМ ИЛИ НЕТ – ЭТО ЗАДАЧА НА РАЗДЕЛЕНИЕ ПО ДВУМ КЛАССАМ.

Японские сладости или нет

ЖЕЛЕ АММИЦУ (ПОЗИТИВНОЕ ЗНАЧЕНИЕ)

ТОРТ (НЕГАТИВНОЕ ЗНАЧЕНИЕ)

Все остальное

Японская сладость

ПОКА ЧТО ДЛЯ ПРОСТОТЫ ПРЕДПОЛОЖИМ, ЧТО ВВОДЯТСЯ ТОЛЬКО ВЕКТОРЫ С ЧИСЛЕННЫМИ ЗНАЧЕНИЯМИ.

ЕСЛИ ВВОДЯТСЯ ДВУХМЕРНЫЕ ВЕКТОРЫ, ТО ДАННЫЕ МОЖНО РАСПОЛОЖИТЬ НА ПЛОСКОСТИ ВОТ ТАКИМ ОБРАЗОМ:

x_2

x_1

Позитивные значения

Негативные значения

ПИШЕТ

В ЗАВИСИМОСТИ ОТ КЛАССА ОБОЗНАЧИМ ИХ ЧЕРНЫМИ ИЛИ БЕЛЫМИ ТОЧКАМИ.

НЕ ВИДИШЬ? МОЖЕТ, ТЕБЕ ОТСЮДА НЕ ПОНЯТНО.

ЧТО? НЕТ! Я ВСЕ ВИЖУ, И МНЕ ВСЕ ПОНЯТНО!

МОЖЕТ МНЕ ПРОЩЕ ВСТАТЬ И ТАК ОБЪЯСНЯТЬ...

поднимается

НЕТ-НЕТ-НЕТ! И ТАК ПОНИМАЮ!

АААААААА

ИЛИ Я СЯДУ РЯДОМ...

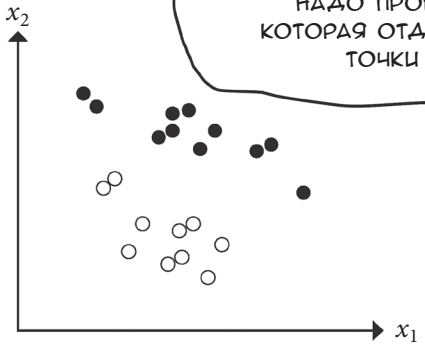
Д-Д-Д-Д-ДД. Я ПЕРЕСЯДУ!

БЛИН!.. ТАК БЛИЗКО!

ИТАК, ПОСМОТРИ НА РИСУНОК, ЧТО ТЫ ВИДИШЬ?

НУ...

КАК И В СЛУЧАЕ РЕГРЕССИИ, НАДО ПРОВЕСТИ ПРЯМУЮ, КОТОРАЯ ОТДЕЛИЛА БЫ ЧЕРНЫЕ ТОЧКИ ОТ БЕЛЫХ?



ИМЕННО! ПРОДОЛЖИМ МЫСЛИТЬ В ТОМ ЖЕ НАПРАВЛЕНИИ.

2.3. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

ИНЫМИ СЛОВАМИ, ЛОГИСТИЧЕСКУЮ РЕГРЕССИЮ МОЖНО ПРЕДСТАВИТЬ КАК РАСШИРЕНИЕ ЗАДАЧИ РЕГРЕССИИ. ПОГОВОРИМ О ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ!

Логистическая регрессия – метод нахождения функции на основании ввода взвешенной суммы данных; если элемент данных принадлежит к положительному классу, то выход формулы регрессии близок к 1, а если к отрицательному, то он будет близок к 0.

ДА!

В СЛУЧАЕ ЗАДАЧИ БИНАРНОЙ КЛАССИФИКАЦИИ, КАК СЕЙЧАС, ПРИЗНАКОВОМУ ОПИСАНИЮ ОБЪЕКТА $\boldsymbol{x} = (x_1, \ldots x_d)^T$ СООТВЕТСТВУЕТ ВЗВЕШЕННАЯ СУММА ВСЕХ ПРИЗНАКОВ $w_1x_1 + \ldots + w_dx_d$.

НЕОБХОДИМО СКОРРЕКТИРОВАТЬ ВЕС ТАК, ЧТОБЫ ДЛЯ ПОЛОЖИТЕЛЬНЫХ ПРИМЕРОВ ФУНКЦИЯ ПРИНИМАЛА ЗНАЧЕНИЯ, БЛИЗКИЕ К 1, А ДЛЯ ОТРИЦАТЕЛЬНЫХ – БЛИЗКИЕ К 0.

● черный кружок × вес = близко к 1

○ белый кружок × вес = близко к 0

Распределение весов

ИНЫМИ СЛОВАМИ, ЕСЛИ НАСТРОИТЬ ВЕС ТАК, ЧТО РЕЗУЛЬТАТ ФОРМУЛЫ РЕГРЕССИИ БУДЕТ РАВЕН 1 ДЛЯ ПОЛОЖИТЕЛЬНЫХ ПРИМЕРОВ И 0 ДЛЯ ОТРИЦАТЕЛЬНЫХ, ТО НЕВОЗМОЖНО РЕШИТЬ ЭТО УРАВНЕНИЕ, ЕСЛИ $x = 0$; ПОЭТОМУ В КАЧЕСТВЕ ПАРАМЕТРА ДОБАВЛЯЕТСЯ ПОСТОЯННАЯ w_0.

$$\hat{g}(\boldsymbol{x}) = w_0 + w_1x_1 + \cdots + w_dx_d = w_0 + \boldsymbol{w}^T\boldsymbol{x}$$

СКРИП

И ПОЛУЧАЕТСЯ ТАКАЯ ФОРМУЛА:

$$\hat{g}(\boldsymbol{x}) = w_0 + w_1 x_1 + \cdots + w_d x_d = w_0 + \boldsymbol{w}^T \boldsymbol{x}$$

здесь $\boldsymbol{w}^T\boldsymbol{x}$ – внутреннее произведение вектора $\boldsymbol{w}$ на вектор $\boldsymbol{x}$. это можно определить как сумму произведений соответствующих элементов.

поскольку для гиперплоскости $\hat{g}(x) = 0$, d-мерную гиперплоскость можно выразить уравнением. если плоскость ведет себя, как указано выше, то что будет после того, как все определено?

Поскольку трудно оценить вероятность принадлежности x к положительному классу, надо сделать так, чтобы разброс результатов на выходе функции $\hat{g}(x)$ колебался от 0 до 1; если x принадлежит к положительному классу, то результат должен быть близок к 1, а если к отрицательному, то близок к 0.

И как это можно сделать?

Если мы возьмем преобразованную функцию $p(+|x) = \frac{1}{1 + e^{-(w_0 + \boldsymbol{w}^T \boldsymbol{x})}}$, то по формуле ниже можно получить вероятность принадлежности x к положительному классу.

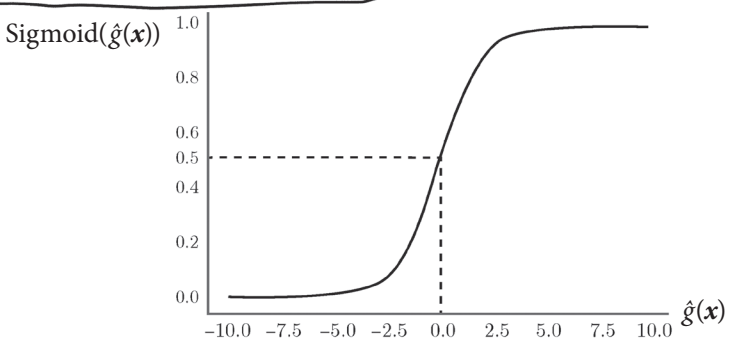
$$p(+|x) = \frac{1}{1 + e^{-(w_0 + \boldsymbol{w}^T \boldsymbol{x})}}$$

А какова вероятность того, что x принадлежит к отрицательному классу?

Если вычесть вероятность положительного класса из 1, то $p(-|x) = 1 - p(+|x)$?

Именно!

Вот график этой функции.



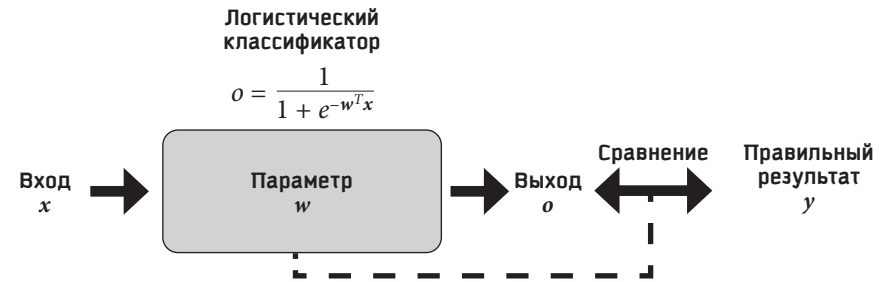
Если посмотреть на график, ясно, что какое бы значение ни приняла функция $\hat{g}(x) = w_0 + \boldsymbol{w}^T \boldsymbol{x}$, результат будет находиться в диапазоне от 0 до 1. Это – сигмоидная функция.

Если $\hat{g}(x) = 0$, то функция примет значение 0,5.

Это функция, показывающая значение вероятности.

Далее поговорим о том, как провести обучение логистической классификации. Логистический классификатор можно рассматривать как вероятностную модель с весовым параметром $\boldsymbol{w}$.

Здесь и далее для простоты объяснения $\boldsymbol{w}$ будет включать в себя w_0.



Пусть в данных для обучения D этой модели при входе $\boldsymbol{x}_i$, а на выходе o_i. Желаемый выход обозначим положительным исходом y_i. Предположим, что у нас задача бинарной классификации, $y_i = 1$, если данные принадлежат к положительному классу, а если к отрицательному, то $y_i = 0$.

Чтобы правильно провести обучение созданной модели и оценить значение, необходимо уравнение правдоподобия, которое приведено ниже. $\prod$ обозначает произведение.

$$P(D \mid \boldsymbol{w}) = \prod_{\boldsymbol{x}_i \in D} o_i^{y_i} (1 - o_i)^{(1-y_i)}.$$

$o_i^{y_i}(1 - o_i)^{(1-y_i)}$ принимает значение o_i, если элемент данных принадлежит к положительному классу, и тогда ($y_i = 1$), и значение $(1 - o_i)$, если к отрицательному ($y_i = 0$). Иными словами, если настроить веса $\boldsymbol{w}$ так, что при положительном результате выходное значение o_i будет близко к 1, а при отрицательном результате o_i будет близко к 0, то во всех данных величина произведения $P(D|\boldsymbol{w})$ будет увеличиваться в зависимости от данных.

При расчете значения максимального правдоподобия для простоты расчета используется логарифмическая функция правдоподобия.

$$L(D) = \log P(D \mid \boldsymbol{w}) = \sum_{\boldsymbol{x}_i \in D} \{y_i \log o_i + (1 - y_i) \log(1 - o_i)\}.$$

Чтобы представить ясней задачу оптимизации, в дальнейшем будем рассматривать задачу минимизации функции ошибки $E(\boldsymbol{w})$, которая может быть определена как логарифмическая функция со знаком минус.

$E(\boldsymbol{w}) = -\log P(D|\boldsymbol{w})$.

Продифференцировав, найдем предельное значение $\boldsymbol{w}$. Поскольку модель – логистический классификатор, выход o_i будет представлять собой сигмоидную функцию.

$$S(z) = \frac{1}{1 + e^{-z}}.$$

Производная сигмоидной функции будет выглядеть так:

$S'(z) = S(z) \cdot (1 - S(z))$.

Поскольку на выходе модели есть функция веса w, при ее изменении меняется величина ошибки. Решение таких задач можно найти методом градиентного спуска. Метод градиентного спуска – это метод сходимости к оптимальному решению путем многократного постепенного уменьшения параметров в направлении градиента минимизируемой функции.

В этом случае мы немного меняем параметр $\boldsymbol{w}$, чтобы найти направление наискорейшего спуска функции ошибки $E(\boldsymbol{w})$. Это «немного» обозначается коэффициентом обучаемости η. Используя метод градиентного спуска, можно вывести новую формулу веса.

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \frac{\partial E(w)}{\partial w}.$$

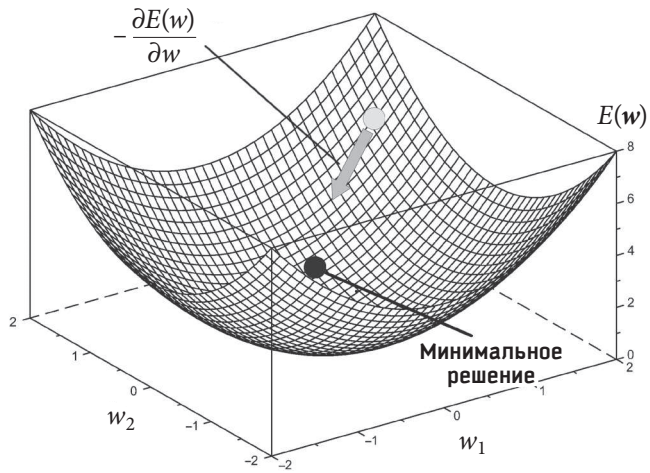
Затем вычислим направление градиента функции ошибки $E(\boldsymbol{w})$ по формуле ниже:

$$\frac{\partial E(w)}{\partial w} = -\sum_{\boldsymbol{x}_i \in D}\left(\frac{y_i}{o_i} - \frac{1-y_i}{1-o_i}\right) o_i(1-o_i)\boldsymbol{x}_i = -\sum_{\boldsymbol{x}_i \in D}(y_i - o_i)\boldsymbol{x}_i.$$

Следовательно, новая формула веса будет выглядеть так:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \eta \sum_{\boldsymbol{x}_i \in D}(y_i - o_i)\boldsymbol{x}_i.$$

Когда новое значение веса будет ниже заранее определенного значения, метод градиентного спуска закончен.



Метод, при котором градиент вычисляют на основании всех данных для обучения D, называется **пакетным** (batch method). Если из D выбирают данные определенной величины и вычисляют отдельный градиент по ним, то это называется **мини-пакетный метод** (mini batch), а метод, когда из D случайно выбирают элемент данных и вычисляют градиент для него, называется **методом стохастического градиентного спуска**.

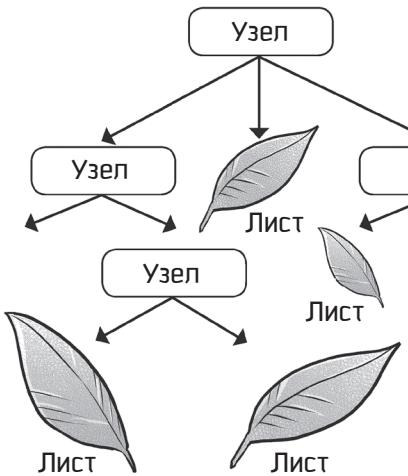
2.4. КЛАССИФИКАЦИЯ ПО РЕШАЮЩЕМУ ДЕРЕВУ

ДАЛЕЕ – КЛАССИФИКАЦИЯ ПУТЕМ **РЕШАЮЩЕГО ДЕРЕВА**... ВОЗМОЖНО, ЭТОТ СПОСОБ ПОДХОДИТ ДЛЯ ОПРЕДЕЛЕНИЯ ТЕХ, КТО РИСКУЕТ ЗАБОЛЕТЬ ДИАБЕТОМ.

ДЕРЕВО? КОТОРОЕ РАСТЕТ?

ДА, ТОЛЬКО КОРНИ У НЕГО БУДУТ СВЕРХУ, А ЛИСТЬЯ – СНИЗУ.

Структура решающего дерева



Решающее дерево состоит из узлов (веток), в которых данные классифицируются, и листьев, в которых выводится результат классификации.

МОЖНО ПРЕОБРАЗОВАТЬ ДЕРЕВО В ЭКВИВАЛЕНТНОЕ ЛОГИЧЕСКОЕ ВЫРАЖЕНИЕ, КОМБИНИРУЯ ЗНАЧЕНИЯ ВЕТВЕЙ УЗЛОВ, КОТОРЫЕ ВЕДУТ К ПОЛОЖИТЕЛЬНОМУ РЕЗУЛЬТАТУ С УСЛОВИЕМ *и*, А ВСЕ ОСТАЛЬНОЕ – С УСЛОВИЕМ *или*.

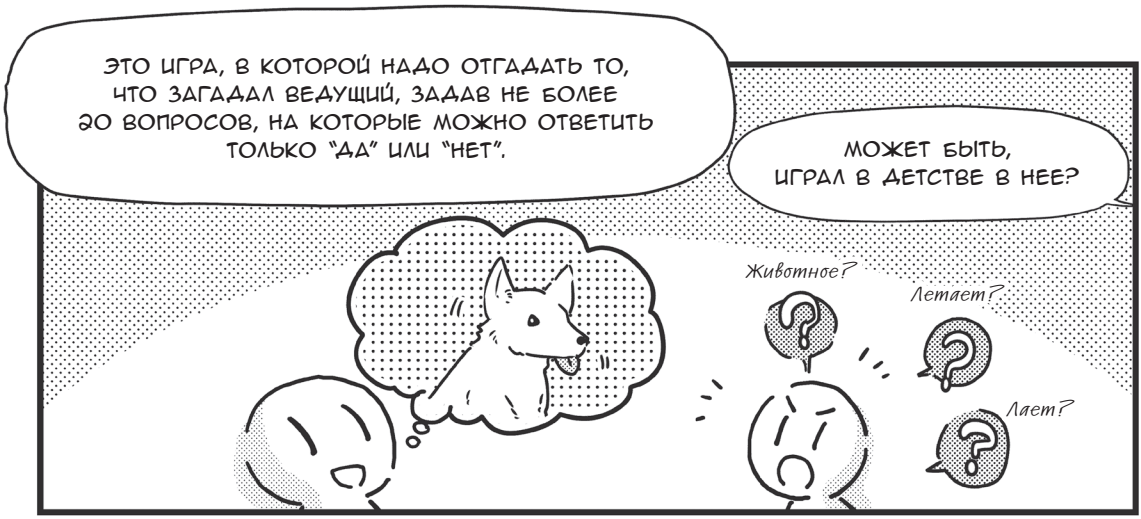
ОГО!

ОДНАКО ПОСКОЛЬКУ ПО СТРУКТУРЕ ДЕРЕВА ЛЕГЧЕ НАГЛЯДНО ПРЕДСТАВИТЬ РЕЗУЛЬТАТ ОБУЧЕНИЯ, ТО ЭТОТ МЕТОД ПРЕДПОЧТИТЕЛЕН.

РЕШАЮЩЕЕ ДЕРЕВО ПРОЩЕ ПОНЯТЬ, ЧЕМ ЛОГИЧЕСКИЕ СХЕМЫ.

ОБЪЯСНИТЬ, КАК ОНО РАБОТАЕТ, МОЖНО НА ПРИМЕРЕ ИГРЫ В "ДАНЕТКИ".

ДАНЕТКИ?


ЭТО ИГРА, В КОТОРОЙ НАДО ОТГАДАТЬ ТО, ЧТО ЗАГАДАЛ ВЕДУЩИЙ, ЗАДАВ НЕ БОЛЕЕ 20 ВОПРОСОВ, НА КОТОРЫЕ МОЖНО ОТВЕТИТЬ ТОЛЬКО "ДА" ИЛИ "НЕТ".
МОЖЕТ БЫТЬ, ИГРАЛ В ДЕТСТВЕ В НЕЕ?
Животное?
Летает?
Лает?

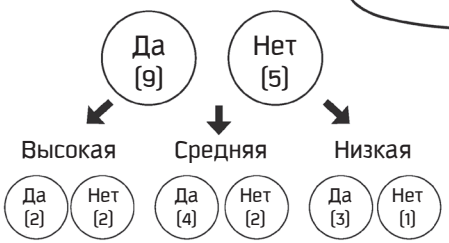

СЕКРЕТ УСПЕХА В ТОМ, ЧТОБЫ СНАЧАЛА НЕ ЗАДАВАТЬ СЛИШКОМ КОНКРЕТНЫЕ ВОПРОСЫ.
ЕСЛИ ПЕРВЫЕ ВОПРОСЫ БУДУТ КОНКРЕТНЫМИ, МОЖНО СУЗИТЬ СПИСОК КАНДИДАТОВ, НО В БОЛЬШИНСТВЕ СЛУЧАЕВ НИКАКОЙ КОНКРЕТНОЙ ИНФОРМАЦИИ УЗНАТЬ НЕ ПОЛУЧИТСЯ.
Оно ядовитое?


СНАЧАЛА НАДО ЗАДАВАТЬ ОБЩИЕ ВОПРОСЫ, ЧТОБЫ ПОНЯТЬ, ЧТО ЭТО?
ИМЕННО!


РЕШАЮЩЕЕ ДЕРЕВО СТРОИТСЯ ИМЕННО НА ОСНОВЕ ЭТОГО СЕКРЕТА – ВСЕ ВОПРОСЫ, НА КОТОРЫЕ ДАЕТСЯ ОТВЕТ, ЯВЛЯЮТСЯ УЗЛАМИ ДАННОГО ДЕРЕВА.
В ОСНОВНОМ ДЛЯ ПОСТРОЙКИ ТАКОГО ДЕРЕВА ИСПОЛЬЗУЕТСЯ **АЛГОРИТМ ID3**.

	Погода	Температура	Влажность	Ветер	Играем?
1	Ясно	Высокая	Высокая	Нет	Нет
2	Ясно	Высокая	Высокая	Да	Нет
3	Облачно	Высокая	Высокая	Нет	Да
4	Дождь	Средняя	Высокая	Нет	Да
5	Дождь	Низкая	Стандартная	Нет	Да
6	Дождь	Низкая	Стандартная	Да	Нет
7	Облачно	Низкая	Стандартная	Да	Да
8	Ясно	Средняя	Высокая	Нет	Нет
9	Ясно	Низкая	Стандартная	Нет	Да
10	Дождь	Средняя	Стандартная	Нет	Да
11	Ясно	Средняя	Стандартная	Да	Да
12	Облачно	Средняя	Высокая	Да	Да
13	Облачно	Высокая	Стандартная	Нет	Да
14	Дождь	Средняя	Высокая	Да	Нет

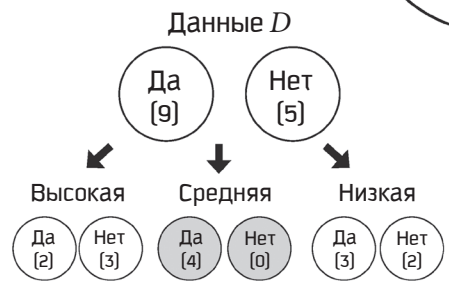
Допустим, мы зададим вопрос про влажность. В зависимости от ответа данные будут делиться так, как показано ниже:



Если мы сначала спросим про влажность, то какой бы ответ ни был выбран, условия для задания второго вопроса не будут отличаться от предыдущих.

А если мы выберем погоду?

В случае облачной погоды все результаты положительные.



Если мы сначала спросим про погоду, то в случае облачной погоды (основываясь на данных) все результаты будут положительными.

Если выберем только ясную погоду, данные будут такими:

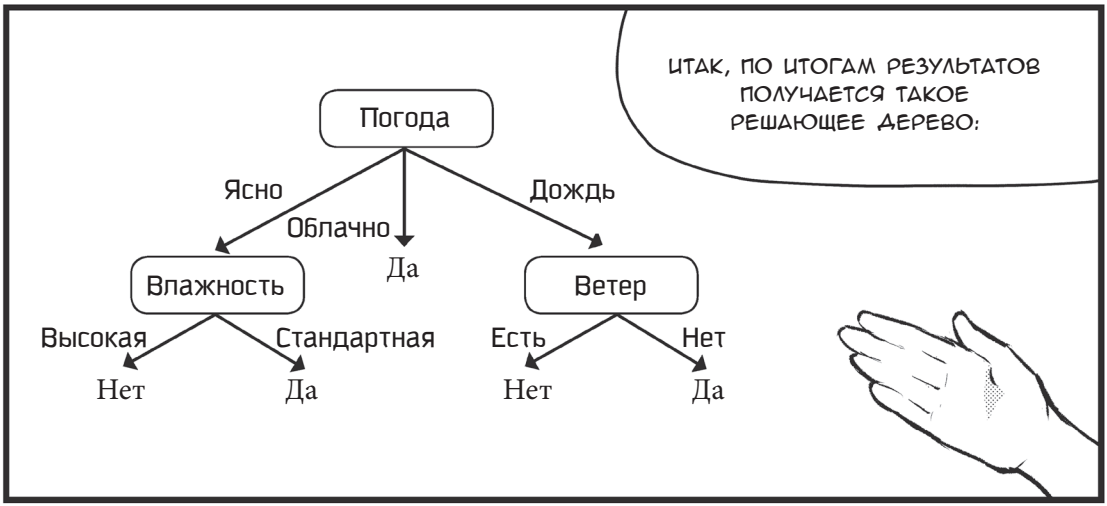
	Погода	Температура	Влажность	Ветер	Играем?
1	Ясно	Высокая	Высокая	Нет	Нет
2	Ясно	Высокая	Высокая	Да	Нет
8	Ясно	Средняя	Высокая	Нет	Нет
9	Ясно	Низкая	Стандартная	Нет	Да
11	Ясно	Средняя	Стандартная	Да	Да

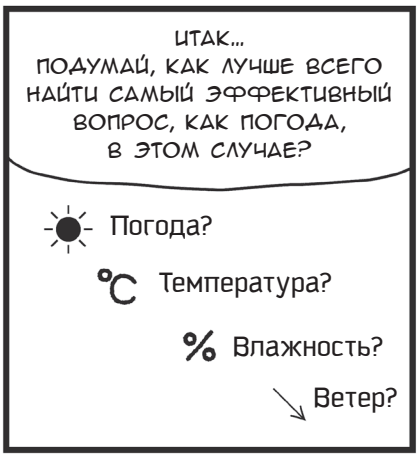
А если теперь спросить про влажность, то все ответы будут "да", если влажность стандартная.

А если мы выберем дождь, посмотрим, что получится...

	Погода	Температура	Влажность	Ветер	Играем?
4	Дождь	Средняя	Высокая	Нет	Да
5	Дождь	Низкая	Стандартная	Нет	Да
6	Дождь	Низкая	Стандартная	Да	Нет
10	Дождь	Средняя	Стандартная	Нет	Да
14	Дождь	Средняя	Высокая	Да	Нет

Значит, если дождь, надо смотреть, есть или нет ветер.

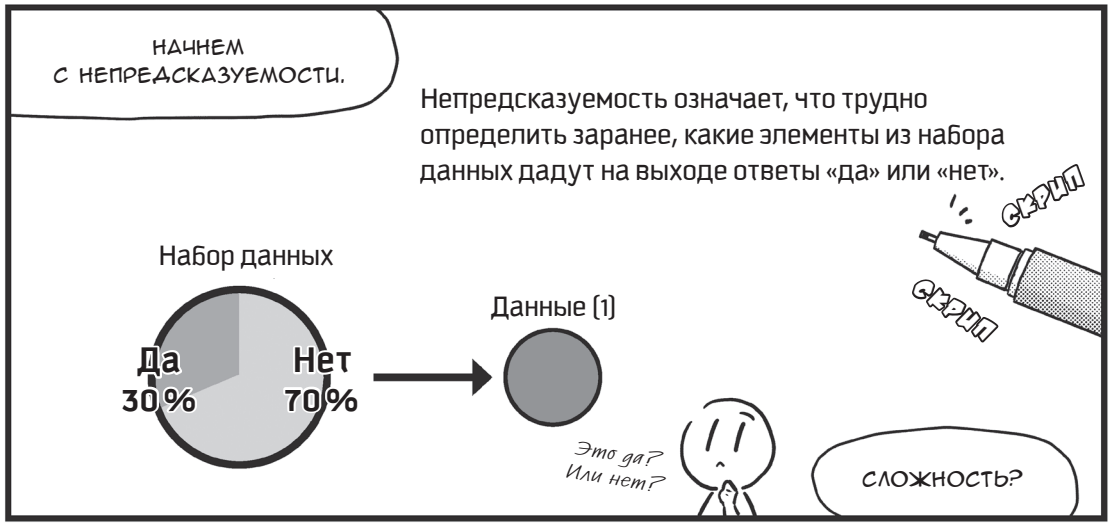

ИТАК, ПО ИТОГАМ РЕЗУЛЬТАТОВ ПОЛУЧАЕТСЯ ТАКОЕ РЕШАЮЩЕЕ ДЕРЕВО:
Погода
Ясно
Облачно
Дождь
Да
Влажность
Ветер
Высокая
Стандартная
Есть
Нет
Нет
Да
Нет
Да


ИТАК...
ПОДУМАЙ, КАК ЛУЧШЕ ВСЕГО НАЙТИ САМЫЙ ЭФФЕКТИВНЫЙ ВОПРОС, КАК ПОГОДА, В ЭТОМ СЛУЧАЕ?
Погода?
Температура?
Влажность?
Ветер?


ХОРОШИЙ ВОПРОС...

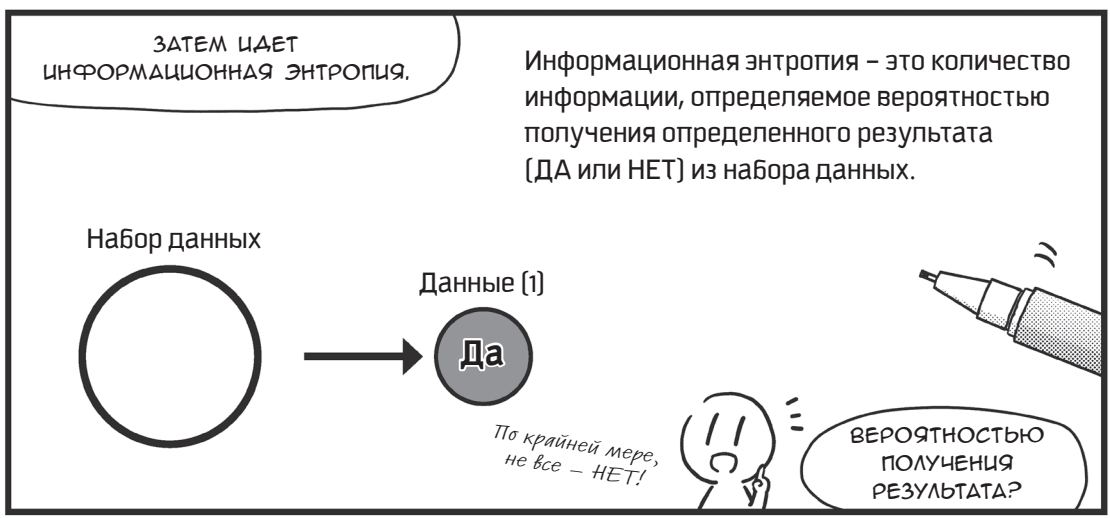

КАЖЕТСЯ, НАДО СРАВНИТЬ ЧИСЛОВЫЕ ВЕЛИЧИНЫ, КОТОРЫЕ МОЖНО ОЦЕНИТЬ.
ИМЕННО! ЕСЛИ МЫ ПЕРЕВЕДЕМ ПАРАМЕТРЫ В ЧИСЛА, ТО МОЖЕМ СДЕЛАТЬ СРАВНЕНИЕ.


КЛЮЧЕВЫЕ СЛОВА В ДАННОМ СЛУЧАЕ –
НЕПРЕДСКА-ЗУЕМОСТЬ И
НЕПРЕДСКА-ЗУЕМОСТЬ
ИНФОРМАЦИОННАЯ ЭНТРОПИЯ.
ИНФОРМАЦИОННАЯ ЭНТРОПИЯ
ЧТО ВСЕ ЭТО ЗНАЧИТ?


НАЧНЕМ С НЕПРЕДСКАЗУЕМОСТИ.
Непредсказуемость означает, что трудно определить заранее, какие элементы из набора данных дадут на выходе ответы «да» или «нет».
СКРИП
СКРИП
Набор данных
Да 30 %
Нет 70 %
Данные (1)
Это да? Или нет?
СЛОЖНОСТЬ?

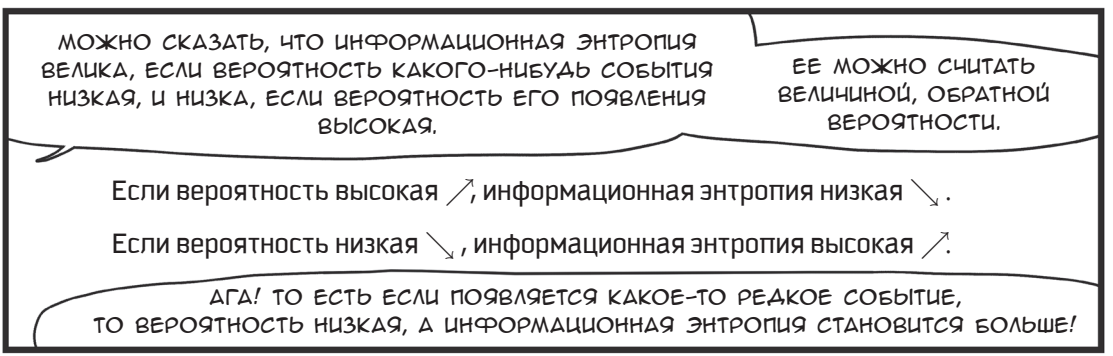

САМЫЙ ТРУДНЫЙ СЛУЧАЙ – ЭТО КОГДА ПОЛОВИНА ОТВЕТОВ "ДА", А ПОЛОВИНА – "НЕТ". А КАКОЙ ТОГДА САМЫЙ ЛЕГКИЙ СЛУЧАЙ?
Нет 50 %
Да 50 %
Нет 100 %
Да 100 %
КОГДА ВСЕ ОТВЕТЫ "ДА" ИЛИ ВСЕ ОТВЕТЫ "НЕТ".


ДА... ИНЫМИ СЛОВАМИ, ВСЕ ДАННЫЕ МОЖНО ОТНЕСТИ К ОДНОМУ И ТОМУ ЖЕ КЛАССУ.

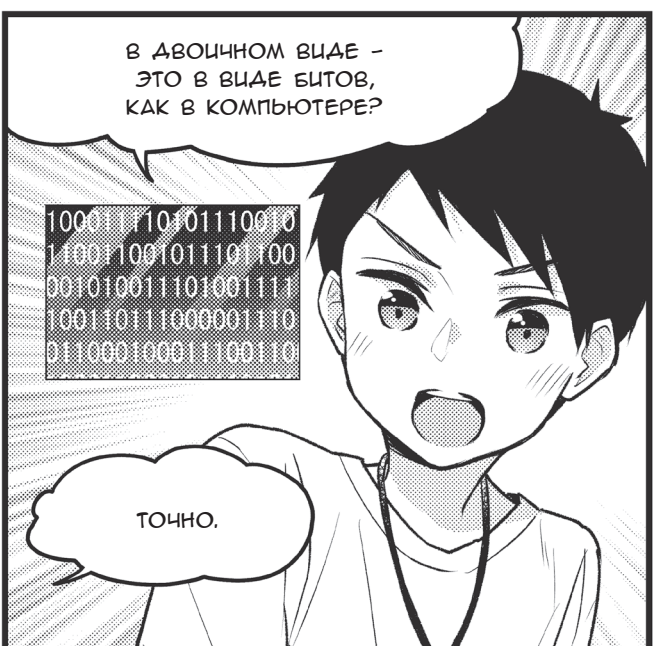

ЗАТЕМ ИДЕТ ИНФОРМАЦИОННАЯ ЭНТРОПИЯ.
Информационная энтропия – это количество информации, определяемое вероятностью получения определенного результата (ДА или НЕТ) из набора данных.
Набор данных
Данные (1)
Да
По крайней мере, не все – НЕТ!
ВЕРОЯТНОСТЬЮ ПОЛУЧЕНИЯ РЕЗУЛЬТАТА?


НАПРИМЕР, ЕСЛИ МЫ ПОЛУЧИЛИ ОТВЕТ "ДА" ИЗ НАБОРА ДАННЫХ, ГДЕ ВСЕ ДАННЫЕ – "ДА", МЫ НЕ ПОЛУЧИЛИ НИКАКОЙ ИНФОРМАЦИИ.
Да!
Да!
Да!
НАВЕРНОЕ, ТАК.


А ЕСЛИ ИЗ 14 ЭЛЕМЕНТОВ ДАННЫХ 13 БУДУТ "ДА", А ОДИН – "НЕТ"?
Да!
Да!
Нет!
У НАС БУДЕТ ИНФОРМАЦИЯ О ТОМ, ЧТО СЛУЧИЛОСЬ ЧТО-ТО НЕОБЫЧНОЕ.


МОЖНО СКАЗАТЬ, ЧТО ИНФОРМАЦИОННАЯ ЭНТРОПИЯ ВЕЛИКА, ЕСЛИ ВЕРОЯТНОСТЬ КАКОГО-НИБУДЬ СОБЫТИЯ НИЗКАЯ, И НИЗКА, ЕСЛИ ВЕРОЯТНОСТЬ ЕГО ПОЯВЛЕНИЯ ВЫСОКАЯ.
ЕЕ МОЖНО СЧИТАТЬ ВЕЛИЧИНОЙ, ОБРАТНОЙ ВЕРОЯТНОСТИ.
Если вероятность высокая ↗, информационная энтропия низкая ↘.
Если вероятность низкая ↘, информационная энтропия высокая ↗.
АГА! ТО ЕСТЬ ЕСЛИ ПОЯВЛЯЕТСЯ КАКОЕ-ТО РЕДКОЕ СОБЫТИЕ, ТО ВЕРОЯТНОСТЬ НИЗКАЯ, А ИНФОРМАЦИОННАЯ ЭНТРОПИЯ СТАНОВИТСЯ БОЛЬШЕ!


ДА... ЕСЛИ ВЫЧИСЛИТЬ ЛОГАРИФМ ПО ОСНОВАНИЮ 2 ДЛЯ ВЕРОЯТНОСТИ, Т. Е. ОБРАТНОЙ ВЕЛИЧИНЫ ЭНТРОПИИ, МОЖНО УЗНАТЬ КОЛИЧЕСТВО ЦИФР, НЕОБХОДИМОЕ ДЛЯ ПРЕДСТАВЛЕНИЯ ЭТОЙ ИНФОРМАЦИИ В ДВОИЧНОМ ВИДЕ.


В ДВОИЧНОМ ВИДЕ – ЭТО В ВИДЕ БИТОВ, КАК В КОМПЬЮТЕРЕ?
ТОЧНО.

А ТЕПЕРЬ РАССМОТРИМ ФОРМУЛУ НЕПРЕДСКАЗУЕМОСТИ ДАННЫХ.

ПОЖАЛУЙСТА.

ЕЕ МОЖНО НАЙТИ, ЕСЛИ МЫ СЛОЖИМ ПРОИЗВЕДЕНИЯ ИНФОРМАЦИОННОЙ ЭНТРОПИИ КАЖДОГО КЛАССА И СООТНОШЕНИЕ КАЖДОГО КЛАССА С ОБЩИМИ ДАННЫМИ, И ЗАПИСАТЬ СЛЕДУЮЩЕЙ ФОРМУЛОЙ:

$$E(D) = -P_{\text{Да}}\log_2 P_{\text{Да}} - P_{\text{Нет}}\log_2 P_{\text{Нет}}$$

КОГДА МЫ ЗАДАЕМ ВОПРОС, ТО ДАННЫЕ ВЕДЬ МОЖНО РАЗДЕЛИТЬ В ЗАВИСИМОСТИ ОТ ОТВЕТА? ЗАТЕМ МОЖНО НАЙТИ НЕПРЕДСКАЗУЕМОСТЬ ПО ФОРМУЛЕ, КОТОРАЯ ПРИВЕДЕНА ВЫШЕ.

МОЖНО ОПРЕДЕЛИТЬ ЗНАЧЕНИЕ СНИЖЕНИЯ НЕПРЕДСКАЗУЕМОСТИ КАК ЗНАЧЕНИЕ ИНФОРМАЦИОННОГО ВЫИГРЫША, И ТАМ, ГДЕ ЭТО ЗНАЧЕНИЕ ВЫШЕ ВСЕГО, НАХОДИТСЯ ВОПРОС, КОТОРЫЙ ПОМОЖЕТ УМЕНЬШИТЬ ВЕРОЯТНОСТЬ РАЗНЫХ ОТВЕТОВ.

А ТЕПЕРЬ РАССЧИТАЕМ НАШИ ДАННЫЕ ПО ГОЛЬФУ!

Следуя шагам 1–5, рассчитаем непредсказуемость и соотношение прироста информации в данных для гольфа.

Шаг 1

Так как в наборе данных D ответов «да» 9, а ответов «нет» – 5, то рассчитаем непредсказуемость по формуле ниже.

$$E(D) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = -0.643 \times (-0.637) - 0.357 \times (-1.495) = 0.94.$$

Шаг 2

Найдем непредсказуемость данных соответственно для ясной погоды, облачной и дождя.

$$E(\text{ясно}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = -0.4 \times (-1.32) - 0.6 \times (-0.74) = 0.971.$$

$$E(\text{облачно}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0 - 0 = 0.$$

$$E(\text{дождь}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = -0.6 \times (-0.74) - 0.4 \times (-1.32) = 0.971.$$

Шаг 3

Возьмем эти величины в качестве весов к данным и рассчитаем непредсказуемость после разделения.

$$\frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{4}{14} \times 0.971 = 0.694.$$

Шаг 4

Вычтя из непредсказуемости изначальных данных величину непредсказуемости данных после разделения, найдем отношение информационного выигрыша к внутренней информации, или же Gain.

Gain(*D*, погода) = 0.94 – 0.694 = 0.246.

Шаг 5

Таким же методом рассчитаем информационный выигрыш для других данных.

Gain(*D*, температура) = 0.029.

Gain(*D*, влажность) = 0.151.

Gain(*D*, ветер) = 0.048.

Следовательно, если первым вопросом для деления данных должен стать вопрос о погоде, непредсказуемость будет самой большой, а потом будет уменьшаться. После разделения данных можно использовать оставшиеся признаки в том же порядке.

В качестве метода расчета непредсказуемости данных вместо вышеуказанного способа можно воспользоваться коэффициентом Джини.

$$\text{Gini}(D) = 1 - P_{\text{Да}}^2 - P_{\text{Нет}}^2$$


ТАК МОЖНО ЧИСЛЕННО СРАВНИТЬ ВЕЛИЧИНЫ ЭФФЕКТИВНЫХ ВОПРОСОВ?


В ОСНОВЕ МЕТОДА, КОТОРЫМ МЫ СЕГОДНЯ ПОЛЬЗОВАЛИСЬ, ЛЕЖИТ "БРИТВА ОККАМА", КОТОРАЯ ГЛАСИТ: "ВЫБИРАЙ САМУЮ ПРОСТУЮ ГИПОТЕЗУ ДЛЯ ПРИМЕНЕНИЯ К ДАННЫМ".
БРИТВА ОККАМА
САМУЮ ПРОСТУЮ?


ЕСЛИ ГИПОТЕЗА СЛИШКОМ СЛОЖНАЯ, ТО ПОЛУЧЕННЫЕ ДАННЫЕ МОЖНО ОБЪЯСНИТЬ СЛУЧАЙНОСТЬЮ, НЕ ТАК ЛИ? А ЕСЛИ ОНА ПРОСТАЯ, ТО ВЕРОЯТНОСТЬ СЛУЧАЙНОГО ОБЪЯСНЕНИЯ ДАННЫХ ПАДАЕТ.
Простая (короткая) гипотеза
Сложная (длинная) гипотеза


ОДНАКО ЕСЛИ, СЛЕДУЯ ЭТОМУ МЕТОДУ, ПОСТРОИТЬ РЕШАЮЩЕЕ ДЕРЕВО, В КОТОРОМ НЕ БУДЕТ ОШИБОК, ТО МОЖНО ДОБИТЬСЯ ПЕРЕОБУЧЕНИЯ –
КОГДА ДЕРЕВО БУДЕТ СЛИШКОМ ХОРОШО ПОДХОДИТЬ К ДАННЫМ ДЛЯ ОБУЧЕНИЯ.
ПЕРЕ-ОБУЧЕНИЯ?


ТЫ ХОЧЕШЬ СКАЗАТЬ, ЧТО ЕСЛИ МЫ ПРИ ОБУЧЕНИИ ПОЛУЧИЛИ НЕБОЛЬШОЕ РЕШАЮЩЕЕ ДЕРЕВО, ТО ОНО И БУДЕТ ИСПОЛЬЗОВАТЬСЯ ДАЛЬШЕ?
Может использоваться!
ЕСЛИ МЫ СМОГЛИ ПРОВЕСТИ ОБУЧЕНИЕ НЕБОЛЬШОГО ДЕРЕВА, ПОСТРОЕННОГО С УЧЕТОМ ID3-АЛГОРИТМА, ТО ВЕРОЯТНОСТЬ СЛУЧАЙНОСТЕЙ БУДЕТ СНИЖЕНА. ДРУГИМИ СЛОВАМИ, СЛУЧАЙНОСТИ НЕ БУДЕТ, БУДЕТ ЗАКОНОМЕРНОСТЬ.

ДО ЭТОГО МЫ СТРОИЛИ РЕШАЮЩЕЕ ДЕРЕВО, КЛАССИФИЦИРУЯ ПО КАТЕГОРИЯМ, А В СЛУЧАЕ ЧИСЛЕННЫХ ПРИЗНАКОВ ПОПРОБУЕМ ОБУЧИТЬ ДЕРЕВО ПО МОДЕЛИ **ДИСКРЕТИЗАЦИИ**, КОТОРАЯ РАЗДЕЛЯЕТ РЯДЫ ЧИСЛОВЫХ ЗНАЧЕНИЙ НА НЕСКОЛЬКО ГРУПП.

МОЖНО РАЗДЕЛЯТЬ И ЧИСЛЕННЫЕ ПРИЗНАКИ.

Классификация по категориям

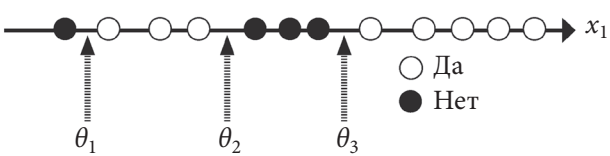
	Погода	Температура	Влажность	Ветер	Играем?
1	Ясно	Высокая	Высокая	Нет	Нет
2	Ясно	Высокая	Высокая	Да	Нет
3	Облачно	Высокая	Высокая	Нет	Да
4	Дождь	Средняя	Высокая	Нет	Да
5	Дождь	Низкая	Стандартная	Нет	Да

Классификация по численным признакам

ID	Радиус	Текстура	Окружность	Опухоль
44	13.17	21.81	85.42	Злокачественная
45	18.65	17.60	123.7	Злокачественная
46	8.20	16.84	51.71	Доброкачественаая
47	13.17	18.66	85.98	Злокачественная
48	12.02	14.63	78.04	Доброкачественаая

ПОСКОЛЬКУ МЫ ХОТИМ НАЙТИ МЕСТА, ГДЕ НЕПРЕДСКАЗУЕМОСТЬ НИЗКАЯ, МЫ НЕ РАЗДЕЛЯЕМ ОДИНАКОВЫЕ КЛАССЫ.

КОГДА МЫ ИЩЕМ ГРАНИЦУ КЛАССА, ОНА БУДЕТ ПОКАЗАНА ВЕРТИКАЛЬНОЙ СТРЕЛКОЙ НА РИСУНКЕ. ЭТО ПОГРАНИЧНОЕ ЗНАЧЕНИЕ БУДЕТ СРЕДНИМ ОТ ЗНАЧЕНИЙ ДО И ПОСЛЕ НЕГО.



ВЫБЕРЕМ МЕСТО, ГДЕ НАИБОЛЕЕ ВЫСОК ИНФОРМАЦИОННЫЙ ВЫИГРЫШ. ВЫПОЛНЯЯ ТЕ ЖЕ ВЫЧИСЛЕНИЯ, ЧТО И В СЛУЧАЕ КАТЕГОРИЙ, МЫ ОБНАРУЖИМ, ЧТО ПРИ РАЗДЕЛЕНИИ С ПОРОГОМ θ_3 ИНФОРМАЦИОННЫЙ ВЫИГРЫШ НАИБОЛЕЕ ВЫСОК.

Попробуем построить модель логистической классификации и решающее дерево.

```
from sklearn.datasets import load_breast_cancer
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifi er
```

В качестве данных возьмем обучающую выборку breast_cancer, которая показывает, доброкачественная или злокачественная опухоль.

```
breast_cancer = load_breast_cancer()
X = breast_cancer.data
y = breast_cancer.target
```

В scikit-learn и для регрессии, и для классификации используются в основном экземпляры класса, и обучение ведется путем метода fit. Сначала – логистическая классификация.

```
clf1 = LogisticRegression()
clf1.fi t(X,y)
```

Находим те же коэффициенты, что и при регрессии.

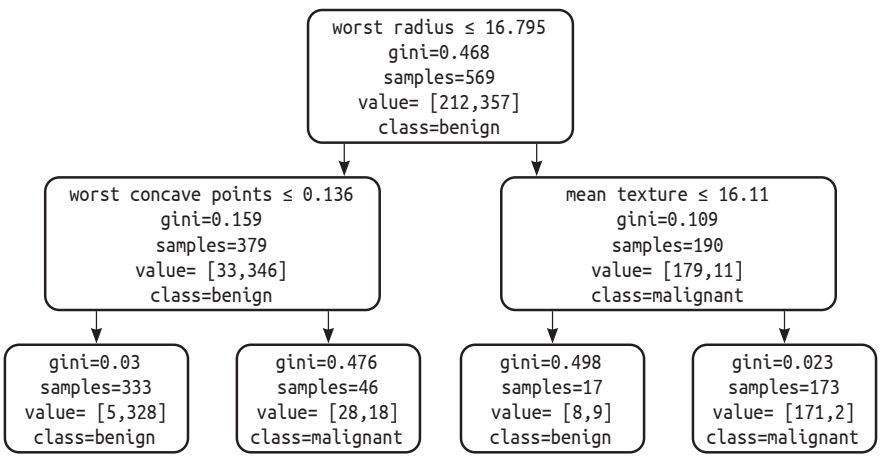
```
for f, w in zip(breast_cancer.feature_names, clf1.coef_[0]) :
    print("{0:<23}: {1:6.2f}". format(f, w))
```

```
mean radius            :   2.10
mean texture           :   0.12
mean perimeter         :  -0.06
...
worst concave points   :  -0.65
worst symmetry         :  -0.69
worst fractal dimension:  -0.11
```

Некоторые коэффициенты, имеющие большие положительные значения, влияют на положительный результат. Большие отрицательные значения влияют на отрицательный результат. Строим решающее дерево тем же методом, что и раньше.

```
clf2 = DecisionTreeClassifi er(max_depth=2)
clf2.fi t(X, y)
```

В результате получится вот такое дерево. На вершине окажется параметр, показывающий средний радиус опухоли. Здесь происходит деление данных по этому параметру, радиус больше или меньше 16.795. Если он меньше, то дальше деление происходит по параметру «вмятины», и если он меньше 0.136, то опухоль доброкачественная, а если больше, то злокачественная. С другой стороны, если средний радиус опухоли больше, чем 16.795, то далее разделение происходит по параметру «текстура», и если он меньше 16.11, то опухоль доброкачественная, а если больше – то злокачественная.



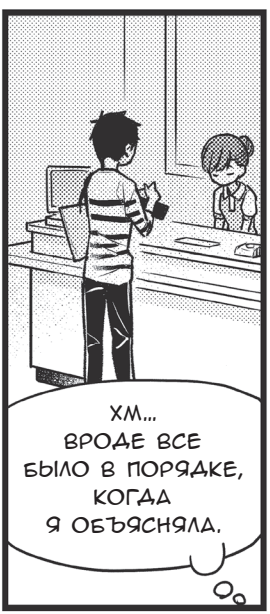

ВИДИШЬ НА КОМПЬЮТЕРЕ?
А... ДА.
ПОДОДВИГАЕТ


ААДА, КАК БЛИЗКО!
НАДО СДЕРЖИВАТЬСЯ!
ВОЛНУЕТСЯ
ТАК, О КЛАССИФИКАЦИИ МЫ ПОГОВОРИЛИ, А ЗАТЕМ ИДЕТ ОЦЕНКА...


НЕТ, СЭМПАЙ, ПОТОМ.
ЧТО?
ОТЧАЯННО


Я ДУМАЮ, Я САМ СМОГУ ИСПРАВИТЬ ДАННЫЕ И ПОСТРОИТЬ РЕШАЮЩЕЕ ДЕРЕВО, Я НЕ ПРОГРАММИСТ, НО НАЙДУ КОГО-НИБУДЬ...
НО...
ВСЕ В ПОРЯДКЕ!


СПАСИБО БОЛЬШОЕ!
ЧЕК
УБЕГАЕТ
А?


ХМ... ВРОДЕ ВСЕ БЫЛО В ПОРЯДКЕ, КОГДА Я ОБЪЯСНЯЛА.


Что?. Меня младший товарищ угостил?
АЗИНК
Спасибо, приходите еще
Подставка для счета


Понедельник
Отдел здравоохранения и благосостояния
ЧТО?
РАЗРАБОТЧИК ИГР?
НУ, НАШ СИСАДМИН КУДЗЁ МАКОТО. ОН ПРОГРАММИСТ И ЕЩЕ В ШКОЛЕ СДЕЛАЛ ОЧЕНЬ ПОПУЛЯРНУЮ ИГРУ.
ОГО...
ОН НЕМНОГО ЛЕГКОМЫСЛЕННО ОТНОСИТСЯ К РАБОЧИМ ОБЯЗАННОСТЯМ, И ГОВОРЯТ, ЧТО ОН РАЗРАБАТЫВАЕТ ИГРЫ ПРЯМО НА РАБОТЕ...
НИЧЕГО СЕБЕ ЧЕЛОВЕК.
СПАСИБО ЗА ИНФОРМАЦИЮ!
УДАЧИ!
Отдел поддержки оборудования
ОН НАВЕРНЯКА ТУТ.


ПРОСТИТЕ! Я КИЁХАРА ИЗ ОТДЕЛА ЗДРАВООХРАНЕНИЯ. КУДЗЁ-САН, ВЫ ТУТ?
КЛИК
КЛИК
Я – КУДЗЁ, ЧТО НАДО?
Я ЗНАЮ, ЧТО ВЫ ХОРОШО ПРОГРАММИРУЕТЕ, И ХОТЕЛ ПОПРОСИТЬ КОЕ О ЧЕМ...
КЛИК
КЛИК
УДИВЛЯЕТСЯ
КОЕ О ЧЕМ?.. ПРОГРАММУ ДЛЯ ОТДЕЛА ЗДРАВООХРАНЕНИЯ НАПИСАТЬ, ЧТО ЛИ?


ОЧЕНЬ УЖ ОН СУРОВЫЙ...
НУ, НА САМОМ ДЕЛЕ...
...ТАКИЕ ДЕЛА.
НУ?
НЕ ЛЮБЛЮ PYTHON.
ПОВОРАЧИВАЕТСЯ ОБРАТНО
ЧТО?
ТЫ Ж ИНФОРМАТИКЕ УЧИЛСЯ В УНИВЕРСИТЕТЕ, ПРАВДА? А ПРОГРАММУ НАПИСАТЬ НЕ МОЖЕШЬ.
Д-ДА... НЕ МОГУ.
ДЛЯ МЕНЯ ЭТО ОЧЕНЬ ТРУДНО, ДЕЙСТВИТЕЛЬНО. ИЗВИНИТЕ, ЧТО ОТВЛЕК ОТ РАБОТЫ.
РАЗОЧАРОВАННО
НЕ ЛЮБЛЮ, ПОТОМУ ЧТО ЭТО СЛИШКОМ ПРОСТО.
ЛАДНО, СДЕЛАЮ!


ПРАВДА?
ДА.
ВООДУШЕВЛЕННО
ЗА ЭТО УГОСТИШЬ МЕНЯ ГДЕ-НИБУДЬ, УГОВОР?
КОНЕЧНО! ОГРОМНОЕ СПАСИБО!
УРА-А! МОЖЕТ БЫТЬ,
И САЙТ ПОЛУЧИТСЯ СДЕЛАТЬ!
РАДУЕТСЯ
Пора переставать делать вид,
что я занят на работе...
Может, кормить будут.
ЖЕЛУДОК
Все ли хорошо?

В кабинете у Саяка (3)

Математическое повторение (2)

То есть тебя угостил Киёхара-кун? Что-то как-то странно.

Именно! Он меня, видимо, за сэмпая не держит!

Наверное, хм… (Бедный Киёхара-кун! Не заметил этого!)

О чем сегодня поговорим? Было что-то непонятное?

А что такое *e* на стр. 51?

e? А, *e* в сигмоидной функции? Это число Непера, которое используется для основания натуральных логарифмов. Это бесконечная дробь, которая равняется 2.71828.

Логарифм – это показатель степени, в которую надо возвести основание, чтобы получилось число. Если основание равно двум, то понятно, почему в решающем дереве встречаются двоичные разряды, но что за странное основание e, и почему они натуральные?

При дифференцировании функции e^x получается e^x, а при дифференцировании $\log e^x$ получается $1/x$, поэтому это очень удобная штука. На самом деле должно быть наоборот, e – это число, которое обладает таким свойством.

Далее – производная вектора на стр. 53.

У… Тебя не проведешь! Давай расскажу!

Если в функции ошибки $E(w)$ изменить величину веса модели w, значение тоже изменится. Так как существует несколько весов, то функция ошибки становится функцией нескольких переменных. Если мы выразим сумму весов в виде вектора, то в функции ошибки появляется аргумент в виде вектора.

Ага, пока понимаю.

А теперь найдем частную производную по вектору. ∂ (закругленная *d*) – это обозначение дифференцирования по одной из переменных. Например, в формуле $\frac{\partial E}{\partial w_0}$ ∂ означает, что в формуле *E* нужно найти производную по переменной w_0.

$$\nabla E = \frac{\partial E}{\partial \boldsymbol{w}} = \left(\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \cdots, \frac{\partial E}{\partial w_d} \right)^T.$$

Тогда нам нужно дифференцировать вектор!

Именно. Это так называемый вектор градиента.

Понятно. А если производная функции одной переменной определяется углом наклона касательной, то производная функции нескольких переменных – это ее градиент?

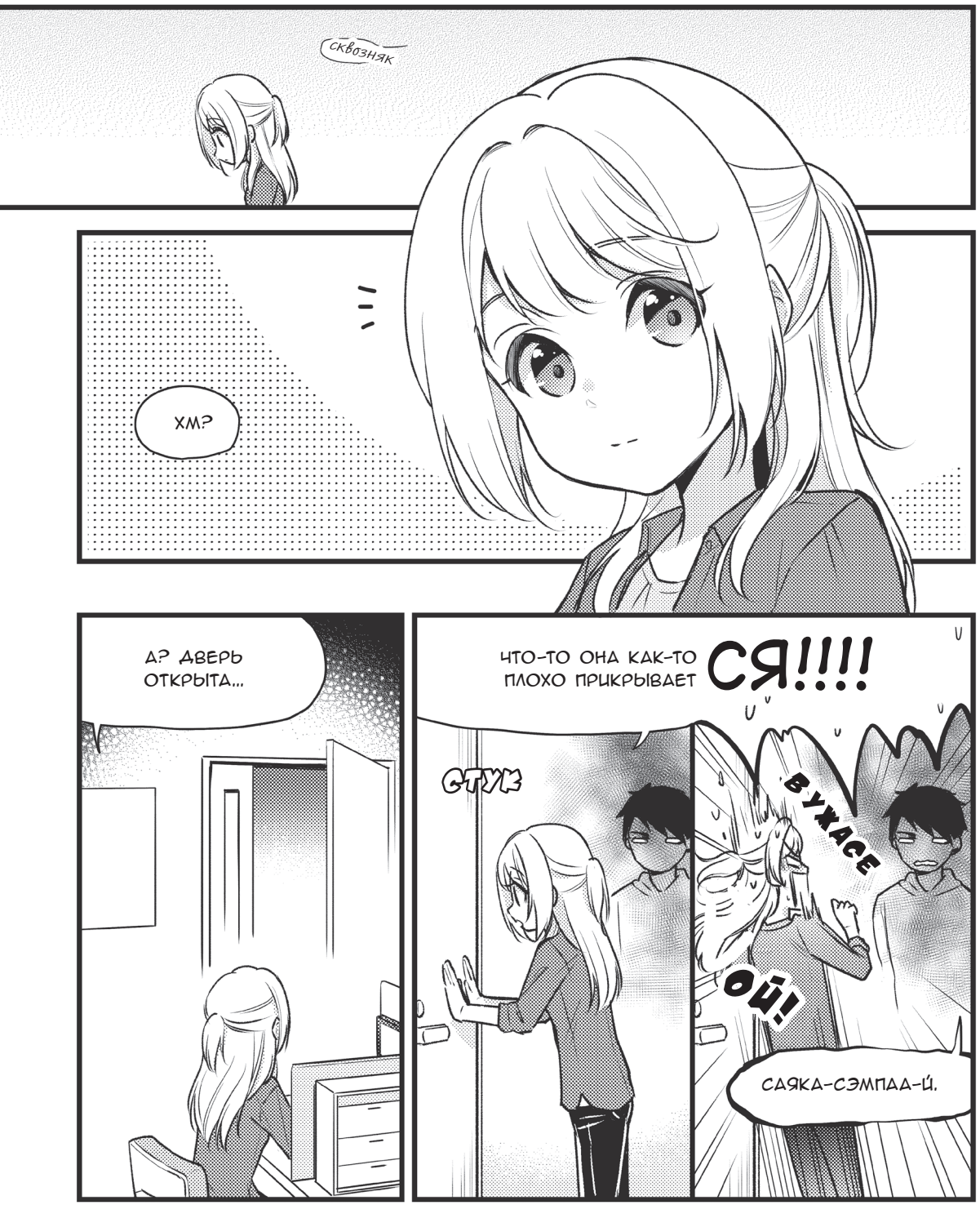
Да. Как на картинке на стр. 54. Текущий вес обозначается точкой на склоне, и если спускаться в обратном направлении (вниз по склону), то можно немного приблизиться к минимуму функции ошибки.

Ага. Тогда все! А Киёхара-кун собирается делать сайт?

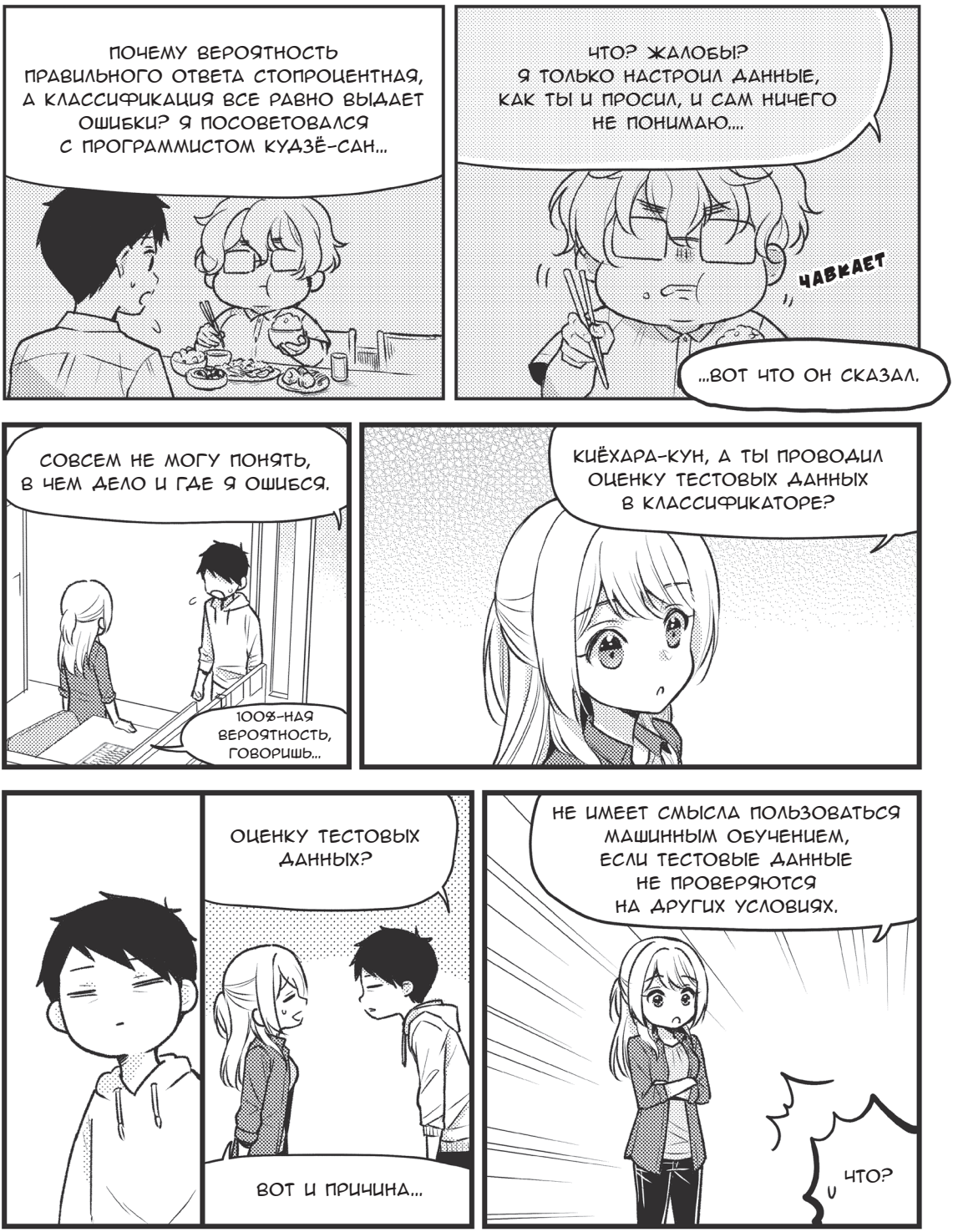
Я волнуюсь за него. Он же сбежал на полпути...

ГЛАВА 3

ОЦЕНКА РЕЗУЛЬТАТОВ


СКВОЗНЯК
ХМ?
ЧТО-ТО ОНА КАК-ТО ПЛОХО ПРИКРЫВАЕТ СЯ!!!!
А? ДВЕРЬ ОТКРЫТА...
СТУК
В УХАСЕ
ОЙ!
САЯКА-СЭМПАД-Й.


ДА ТЫ МЕНЯ ОПЯТЬ НАПУГАЛ!
ИЗВИНИТЕ...
СКОЛЬКО МО-О-О-ОЖНО-О-О-О!
Я ЖЕ ПРОСИЛА ПОТИШЕ, НО ЭТО СЛИШКОМ ТИХО!
ИЗВИНИТЕ...
ЧТО СЛУЧИЛОСЬ? Я ТРИ МЕСЯЦА ПОСЛЕ РЕСТОРАНА ТЕБЯ НЕ ВИДЕЛА, ДУМАЛА, ВСЕ ХОРОШО.
.....
ДА, НО...
Я НАШЕЛ ПРОГРАММИСТА, И ВМЕСТЕ МЫ СДЕЛАЛИ САЙТ С ИСПОЛЬЗОВАНИЕМ РЕШАЮЩЕГО ДЕРЕВА, КОТОРЫЙ С ВЕРОЯТНОСТЬЮ 100 % ОПРЕДЕЛЯЕТ СТЕПЕНЬ РИСКА ЗАБОЛЕВАНИЯ ДИАБЕТОМ.
Возможен ли у вас диабет?
ДА
или
НЕТ
Выберите нужный ответ
НАЧАТЬ
НО НА САЙТ МНОГО ЖАЛУЮТСЯ.
УЖЕ ТРОЕ ЛЮДЕЙ НЕ ИЗ ГРУППЫ РИСКА ПРОДОЛЖИЛИ ВЕСТИ ОБЫЧНЫЙ ОБРАЗ ЖИЗНИ И ОКАЗАЛИСЬ С ДИАБЕТОМ.
А ТЕ, КТО БЫЛ В ГРУППЕ РИСКА, ИДУТ В БОЛЬНИЦУ, И ИМ ГОВОРЯТ, ЧТО ВСЕ НОРМАЛЬНО.


ПОЧЕМУ ВЕРОЯТНОСТЬ ПРАВИЛЬНОГО ОТВЕТА СТОПРОЦЕНТНАЯ, А КЛАССИФИКАЦИЯ ВСЕ РАВНО ВЫДАЕТ ОШИБКИ? Я ПОСОВЕТОВАЛСЯ С ПРОГРАММИСТОМ КУДЗЁ-САН...
ЧТО? ЖАЛОБЫ? Я ТОЛЬКО НАСТРОИЛ ДАННЫЕ, КАК ТЫ И ПРОСИЛ, И САМ НИЧЕГО НЕ ПОНИМАЮ....
ЧАВКАЕТ
...ВОТ ЧТО ОН СКАЗАЛ.
СОВСЕМ НЕ МОГУ ПОНЯТЬ, В ЧЕМ ДЕЛО И ГДЕ Я ОШИБСЯ.
100%-НАЯ ВЕРОЯТНОСТЬ, ГОВОРИШЬ...
КИЁХАРА-КУН, А ТЫ ПРОВОДИЛ ОЦЕНКУ ТЕСТОВЫХ ДАННЫХ В КЛАССИФИКАТОРЕ?
ОЦЕНКУ ТЕСТОВЫХ ДАННЫХ?
ВОТ И ПРИЧИНА...
НЕ ИМЕЕТ СМЫСЛА ПОЛЬЗОВАТЬСЯ МАШИННЫМ ОБУЧЕНИЕМ, ЕСЛИ ТЕСТОВЫЕ ДАННЫЕ НЕ ПРОВЕРЯЮТСЯ НА ДРУГИХ УСЛОВИЯХ.
ЧТО?

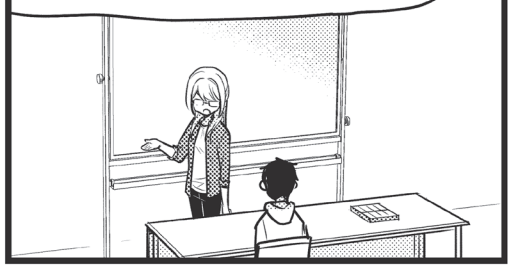

НАДО БЫЛО ТЕБЕ РАССКАЗАТЬ ОБ ОЦЕНКЕ ТЕСТОВЫХ ДАННЫХ... ЭТО МОЯ ВИНА...
НЕТ-НЕТ, ЭТО НЕ ВАША ВИНА, А МОЯ.
У ТЕБЯ ЕСТЬ ВРЕМЯ, КИЁХАРА-КУН?
Я РАССКАЖУ ТЕБЕ ОБ ЭТОМ.
ДА.
ПОЖАЛУЙСТА.

3.1. БЕЗ ПРОВЕРКИ ТЕСТОВЫХ ДАННЫХ НИКАК НЕЛЬЗЯ

ГОВОРИШЬ, КЛАССИФИКАТОР С ТОЧНОСТЬЮ 100 %?

ДА, ЕСЛИ ВЕРИТЬ ДАННЫМ ДЛЯ ОБУЧЕНИЯ.

НО ЕСЛИ В ДАННЫХ ДЛЯ ОБУЧЕНИЯ ВЫСОКАЯ ВЕРОЯТНОСТЬ ПРАВИЛЬНОГО ОТВЕТА, ТО В ЭТОМ НЕТ СМЫСЛА.



МОЖНО ПОСТРОИТЬ ДЕРЕВО С ТОЧНОСТЬЮ РЕШЕНИЙ В 100 % В СЛУЧАЕ, ЕСЛИ НЕТ ОГРАНИЧЕНИЙ НА РАЗМЕР ДЕРЕВА И ЕСЛИ НЕТ ПРОТИВОРЕЧИЯ В ДАННЫХ, ТО ЕСТЬ КОГДА ОДИНАКОВЫЕ ПРИЗНАКОВЫЕ ОПИСАНИЯ ОТНЕСЕНЫ К РАЗНЫМ КЛАССАМ.

НО ЕСЛИ ПРИ ИСПОЛЬЗОВАНИИ ДАННЫХ ИЗ ОБУЧАЮЩЕЙ ВЫБОРКИ ТОЧНОСТЬ РЕШЕНИЯ 100 %, ТО ПОЧЕМУ ЭТО ПЛОХО?

В ТАКИХ СИСТЕМАХ, ЕСЛИ В ДАННЫХ ДЛЯ ОБУЧЕНИЯ ВСЕ ЧЕТКО НАСТРОЕНО, ВЕЛИКА ВЕРОЯТНОСТЬ ТОГО, ЧТО НОВЫЕ ДАННЫЕ БУДУТ ИНТЕРПРЕТИРОВАТЬСЯ НЕПРАВИЛЬНО?

КОРОЧЕ, ЧТО ЛУЧШЕ ДЕЛАТЬ?

3.2. ОБУЧАЮЩАЯ, ТЕСТОВАЯ И КОНТРОЛЬНАЯ ВЫБОРКИ


ЕСЛИ НЕЛЬЗЯ ПОЛУЧИТЬ ДОСТАТОЧНО НЕИЗВЕСТНЫХ ДАННЫХ, ТО И ТЕСТИРОВАНИЕ НЕВОЗМОЖНО?
ЕСЛИ ИХ НЕ ХВАТАЕТ... Я ОБ ЭТОМ.


ТАК... ТОЛЬКО В СЛУЧАЕ, ЕСЛИ ЕСТЬ ДОСТАТОЧНО ДАННЫХ И ДЛЯ ОБУЧЕНИЯ, И ДЛЯ ТЕСТОВОЙ ВЫБОРКИ, МОЖНО ИСПОЛЬЗОВАТЬ ЭТОТ МЕТОД.
НО ДАЖЕ ЕСЛИ ДАННЫХ ДОСТАТОЧНО, ТО НЕ ВСЕГДА ЭТО ХОРОШИЙ МЕТОД.
ПОЧЕМУ?


У НАС ЕСТЬ ГИПЕРПАРАМЕТР, КОТОРЫЙ ВЛИЯЕТ НА РЕЗУЛЬТАТЫ ОБУЧЕНИЯ? ВЕС ДОПОЛНИТЕЛЬНОГО ЧЛЕНА ПРИ ЛИНЕЙНОЙ РЕГРЕССИИ ИЛИ ЖЕ ТОЛЩИНА ДЕРЕВА ПРИ ПОСТРОЕНИИ РЕШАЮЩЕГО ДЕРЕВА?
НУЖНА ПРОВЕРКА АДЕКВАТНОСТИ ВЕЛИЧИНЫ ГИПЕРПАРАМЕТРА, КОТОРЫЙ ИСПОЛЬЗУЕТСЯ ПРИ ОБУЧЕНИИ.
ЕСЛИ ИСПОЛЬЗОВАТЬ ТЕСТОВУЮ ВЫБОРКУ ДЛЯ ОЦЕНКИ, ТО ДАННЫЕ ИЗ ТЕСТОВОЙ ВЫБОРКИ НЕ СМОГУТ РАССМАТРИВАТЬСЯ КАК НЕИЗВЕСТНЫЕ ДАННЫЕ.
Проверь!
ТЕСТОВАЯ ВЫБОРКА
ОБУЧАЮЩАЯ ВЫБОРКА
ОБУЧАЮЩАЯ ВЫБОРКА
Но я же знаю данные из тестовой выборки
ДЕЙСТВИТЕЛЬНО.


МЕТОД ПРОВЕРКИ НА ЗАРЕЗЕРВИРОВАННЫХ ДАННЫХ ДЛЯ ОЦЕНКИ ЭФФЕКТИВНОСТИ ЗАКЛЮЧАЕТСЯ В ТОМ, ЧТО ДАННЫЕ ДЕЛЯТ НА ТРИ ГРУППЫ – ОБУЧАЮЩУЮ, ТЕСТОВУЮ И КОНТРОЛЬНУЮ ВЫБОРКИ.
ВСЕ ДАННЫЕ
В
Ж
И
Х!
ТЕСТОВАЯ ВЫБОРКА
ОБУЧАЮЩАЯ ВЫБОРКА
ПРОВЕРОЧНАЯ ВЫБОРКА
Ничего себе
АПЛОДИСМЕНТЫ

А ЧТОБЫ ДЕЛИТЬ ДАННЫЕ НА ТРИ ЧАСТИ, ИХ ДОЛЖНО БЫТЬ МНОГО?

КОГДА ДАННЫХ МАЛО, ИСПОЛЬЗУЕТСЯ **МЕТОД ПЕРЕКРЕСТНОЙ ПРОВЕРКИ (КРОСС-ВАЛИДАЦИИ).**

ФУХ!

ОБУЧЕНИЕ

ТЕСТ

ПРОВЕРКА

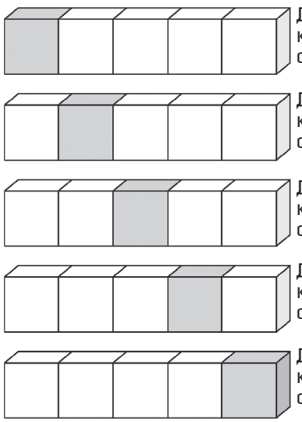
ЧТО ЭТО?

С МЕТОДОМ ПРОВЕРКИ НА ЗАРЕЗЕРВИРОВАННЫХ ДАННЫХ ЕЕ ОБЪЕДИНЯЕТ ДЕЛЕНИЕ ДАННЫХ НА ВЫБОРКУ ДЛЯ ОБУЧЕНИЯ И ТЕСТОВУЮ ВЫБОРКУ, ОДНАКО СУТЬ В ТОМ, ЧТО ВСЕ ДАННЫЕ ИСПОЛЬЗУЮТСЯ В КАЧЕСТВЕ ТЕСТОВОЙ ВЫБОРКИ.

НАДЕВАЕТ ОЧКИ

ВСЕ ДАННЫЕ? ЭТО КАК?

3.3. МЕТОД ПЕРЕКРЕСТНОЙ ПРОВЕРКИ (КРОСС-ВАЛИДАЦИИ)



ПРЕДСТАВИМ, ЧТО ДАННЫЕ ДЕЛЯТСЯ, КАК НА КАРТИНКЕ, НА **ТЕСТОВЫЕ БЛОКИ**. ДЛЯ НИХ ОБЫЧНО БЕРЕТСЯ 10 % ДАННЫХ.

ПОЧЕМУ 10 %?

ЕСЛИ ДАННЫЕ ДЕЛЯТСЯ НА 10 ЧАСТЕЙ, ТО СИСТЕМА ИСПОЛЬЗУЕТ 90 % ДЛЯ ОБУЧЕНИЯ, А ВСЕ ДАННЫЕ ДЛЯ ТЕСТИРОВАНИЯ. ЭФФЕКТИВНОСТЬ ОБУЧЕНИЯ В ТАКОЙ МОДЕЛИ БЛИЗКА К ВЕРХНЕМУ ПРЕДЕЛУ, ТАК КАК ТЕСТИРОВАНИЕ ПРОИЗВОДИТСЯ ДАННЫМИ, МАКСИМАЛЬНО ПРИБЛИЖЕННЫМИ К РАСПРЕДЕЛЕНИЮ НЕИЗВЕСТНЫХ ДАННЫХ.

АГА.


Метод, при котором для оценки используется один блок данных, называется **контролем по отдельным объектам**.
А разве у кросс-валидации нет недостатков?


При использовании этого метода оценка эффективности может занять некоторое время, если количество тестовых блоков m велико.
m
А если это число m слишком низкое, то система не сможет обучиться.

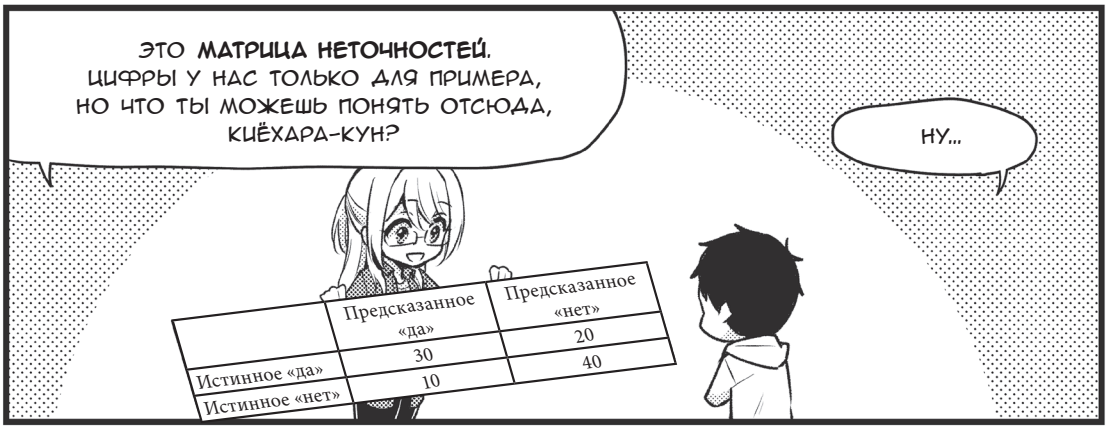

Однако поскольку метод перекрестной проверки используется, когда данных не так уж и много и когда хватает одного раунда для оценки, то ему можно доверять. Контроль по отдельным элементам тоже можно использовать, если тасовать фрагменты и проделать проверку несколько раз.

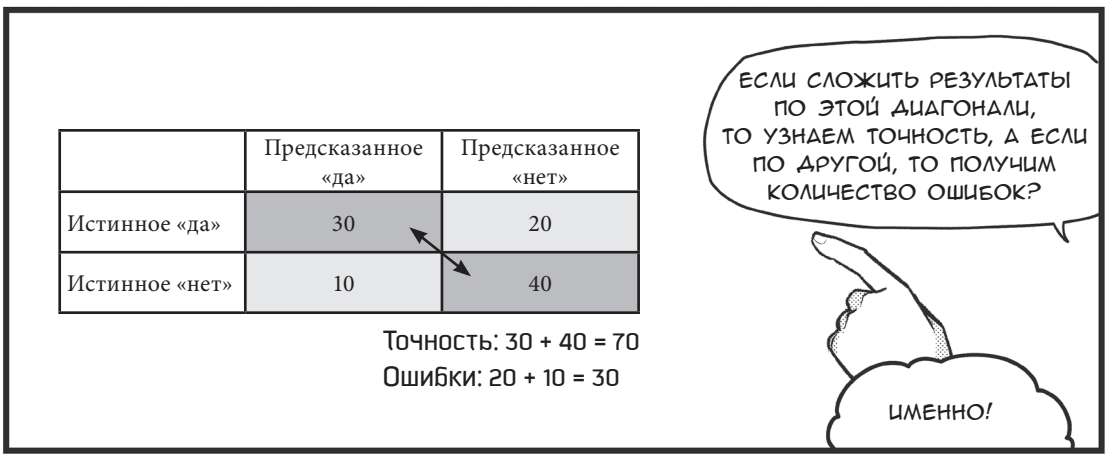

А теперь рассмотрим эти методы оценки в деталях, с числовыми величинами.
Пожалуйста!

3.4. Доля правильно предсказанных объектов, точность, полнота и F-мера

	Предсказанное «да»	Предсказанное «нет»
Истинное «да»	30	20
Истинное «нет»	10	40

Если мы объединим истинный класс, разделенный на истинное "да" и истинное "нет", с предсказанным классом, который тоже разделен на "да" и "нет", то получим четыре группы. Истинное "да" и истинное "нет" – это результаты ответов "да" или "нет", а прогнозы по классификатору будут делиться на предсказанное "да" и предсказанное "нет".

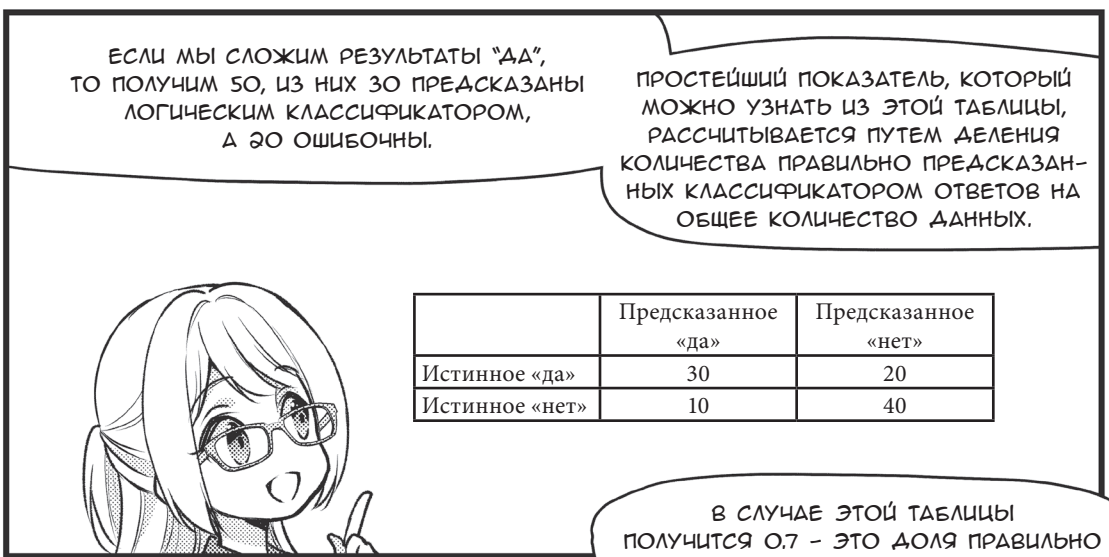




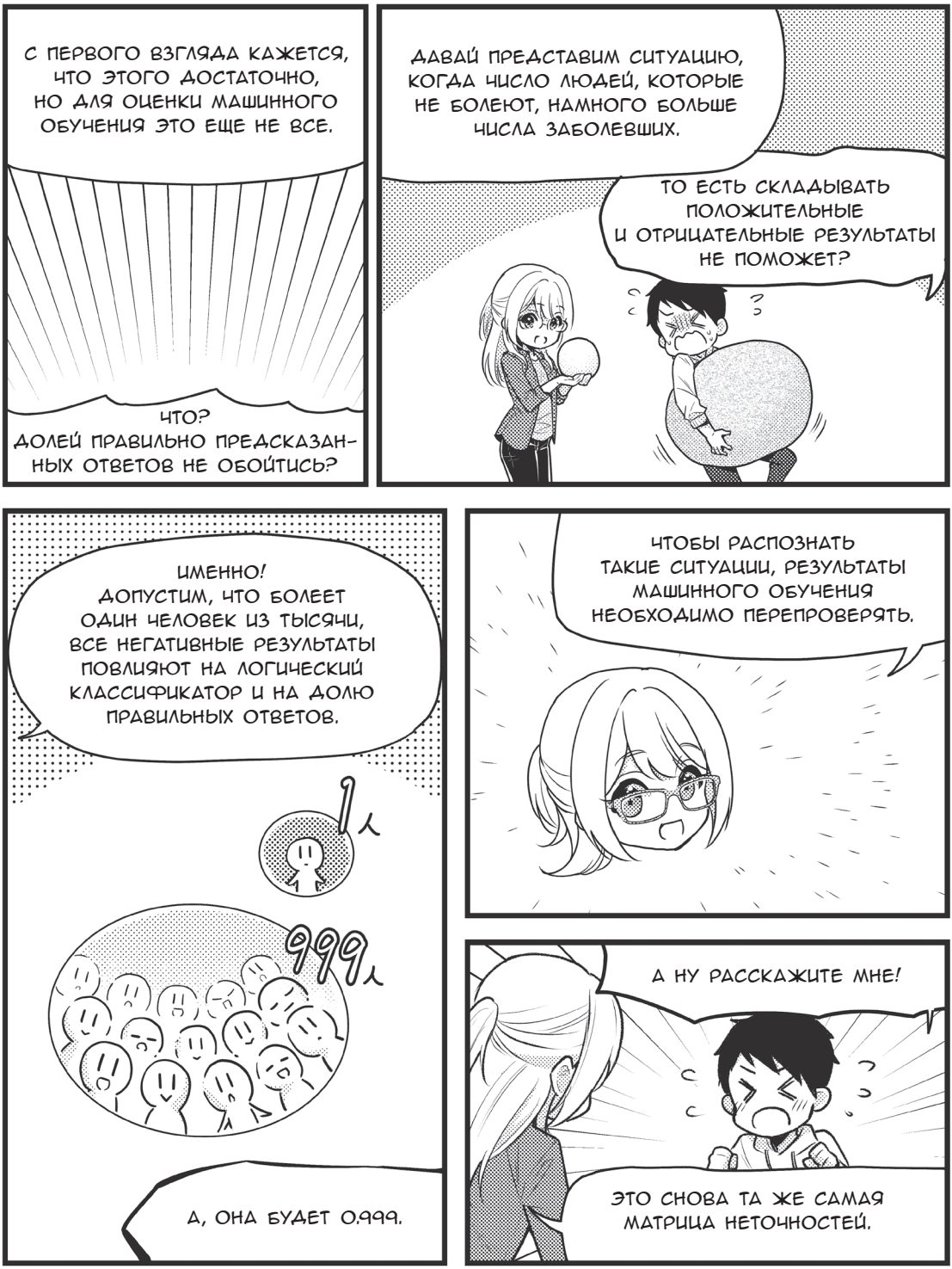
	Предсказанное «да»	Предсказанное «нет»
Истинное «да»	30	20
Истинное «нет»	10	40

Точность: 30 + 40 = 70

Ошибки: 20 + 10 = 30



	Предсказанное «да»	Предсказанное «нет»
Истинное «да»	30	20
Истинное «нет»	10	40


С ПЕРВОГО ВЗГЛЯДА КАЖЕТСЯ, ЧТО ЭТОГО ДОСТАТОЧНО, НО ДЛЯ ОЦЕНКИ МАШИННОГО ОБУЧЕНИЯ ЭТО ЕЩЕ НЕ ВСЕ.
ЧТО? ДОЛЕЙ ПРАВИЛЬНО ПРЕДСКАЗАННЫХ ОТВЕТОВ НЕ ОБОЙТИСЬ?
ДАВАЙ ПРЕДСТАВИМ СИТУАЦИЮ, КОГДА ЧИСЛО ЛЮДЕЙ, КОТОРЫЕ НЕ БОЛЕЮТ, НАМНОГО БОЛЬШЕ ЧИСЛА ЗАБОЛЕВШИХ.
ТО ЕСТЬ СКЛАДЫВАТЬ ПОЛОЖИТЕЛЬНЫЕ И ОТРИЦАТЕЛЬНЫЕ РЕЗУЛЬТАТЫ НЕ ПОМОЖЕТ?
ИМЕННО! ДОПУСТИМ, ЧТО БОЛЕЕТ ОДИН ЧЕЛОВЕК ИЗ ТЫСЯЧИ, ВСЕ НЕГАТИВНЫЕ РЕЗУЛЬТАТЫ ПОВЛИЯЮТ НА ЛОГИЧЕСКИЙ КЛАССИФИКАТОР И НА ДОЛЮ ПРАВИЛЬНЫХ ОТВЕТОВ.
1人
999人
А, ОНА БУДЕТ 0.999.
ЧТОБЫ РАСПОЗНАТЬ ТАКИЕ СИТУАЦИИ, РЕЗУЛЬТАТЫ МАШИННОГО ОБУЧЕНИЯ НЕОБХОДИМО ПЕРЕПРОВЕРЯТЬ.
А НУ РАССКАЖИТЕ МНЕ!
ЭТО СНОВА ТА ЖЕ САМАЯ МАТРИЦА НЕТОЧНОСТЕЙ.

	Предсказанное «да»	Предсказанное «нет»
Истинное «да»	Истинное «за» (ИЗ)	Ложное «против» (ЛП)
Истинное «нет»	Ложное «за» (ЛЗ)	Истинное «против» (ИП)

НАПРИМЕР, ЛЕВЫЙ ВЕРХНИЙ ЭЛЕМЕНТ НАЗЫВАЕТСЯ ИСТИННЫМ "ЗА", ПОТОМУ ЧТО КЛАССИФИКАТОР ПРЕДСКАЗАЛ "ДА" И ЭТО СОВПАЛО С ИСТИННЫМ РЕЗУЛЬТАТОМ. ДЛЯ СОКРАЩЕНИЯ ИСПОЛЬЗУЕТСЯ АББРЕВИАТУРА ИЗ.

А "ЛОЖНОЕ ПРОТИВ" – ЭТО КОГДА БЫЛО ПРЕДСКАЗАНО "НЕТ", НО РЕЗУЛЬТАТ НА САМОМ ДЕЛЕ "ДА".

ТАКИМ ОБРАЗОМ, ДОЛЯ ПРАВИЛЬНО ПРЕДСКАЗАННЫХ ДАННЫХ БУДЕТ ВЫЧИСЛЯТЬСЯ ПО ФОРМУЛЕ НИЖЕ:

$$\text{Доля правильно предсказанных данных} = \frac{\text{ИЗ} + \text{ИП}}{\text{ИЗ} + \text{ЛП} + \text{ЛЗ} + \text{ИП}}$$

НУЖНО РАЗДЕЛИТЬ КОЛИЧЕСТВО ПРАВИЛЬНЫХ ОТВЕТОВ НА ОБЩЕЕ КОЛИЧЕСТВО ДАННЫХ.

ПОГОВОРИМ О **ТОЧНОСТИ**. ОНА ПОКАЗЫВАЕТ, МОЖНО ЛИ ДОВЕРЯТЬ КЛАССИФИКАТОРУ ПРИ ДЕЛЕНИИ ДАННЫХ.

ТОЧНОСТЬ ВЫЧИСЛЯЕТСЯ ПО ПРЕДСТАВЛЕННОЙ ФОРМУЛЕ.

ДОПУСТИМ, МЫ ХОТИМ УЗНАТЬ, НАСКОЛЬКО ТОЧНО МЫ ОПРЕДЕЛИЛИ БОЛЕЗНЬ.

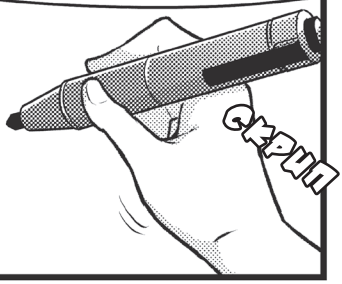
$$\text{Точность} = \frac{\text{ИЗ}}{\text{ИЗ} + \text{ЛЗ}}$$

МЫ ДЕЛИМ КОЛИЧЕСТВО ИСТИННЫХ "ЗА" НА СУММУ "ДА", ПРЕДСКАЗАННЫХ КЛАССИФИКАТОРОМ.

После точности идет **полнота** данных. Она указывает, насколько адекватно оцениваются истинные "да". Например, мы можем узнать, насколько адекватна выборка на основании количества на самом деле заболевших людей.

Чтобы найти полноту, надо разделить количество истинных "за", правильно предсказанных классификатором, на общую сумму истинных "да".

$$\text{Полнота} = \frac{\text{ИЗ}}{\text{ИЗ} + \text{ЛП}}$$



Именно!

Точность и полнота находятся в таких отношениях: если использовать один из этих параметров, другой теряет значимость, и наоборот.

Это как?

Допустим, точность классификатора, который выдает положительный результат, будет выше, если мы заранее знаем, что в наших данных есть большое количество людей, заболевших той или иной болезнью.

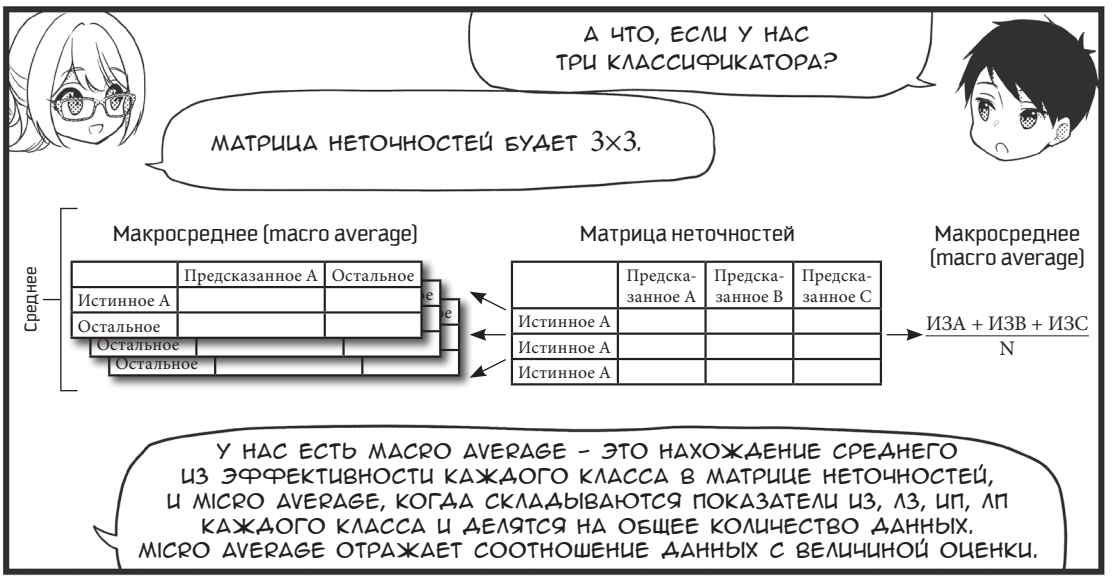
Хм. А если у человека лишь незначительный симптом и мы точно не знаем, болен он или нет, то болезнь и проглядеть можно?

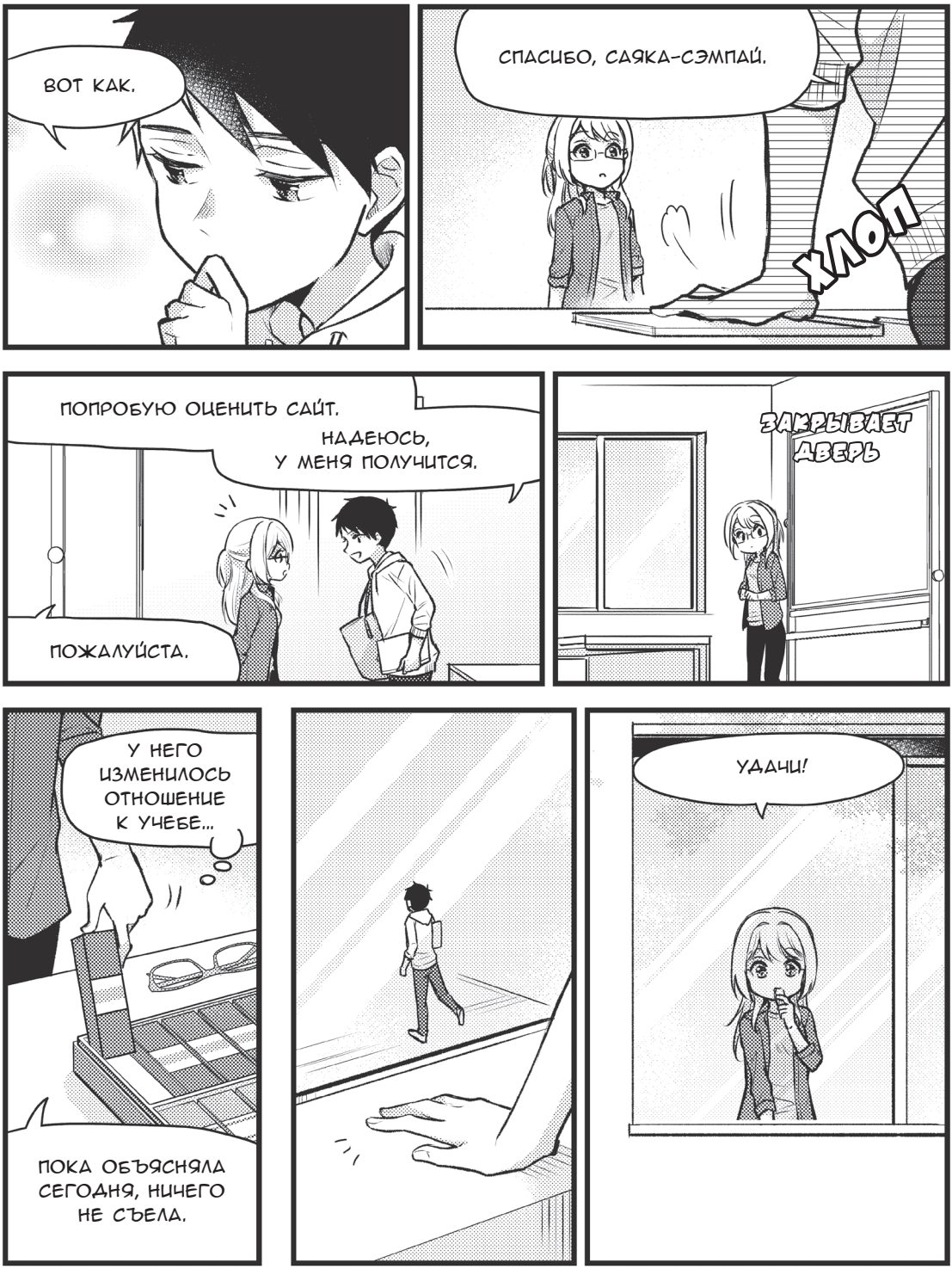
Да! Потому что полнота данных в этом случае низкая.

И наоборот, если в приоритете высокая полнота данных, то при малейшем сомнении классификатор будет выдавать результаты "да".

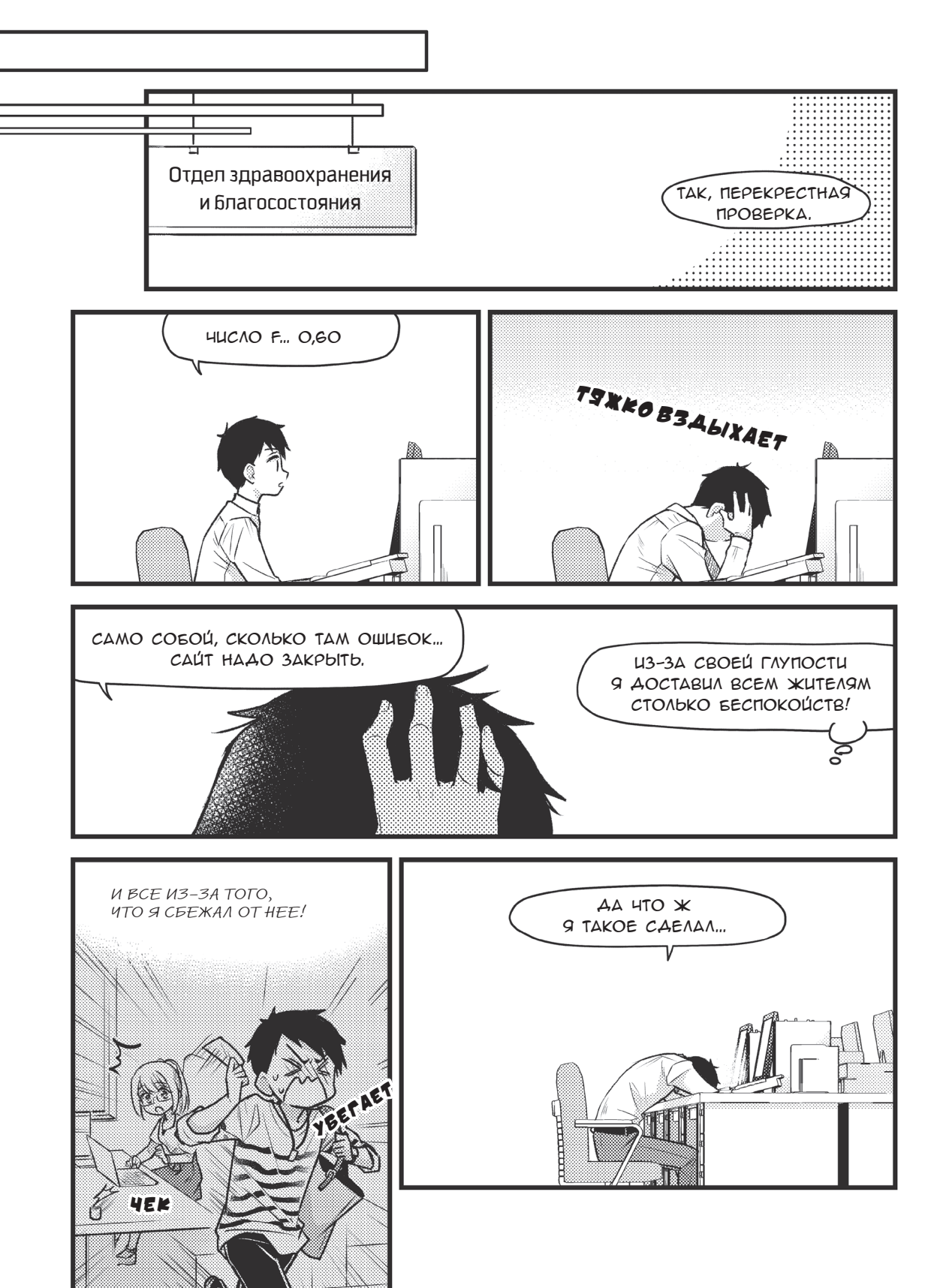
Что?

$$F\text{-мера} = 2 \times \frac{\text{Точность} \times \text{Полнота}}{\text{Точность} + \text{Полнота}}$$





З.Ч. ДОЛЯ ПРАВИЛЬНО ПРЕДСКАЗАННЫХ ОБЪЕКТОВ, ТОЧНОСТЬ, ПОЛНОТА И F-МЕРА


Отдел здравоохранения
и благосостояния
ТАК, ПЕРЕКРЕСТНАЯ ПРОВЕРКА.
ЧИСЛО F... 0,60
ТЯЖКО ВЗДЫХАЕТ
САМО СОБОЙ, СКОЛЬКО ТАМ ОШИБОК... САЙТ НАДО ЗАКРЫТЬ.
ИЗ-ЗА СВОЕЙ ГЛУПОСТИ Я ДОСТАВИЛ ВСЕМ ЖИТЕЛЯМ СТОЛЬКО БЕСПОКОЙСТВ!
ДА ЧТО Ж Я ТАКОЕ СДЕЛАЛ...
И ВСЕ ИЗ-ЗА ТОГО, ЧТО Я СБЕЖАЛ ОТ НЕЕ!
ЧЕК
УБЕГАЕТ

В кабинете у Саяка (4)

Математическое повторение (3)

Сегодняшняя беседа была довольно проста с точки зрения математики, однако она важна для машинного обучения. Можно скопипастить полученные данные, но нехорошо, если возникает ситуация «все сделал, но не понимаю, что получилось».

Я немного не понимаю число *F*. Это всего лишь среднее от точности и аккуратности?

Обычное среднее – это арифметическое среднее. Если у нас есть числа a и b, то оно определяется по формуле $(a + b)/2$. Так определяют средний балл теста, температуру и т. п., но для прогнозов используется другой метод.

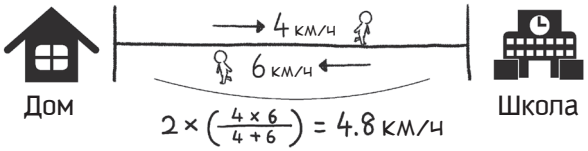
Что касается точности или полноты, то среднее находится другим способом.

Почему не используется среднее арифметическое?

Ну, давай рассмотрим пример из повседневной жизни – скорость. Как она определяется?

Расстояние, поделенное на время!

Да. Возьмем такой пример.



Вот когда ты идешь в школу, скорость 4 км/ч.

А когда ты возвращаешься из школы, то 6 км/ч. Расстояние одинаковое, верно? Но какая будет средняя скорость?

Расстояние не указано, поэтому проведем вычисления так. В одну сторону расстояние будет равно x, тогда туда и обратно – $2x$. Время туда будет равно $x/4$, а обратно – $x/6$, и средняя скорость будет рассчитана по формуле:

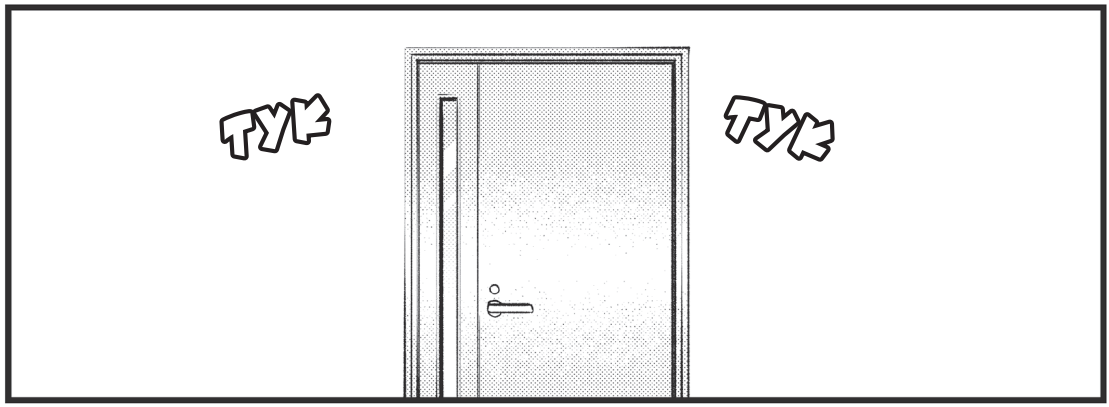
$$\frac{2x}{\frac{x}{4}+\frac{x}{4}} = 2 \times \frac{x}{\frac{x(4+6)}{4 \times 6}} = 2 \times \frac{4 \times 6}{4+6} = \frac{48}{10} = 4.8$$

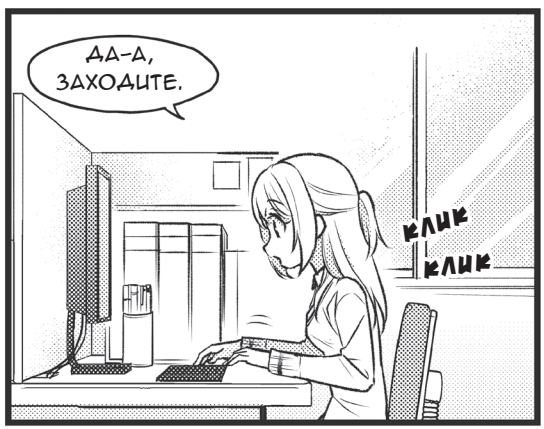
Ага, x исчез. То есть не 5 км/ч, а 4,8 км/ч.

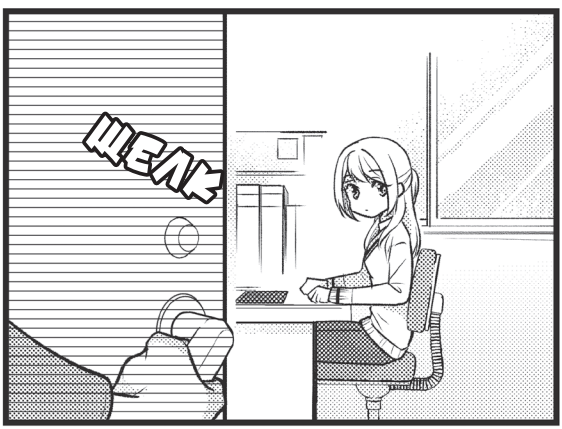
Да. Эта формула $2 \times \left(\frac{4 \times 6}{4+6}\right)$ используется и для нахождения F-меры. Это среднее гармоническое значение.

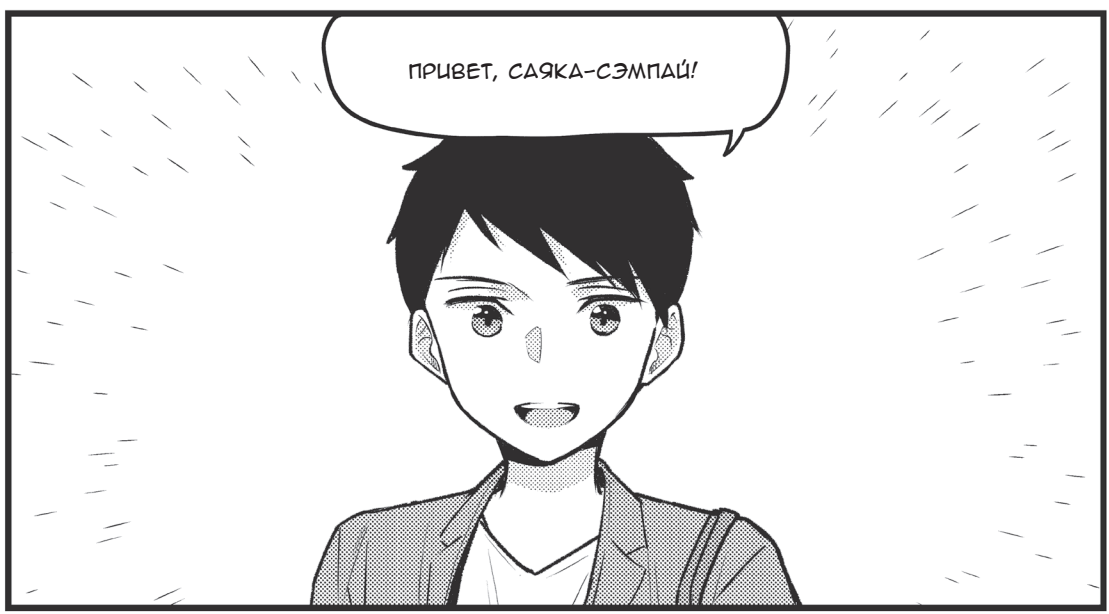
ГЛАВА 4

ГЛУБОКОЕ ОБУЧЕНИЕ


ТУК
ТУК


ДА-А, ЗАХОДИТЕ.
КЛИК
КЛИК


ЩЕЛК


ПРИВЕТ, САЯКА-СЭМПАЙ!

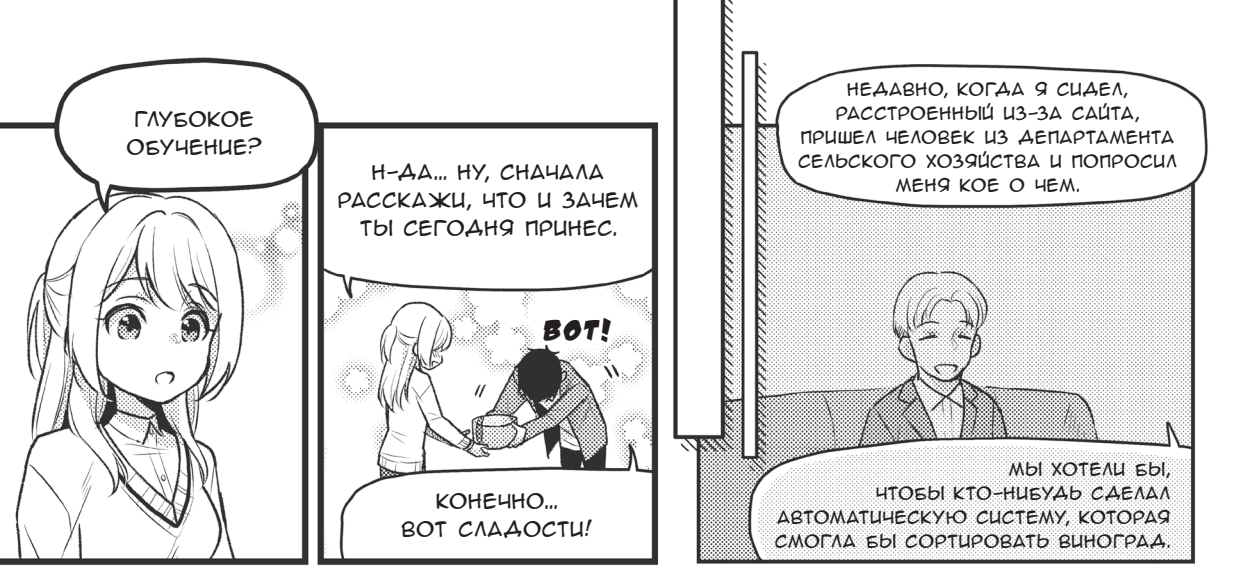

КИЁХАРА-КУН!
УДИВЛЕННО


ТЫ ЧЕГО ТАК ОБЫЧНО ЗАХОДИШЬ?
Ты же всегда врываешься в лабораторию.
Ну...
НУ, Я ТАК ТОЖЕ МОГУ.


С ТОГО РАЗА ДВЕ НЕДЕЛИ ПРОШЛО, И ТЕБЯ ОПЯТЬ ЧТО-ТО БЕСПОКОИТ?
ИЗВИНИТЕ, ЧТО НЕ ПРЕДУПРЕДИЛ ЗАРАНЕЕ! ХОТЕЛ СКАЗАТЬ, ЧТО САЙТ ГОТОВ.

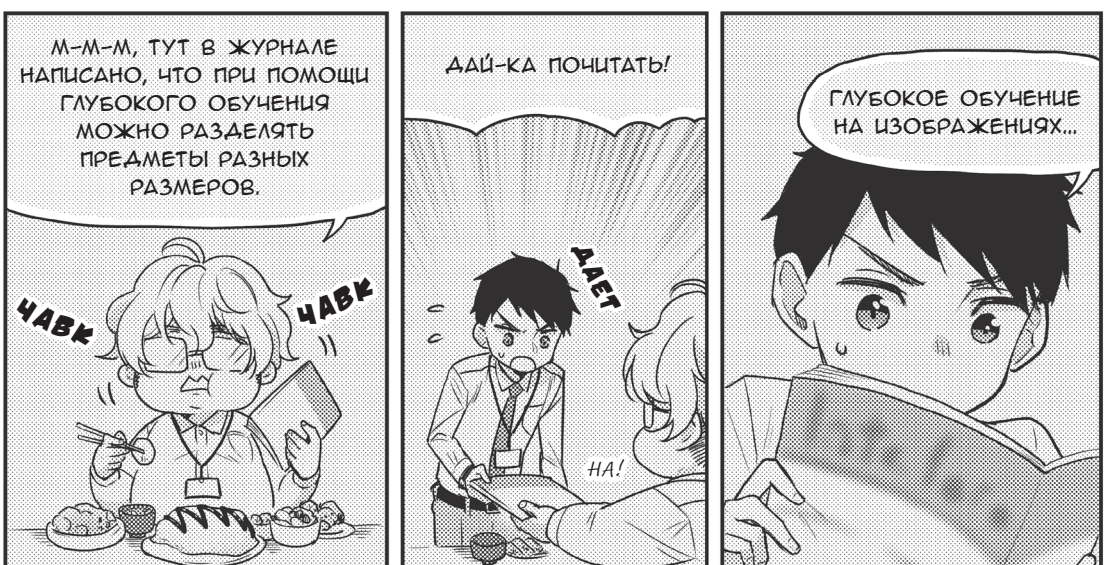

ТАК-ТАК, ТЫ НА ЧТО-ТО НАМЕКАЕШЬ?
ДА, СЭМПАЙ, У МЕНЯ К ВАМ ЕЩЕ ОДНА ПРОСЬБА, СОВСЕМ ДРУГАЯ.


СЭМПАЙ!
РАССКАЖЕТЕ МНЕ ПРО ГЛУБОКОЕ ОБУЧЕНИЕ?
Глубокое обучение (deep learning)
Один из способов машинного обучения с использованием нейронных сетей. В последнее время привлекает к себе внимание благодаря высокой эффективности в сферах распознавания изображений, звуков и естественной речи.


ГЛУБОКОЕ ОБУЧЕНИЕ?
Н-ДА... НУ, СНАЧАЛА РАССКАЖИ, ЧТО И ЗАЧЕМ ТЫ СЕГОДНЯ ПРИНЕС.
ВОТ!
КОНЕЧНО... ВОТ СЛАДОСТИ!
НЕДАВНО, КОГДА Я СИДЕЛ, РАССТРОЕННЫЙ ИЗ-ЗА САЙТА, ПРИШЕЛ ЧЕЛОВЕК ИЗ ДЕПАРТАМЕНТА СЕЛЬСКОГО ХОЗЯЙСТВА И ПОПРОСИЛ МЕНЯ КОЕ О ЧЕМ.
МЫ ХОТЕЛИ БЫ, ЧТОБЫ КТО-НИБУДЬ СДЕЛАЛ АВТОМАТИЧЕСКУЮ СИСТЕМУ, КОТОРАЯ СМОГЛА БЫ СОРТИРОВАТЬ ВИНОГРАД.


МЕСТНАЯ КОМПАНИЯ ПО ИЗГОТОВЛЕНИЮ ЭЛЕКТРОНИКИ СДЕЛАЛА ПРОГРАММУ ДЛЯ АВТОМАТИЧЕСКОЙ УПАКОВКИ ВИНОГРАДА НА КОНВЕЙЕРЕ, НО СОРТИРОВАТЬ ВИНОГРАД ОНА НЕ УМЕЕТ.
ВОТ ОНО ЧТО...


МОИ ПРОШЛЫЕ НЕУДАЧИ ПРИВЕЛИ МЕНЯ К ТОМУ, ЧТО Я РЕШИЛ ПОСОВЕТОВАТЬСЯ С КУДЗЁ-САН.
Столовая
Меню обеда


М-М-М, ТУТ В ЖУРНАЛЕ НАПИСАНО, ЧТО ПРИ ПОМОЩИ ГЛУБОКОГО ОБУЧЕНИЯ МОЖНО РАЗДЕЛЯТЬ ПРЕДМЕТЫ РАЗНЫХ РАЗМЕРОВ.
ЧАВК
ЧАВК
ДАЙ-КА ПОЧИТАТЬ!
ДАЕТ
НА!
ГЛУБОКОЕ ОБУЧЕНИЕ НА ИЗОБРАЖЕНИЯХ...


ЭТО БЫЛО КАК РАЗ ТО, ЧТО МНЕ НУЖНО, И Я ПОБЛАГОДАРИЛ КУДЗЁ-САН ЗА ЭТУ СТАТЬЮ.
НА САМОМ ДЕЛЕ Я ИГРАЛ ТУТ С КАКИМ-ТО СПЕЦОМ ПО МАШИННОМУ ОБУЧЕНИЮ, ПОГОВОРИЛ С НИМ, И ОН ПОСОВЕТОВАЛ МНЕ ЭТУ КНИГУ.


НО ЭТО... У ТЕБЯ НИЧЕГО НЕ ПОЛУЧАЕТСЯ, А ТЫ ВСЕ СТАРАЕШЬСЯ?
ДА! Я ЖЕ ХОЧУ ПОМОЧЬ ФЕРМЕРАМ СОРТИРОВАТЬ ВИНОГРАД.

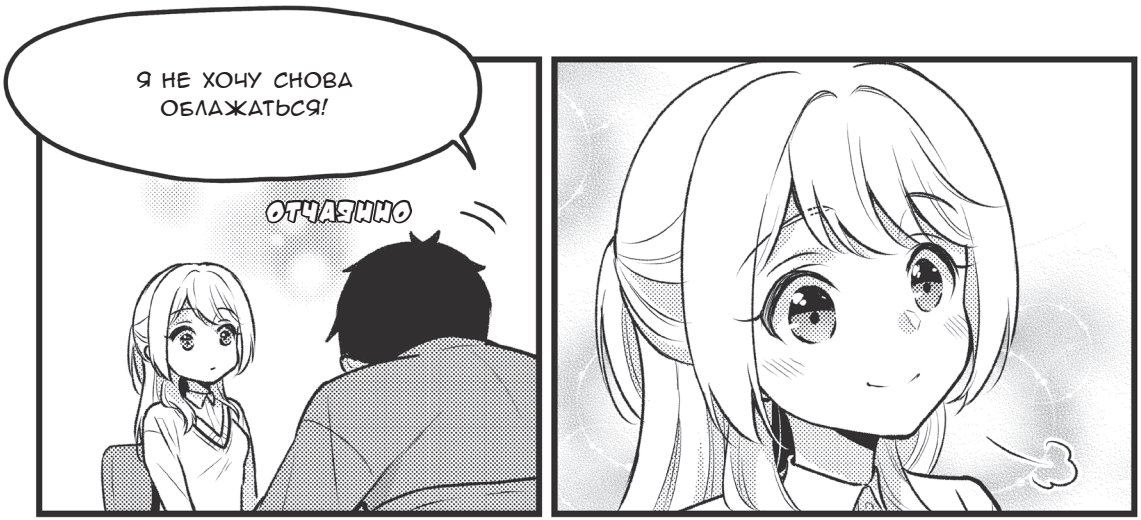

ДОБРАЯ ДУША.
ХЛЮП


ТЫ ЧТО-ТО СКАЗАЛ?
НЕ, ЕСЛИ ПОМОЩЬ С ПРОГРАММОЙ ПОНАДОБИТСЯ, ЗОВИ!
СПАСИБО!


Я ЧИТАЛА ЭТУ ЗАМЕТКУ. ХОРОШИЙ ЧЕЛОВЕК ЭТОТ КУДЗЁ-САН.
Я СНАЧАЛА ПОДУМАЛ, ЧТО С НИМ ТРУДНО ПОЛАДИТЬ, НО ОН ВСЕГДА ВЫРУЧИТ.
ХРУМ
ХРУМ


А ТЫ ХОЧЕШЬ, ЧТОБЫ Я ТЕБЕ ПОМОГАЛА? НАМИГОЭ-СЭНСЭЙ УЖЕ ВЕРНУЛСЯ.

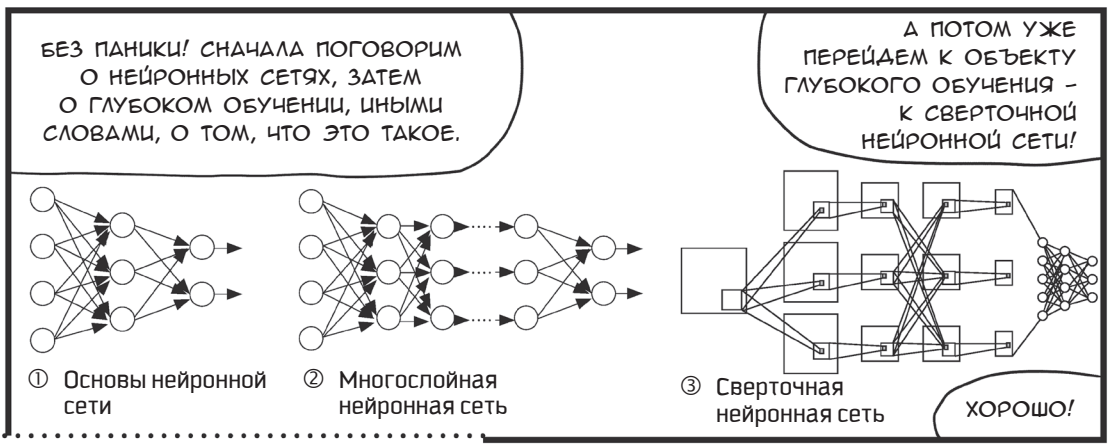

НУ, Я ДУМАЛ, САЯКА-СЭМПАЙ, ВЫ МНЕ ПОМОЖЕТЕ.


Я НЕ ХОЧУ СНОВА ОБЛАЖАТЬСЯ!
ОТЧАЯННО

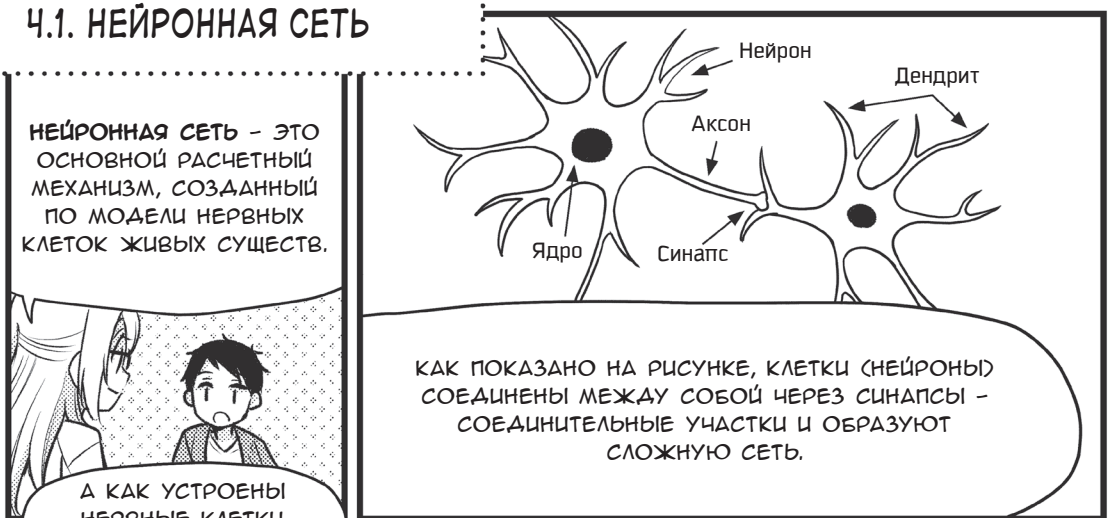

ПУСТЬ В ПРОШЛЫЙ РАЗ У ТЕБЯ НЕ ПОЛУЧИЛОСЬ, НО В ЭТОТ ПОЛУЧИТСЯ! Я ТЕБЕ ВСЕ РАССКАЖУ!
ДА! СПАСИБО ВАМ!


ИТАК, РАЗБЕРЕМ СЕГОДНЯ, КАК СОЗДАТЬ АВТОМАТИЗИРОВАННУЮ СИСТЕМУ ПО СОРТИРОВКЕ ВИНОГРАДА.


КАК МЫ МОЖЕМ ЕГО СОРТИРОВАТЬ?
ПО РАЗМЕРУ, ФОРМЕ, ЦВЕТУ, ВЕЛИЧИНЕ СЕМЕЧЕК И ФОРМЕ ГРОЗДИ.
Форма
Величина семечек
Размер
Цвет
Форма грозди



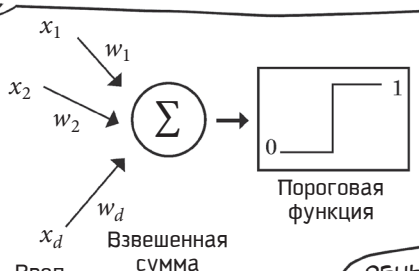
Ч.1. НЕЙРОННАЯ СЕТЬ



КАЖДЫЙ НЕЙРОН ПОЛУЧАЕТ ПОЛОЖИТЕЛЬНО ИЛИ ОТРИЦАТЕЛЬНО ЗАРЯЖЕННЫЙ СИГНАЛ ОТ СВЯЗАННЫХ С НИМ НЕЙРОНОВ.

КОГДА СУММА ЗАРЯДОВ ДОСТИГАЕТ ОПРЕДЕЛЕННОГО ЗНАЧЕНИЯ, ОН НАЧИНАЕТ ГЕНЕРИРОВАТЬ ЭЛЕКТРИЧЕСКИЙ СИГНАЛ.

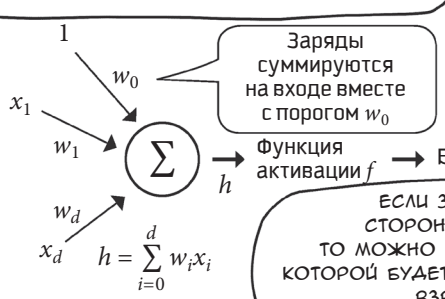
ТАКИЕ НЕЙРОНЫ, КАК НА КАРТИНКЕ, НАЗЫВАЮТСЯ **ПОРОГОВЫМИ ЛОГИЧЕСКИМИ ЭЛЕМЕНТАМИ**.



ПОРОГОВЫМИ?

РЕЧЬ ИДЕТ О ПРЕДЕЛЕ, ИЛИ ГРАНИЦЕ.

ФУНКЦИЯ $f(h)$ ОТ СУММЫ ЗАРЯДОВ НА ВХОДЕ h НАЗЫВАЕТСЯ **ФУНКЦИЕЙ АКТИВАЦИИ**.



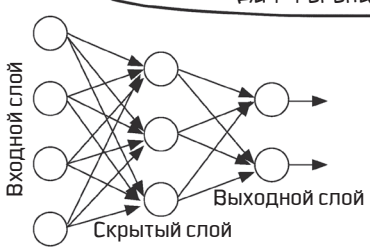
Заряды суммируются на входе вместе с порогом w_0

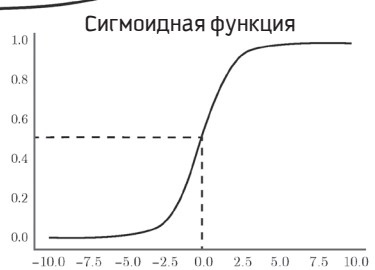
ЕСЛИ ЗАДАТЬ ЗНАЧЕНИЕ ПОРОГА w_0 И ДОБАВИТЬ НА ВХОДНОЙ СТОРОНЕ ВХОД, ВЕЛИЧИНА КОТОРОГО ВСЕГДА БУДЕТ РАВНА w_0, ТО МОЖНО ИСПОЛЬЗОВАТЬ ПРОСТУЮ ФУНКЦИЮ АКТИВАЦИИ, ВЫХОД КОТОРОЙ БУДЕТ РАВЕН 1 ИЛИ 0 В ЗАВИСИМОСТИ ОТ ТОГО, ЯВЛЯЕТСЯ ЛИ ВЗВЕШЕННАЯ СУММА ПОЛОЖИТЕЛЬНОЙ ИЛИ ОТРИЦАТЕЛЬНОЙ.

СОЕДИНЕНИЕ МНОЖЕСТВА СЛОЕВ ТАКИХ ЭЛЕМЕНТОВ НАЗЫВАЕТСЯ **НЕЙРОННОЙ СЕТЬЮ**. КОГДА МНОЖЕСТВО СЛОЕВ ЭЛЕМЕНТОВ ОБРАЗУЕТ НЕЙРОННУЮ СЕТЬ, ТО ПОРОГОВАЯ ФУНКЦИЯ, ФУНКЦИЯ АКТИВАЦИИ, ВЕДЕТ СЕБЯ КАК СИГМОИДНАЯ ФУНКЦИЯ. МЫ ЕЕ ВСТРЕЧАЛИ, КОГДА ГОВОРИЛИ О ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ.

А ПОЧЕМУ ИСПОЛЬЗУЕТСЯ СИГМОИДНАЯ ФУНКЦИЯ?

ПРИ ОБУЧЕНИИ ФУНКЦИЮ АКТИВАЦИИ НЕОБХОДИМО ДИФФЕРЕНЦИРОВАТЬ, А ПОРОГОВУЮ ФУНКЦИЮ ДИФФЕРЕНЦИРОВАТЬ НЕЛЬЗЯ.





ЭТО **НЕЙРОННАЯ СЕТЬ С ПРЯМЫМ РАСПРОСТРАНЕНИЕМ.**

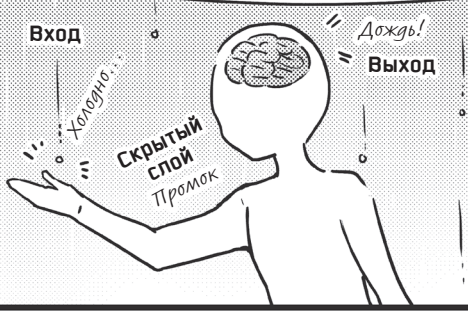
МОДЕЛЬ, СИГНАЛ В КОТОРОЙ МОЖЕТ ТОЛЬКО ИДТИ ВПЕРЕД.

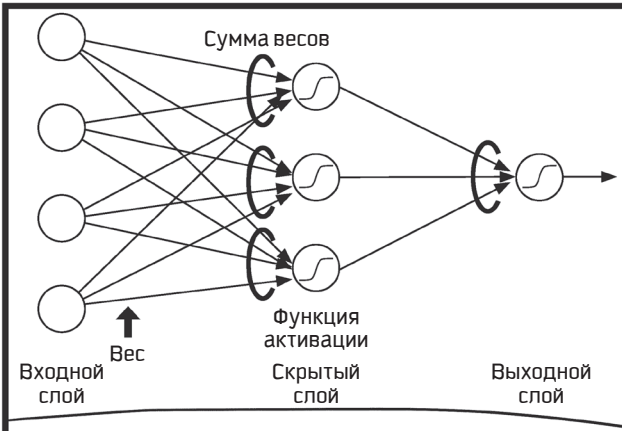
ДА.
ВСЕ ЭЛЕМЕНТЫ СОЕДИНЕНЫ С СОСЕДНИМИ СЛОЯМИ, НО БЕЗ ОБРАТНОЙ СВЯЗИ. СИГНАЛ ИДЕТ ОТ ВХОДА К ВЫХОДУ В ПРЯМОМ НАПРАВЛЕНИИ.

ЧЕРЕЗ ТРИ СЛОЯ?

ПРИ ЧИСЛЕННЫХ РАСЧЕТАХ МОЖНО ИСПОЛЬЗОВАТЬ ТОЛЬКО ДВА, СКРЫТЫЙ СЛОЙ И ВЫХОДНОЙ.

ЕСЛИ ПРОВЕСТИ АНАЛОГИЮ С ЖИВЫМИ ОРГАНИЗМАМИ, ТО **ВХОДНОЙ СЛОЙ** – ЭТО КЛЕТКИ, КОТОРЫЕ РЕАГИРУЮТ НА ВНЕШНИЕ РАЗДРАЖИТЕЛИ, **СКРЫТЫЙ СЛОЙ** – КЛЕТКИ, КОТОРЫЕ ПЕРЕДАЮТ СИГНАЛЫ, А **ВЫХОДНОЙ СЛОЙ** – КЛЕТКИ МОЗГА, КОТОРЫЕ КЛАССИФИЦИРУЮТ СИГНАЛЫ.





ВХОДНОЙ СИГНАЛ ВЫХОДИТ ИЗ ВХОДНОГО СЛОЯ БЕЗ ИЗМЕНЕНИЙ, ЕГО ЗАРЯДЫ СУММИРУЮТСЯ И ПОСТУПАЮТ НА СКРЫТЫЙ СЛОЙ. НА СКРЫТОМ СЛОЕ СУММА ЗАРЯДОВ, ПОСТУПИВШИХ СО ВСЕХ ВХОДНЫХ СЛОЕВ, ОБРАБАТЫВАЕТСЯ С ПОМОЩЬЮ ФУНКЦИИ АКТИВАЦИИ. В СЛУЧАЕ ЕСЛИ В НЕЙРОННОЙ СЕТИ С ПРЯМЫМ РАСПРОСТРАНЕНИЕМ РЕШАЕТСЯ ЗАДАЧА БИНАРНОЙ КЛАССИФИКАЦИИ, НА ВЫХОДНОМ СЛОЕ ОДИН ЭЛЕМЕНТ.
КАК И В СЛУЧАЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ, ВЫХОДНОЕ ЗНАЧЕНИЕ ЭТОГО СЛОЯ МОЖНО РАССМАТРИВАТЬ КАК ВЕРОЯТНОСТЬ ТОГО, ЧТО ВХОДНОЙ СИГНАЛ ОТНОСИТСЯ К ПОЛОЖИТЕЛЬНОМУ КЛАССУ.

ЕСЛИ ЖЕ КЛАССИФИКАЦИЯ ВЕДЕТСЯ ПО НЕСКОЛЬКИМ КАТЕГОРИЯМ, ТО КОЛИЧЕСТВО ВЫХОДНЫХ ЭЛЕМЕНТОВ СОВПАДАЕТ С КОЛИЧЕСТВОМ КЛАССОВ. В ТАКОМ СЛУЧАЕ ВЕРОЯТНО, ЧТО БОЛЬШОЕ КОЛИЧЕСТВО ЭЛЕМЕНТОВ НА ВЫХОДНОМ СЛОЕ ВЫДАСТ ЗНАЧЕНИЕ, БЛИЗКОЕ К 1.

ТОГДА В КАЧЕСТВЕ ФУНКЦИИ АКТИВАЦИИ $f(h)$ ИСПОЛЬЗУЕТСЯ НЕ СИГМОИДА, А **ФУНКЦИЯ SOFTMAX**.

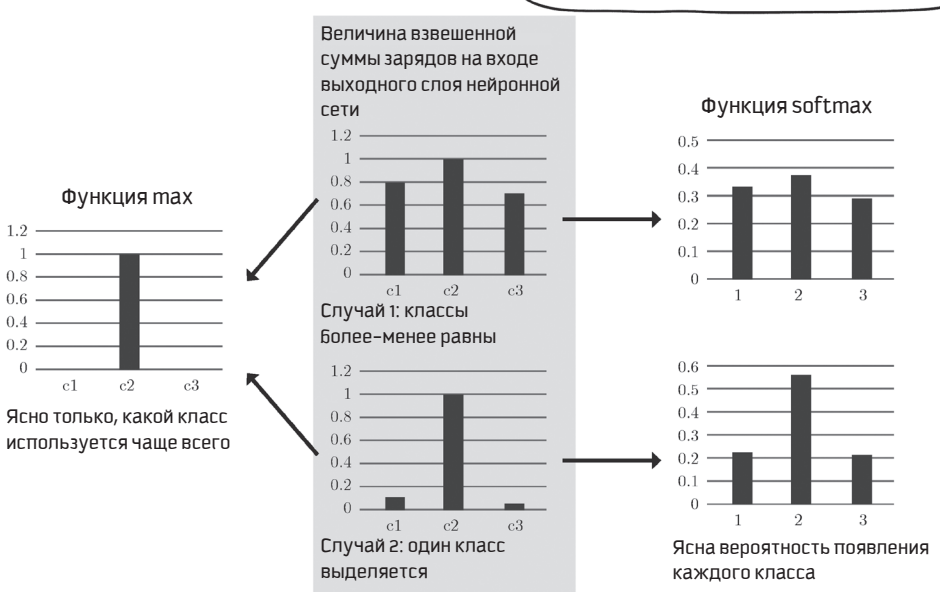
$$g_k = \frac{\exp(h_k)}{\sum_{j=1}^{c} \exp(h_j)}$$

ЗДЕСЬ h_k – СУММА ЗАРЯДОВ НА ВЫХОДЕ ИЗ СКРЫТОГО СЛОЯ, КОТОРАЯ СООТВЕТСТВУЕТ КОЛИЧЕСТВУ ЭЛЕМЕНТОВ ВЫХОДНОГО СЛОЯ И КОЛИЧЕСТВУ КЛАССОВ k.

КАК ВЫ ДУМАЕТЕ, МОЖНО ПРИМЕНИТЬ СИГМОИДНУЮ ФУНКЦИЮ К ВЗВЕШЕННОЙ СУММЕ ЗАРЯДОВ, КОТОРАЯ ПРИНИМАЕТ ЗНАЧЕНИЯ ОТ 0 ДО 1, А ЗАТЕМ ВЗЯТЬ МАКСИМАЛЬНОЕ ЗНАЧЕНИЕ?

ЕСЛИ В КАЧЕСТВЕ ФУНКЦИИ АКТИВАЦИИ ИСПОЛЬЗУЕТСЯ SOFTMAX, ТО ПРИ СЛОЖЕНИИ ВЫХОДОВ ВСЕХ ЭЛЕМЕНТОВ ВЫХОДНОГО СЛОЯ g_k ПОЛУЧИТСЯ 1, И ПОЭТОМУ МОЖНО ГОВОРИТЬ О ВЕРОЯТНОСТИ.

И ПОНЯТЬ ВЕРОЯТНОСТЬ ИСПОЛЬЗОВАНИЯ КАЖДОГО КЛАССА.



Ч.2. ОБУЧЕНИЕ МЕТОДОМ ОБРАТНОГО РАСПРОСТРАНЕНИЯ ОШИБОК

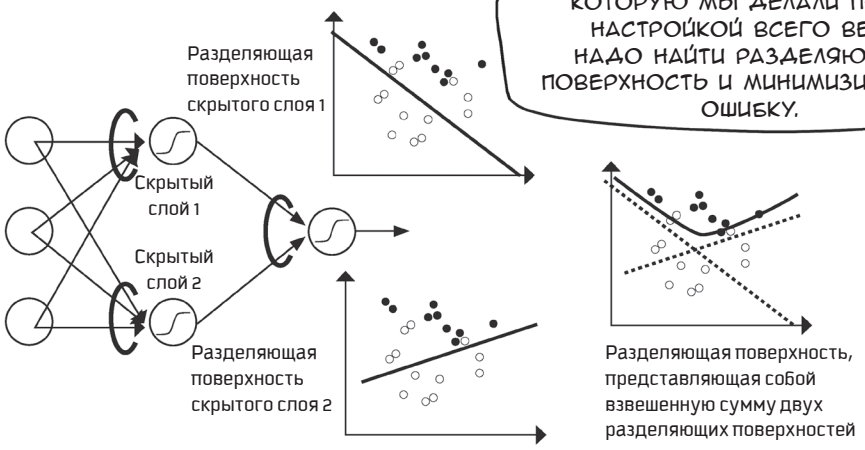
КОГДА МЫ ГОВОРИМ ОБ "ОБУЧЕНИИ" В НЕЙРОННЫХ СЕТЯХ С ПРЯМЫМ РАСПРОСТРАНЕНИЕМ, МЫ ГОВОРИМ О "ВЗВЕШЕННОЙ СУММЕ ЗАРЯДОВ ВХОДНОГО СИГНАЛА".

ПО ПОЛУЧЕННЫМ ДАННЫМ МЫ ОПРЕДЕЛЯЕМ И РЕГУЛИРУЕМ СУММУ ЗАРЯДОВ КАЖДОГО ЭЛЕМЕНТА.

КАК ЭТО ПРОИСХОДИТ?

КАЖДЫЙ ЭЛЕМЕНТ ВЫПОЛНЯЕТ НЕЛИНЕЙНОЕ ПРЕОБРАЗОВАНИЕ С ПОМОЩЬЮ СИГМОИДНОЙ ФУНКЦИИ. ЕСЛИ СЛОЖИТЬ СУММЫ ЗАРЯДОВ (ВЕСОВ), ТО ПОЛУЧИТСЯ НЕЛИНЕЙНЫЙ КЛАССИФИКАТОР ПРИЗНАКОВОГО ПРОСТРАНСТВА.

ЧТОБЫ УМЕНЬШИТЬ КОЛИЧЕСТВО ОШИБОК КЛАССИФИКАЦИИ В ЭТОМ НЕЛИНЕЙНОМ КЛАССИФИКАТОРЕ, КАК И В СЛУЧАЕ КЛАССИФИКАЦИИ, КОТОРУЮ МЫ ДЕЛАЛИ ПЕРЕД НАСТРОЙКОЙ ВСЕГО ВЕСА, НАДО НАЙТИ РАЗДЕЛЯЮЩУЮ ПОВЕРХНОСТЬ И МИНИМИЗИРОВАТЬ ОШИБКУ.



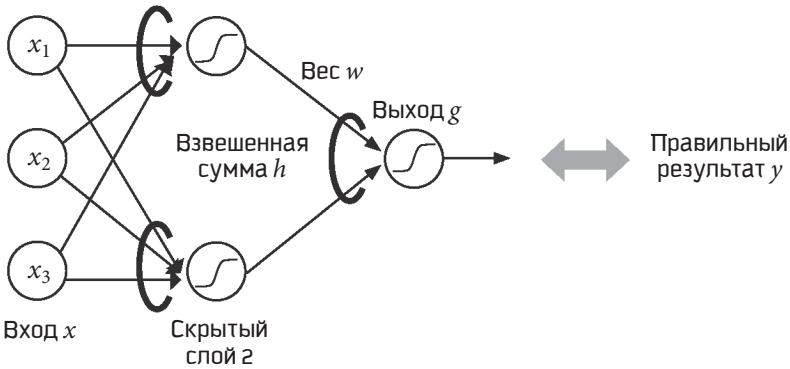
НАСТРОЙКА ВЕСА ОТ СКРЫТОГО СЛОЯ К ВЫХОДНОМУ ПРОИЗВОДИТСЯ С ПОМОЩЬЮ ВЫХОДА НЕЙРОННОЙ СЕТИ И **ОБУЧАЮЩЕГО СИГНАЛА**. ИНЫМИ СЛОВАМИ, ОБУЧЕНИЕ ВЕДЕТСЯ ПУТЕМ СРАВНЕНИЯ С ПРАВИЛЬНЫМ РЕЗУЛЬТАТОМ И ОЦЕНКИ ОШИБКИ.

А ЕСЛИ ОБУЧАЮЩЕГО СИГНАЛА НА СКРЫТОМ СЛОЕ НЕТ, ТО И ОШИБКИ, ПОЛУЧАЕТСЯ, СРАВНИВАТЬ НЕ С ЧЕМ.

УГУ

ИМЕННО! В ТАКОМ СЛУЧАЕ В МНОГОСЛОЙНОЙ СЕТИ ИДЕТ ОБУЧЕНИЕ ПУТЕМ **МЕТОДА ОБРАТНОГО РАСПРОСТРАНЕНИЯ ОШИБОК**.

Поговорим о методе обратного распространения ошибок. Пусть обучение ведется в нейронной сети с архитектурой, как на рисунке ниже:



В качестве данных для обучения возьмем пары: признаковое описание объекта x и результат y. Обозначим набор данных D, и в нем есть i-я пара (x_i, y_i). Функция ошибки может быть определена по-разному, но в данном случае будем минимизировать ее, вычисляя квадрат ошибки.

$$E(\boldsymbol{w}) \equiv \frac{1}{2} \sum_{\boldsymbol{x}_i \in D} (g_i - y_i)^2. \tag{4.1}$$

Здесь $\boldsymbol{w}$ обозначает всю сумму весов нейронной сети. Используя описанный в главе 2 метод градиентного спуска, найдем один вес w из весов $\boldsymbol{w}$ и, регулируя его величину, снизим ошибку.

$$w \leftarrow w - \eta \frac{\partial E(\boldsymbol{w})}{\partial w}. \tag{4.2}$$

Далее найдем частную производную по w функции ошибки $E(\boldsymbol{w})$, используя метод градиентного спуска. В этом случае изменение веса w изменит и выход g функции активации. Используя формулу дифференцирования сложной функции, получим следующее выражение:

$$\frac{\partial E(\boldsymbol{w})}{\partial w} = \frac{\partial E(\boldsymbol{w})}{\partial h} \frac{\partial h}{\partial w}. \tag{4.3}$$

После вычисления второго множителя в правой части формулы (4.3) из определения суммы весов h можно получить значение выхода предыдущего слоя в сочетании с весом w. Первый множитель можно записать, как показано ниже, с использованием производной. После вычислений у нас получится величина ошибки ε.

$$\varepsilon = \frac{\partial E(\boldsymbol{w})}{\partial h} = \frac{\partial E(\boldsymbol{w})}{\partial g}\frac{\partial g}{\partial h}. \tag{4.4}$$

В формуле 4.4 второй множитель справа – это производная функции активации. Поскольку в качестве функции активации используется сигмоида, то это $g(1 - g)$. Значение первого множителя справа зависит от того, между какими слоями вес определяется. Если w – это вес от скрытого слоя к выходному, то первый множитель – это производная функции ошибки.

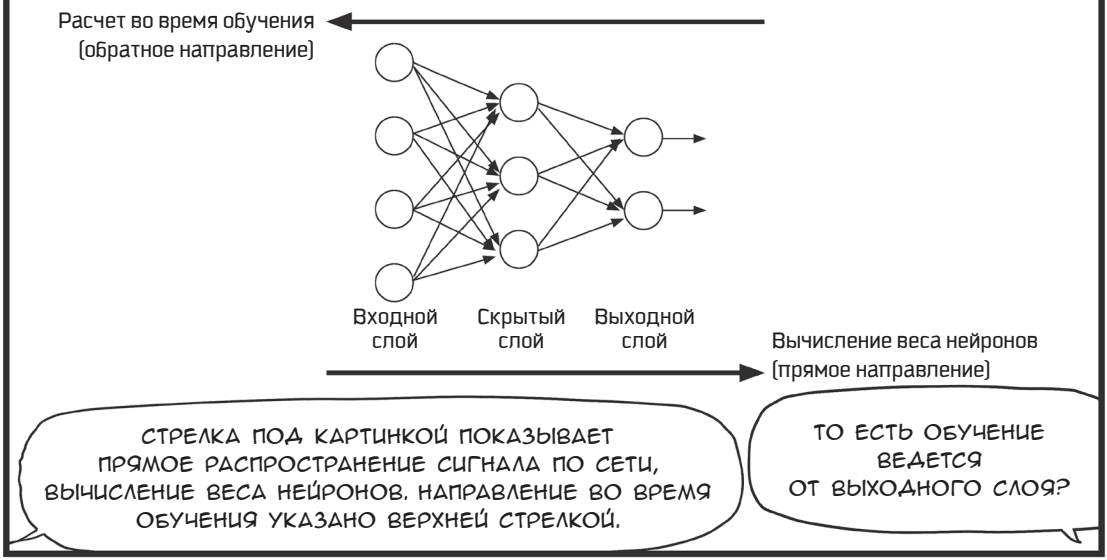
$$\frac{\partial E(\boldsymbol{w})}{\partial g} = g - y. \tag{4.5}$$

С другой стороны, если w – это вес от входного слоя к скрытому, то g в первом множителе – это выход скрытого слоя, и его величина будет влиять на выход выходного слоя h. Обычно выходных слоев много, и их обозначают h_j, где j – номер слоя, и первый множитель уравнения (4.4) примет такой вид:

$$\frac{\partial E(\boldsymbol{w})}{\partial g} = \sum_j \frac{\partial E(\boldsymbol{w})}{\partial h_j}\frac{\partial h_j}{\partial g} = \sum_j \varepsilon_j w_j. \tag{4.5}$$

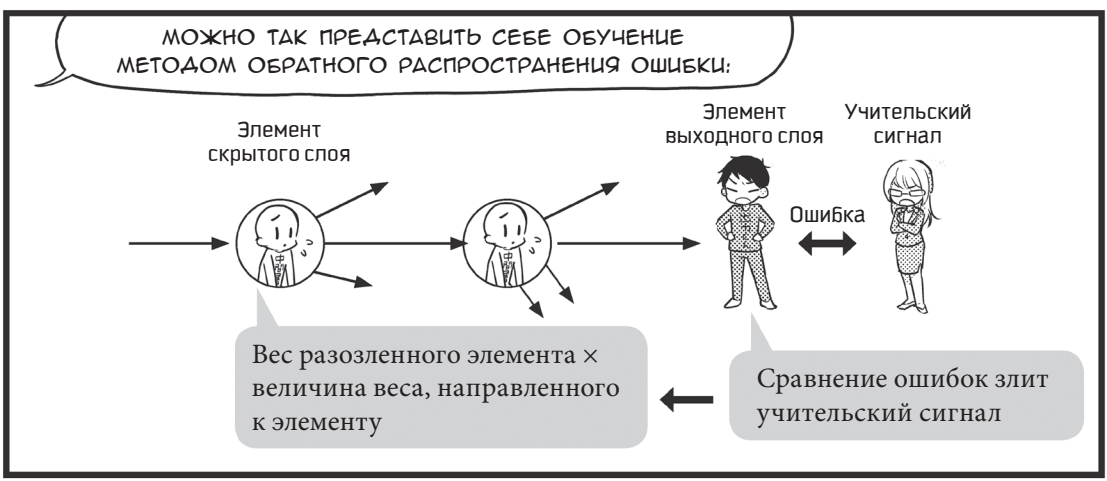
Мы используем ε из формулы (4.4), ε_j – это ошибка от скрытого слоя j, суть метода в том, что эта величина используется для корректировки суммы весов от входного слоя к скрытому. После обобщения ε можно вычислить и по формулам, приведенным ниже:

$$\varepsilon = \begin{cases} (g - y)g(1 - g) & \textbf{в случае от скрытого слоя к выходному} \\ \sum_j \varepsilon_j w_j g(1 - g) & \textbf{в случае от начального слоя к скрытому} \end{cases}$$


Расчет во время обучения
(обратное направление)
Входной слой
Скрытый слой
Выходной слой
Вычисление веса нейронов
(прямое направление)
СТРЕЛКА ПОД КАРТИНКОЙ ПОКАЗЫВАЕТ ПРЯМОЕ РАСПРОСТРАНЕНИЕ СИГНАЛА ПО СЕТИ, ВЫЧИСЛЕНИЕ ВЕСА НЕЙРОНОВ. НАПРАВЛЕНИЕ ВО ВРЕМЯ ОБУЧЕНИЯ УКАЗАНО ВЕРХНЕЙ СТРЕЛКОЙ.
ТО ЕСТЬ ОБУЧЕНИЕ ВЕДЕТСЯ ОТ ВЫХОДНОГО СЛОЯ?


МОЖНО ПРЕДСТАВИТЬ, ЧТО ВЫХОДНОЙ СЛОЙ ОТПРАВЛЯЕТ С ВЫХОДА НАЗАД СИГНАЛ ОБУЧЕНИЯ, И ЕСЛИ ОШИБКА ВЕЛИКА, ТО ОН ЗЛИТСЯ.
СЕРДИТСЯ
СЕРДИТСЯ
РАЗ ОДНИ ОШИБКИ, ТО ОН, КОНЕЧНО, БУДЕТ СЕРДИТ.


ЗАТЕМ РАЗОЗЛЕННЫЕ ВЫХОДНЫМ СЛОЕМ СИГНАЛЫ КОРРЕКТИРУЮТ ВЕЛИЧИНУ ВЕСОВ НА СРЕДНЕМ СЛОЕ И РУГАЮТ ЕГО.
СЕРДИТСЯ
СЕРДИТСЯ
КАК В ФИРМЕ....


МОЖНО ТАК ПРЕДСТАВИТЬ СЕБЕ ОБУЧЕНИЕ МЕТОДОМ ОБРАТНОГО РАСПРОСТРАНЕНИЯ ОШИБКИ:
Элемент скрытого слоя
Элемент выходного слоя
Учительский сигнал
Ошибка
Вес разозленного элемента × величина веса, направленного к элементу
Сравнение ошибок злит учительский сигнал

Ч.3. ВЫЗОВЫ ГЛУБОКОГО ОБУЧЕНИЯ

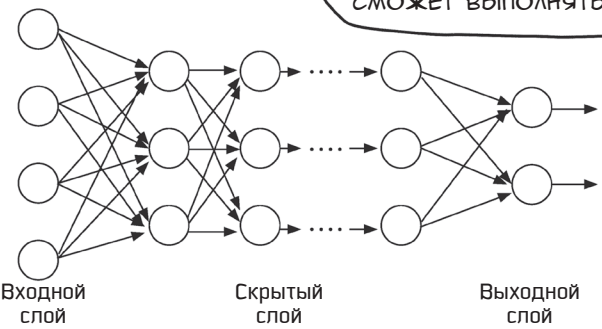
С чем легко справляются глубокие нейронные сети:

- распознавание звука;
- распознавание изображений;
- распознавание естественной речи.

4.3.1. Проблема глубокой нейронной сети

Так, для начала – проблема...

Казалось бы, что если увеличить количество скрытых слоев в нейронной сети с прямым обучением, как на картинке, то ее эффективность тоже вырастет, и она сможет выполнять разные задачи.



Однако в обучении многослойной нейронной сети методом обратного распространения ошибок при возвращении к корректировке величины веса возникает

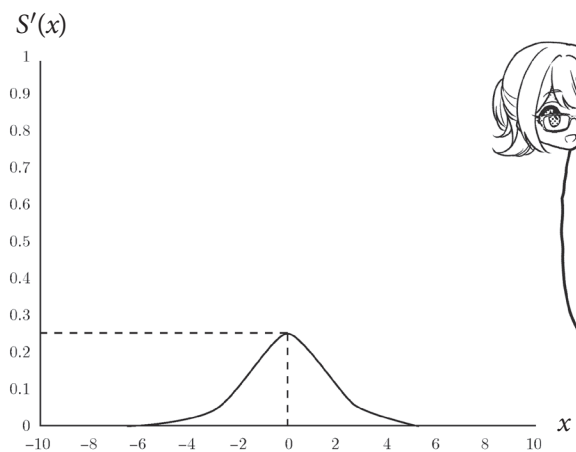
проблема исчезающего градиента...

Проблема исчезающего градиента?

Формула для вычисления изменения величины ошибки на каждом этапе представляет собой производную сигмоидной функции:

$$S(x) = S(x)(1 - S(x)).$$

Посмотри на график, как она себя ведет в реальной жизни.



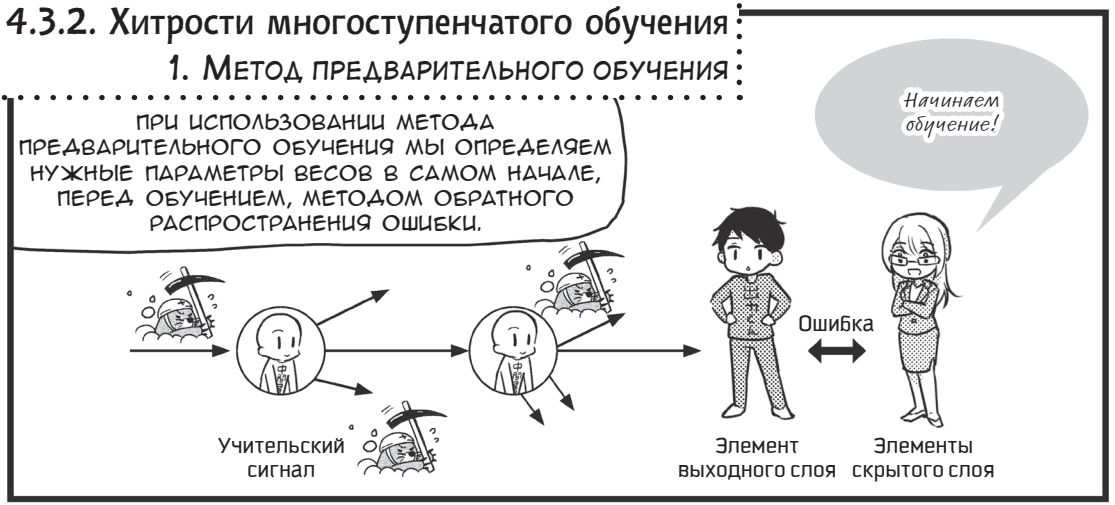
Максимальный градиент – 0,25.

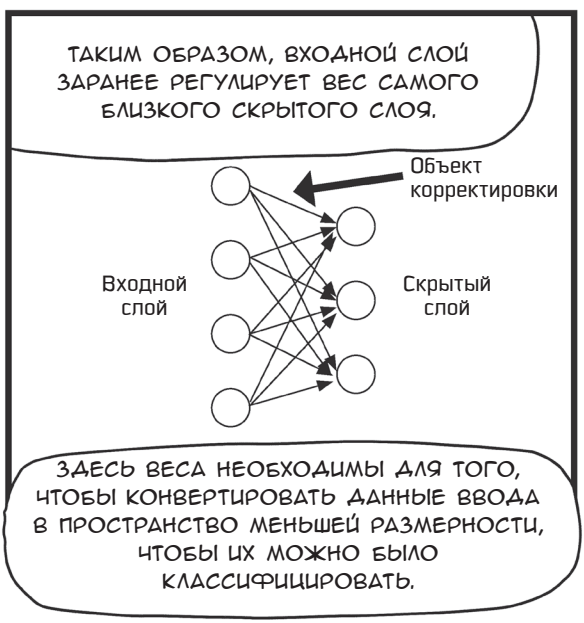
Да.
Это сравнительно высокое значение появляется, только когда входные сигналы равны 0 или близки к нему. А в остальных случаях градиент близится к 0.

Ого, тогда обучение не продвинется слишком далеко от входного слоя.

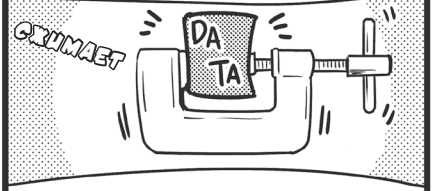
4.3.2. Хитрости многоступенчатого обучения

1. Метод предварительного обучения

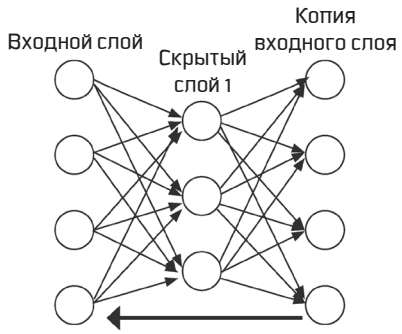




ТОГДА ПРЕДСТАВЬ, ЧТО ПРОБЛЕМА В ТОМ, ЧТО "НАДО СЖАТЬ ПРИЗНАКОВОЕ ОПИСАНИЕ ОБЪЕКТА, ЧТОБЫ ОНО УЛОЖИЛОСЬ В НЕБОЛЬШОЕ КОЛИЧЕСТВО ЕДИНИЦ И С МИНИМАЛЬНЫМИ ПОТЕРЯМИ, НАСКОЛЬКО ЭТО ВОЗМОЖНО".



ЭТО КАК?

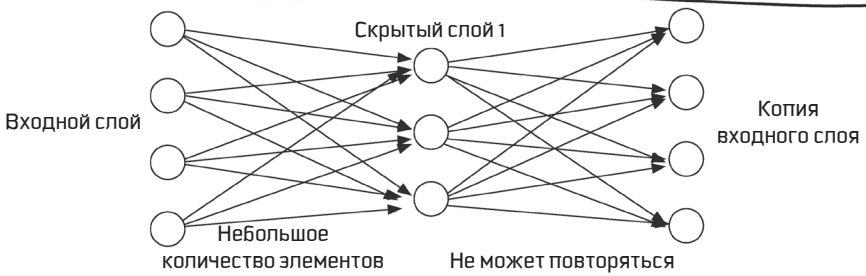


Обучение методом обратного распространения ошибок

ПРЕЖДЕ ВСЕГО СОЗДАЕТСЯ КОПИЯ ЭЛЕМЕНТОВ ВХОДНОГО СЛОЯ НАД СКРЫТЫМ СЛОЕМ, И ОНА СТАНОВИТСЯ ВЫХОДНЫМ СЛОЕМ. ЗАТЕМ ИНФОРМАЦИЯ ИЗ ВХОДНОГО СЛОЯ СНОВА ПОСТУПАЕТ НА ВЫХОДНОЙ И ПРОХОДИТ ОБУЧЕНИЕ. ЭТО ТАК НАЗЫВАЕМЫЙ АВТОКОДИРОВЩИК.

ТО ЕСТЬ КОПИЯ ВХОДНОГО СЛОЯ СТАНОВИТСЯ ВЫХОДНЫМ СЛОЕМ? ОНА ТАКАЯ ЖЕ, КАК И ВХОДНОЙ СЛОЙ?

В ЦЕЛОМ КОЛИЧЕСТВО ЭЛЕМЕНТОВ СКРЫТОГО СЛОЯ ПО СРАВНЕНИЮ С КОЛИЧЕСТВОМ ЭЛЕМЕНТОВ ВХОДНОГО СЛОЯ НЕВЕЛИКО, И ПОЭТОМУ ОНИ КОПИРУЮТ ИНФОРМАЦИЮ ВХОДНОГО СЛОЯ, ЧТОБЫ ВЫХОДНОЙ СЛОЙ НЕ ПОВТОРЯЛСЯ.

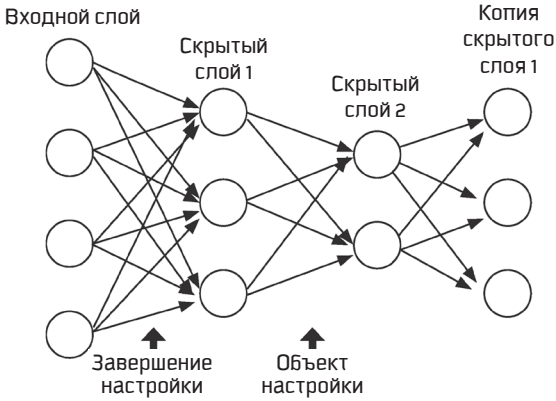


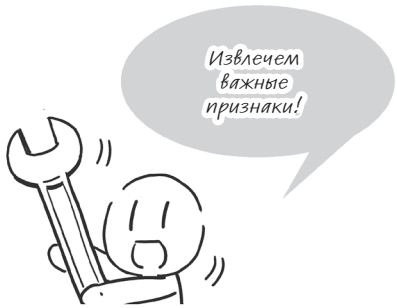
ВОТ КАК.

БЕЗ ЭТОГО НЕЛЬЗЯ ИЗВЛЕЧЬ ПРИЗНАКИ ДАННЫХ?

СКРЫТЫЙ СЛОЙ ПОЛУЧАЕТ ЗАДАЧУ "ИЗВЛЕЧЬ ИНФОРМАЦИЮ ИЗ ПРОСТРАНСТВА ДАННЫХ, СЖАТОГО ДО БОЛЕЕ НИЗКОГО ИЗМЕРЕНИЯ".

После того как настроены веса между входным слоем и скрытым слоем 1, они фиксируются, затем обучение ведется между скрытым слоем 1 и скрытым слоем 2, и так повторяется вплоть до выходного слоя.





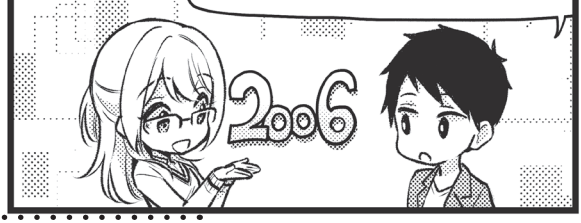
Если не извлекать признаки, то, поскольку количество элементов (узлов) уменьшается по мере удаления от входного слоя, информация не сохраняется.

Метод предварительного обучения позволил решить проблему исчезающего градиента и продолжать извлекать абстрактные информационные выражения, сохраняя информацию.

Метод предварительного обучения

После разработки этого метода в 2006 году стала возможной разработка глубоких нейронных сетей.

Действительно, прорыв...

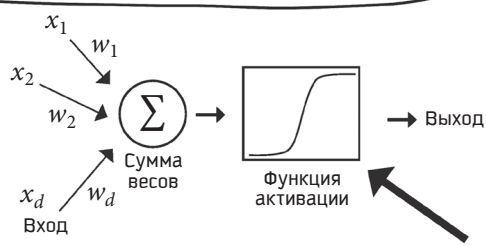


4.3.3. Хитрости многоступенчатого обучения

2. Функция активации

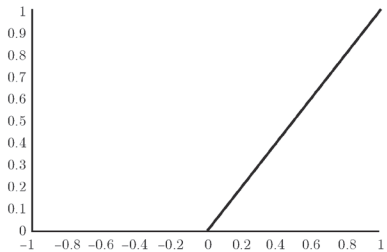
Другой подход к проблеме исчезающего градиента – это **метод настройки функции активации**.

Надо обратить внимание, как функция активации обращается с суммой весов.

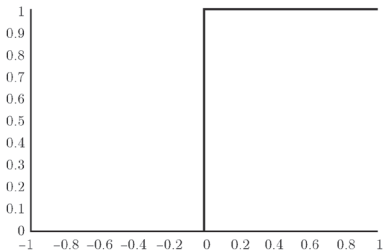


ВМЕСТО СИГМОИДЫ, О КОТОРОЙ МЫ ГОВОРИЛИ, ИСПОЛЬЗУЕТСЯ **УСЕЧЕННАЯ ЛИНЕЙНАЯ ФУНКЦИЯ**, ГДЕ $f(x) = \max(0, x)$. ЭЛЕМЕНТ, КОТОРЫЙ ИСПОЛЬЗУЕТ ЭТУ ФУНКЦИЮ, НАЗЫВАЕТСЯ **БЛОКОМ ЛИНЕЙНОЙ РЕКТИФИКАЦИИ (RELU)**.

УСЕЧЕННАЯ ЛИНЕЙНАЯ ФУНКЦИЯ – ЭТО ЛИНЕЙНАЯ ФУНКЦИЯ, КОТОРАЯ ВОЗВРАЩАЕТ ЗНАЧЕНИЕ x, ЕСЛИ x ПОЛОЖИТЕЛЬНО, И 0 В ПРОТИВНОМ СЛУЧАЕ.



Усеченная линейная функция



Производная усеченной линейной функции

В ОТЛИЧИЕ ОТ СИГМОИДЫ, ГРАДИЕНТ РАВЕН 1.

ПРИ ИСПОЛЬЗОВАНИИ БЛОКА ЛИНЕЙНОЙ РЕКТИФИКАЦИИ ГРАДИЕНТ РАВЕН 1, И ОШИБКА НЕ ИСЧЕЗАЕТ. ОДНАКО ВЫХОД МНОГИХ ЭЛЕМЕНТОВ РАВЕН 0, НЕЙРОННАЯ СЕТЬ МОЖЕТ С БОЛЬШОЙ СКОРОСТЬЮ СЧИТАТЬ ГРАДИЕНТЫ, И ОБУЧЕНИЕ МОЖНО ВЕСТИ БЕЗ ПРЕДВАРИТЕЛЬНОГО ОБУЧЕНИЯ.

КРУТО, ЧТО МОЖНО ОБОЙТИСЬ БЕЗ ПРЕДВАРИТЕЛЬНОГО ОБУЧЕНИЯ.

НЕТ

Предварительное обучение

ФУХ

4.3.4. Хитрости многоступенчатого обучения

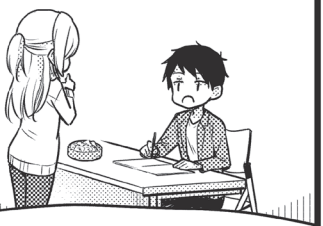
3. Как избежать переобучения

В глубоком обучении, кроме проблемы исчезающего градиента, есть еще и проблема переобучения.

Если в модели есть параметры с большими весами, они могут слишком хорошо объяснять примеры в данных для обучения.

Чтобы избежать переобучения, можно использовать **метод прореживания** (**dropout**), при котором переобучение происходит реже, а универсальность системы повышается.

А как работает этот метод?

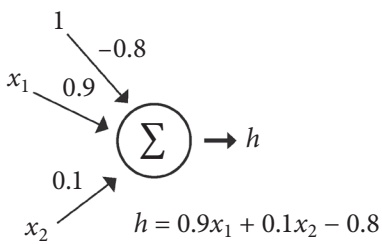


Обучение ведется путем исключения случайных нейронов.

Обучение путем исключения случайных элементов?

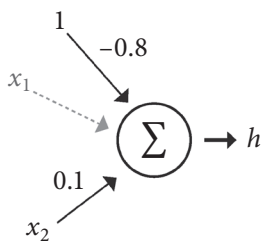
Первым делом в каждом слое случайным образом убираются элементы в зависимости от p.

Например, при $p = 0{,}5$ в нейронной сети работает половина элементов.



Хотя оба элемента – и x_1, и x_2 – важны, небольшая разница во время обучения приводит к большой разнице в весах в окончательном результате. Если величина x_1 в неизвестных данных будет даже ненамного меньше, это может привести к ошибочной идентификации.

Исключение

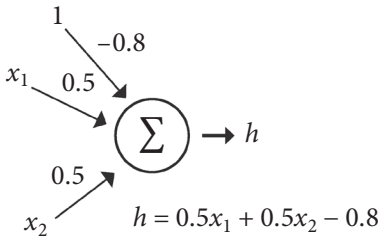


Поскольку любой входной элемент во время обучения может исчезнуть с определенной вероятностью, значения весов выровнены так, что только один из них даст правильный ответ.

Значения на входе в неизвестных данных могут немного меняться, но это не важно.

А затем обучение ведется методом обратного распространения ошибок на одном "мини-пакете" данных.

Когда логистическая классификация ведется уже обученной нейронной сетью, вес увеличивается в p раз. Производится обучение нескольких блоков нейронной сети, результаты усредняются.



Но почему при использовании метода прореживания не происходит переобучения?

Существует теория, что, во-первых, благодаря снижению степени свободы происходит регуляризация, а во-вторых, градиент остается таким же, даже несмотря на то, что распределение взвешенной суммы весов h на входе на узел растет.

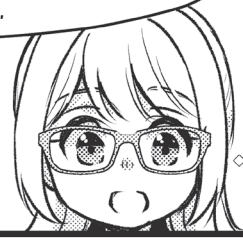
Конечно, до сих пор существуют разногласия среди ученых, но, как бы то ни было, если избежать многократного ввода одних и тех же данных в нейронных сетях с такой же структурой, то можно избежать переобучения.

Нельзя, чтобы глубокая сеть все запоминала!

4.3.5. Нейронные сети со специализированной структурой

Мы говорили о хитростях, которые помогают решить проблемы многослойных нейронных сетей, а теперь поговорим о еще одном методе обучения.

И что это за метод?

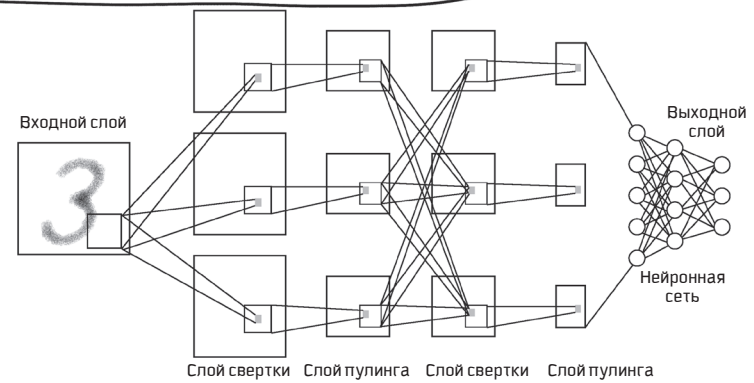


ЭТО МЕТОД СПЕЦИАЛИЗАЦИИ ЗАДАНИЙ В СТРУКТУРЕ НЕЙРОННОЙ СЕТИ. ПРИМЕРОМ СПЕЦИАЛИЗИРОВАННОЙ НЕЙРОННОЙ СЕТИ СЛУЖИТ **СВЕРТОЧНАЯ НЕЙРОННАЯ СЕТЬ**, КОТОРАЯ ИСПОЛЬЗУЕТСЯ ПРИ РАСПОЗНАВАНИИ ИЗОБРАЖЕНИЙ.

РАСПОЗНАВАНИЕ ЛИЦ

НЕДАВНО ОНА СТАЛА ОТЛИЧНО РАСПОЗНАВАТЬ ИЗОБРАЖЕНИЯ.

НА РИСУНКЕ ПОКАЗАНА НЕЙРОННАЯ СЕТЬ, ГДЕ ЧЕРЕДУЮТСЯ СЛОИ СВЕРТКИ И СЛОИ ПУЛИНГА. ПОСЛЕДНИЙ СЛОЙ ПУЛИНГА ПОЛУЧАЕТ ВЫВОД, И НА НЕМ РАЗМЕЩАЕТСЯ ОБЫЧНАЯ НЕЙРОННАЯ СЕТЬ.



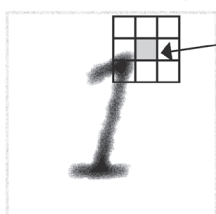
РАССМОТРИМ, КАК РАБОТАЕТ СВЕРТОЧНЫЙ СЛОЙ ПРИ ПОМОЩИ НАЛОЖЕНИЯ ФИЛЬТРА.

ФИЛЬТРА?

ПРИ НАЛОЖЕНИИ ФИЛЬТРА ВЫДЕЛЯЕТСЯ ПАТТЕРН, НАПРИМЕР ФИЛЬТР, КАК НА ИЗОБРАЖЕНИИ, ПРЕДСТАВЛЯЮЩИЙ СОБОЙ МАЛЕНЬКУЮ КАРТИНКУ 3×3, ТАК?

ТАК.

Развертка изображения по пикселям



Данные изображения x

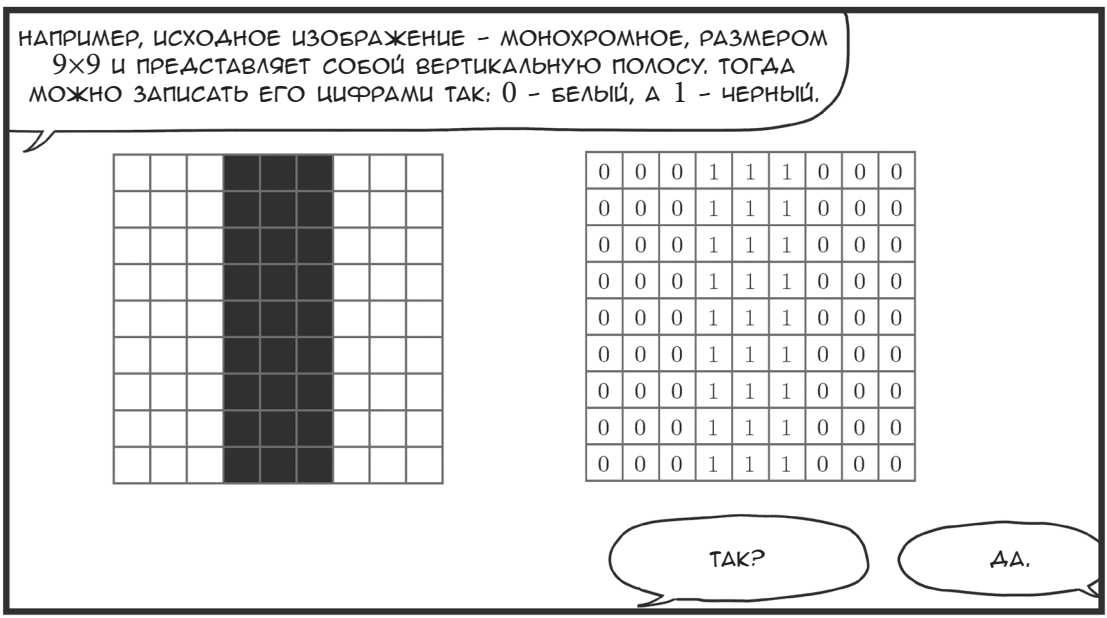
Значение изображения конвертируется в

$$\sum_{p=0}^{2}\sum_{q=0}^{2} x_{i+p,j+q} h_{pq}$$

−1	0	1
−1	0	1
−1	0	1

Фильтр выделения края h


ФИЛЬР ИЗОБРАЖЕНИЯ НАКЛАДЫВАЕТСЯ НА КАЖДЫЙ ПИКСЕЛЬ ЕГО РАЗВЕРТКИ, И ИЗ ИЗОБРАЖЕНИЯ ВЫДЕЛЯЕТСЯ ПАТТЕРН. В ЭТОМ ПРИМЕРЕ ФИЛЬТР ВЫДЕЛЕНИЯ КРАЯ БУДЕТ РЕАГИРОВАТЬ НА ИЗМЕНЕНИЕ ЦВЕТА В ВЕРТИКАЛЬНОМ НАПРАВЛЕНИИ.
В ВЕРТИКАЛЬНОМ НАПРАВЛЕНИИ? НО Я ВИЖУ ТОЛЬКО РЯДЫ ЦИФР: −1, 1 И 0.
-1 0 1
-1 0 1
-1 0 1


НАПРИМЕР, ИСХОДНОЕ ИЗОБРАЖЕНИЕ – МОНОХРОМНОЕ, РАЗМЕРОМ 9×9 И ПРЕДСТАВЛЯЕТ СОБОЙ ВЕРТИКАЛЬНУЮ ПОЛОСУ. ТОГДА МОЖНО ЗАПИСАТЬ ЕГО ЦИФРАМИ ТАК: 0 – БЕЛЫЙ, А 1 – ЧЕРНЫЙ.
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
0 0 0 1 1 1 0 0 0
ТАК?
ДА.


НА ИЗОБРАЖЕНИЕ НАКЛАДЫВАЕТСЯ ФИЛЬТР, И ИЩЕТСЯ ПАТТЕРН. НАПРИМЕР, ЕСЛИ ФИЛЬТР НАКЛАДЫВАЕТСЯ НА КАРТИНКУ СЛЕВА, ТО ЗНАЧЕНИЯ ПЕРЕМНОЖАЮТСЯ И СТАНОВЯТСЯ РАВНЫМИ 0.
0 × −1 + 0 × 0 + 0 × 1⋯
0 0 0
0 0 0
0 0 0
Изображение
Наложение
*
Плоское изображение
−1 0 1
−1 0 1
−1 0 1
Фильтр
= 0
ВИЖУ!

А ЧТО, ЕСЛИ ПОМЕНЯТЬ ЗНАЧЕНИЯ В ЦЕНТРАЛЬНОЙ ВЕРТИКАЛЬНОЙ ПОЛОСЕ ФИЛЬТРА?

0	1	1
0	1	1
0	1	1

Изображение

Наложение

*

Изображение с вертикальной полосой

−1	0	1
−1	0	1
−1	0	1

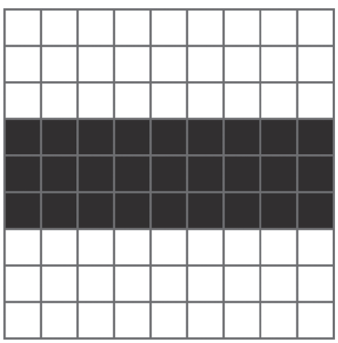
Фильтр

ТОГДА ПРИ УМНОЖЕНИИ И СЛОЖЕНИИ ПОЛУЧИТСЯ 3.

3 ЖЕ БОЛЬШЕ, ЧЕМ 0? ПРИ НАЛОЖЕНИИ ФИЛЬТРА ТОЛЬКО ВЕРТИКАЛЬНАЯ ПОЛОСА БУДЕТ ИМЕТЬ БОЛЬШИЕ ЗНАЧЕНИЯ. ЗНАЧИТ, МОЖНО ВЫДЕЛИТЬ ПАТТЕРН.

ПОНЯТНО.

А ЧТО ПОЛУЧИТСЯ, ЕСЛИ МЫ ВОЗЬМЕМ ИЗОБРАЖЕНИЕ С ГОРИЗОНТАЛЬНОЙ ПОЛОСОЙ И НАЛОЖИМ НА НЕГО ВЕРТИКАЛЬНЫЙ ФИЛЬТР ВЫДЕЛЕНИЯ КРАЯ?



0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

НУ, ЕСЛИ ПЕРЕМНОЖИТЬ,

0	0	0
1	1	1
1	1	1

*

-1	0	1
-1	0	1
-1	0	1

БУДЕТ 0.

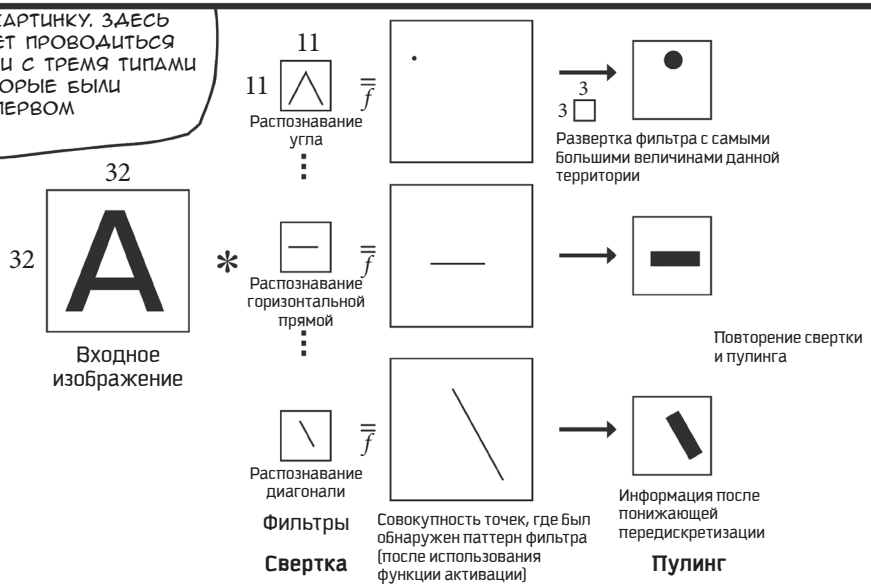
ДА!

Даже если у нас будет горизонтальная полоса, то все равно при перемножении и сложении участков с одинаковой мощностью (–1 справа и –1 слева) получится 0. Иными словами, фильтр выделения края не сможет выделить паттерн.

А если хотим выделить другой признак, то нам лучше поменять фильтр?

Да. На первом слое свертки как раз и определяется количество фильтров того же размера, что и изображение на входе, необходимых для извлечения признаков.

Посмотри на картинку. Здесь операция будет проводиться в соответствии с тремя типами фильтров, которые были отобраны на первом слое свертки.

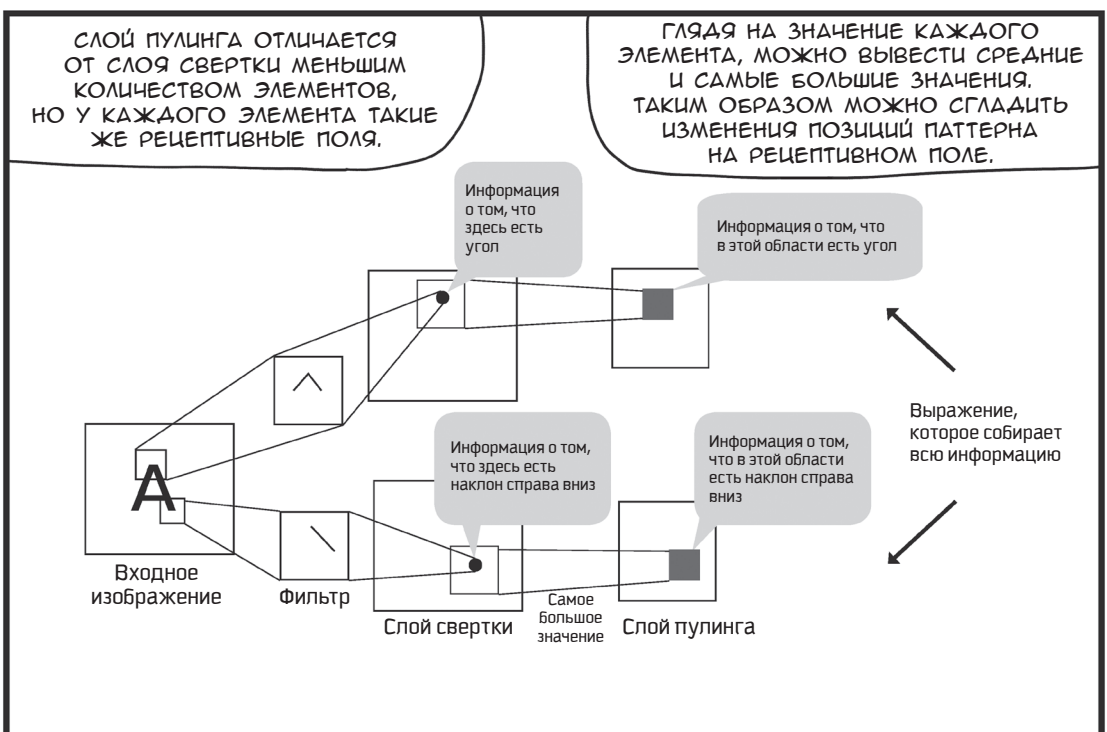


Чтобы распознать букву A, используются фильтры угла, горизонтали и диагонали – всего три фильтра.

Каждый элемент сверточного слоя связан только с частью входного изображения, и вес его распределяется между всеми нейронами. Область этой связи соответствует размеру фильтра и называется **рецептивным полем**.

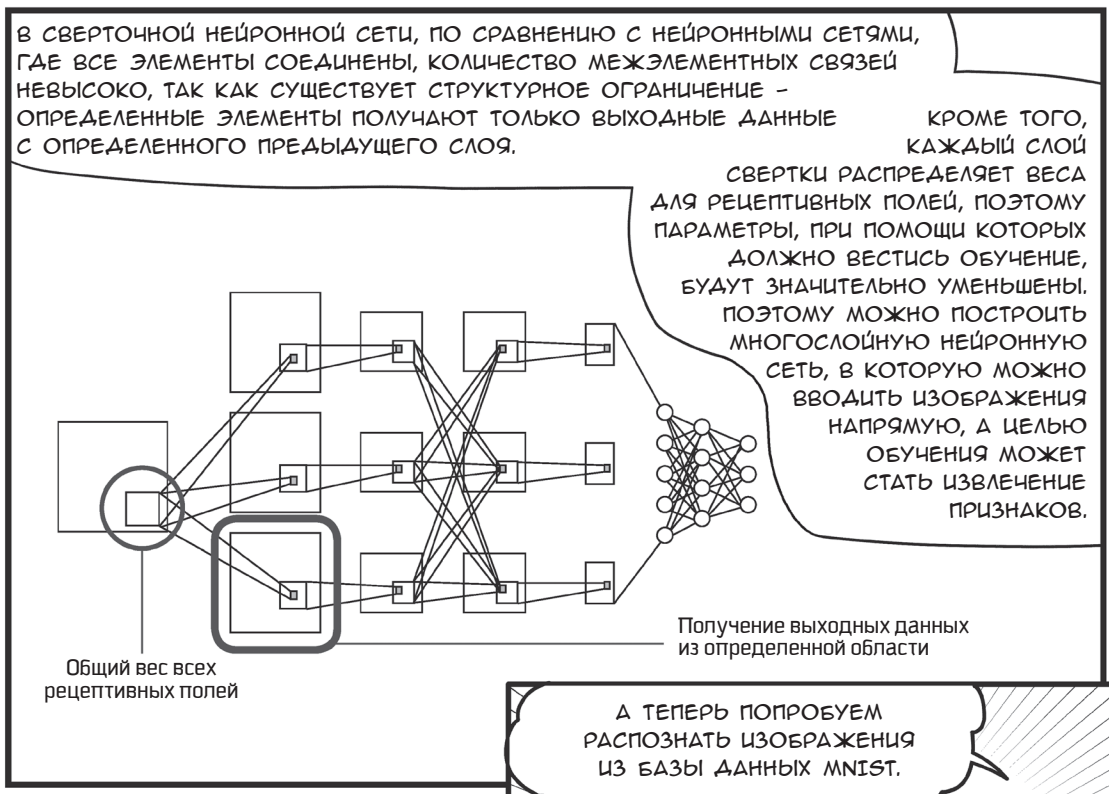

То есть фильтры точно так же выделяют паттерны.
Да. Но на следующем слое пулинга необходимо сжать полученные данные, отрегулировав признаки.
Объединенные данные фильтра
Информация после сжатия (понижающей передискретизации)
Пулинг
А почему?


Например, если необходимо распознать рукописный текст, ведь все пишут по-разному, не так ли?
Почерк у всех людей разный.
Поэтому в слое пулинга убираются мелкие детали информации, чтобы уменьшить различия.
Это как?


Слой пулинга отличается от слоя свертки меньшим количеством элементов, но у каждого элемента такие же рецептивные поля.
Глядя на значение каждого элемента, можно вывести средние и самые большие значения. Таким образом можно сгладить изменения позиций паттерна на рецептивном поле.
Информация о том, что здесь есть угол
Информация о том, что в этой области есть угол
Выражение, которое собирает всю информацию
Информация о том, что здесь есть наклон справа вниз
Информация о том, что в этой области есть наклон справа вниз
Входное изображение
Фильтр
Слой свертки
Самое большое значение
Слой пулинга


ЧТОБЫ ТЫ ПОНЯЛ, КАК РАБОТАЕТ СВЕРТОЧНАЯ НЕЙРОННАЯ СЕТЬ, ПРОДЕМОНСТРИРУЕМ ВЫСОКОЭФФЕКТИВНОЕ РАСПОЗНАВАНИЕ КАРТИНОК.


ДЛЯ ЭТОГО НЕОБХОДИМО БОЛЬШОЕ КОЛИЧЕСТВО ФОТОГРАФИЙ. И ГДЕ МОЖНО ЕГО НАЙТИ?
ДЕЙСТВИТЕЛЬНО.


В СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ, ПО СРАВНЕНИЮ С НЕЙРОННЫМИ СЕТЯМИ, ГДЕ ВСЕ ЭЛЕМЕНТЫ СОЕДИНЕНЫ, КОЛИЧЕСТВО МЕЖЭЛЕМЕНТНЫХ СВЯЗЕЙ НЕВЫСОКО, ТАК КАК СУЩЕСТВУЕТ СТРУКТУРНОЕ ОГРАНИЧЕНИЕ – ОПРЕДЕЛЕННЫЕ ЭЛЕМЕНТЫ ПОЛУЧАЮТ ТОЛЬКО ВЫХОДНЫЕ ДАННЫЕ С ОПРЕДЕЛЕННОГО ПРЕДЫДУЩЕГО СЛОЯ.
КРОМЕ ТОГО, КАЖДЫЙ СЛОЙ СВЕРТКИ РАСПРЕДЕЛЯЕТ ВЕСА ДЛЯ РЕЦЕПТИВНЫХ ПОЛЕЙ, ПОЭТОМУ ПАРАМЕТРЫ, ПРИ ПОМОЩИ КОТОРЫХ ДОЛЖНО ВЕСТИСЬ ОБУЧЕНИЕ, БУДУТ ЗНАЧИТЕЛЬНО УМЕНЬШЕНЫ. ПОЭТОМУ МОЖНО ПОСТРОИТЬ МНОГОСЛОЙНУЮ НЕЙРОННУЮ СЕТЬ, В КОТОРУЮ МОЖНО ВВОДИТЬ ИЗОБРАЖЕНИЯ НАПРЯМУЮ, А ЦЕЛЬЮ ОБУЧЕНИЯ МОЖЕТ СТАТЬ ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ.
Получение выходных данных из определенной области
Общий вес всех рецептивных полей


А ТЕПЕРЬ ПОПРОБУЕМ РАСПОЗНАТЬ ИЗОБРАЖЕНИЯ ИЗ БАЗЫ ДАННЫХ MNIST.
ДА!

Для кодирования глубокого обучения в Python лучше всего использовать библиотеки.

Здесь мы будем использовать библиотеку Keras. Keras представляет собой надстройку над фреймворками Deeplearning4j, TensorFlow и Theano. Поскольку она хорошо описана, кодирование для задач глубокого обучения будет несложным.

```
import keras
```

MNIST – база данных рукописных изображений символов. Одно изображение представляет собой квадрат размером 28×28 пикселей со значением насыщенности от 0 до 255.

Мы используем 60 000 изображений для обучения и 10 000 для оценки. MNIST автоматически загрузится в Keras, а для разделения изображения на данные для обучения и для оценки используем метод numpy array.

```
from keras.datasets import mnist
(X_train, y_train), (X_test, y_test) = mnist.load_data()
```

Затем немного обработаем данные. Сначала преобразуем вход в стандартную систему распознавания изображений сверточной нейронной сети.

Поскольку обычно распознаются цветные изображения, в одном изображении будет трехмерный тензор (количество пикселов по вертикали × количество пикселов по горизонтали × значение цвета), а в данных для ввода будет четырехмерный тензор, к которому добавится количество изображений.

Так как изображения в градациях серого, которые мы используем сейчас, будут описаны количеством изображений × количеством пикселей по вертикали × количеством пикселей по горизонтали, четвертое измерение тензора будет представлено «цветом» со значением 1.

Поскольку на входе нейронной сети будет небольшая область значений от 0 до 1, нет необходимости на первом этапе настраивать вес или разряды коэффициентов для обучения. Поскольку максимальная величина пиксела равна 255, после преобразования целочисленного типа в тип с плавающей запятой выполнится операция деления всех данных на 255.

```
img_rows, img_cols = 28, 28

X_train = X_train.reshape(X_train.shape[0], img_rows, img_cols, 1)
X_test = X_test.reshape(X_test.shape[0], img_rows, img_cols, 1)
input_shape = (img_rows, img_cols, 1)

X_train = X_train.astype('fl oat32') / 255
X_test = X_test.astype('fl oat32') /255
```

Далее настроим выход. Правильной меткой будет целое число от 0 до 9, которым обозначаются изображения, поэтому нужен десятимерный вектор, называющийся one-hot.

При использовании унитарного кодирования только один выход равен 1, а оставшиеся – 0. На выходном слое нейронной сети количество классов (на этот раз 10) будет преобразовано в формат, который будет легко подаваться в качестве учительского сигнала.

```
from keras.utils import to_categorical
Y_train = to_categorical(y_train)
Y_test = to_categorical(y_test)
```

А теперь определим структуру сверточной нейронной сети. У нас будет по два слоя свертки (фильтр размером 3×3) и пулинга (размер 2×2), выход будет конвертирован в одномерный вектор и перенаправлен на двухслойную нейронную сеть для классификации. В качестве функции активации на выходном слое используется softmax, а на остальных RELU.

```
from keras.models import Sequential
from keras.layers import Conv2D, MaxPooling2D, Flatten, Dense

n_out = len(Y_train[0]) # 10

model = Sequential()
model.add(Conv2D(16, kernel_size=(3, 3),
                 activation='relu',
                 input_shape=input_shape))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Conv2D(32, (3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Flatten())
model.add(Dense(128, activation='relu'))
model.add(Dense(n_out, activation='softmax'))
model.summary()
```

Применяя метод summary Keras, посмотрим структуру полученной нейросети.

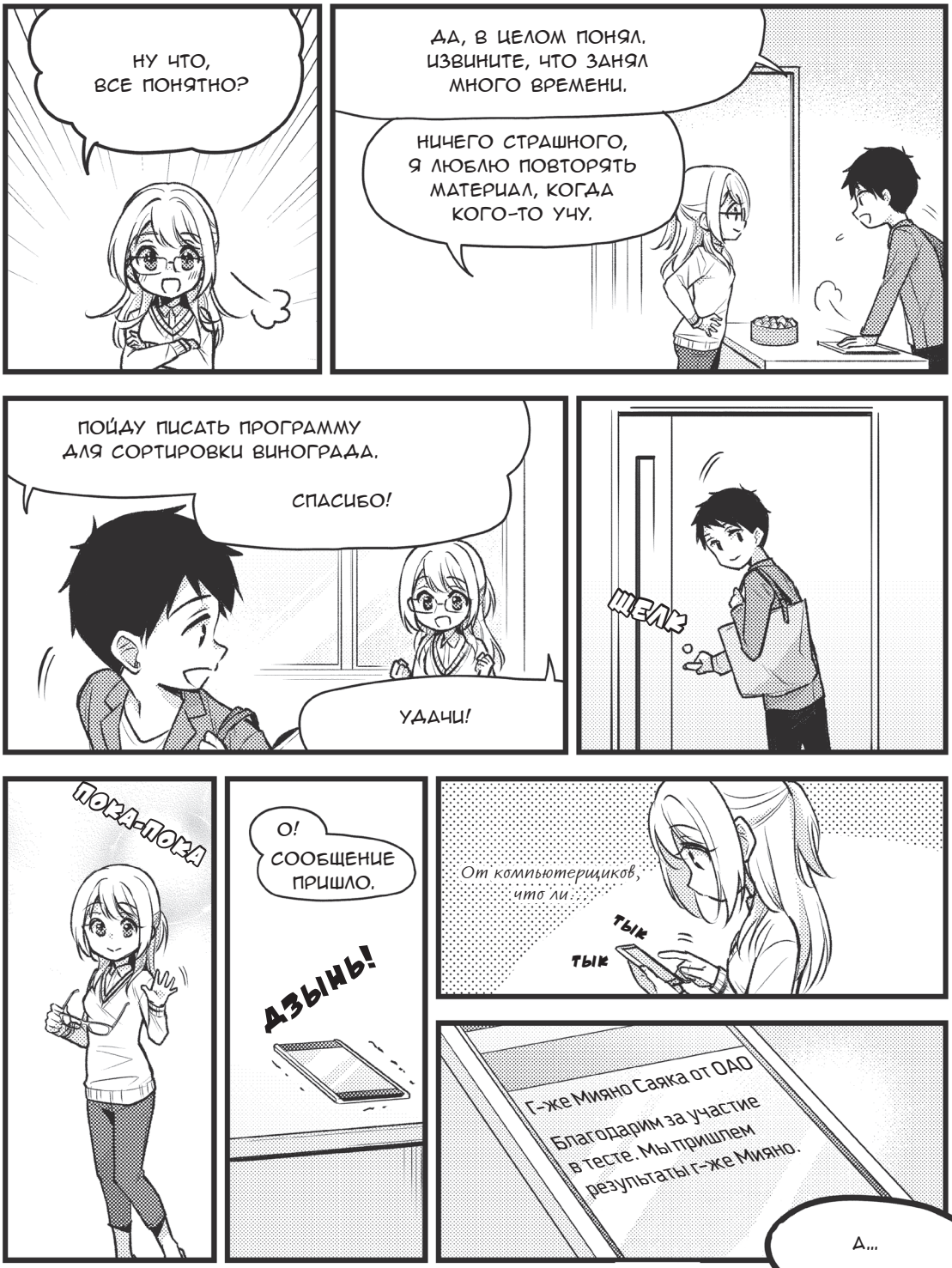
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 26, 26, 16)        160
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 13, 13, 16)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 11, 11, 32)        4640
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 5, 5, 32)          0
_________________________________________________________________
fl atten_1 (Flatten)         (None, 800)               0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               102528
_________________________________________________________________
dense_2 (Dense)              (None, 10)                1290
=================================================================
Total params: 108,618
Trainable params: 108,618
Non-trainable params: 0
_________________________________________________________________
```

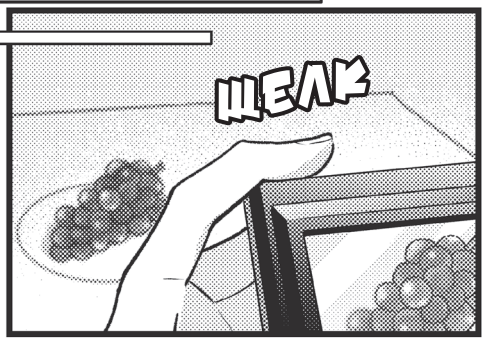
При помощи метода compile найдем categorical cross entropy и rmsprop, а также проведем обучение методом fit.

```
model.compile(loss = 'categorical_crossentropy',
              optimizer = 'rmsprop',
              metrics = ['accuracy'])
model.fi t(X_train, Y_train, epochs=5, batch_size=200)
score = model.evaluate(X_test, Y_test, verbose=0)
print('Test loss:', score[0])
print('Test accuracy:', score[1])

Epoch 1/5
60000/60000 [==============================] - 13s 224us/step - loss: 0.2883 -
acc: 0.9130
Epoch 2/5
60000/60000 [==============================] - 13s 210us/step - loss: 0.0763 -
acc: 0.9765
Epoch 3/5
60000/60000 [==============================] - 14s 239us/step - loss: 0.0516 -
acc: 0.9836
Epoch 4/5
60000/60000 [==============================] - 14s 238us/step - loss: 0.0384 -
acc: 0.9874
Epoch 5/5
60000/60000 [==============================] - 14s 235us/step - loss: 0.0306 -
acc: 0.9906
Test loss: 0.03475515839108266
Test accuracy: 0.9878
```

Точность распознавания – 98,78 %. Она очень высокая.


НУ ЧТО, ВСЕ ПОНЯТНО?
ДА, В ЦЕЛОМ ПОНЯЛ. ИЗВИНИТЕ, ЧТО ЗАНЯЛ МНОГО ВРЕМЕНИ.
НИЧЕГО СТРАШНОГО, Я ЛЮБЛЮ ПОВТОРЯТЬ МАТЕРИАЛ, КОГДА КОГО-ТО УЧУ.
ПОЙДУ ПИСАТЬ ПРОГРАММУ ДЛЯ СОРТИРОВКИ ВИНОГРАДА.
СПАСИБО!
УДАЧИ!
ЩЕЛК
ПОКА-ПОКА
О! СООБЩЕНИЕ ПРИШЛО.
ДЗЫНЬ!
От компьютерщиков, что ли...
ТЫК
ТЫК
Г-же Мияно Саяка от ОАО
Благодарим за участие в тесте. Мы пришлем результаты г-же Мияно.
А...



СЕЙЧАС СИСТЕМА РАСПОЗНАЕТ ВИНОГРАД С ТОЧНОСТЬЮ ДО 98 %.

Попросил прислать данные для обучения – множество фотографий винограда.

Выделил обучающую, тестовую и контрольную выборки.

Сделал систему сортировки – сверточную нейронную сеть на основании данных для обучения.

Оценил при помощи данных для анализа, откорректировал фильтры и элементы.

Наконец, протестировал на тестовой выборке.

ТЕПЕРЬ ДВЕ ОШИБКИ ИЗ 100 ФОТОГРАФИЙ, ЗНАЧИТ, НАДО ПОДНЯТЬ ЭФФЕКТИВНОСТЬ.

УГУ.


Я же недавно тебя угощал!
Как-то маловато за мою помощь...
Это тебя.
Да? Что?
Голос из телефона
Уговори их на 98 %, помощник!
Что?
Это Хасимото!
А-алло!
Это Киёхара из мэрии. Простите, если не вовремя.
Кудзё-ё-ё-ё-ё-ё-ё-ё!
А, Киёхара-сан! Здравствуйте. Вы о сортировке винограда? Все настроили?
Да, но точность составляет 98 %.
98 %?!
Ну... мы бы хотели ее еще увеличить, если можно...
Ха-ха-ха, Киёхара-сан, и 98 % хватит.
Что?
Люди на 5 % ошибаются, так что все в порядке!


СПАСИБО ЗА ВАШУ УСЕРДНУЮ РАБОТУ! ВЫ НАС ОЧЕНЬ ВЫРУЧИЛИ!
ДА НЕТ, СПАСИБО ВАМ...
ТЫК
ТЫК
МОЛОДЕЦ, КИЁХАРА! ПОШЛИ УЖИНАТЬ...
НУ ПОШЛИ...
ЖАЛЬ Я НЕ СМОГ УЛУЧШИТЬ ПРОИЗВОДИТЕЛЬНОСТЬ ДО ПРЕДЕЛА...
У МЕНЯ ЕЩЕ МАЛО ОПЫТА, НО Я УЖЕ МОГУ ПОМОГАТЬ ЛЮДЯМ...
ТЫК
БУДУ СТАРАТЬСЯ.
Как важно, что сегодня я услышал «спасибо»
БЫСТРЕЙ, КИЁХАРА.
ИДУ!

В кабинете у Саяка (5)

Математическое повторение (4)

Киёхара-кун, кажется, исправился. Я слышала, что у него были проблемы с сайтом.

Ну, надо было мне его сначала научить как следует... Но хорошо, что он исправился.

Я только слышала ваш разговор, но не кажется ли тебе, что он уже повзрослел?

Да, когда он учился, положиться на него было нельзя. Но сейчас он, кажется, стал получше...

Я хотела бы с ним встретиться, если можно. Кстати, разве он не симпатичный?

Ну… Да вроде бы и нет, не знаю. Когда можно будет встретиться, я дам знать.

Да, если есть время. (Какая упрямая сестрица. Бедный Киёхара.)

Сегодня мы говорили о глубоком обучении. Но это на практике выглядит страшно.

Если привыкнуть к тензорам, то его код становится простым. Тензор – многомерное пространство на языке программирования.

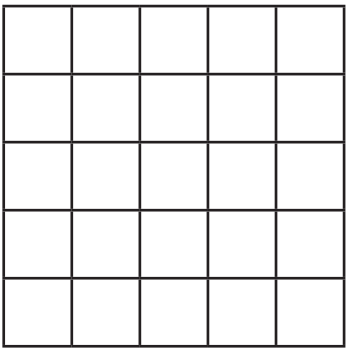
Я знаю, что в программировании есть двухмерное распределение, но кроме этого ничего не слышала.

Ай-тян, ты писала программу по обработке изображений?

Я этим занималась, но программу я не писала.

Ну, представь себе черно-белое изображение. Пусть черный – 0, а белый – 255, а цвета между ними могут принимать значения от 1 до 254. Чем темнее серый, тем меньше число, а чем светлее, тем больше. И как мы выразим данные, которые расположены по горизонтали и вертикали в виде прямоугольника?

В виде двухмерной фигуры из целых чисел?



Ага. Эта двухмерная фигура, которую можно представить в виде матрицы.

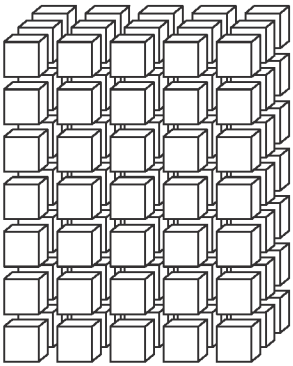
И ее можно записать в фигурных скобках.

А теперь возьмем цветное изображение. Обычно используется формат RGB, где 1 пиксель может иметь значение красного, зеленого или синего цвета. При сложении трех цветов получается еще один цвет, и можно записывать разные выражения.

Это интересно.

Как тогда будут представлены данные?

Поскольку красный, зеленый и синий представлены двумерными массивами одного размера, то если их сгруппировать, получится трехмерный массив?



Да, численное выражение структуры данных в таком случае будет выражено 3D-тензором.

А еще есть 4D-тензор, который состоит из нескольких 3D-изображений...

Как нескольких цветных изображений… И это все данные для обучения?

Именно. Какой тензор будет использоваться для видео?

Если у нас видео, то изображения выстроены по оси времени. И чтобы его обработать, 4D-мерный тензор становится… 5D?

Да!

Ого… тензоры все сложнее и сложнее.

Но в машинном обучении используются только они. В глубоком обучении, чтобы упростить последнюю классификацию, преобразуют тензоры, и все становится легко.

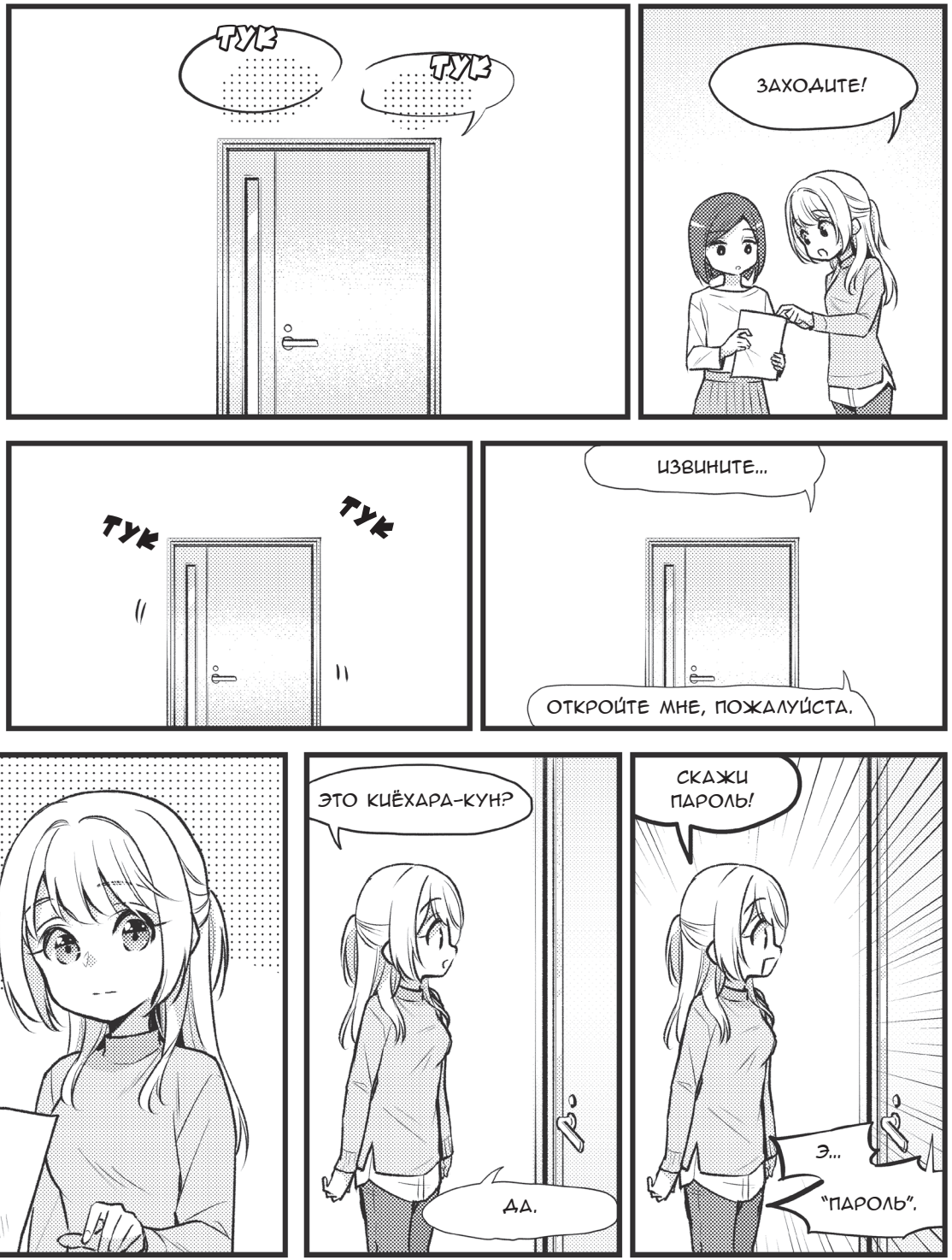
Ага… Кстати, тетя спрашивала, нашла ли ты работу?

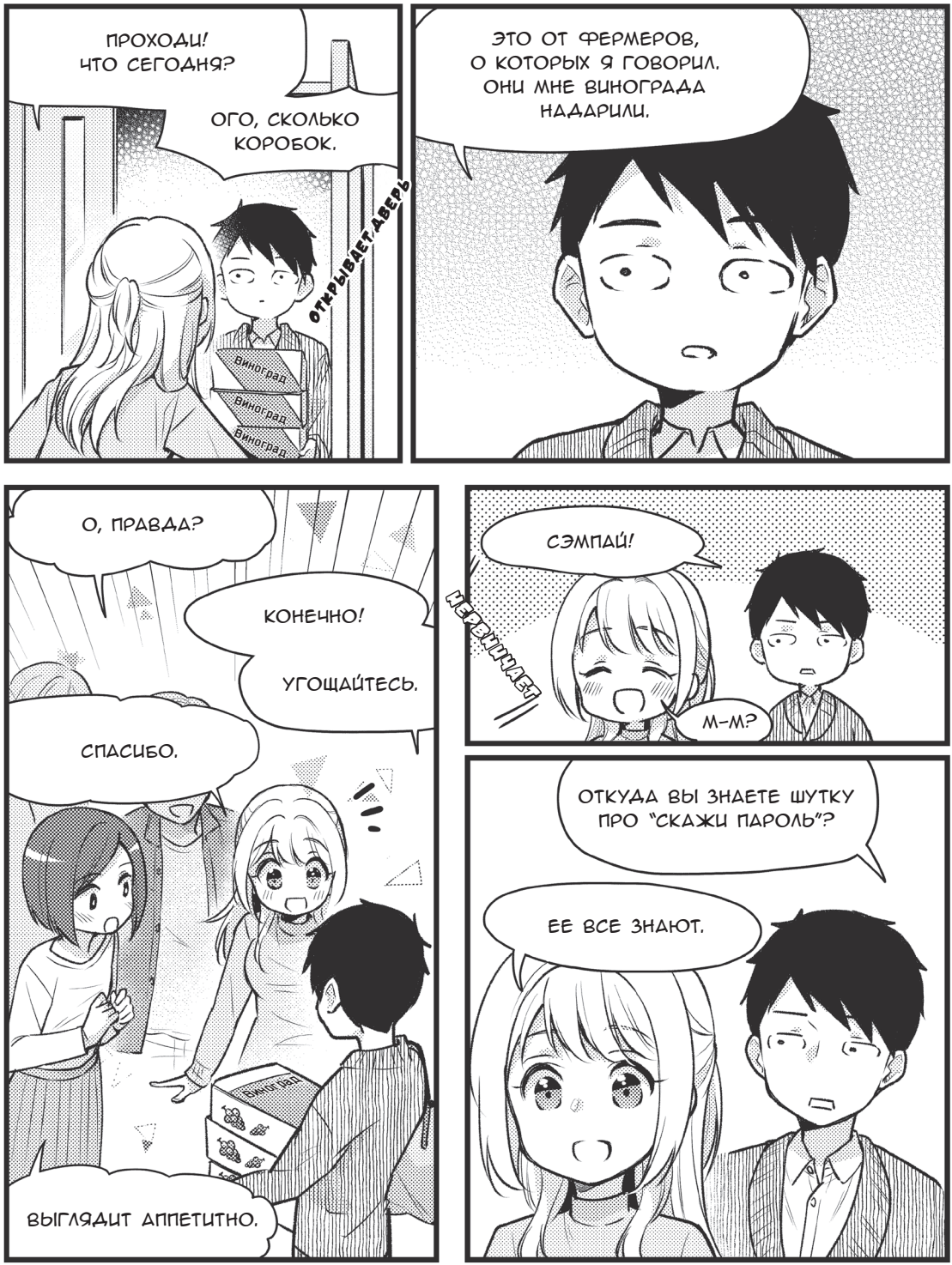
Да вроде… Но пока еще думаю…

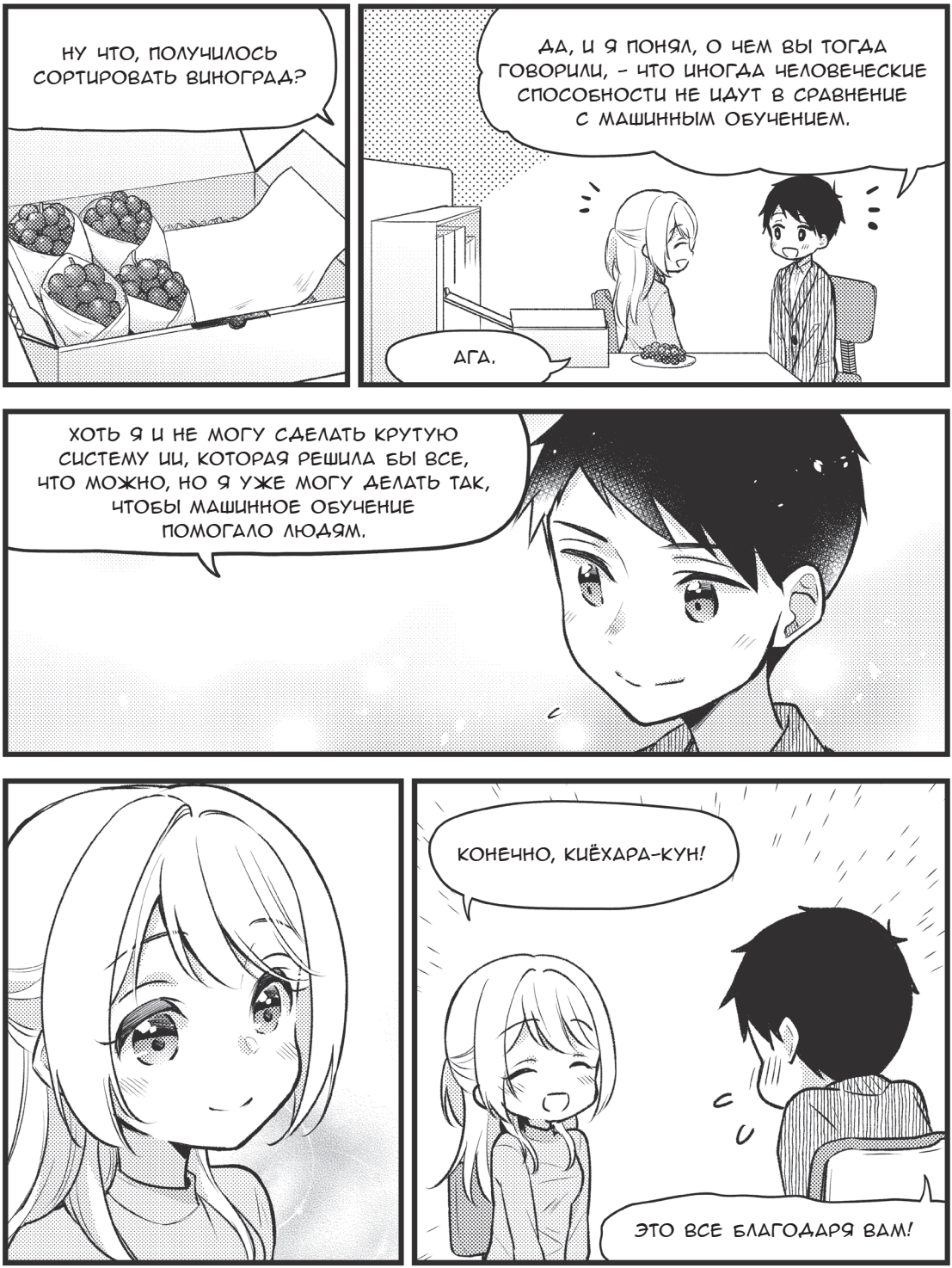
Да… у взрослых много разных дел.

ГЛАВА 5

АНСАМБЛЕВЫЕ МЕТОДЫ


ТУК
ТУК
ЗАХОДИТЕ!
ТУК
ТУК
ИЗВИНИТЕ...
ОТКРОЙТЕ МНЕ, ПОЖАЛУЙСТА.
ЭТО КИЁХАРА-КУН?
ДА.
СКАЖИ ПАРОЛЬ!
Э...
"ПАРОЛЬ".


ПРОХОДИ! ЧТО СЕГОДНЯ?
ОГО, СКОЛЬКО КОРОБОК.
ОТКРЫВАЕТ ДВЕРЬ
Виноград
Виноград
Виноград
ЭТО ОТ ФЕРМЕРОВ, О КОТОРЫХ Я ГОВОРИЛ. ОНИ МНЕ ВИНОГРАДА НАДАРИЛИ.
О, ПРАВДА?
КОНЕЧНО! УГОЩАЙТЕСЬ.
СПАСИБО.
ВЫГЛЯДИТ АППЕТИТНО.
СЭМПАЙ!
ПЕРЕПУГАЛСЯ
М-М?
ОТКУДА ВЫ ЗНАЕТЕ ШУТКУ ПРО "СКАЖИ ПАРОЛЬ"?
ЕЕ ВСЕ ЗНАЮТ.


НУ ЧТО, ПОЛУЧИЛОСЬ СОРТИРОВАТЬ ВИНОГРАД?
ДА, И Я ПОНЯЛ, О ЧЕМ ВЫ ТОГДА ГОВОРИЛИ, – ЧТО ИНОГДА ЧЕЛОВЕЧЕСКИЕ СПОСОБНОСТИ НЕ ИДУТ В СРАВНЕНИЕ С МАШИННЫМ ОБУЧЕНИЕМ.
АГА.
ХОТЬ Я И НЕ МОГУ СДЕЛАТЬ КРУТУЮ СИСТЕМУ ИИ, КОТОРАЯ РЕШИЛА БЫ ВСЕ, ЧТО МОЖНО, НО Я УЖЕ МОГУ ДЕЛАТЬ ТАК, ЧТОБЫ МАШИННОЕ ОБУЧЕНИЕ ПОМОГАЛО ЛЮДЯМ.
КОНЕЧНО, КИЁХАРА-КУН!
ЭТО ВСЕ БЛАГОДАРЯ ВАМ!


А разве ты не занят?
Ну... да... Я думаю, не открыть ли нам снова сайт, который определял вероятность заболеть диабетом, обучив его при помощи глубокого обучения?
М-м-м. Переоткрыть сайт...
А у тебя есть время?
Да.
Не хочешь немного поучиться?
Конечно!
Denisan
Вот опять семейный ресторанчик.
Я из него сбежал в прошлый раз.
Сегодня буду учиться до конца и не сбегу.
Извини, Киёхара! Заставила ждать.
Топ
Топ
А... вы во...
...время.


Э... ВЫ ПЕРЕОДЕЛИСЬ?
ДЛЯ МЕНЯ, ЧТО ЛИ? БЫТЬ НЕ МОЖЕТ...
М-м-милая! милая!
Я СХОДИЛА ПЕРЕОДЕТЬСЯ.
ЗАБЫЛА, ЧТО МЫ ПОТОМ ПООБЕДАЕМ ТУТ С СЕМЬЕЙ.
А, КОНЕЧНО... ДА...
НУ ЧТО, ДАВАЙ Я СЯДУ РЯДОМ С ТОБОЙ.
ОЙ!
Э...
ПОДСАЖИВАЕТСЯ
ЧЕРТ! ВСЯ УВЕРЕННОСТЬ ПРОПАЛА.
ЗАПИСЫВАЕТ
ТАК, МНЕ, ПОЖАЛУЙСТА, ТЫКВЕННЫЙ ПИРОГ, ТОРТ, КАШТАНОВОЕ МОРОЖЕНОЕ...
ТАК... НАЧНЕМ УРОК?
ПОПРАВЛЯЕТ ОЧКИ
ДА!

* Мондзю – бодхисатва, олицетворение высшей мудрости. – *Прим. перев.*

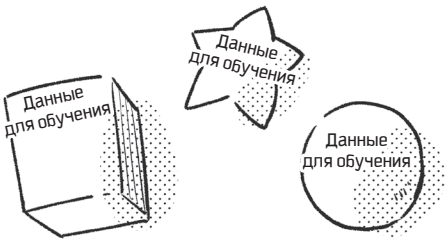

НЕТ!
НЕТ!
НЕТ!
ЧТО КАСАЕТСЯ ЭФФЕКТИВНОСТИ,
ТО НЕТ НИКАКОЙ РАЗНИЦЫ,
ОДИН ИЛИ НЕСКОЛЬКО ЧЕЛОВЕК ТАК СКАЗАЛИ.


ДА! НО СЕКРЕТ АНСАМБЛЕВЫХ МЕТОДОВ В ТОМ,
ЧТО У КАЖДОГО СВОЙ СОБСТВЕННЫЙ АЛГОРИТМ.
БАМ!
БАМ!

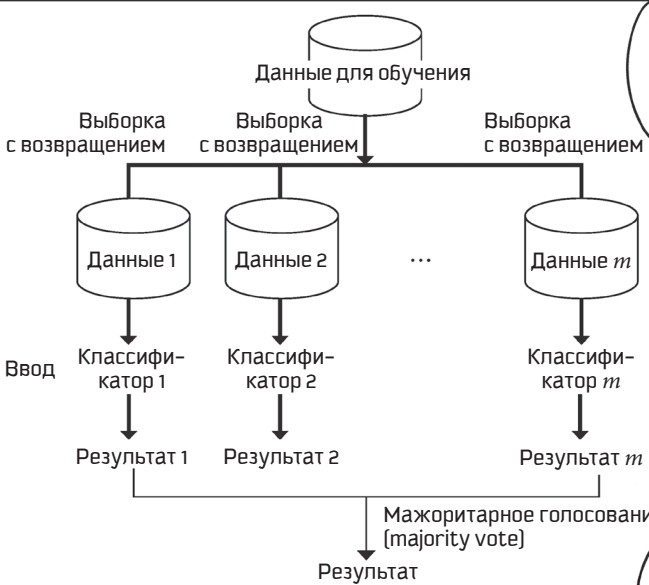
5.1. БЭГГИНГ


ЕСТЬ ТРИ СЕКРЕТА,
КОТОРЫЕ ИСПОЛЬЗУЮТСЯ
В АНСАМБЛЕВЫХ МЕТОДАХ,
И ПЕРВЫЙ ИЗ НИХ –
БЭГГИНГ!
ПЕРВЫЙ МЕТОД
УГУ

ПЕРВАЯ МЫСЛЬ: ЧТОБЫ СДЕЛАТЬ АЛГОРИТМЫ С ОТЛИЧАЮЩИМСЯ ПОВЕДЕНИЕМ, НУЖНО ИСПОЛЬЗОВАТЬ НЕСКОЛЬКО ОТЛИЧАЮЩИХСЯ НАБОРОВ ДАННЫХ ДЛЯ ОБУЧЕНИЯ.



НО, НАВЕРНОЕ, ЭТО ОЧЕНЬ ТРУДНО – ПОДГОТОВИТЬ РАЗНЫЕ ДАННЫЕ ДЛЯ ОБУЧЕНИЯ?



ПРИ БЭГГИНГЕ ИЗ ДАННЫХ ДЛЯ ОБУЧЕНИЯ ДЕЛАЕТСЯ **ВЫБОРКА С ВОЗВРАЩЕНИЕМ**, КОГДА ИЗ ИСХОДНЫХ ДАННЫХ ВЫБИРАЕТСЯ НЕКОТОРОЕ КОЛИЧЕСТВО НАБОРОВ ДАННЫХ ОДИНАКОВОГО РАЗМЕРА. ЗАТЕМ ДЛЯ КАЖДОГО НАБОРА ДАННЫХ СОЗДАЕТСЯ КЛАССИФИКАТОР ПРИ ПОМОЩИ ОДНОГО И ТОГО ЖЕ АЛГОРИТМА.

ВЫБОРКА С ВОЗВРАЩЕНИЕМ?

ПРИ ЭТОМ МЕТОДЕ ДАННЫЕ, ПОПАВШИЕ В ВЫБОРКУ, ЗАПИСЫВАЮТСЯ И ВОЗВРАЩАЮТСЯ. КАКИЕ-ТО ДАННЫЕ МОГУТ ПОПАСТЬ В ВЫБОРКУ МНОГО РАЗ, А КАКИЕ-ТО – НИ РАЗУ.

ДАВАЙ ПОСЧИТАЕМ, НАСКОЛЬКО ПРИ ИСПОЛЬЗОВАНИИ ЭТОГО МЕТОДА БУДУТ ОТЛИЧАТЬСЯ ДАННЫЕ В РАЗНЫХ ВЫБОРКАХ!

ПЕРЕВОРАЧИВАЕТ СТРАНИЦУ

ДА!

Допустим, в наборе данных есть N отдельных элементов. Какова вероятность, что один элемент не попадет в выборку?

Если элементов N, то вероятность того, что он попадет в выборку, – $\frac{1}{N}$.

А вероятность, что не попадет, – $(1 - \frac{1}{N})$.

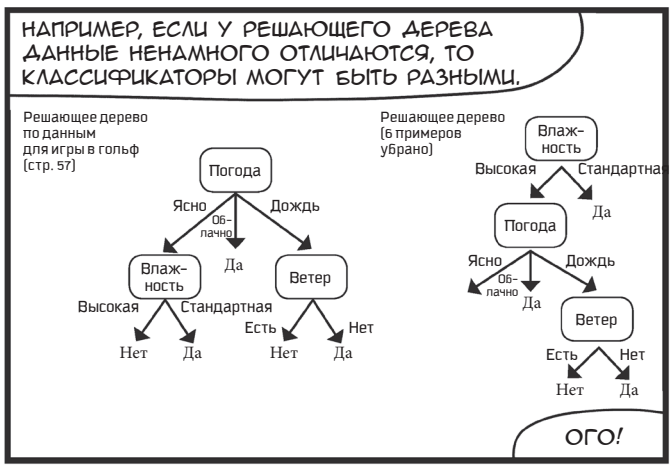
Да. А вероятность того, что данные не попадут в выборку N раз, равна $\left(1 - \frac{1}{N}\right)^N$. Таким образом, рассчитаем вероятность того, что данные не попадут в выборку...

Если $N = 10$, то она равна 0,349.
Если $N = 100$, то 0,366.
Если $N \to \infty$, то вероятность равна $1/e = 0{,}368$.

Число N не слишком влияет на результат.

Именно! При таких расчетах ясно, каким бы N ни было, при выборке с возвращением примерно ⅓ исходных данных не попадут в нее.

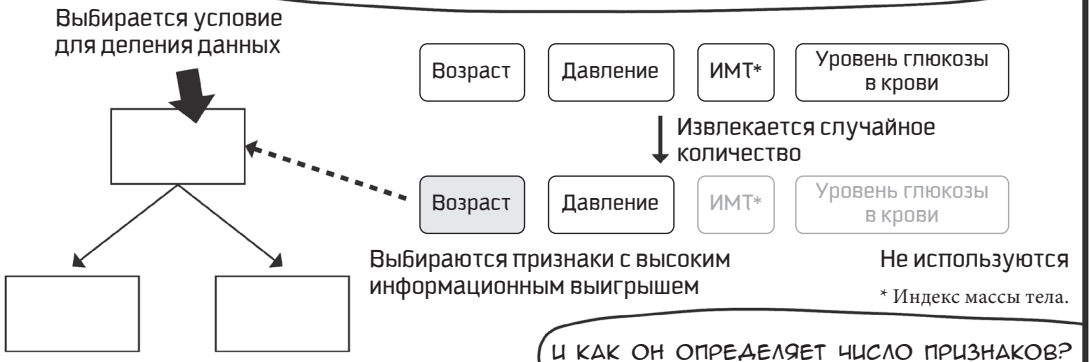
Целая треть данных!



5.2. Случайный лес

СТРОИТСЯ ДЕРЕВО РЕШЕНИЙ ПО КЛАССИФИКАТОРАМ ДЛЯ КАЖДОГО НАБОРА ДАННЫХ.

ВЫБИРАЕТСЯ ЗАРАНЕЕ ОПРЕДЕЛЕННОЕ КОЛИЧЕСТВО ПРИЗНАКОВ ИЗ ВСЕХ, А УЖЕ В НИХ ИЩЕТСЯ ПРИЗНАК С МАКСИМАЛЬНО ЭФФЕКТИВНЫМ РАЗДЕЛЕНИЕМ.



* Индекс массы тела.

И КАК ОН ОПРЕДЕЛЯЕТ ЧИСЛО ПРИЗНАКОВ?

КОЛИЧЕСТВО ПРИЗНАКОВ ДЛЯ ВЫБОРА ИЗ ОБЩЕГО ЧИСЛА ПРИЗНАКОВ ($\sqrt{d}$) – ЭТО ЧАСТО ЛИБО КВАДРАТНЫЙ КОРЕНЬ ИЗ d, ЛИБО $\log_2 d$.

Общее количество признаков d	floor ($\sqrt{d}$)	floor $\log_2 d$
5	2	2
10	3	3
50	7	5
100	10	6

floor(x) – это наибольшее целое число, но не больше x.

ОПЕРАЦИЯ ПРОДОЛЖАЕТСЯ РЕКУРСИВНО, ПОКА В КАЖДОМ ЛИСТЕ НЕ ОСТАНУТСЯ ПРЕДСТАВИТЕЛИ ОДНОГО КЛАССА.

ИТАК, КИЁХАРА-КУН, ВОПРОС! НА ЧТО НУЖНО ОБРАЩАТЬ ВНИМАНИЕ, КОГДА МЫ СТРОИМ РЕШАЮЩЕЕ ДЕРЕВО?

ЩЕЛК

НУ...

ПЕ-ПЕРЕОБУЧЕНИЕ!

ЩЕЛК

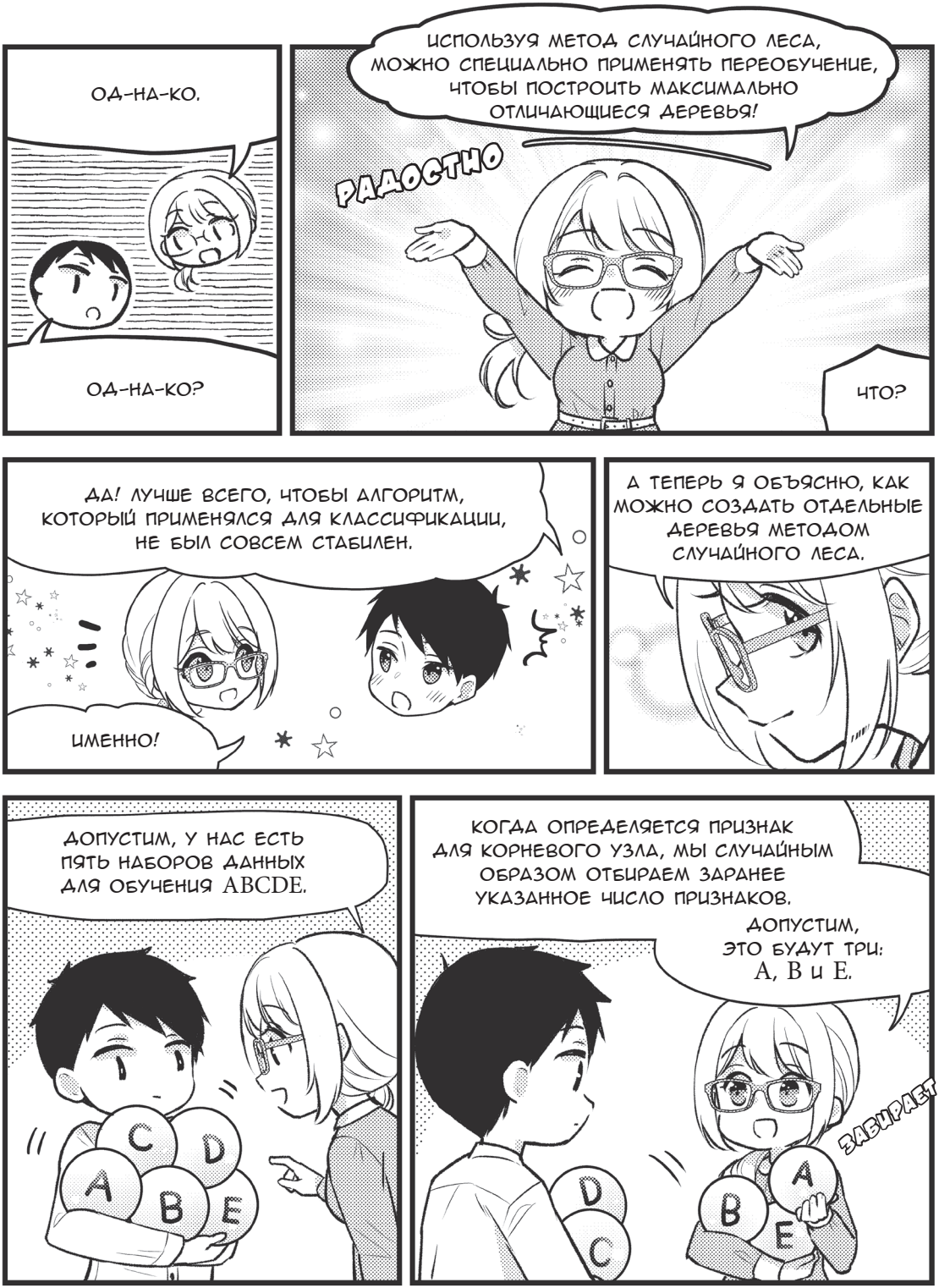
ИМЕННО!

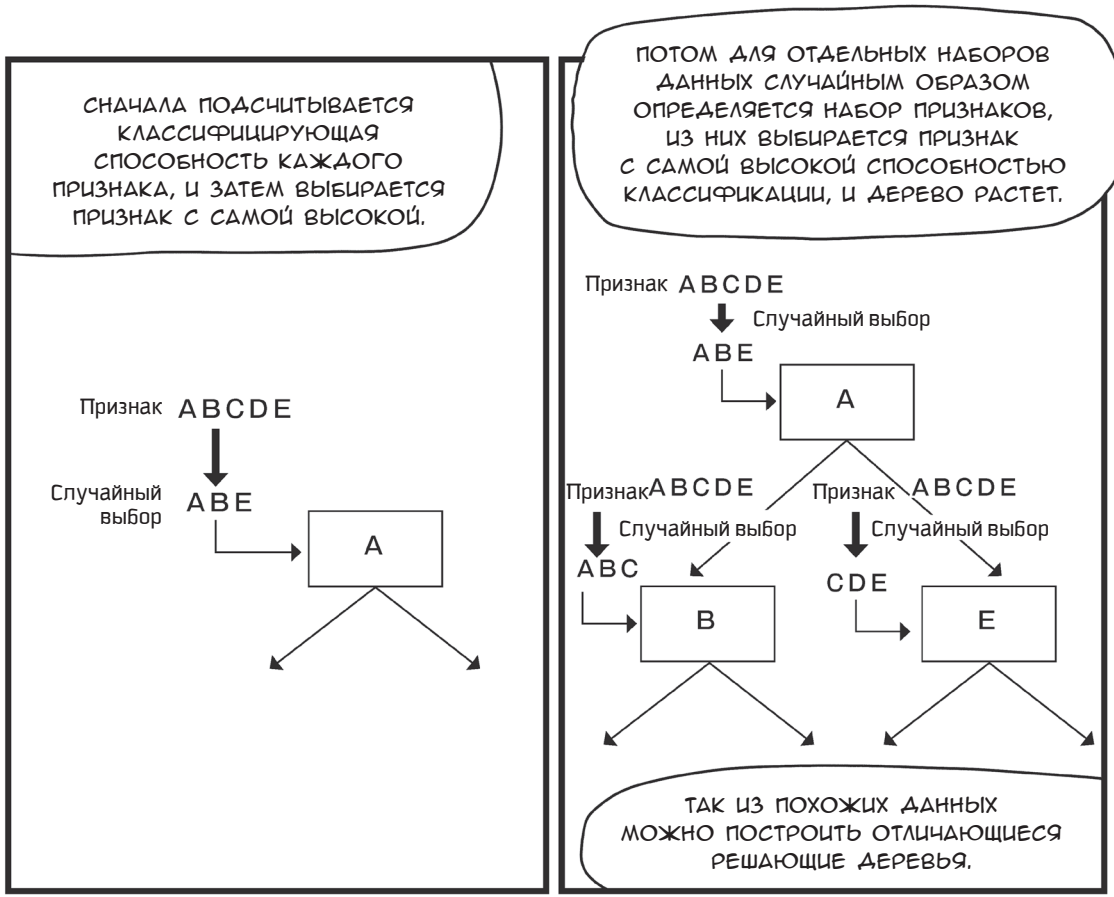
РАНЬШЕ МЫ ГОВОРИЛИ, ЧТО В КАЧЕСТВЕ МЕРЫ ПРОТИВ ПЕРЕОБУЧЕНИЯ МОЖНО ЛИБО ОСТАНОВИТЬ РОСТ, КОГДА КОЛИЧЕСТВО ДАННЫХ В ЛИСТЬЯХ УПАДЕТ НИЖЕ ОПРЕДЕЛЕННОГО УРОВНЯ, ЛИБО ЖЕ ОБРЕЗАТЬ ВЕТВИ, КОГДА ДЕРЕВО РАЗРОСЛОСЬ.

ЧИК

ЧИК

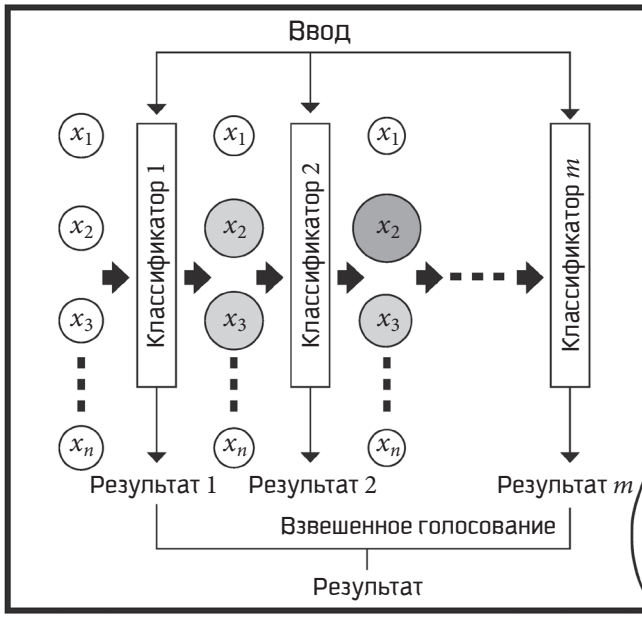
НЕРВНИЧАЕТ


ОД-НА-КО.
ОД-НА-КО?
ИСПОЛЬЗУЯ МЕТОД СЛУЧАЙНОГО ЛЕСА, МОЖНО СПЕЦИАЛЬНО ПРИМЕНЯТЬ ПЕРЕОБУЧЕНИЕ, ЧТОБЫ ПОСТРОИТЬ МАКСИМАЛЬНО ОТЛИЧАЮЩИЕСЯ ДЕРЕВЬЯ!
РАДОСТНО
ЧТО?
ДА! ЛУЧШЕ ВСЕГО, ЧТОБЫ АЛГОРИТМ, КОТОРЫЙ ПРИМЕНЯЛСЯ ДЛЯ КЛАССИФИКАЦИИ, НЕ БЫЛ СОВСЕМ СТАБИЛЕН.
ИМЕННО!
А ТЕПЕРЬ Я ОБЪЯСНЮ, КАК МОЖНО СОЗДАТЬ ОТДЕЛЬНЫЕ ДЕРЕВЬЯ МЕТОДОМ СЛУЧАЙНОГО ЛЕСА.
ДОПУСТИМ, У НАС ЕСТЬ ПЯТЬ НАБОРОВ ДАННЫХ ДЛЯ ОБУЧЕНИЯ ABCDE.
A
B
C
D
E
КОГДА ОПРЕДЕЛЯЕТСЯ ПРИЗНАК ДЛЯ КОРНЕВОГО УЗЛА, МЫ СЛУЧАЙНЫМ ОБРАЗОМ ОТБИРАЕМ ЗАРАНЕЕ УКАЗАННОЕ ЧИСЛО ПРИЗНАКОВ.
ДОПУСТИМ, ЭТО БУДУТ ТРИ: A, B И E.
ЗАБИРАЕТ
A
B
E
D
C


СНАЧАЛА ПОДСЧИТЫВАЕТСЯ КЛАССИФИЦИРУЮЩАЯ СПОСОБНОСТЬ КАЖДОГО ПРИЗНАКА, И ЗАТЕМ ВЫБИРАЕТСЯ ПРИЗНАК С САМОЙ ВЫСОКОЙ.
Признак ABCDE
Случайный выбор
ABE
A
ПОТОМ ДЛЯ ОТДЕЛЬНЫХ НАБОРОВ ДАННЫХ СЛУЧАЙНЫМ ОБРАЗОМ ОПРЕДЕЛЯЕТСЯ НАБОР ПРИЗНАКОВ, ИЗ НИХ ВЫБИРАЕТСЯ ПРИЗНАК С САМОЙ ВЫСОКОЙ СПОСОБНОСТЬЮ КЛАССИФИКАЦИИ, И ДЕРЕВО РАСТЕТ.
Признак ABCDE
Случайный выбор
ABE
A
Признак ABCDE
Случайный выбор
ABC
B
Признак ABCDE
Случайный выбор
CDE
E
ТАК ИЗ ПОХОЖИХ ДАННЫХ МОЖНО ПОСТРОИТЬ ОТЛИЧАЮЩИЕСЯ РЕШАЮЩИЕ ДЕРЕВЬЯ.

5.3. БУСТИНГ


В БЭГГИНГЕ И МЕТОДЕ СЛУЧАЙНОГО ЛЕСА ИЗМЕНЯЮТСЯ ИСПОЛЬЗУЕМЫЕ НАБОРЫ ДАННЫХ, ИЗМЕНЯЮТСЯ УСЛОВИЯ ВЫБОРА КЛАССИФИКАТОРОВ, И ПОЭТОМУ МОЖНО СТРОИТЬ ОТЛИЧАЮЩИЕСЯ КЛАССИФИКАТОРЫ.
КРОМЕ ТОГО, В МЕТОДЕ БУСТИНГА НАБОР КЛАССИФИКАТОРОВ, КОТОРЫЕ ВЕДУТ СЕБЯ ПО-РАЗНОМУ, СОЗДАЕТСЯ ПУТЕМ ПОСЛЕДОВАТЕЛЬНОГО ДОБАВЛЕНИЯ КЛАССИФИКАТОРОВ, КОТОРЫЕ СПЕЦИАЛИЗИРУЮТСЯ НА УМЕНЬШЕНИИ ОШИБОК.
ТРЕТИЙ МЕТОД
КЛАССИФИКАТОР, КОТОРЫЙ СПЕЦИАЛИЗИРУЕТСЯ НА УМЕНЬШЕНИИ КОЛИЧЕСТВА ОШИБОК?
ДА.
ДЛЯ ЭТОГО НУЖНО ОПРЕДЕЛИТЬ ВЕС КАЖДОГО ЭЛЕМЕНТА ДАННЫХ.



ПРЕЖДЕ ВСЕГО СОЗДАЕТСЯ КЛАССИФИКАТОР ДАННЫХ ДЛЯ ОБУЧЕНИЯ С ОДИНАКОВЫМИ ВЕСАМИ. УВЕЛИЧИВАЕМ ВЕС ДАННЫХ, КОТОРЫЕ ОШИБОЧНО КЛАССИФИЦИРОВАНЫ С ПОМОЩЬЮ ЭТОГО КЛАССИФИКАТОРА.

А КАК НАСТРАИВАЕТСЯ ВЕС ДАННЫХ?

ТАК, ЧТОБЫ СУММА ВЕСОВ ПРАВИЛЬНО КЛАССИФИЦИРОВАННЫХ ДАННЫХ И СУММА ВЕСОВ НЕПРАВИЛЬНО КЛАССИФИЦИРОВАННЫХ ДАННЫХ СТАНОВИЛИСЬ ОДИНАКОВЫМИ.

ЗАТЕМ ДЛЯ НАБОРА ДАННЫХ, В КОТОРОМ БЫЛИ ИЗМЕНЕНЫ ВЕСА, ПРОВОДИТСЯ ОБУЧЕНИЕ ПРИ ПОМОЩИ СЛЕДУЮЩЕГО КЛАССИФИКАТОРА.

Остались ошибки...

Оставь их мне!

И ТАК ПОЭТАПНО СОЗДАЮТСЯ НОВЫЕ КЛАССИФИКАТОРЫ. НОВЫЙ КЛАССИФИКАТОР, В ОТЛИЧИЕ ОТ РАНЕЕ СОЗДАННОГО, КЛАССИФИЦИРУЮЩЕГО ОШИБОЧНЫЕ ДАННЫЕ, ВОСПОЛНЯЕТ ЕГО СЛАБЫЕ СТОРОНЫ.

ЭТО МЕТОД **ADABOOST**.

УЧЕБНЫЙ АЛГОРИТМ КЛАССИФИКАТОРА, КОТОРЫЙ ИСПОЛЬЗУЕТСЯ В БУСТИНГЕ, ДОЛЖЕН В ОСНОВНОМ ПРИМЕНЯТЬ ВЕСА ДАННЫХ В КАЧЕСТВЕ КРИТЕРИЯ ДЛЯ СОЗДАНИЯ НОВОГО КЛАССИФИКАТОРА.

ТО ЕСТЬ ЭТОТ МЕТОД НЕЛЬЗЯ ИСПОЛЬЗОВАТЬ, ЕСЛИ НЕ ДУМАТЬ О ВЕСАХ С САМОГО НАЧАЛА.

НЕТ, МОЖНО СПРАВИТЬСЯ, СОЗДАВАЯ НАБОРЫ ДАННЫХ ПУТЕМ ВЫБОРКИ С ВОЗВРАЩЕНИЕМ, ЗАНОВО РАСПРЕДЕЛЯЯ ВЕСА.

УДОБНО! А РЕШЕНИЕ ПРИНИМАЕТСЯ БОЛЬШИНСТВОМ ГОЛОСОВ, КАК В БЭГГИНГЕ?

В СЛУЧАЕ МЕТОДА ADABOOST КЛАССИФИКАТОР СОЗДАЕТСЯ НА ОСНОВАНИИ ОШИБОК НА ПРЕДЫДУЩЕМ ЭТАПЕ, ТАК?

НАДЕЖНОСТЬ

ИНЫМИ СЛОВАМИ, ЕСЛИ В ИСХОДНЫХ ДАННЫХ ДЛЯ ОБУЧЕНИЯ МНОГО ОШИБОК, ТО НАДЕЖНОСТЬ КЛАССИФИКАТОРА, СОЗДАННОГО НА ОСНОВЕ ЭТИХ ДАННЫХ, ДЛЯ НЕИЗВЕСТНЫХ ДАННЫХ НА ВХОДЕ БУДЕТ УВЕЛИЧИВАТЬСЯ.

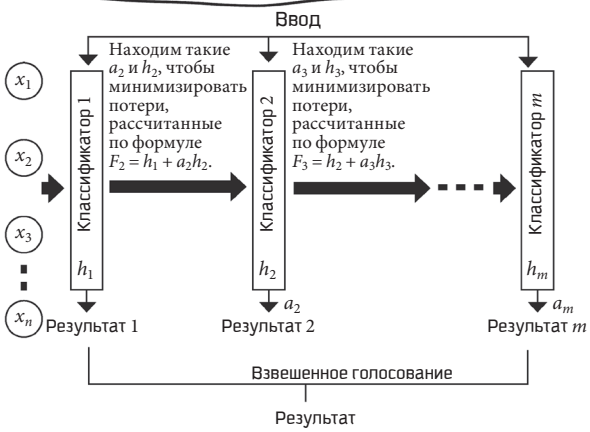
КАК ОПРЕДЕЛЯЕТСЯ РЕЗУЛЬТАТ КЛАССИФИКАЦИИ?

ВЫЧИСЛЯЕТСЯ ВЕС НА ОСНОВАНИИ ВЕЛИЧИНЫ ФУНКЦИИ ОШИБОК ДЛЯ КАЖДОГО КЛАССИФИКАТОРА, ПРОВОДИТСЯ ГОЛОСОВАНИЕ ВЕСОВ, И ПОЛУЧАЕТСЯ РЕЗУЛЬТАТ.

ПОСКОЛЬКУ В ADABOOST К КЛАССИФИКАТОРАМ НА ПРЕДЫДУЩИХ ЭТАПАХ ОДИН ЗА ДРУГИМ ДОБАВЛЯЮТСЯ КЛАССИФИКАТОРЫ, КОТОРЫЕ ПРАВИЛЬНО КЛАССИФИЦИРОВАЛИ ОШИБОЧНЫЕ ДАННЫЕ, ЕГО ЭФФЕКТИВНОСТЬ ПОВЫШАЕТСЯ.

ЕСТЬ СПОСОБ, ПРИ КОТОРОМ В КАЧЕСТВЕ ЕЩЕ ОДНОГО УРАВНЕНИЯ МОЖНО ИСПОЛЬЗОВАТЬ ФУНКЦИЮ ПОТЕРЬ. СОСТАВНОЙ КЛАССИФИКАТОР, КОТОРЫЙ ЯВЛЯЕТСЯ РЕЗУЛЬТАТОМ БУСТИНГА, МОЖЕТ ОПРЕДЕЛИТЬ ФУНКЦИЮ ПОТЕРЬ.

МОЖНО ПРЕДСТАВИТЬ УРАВНЕНИЕ, ПРИ КОТОРОМ ПОСЛЕДУЮЩИЕ КЛАССИФИКАТОРЫ БУДУТ ВЫБИРАТЬСЯ ТАК, ЧТОБЫ МАКСИМАЛЬНО УМЕНЬШИТЬ ФУНКЦИЮ ПОТЕРИ.



БУСТИНГ, ОСНОВАННЫЙ НА ЭТОЙ ИДЕЕ, НАЗЫВАЕТСЯ **ГРАДИЕНТНЫМ БУСТИНГОМ**.

А ТЕПЕРЬ НЕМНОГО ПОКОДИРУЕМ!

Возьмем данные diabetes.arff из инструментов для машинного обучения Weka. Они близки к данным, необходимым для сайта с тестом определения вероятности диабета.

Хотя в scikit-learn есть такой же набор данных, но он нужен для задач регрессии, и объяснить результат будет трудно, поэтому для задач классификации мы используем набор diabetes.arff.

diabetes.arff содержит в себе результаты осмотров и диагнозы. Признаки – возраст, давление, ИМТ и т. п.

Для начала надо скачать diabetes.arff. Можно скачать несколько наборов данных с https://www.cs.waikato.ac.nz/ml/weka/datasets.html и найти там нужный файл.

```
import numpy as np
from scipy.io import arff
from sklearn.ensemble import BaggingClassifi er, RandomForestClassifi er,
AdaBoostClassifi er, GradientBoostingClassifi er
from sklearn.model_selection import cross_val_score
```

Файл формата arff будет прочитан как модуль arff scipy. Поскольку признаковые описания и метки правильных ответов находятся в одной строке, каждый можно сохранить в отдельный массив numpy.

```
data, meta = arff.loadarff('diabetes.arff')
X = np.empty((0,8), np.fl oat)
y = np.empty((0,1), np.str)
```

```
for e in data:
    e2 = list(e)
    X = np.append(X, [e2[0:8]], axis=0)
    y = np.append(y, e2[8:9])
```

В scikit-learn ансамблевые методы используютсяе так же, как и обучение и оценка данных с классификатором. Используя перекрестную проверку по 10 % данных, определим среднюю точность и дисперсию.

```
clf1 = BaggingClassifi er()
scores = cross_val_score(clf1, X, y, cv=10)
print("{0:4.2f} +/- {1:4.2f} %".format(scores.mean() * 100, scores.std() * 100))
73.69 +/- 5.11 %

clf2 = RandomForestClassifi er()
scores = cross_val_score(clf2, X, y, cv=10)
print("{0:4.2f} +/- {1:4.2f} %".format(scores.mean() * 100, scores.std() * 100))
74.72 +/- 5.72 %

clf3 = AdaBoostClassifi er()
scores = cross_val_score(clf3, X, y, cv=10)
print("{0:4.2f} +/- {1:4.2f} %".format(scores.mean() * 100, scores.std() * 100))
75.52 +/- 5.71 %

clf4 = GradientBoostingClassifi er()
scores = cross_val_score(clf4, X, y, cv=10)
print("{0:4.2f} +/- {1:4.2f} %".format(scores.mean() * 100, scores.std() * 100))
76.30 +/- 5.11 %
```

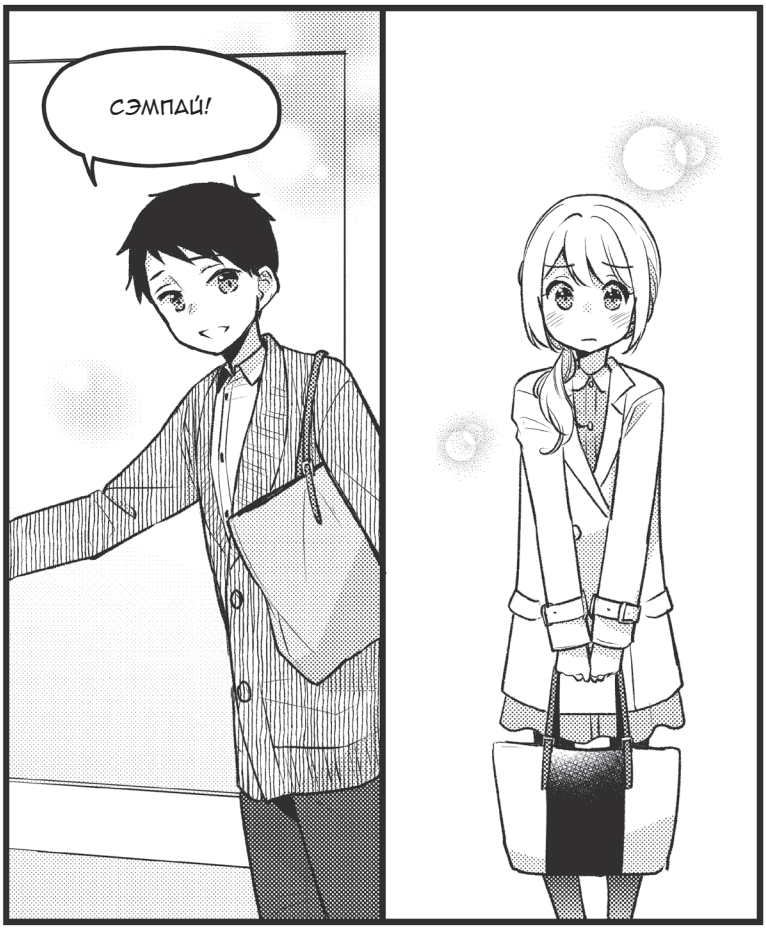
Используя параметры по умолчанию, можно получить хорошие результаты при помощи градиентного бустинга.


НУ, НА СЕГОДНЯ ВСЕ ПРО АНСАМБЛЕВЫЕ МЕТОДЫ.
СНИМАЕТ ОЧКИ
УДАЧИ С ОТКРЫТИЕМ САЙТА!
СПАСИБО!
ЛУЧШЕ БЫ Я В ТОТ РАЗ ВСЕ ДОСЛУШАЛ...
НАВЕРНОЕ, НАДО БЕЖАТЬ, ПОТОМУ ЧТО БОЮСЬ СНОВА УСЛЫШАТЬ, ЧТО Я ЕЙ КАК "БРАТИК".
Я ОЧЕНЬ БЛАГОДАРЕН ВАМ, СЭМПАЙ!
НЕТ-НЕТ, ЭТО Я САМА ВЫЗВАЛАСЬ ТЕБЯ УЧИТЬ!
КСТАТИ, МНЕ НЕДАВНО ПРИШЛО ПИСЬМО...
ПИСЬМО?..
МЕНЯ ВЗЯЛИ НА РАБОТУ!
ЧТО?
НУ ЭТО ЖЕ ПРЕКРАСНО! ПОЗДРАВЛЯЮ!
УСТАЛО
ДА... НО Я ДУМАЛА, ЧТО НЕ ВОЗЬМУТ.
И ПОСЛЕ ВЫПУСКА Я ПЕРЕЕДУ В ТОКИО.
!


КОНЕЧНО, БУДЕТ ТРУДНО ВСТРЕЧАТЬСЯ, И Я НЕ СМОГУ ТЕБЯ БОЛЬШЕ УЧИТЬ.
Я ЖАЛЕЮ ОБ ЭТОМ, ПОТОМУ ЧТО ТЫ ХОРОШИЙ УЧЕНИК.
А ТЕПЕРЬ Я ЗАПЛАЧУ́.
НЕТ УЖ!
ЗАБИРАЕТ СЧЕТ
ЭТО Я ДОЛЖЕН, ПЛАТА ЗА ОБУЧЕНИЕ!
НО ТЫ ЖЕ УЖЕ ЗАПЛАТИЛ ДО ЭТОГО, НЕ ХОЧУ, ЧТОБЫ ТЫ ПОСТОЯННО МЕНЯ УГОЩАЛ!
НЕТ!
ЗАБИРАЕТ СЧЕТ
ТО МЛАДШИЙ ТОВАРИЩ,
ТО БРАТИК,
ТО УЧЕНИК.
ПОЗВОЛЬ МНЕ ЗАПЛАТИТЬ, КАК МУЖЧИНЕ!


ХОТЬ Я И УГОЩАЮ, НО НЕ ХОЧУ ВЫГЛЯДЕТЬ ЛУЧШЕ, ЧЕМ Я ЕСТЬ.
Спасибо!
СТЫ-Ы-Ы-Ы-ДНО!


ОТКРЫВАЕТ ДВЕРЬ


СЭМПАЙ!


СПАСИБО...
БОЛЬШОЕ.
КАЖЕТСЯ, БУДТО БЫ ТЕБЕ НЕ ПОНРАВИЛОСЬ.


ВРОДЕ ЕМУ ГОВОРИЛА, ЧТО ОН КАК БРАТИК?..
ОТПРАЗДНУЕМ МОЮ НОВУЮ РАБОТУ КАК-НИБУДЬ?
НУ ДА, НУ ДА.

В кабинете у Саяка (6)

Математическое повторение (5)

Сегодня вы много говорили о классификаторах и их эффективности, но мне не совсем понятно, что это такое.

Что было непонятно?

Ну, я поняла, что даже если много людей дают одинаковый ответ, они не умнее одного, который ответит так же. Но если мы соберем много людей, умных и не очень, и они будут давать разные ответы, они не поссорятся?

Конечно, могут поссориться. Но чтобы они не поссорились, мы используем голосование по методу большинства, и оно может численно подтвердить наилучший ответ.

Прежде всего пусть для одинаковых данных для обучения мы будем использовать L разных классификаторов.

То есть у нас есть L умных людей.

Ну, можно взять и неумных людей. Хоть это немногим лучше, чем случайные ответы, но математическая формулировка не изменится.

Правда?

Предположим, что процент ошибок классификатора e один и тот же, и ошибки независимы.

Под независимостью ошибок имеется в виду то, что вероятность ошибки классификатора для каждого элемента данных независима. Иными словами, можно предположить, что нет данных, для которых вместе ошибется множество классификаторов.

Основываясь на этой гипотезе, рассмотрим вероятность того, что из L классификаторов m классификаторов ошиблись. Пусть $m = 1$.

То есть если вероятность того, что ошибся один классификатор, равна e, то вероятность того, что остальные $(L - 1)$ классификаторы не ошиблись, равна $(1 - e)^{(L-1)}$.

Так как ошибиться может любой из L классификаторов, то при перемножении получим $Le(1 - e)^{(L-1)}$.

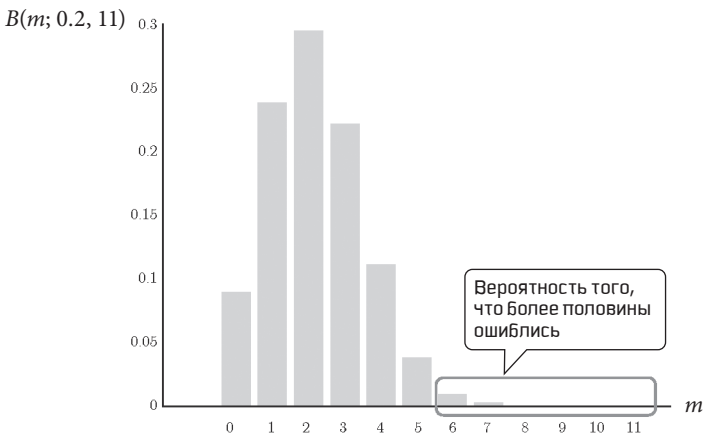
А если ошиблись два?

Тогда число способов выбрать 2 из L будет LC_2, и, значит, у нас будет $LC_2e_2(1 - e)^{(L-2)}$.

Так. Вероятность, что m из L классификаторов, вероятность ошибки каждого из которых e, находится при помощи биномиального распределения $B(m; e, L)$.

$B(m; e, L) =_L C_m e^m (1 - e)^{L-m}$.

Предположим, что количество классификаторов $L = 11$, коэффициент ошибки каждого $e = 0.2$, тогда биноминальное распределение $B(m; 0.2, 11)$ будет представлено на графике ниже.



Если результат классификации определяется большинством голосов, то вероятность полностью неправильного решения будет равняться сумме вероятностей ошибок шести классификаторов и больше. При подсчете получится 1,2 %.

Если вероятность ошибки каждого классификатора по отдельности 20 %, но мы используем все 11 классификаторов, то вероятность ошибки составит всего 1,2 %.

Ого! Ансамблевые методы – это здорово!

Однако и у этого метода есть недостатки. Данные, которые трудно классифицировать даже людям, могут вызвать ошибки у многих классификаторов, и поэтому на самом деле эффективность не так высока.

Но чтобы по возможности разрешить эту проблему, можно делать классификаторы с разным поведением, для чего и нужны ансамблевые методы. Понятно в целом?

В целом да. Жаль, что ты, Сая, уезжаешь в Токио. Там будет тебе одиноко...

Конечно, уезжать из дома всегда немного грустно…

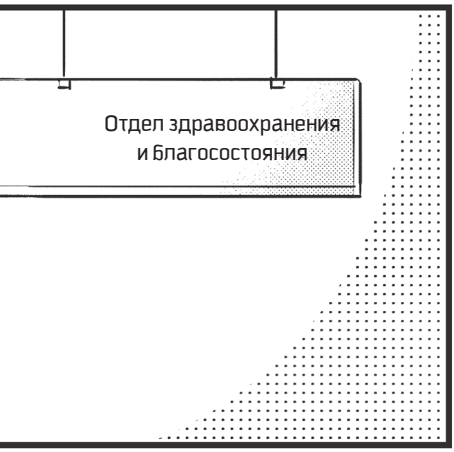
А что ты сказала Киёхара-куну?

Ну… ну что он мне как братик, но это было еще тогда...

И почему ты это вспомнила?

ГЛАВА 6

ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ


Отдел здравоохранения
и благосостояния


ТАК...


КИЁХАРА-КУН, МОЖНО НА МИНУТОЧКУ?
ДА, ДИРЕКТОР.


ЧТО СЛУЧИЛОСЬ?
ВСТАЁТ
ИЗВИНИТЕ, ЧТО ОТВЛЕКАЮ, МОЖНО С ВАМИ НЕМНОГО ПОГОВОРИТЬ?


ХОЧУ ВАС КОЕ О ЧЕМ ПОПРОСИТЬ, КИЁХАРА-КУН. ВЫ ЖЕ ХОРОШО РАЗБИРАЕТЕСЬ В МАШИННОМ ОБУЧЕНИИ?
НУ-У-У-У-У-У-У...


У САЙТА, КОТОРЫЙ ОПРЕДЕЛЯЕТ ВЕРОЯТНОСТЬ ЗАБОЛЕВАНИЯ ДИАБЕТОМ, ХОРОШИЕ ОТЗЫВЫ.
ТАК СКАЗАТЬ
С-СПАСИБО!

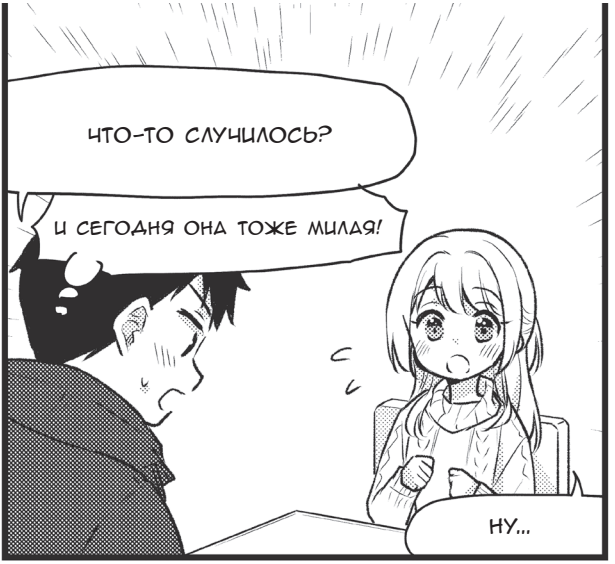

НЕ СТОИТ. В ЭТОМ ГОДУ, КАК ВАМ ИЗВЕСТНО, Я ВЫХОЖУ НА ПЕНСИЮ.
И ЕСТЬ ОДНО ДЕЛО, СЛИШКОМ СЛОЖНОЕ ДЛЯ МОЕГО ПРЕЕМНИКА.
ЧТО ЗА ДЕЛО?

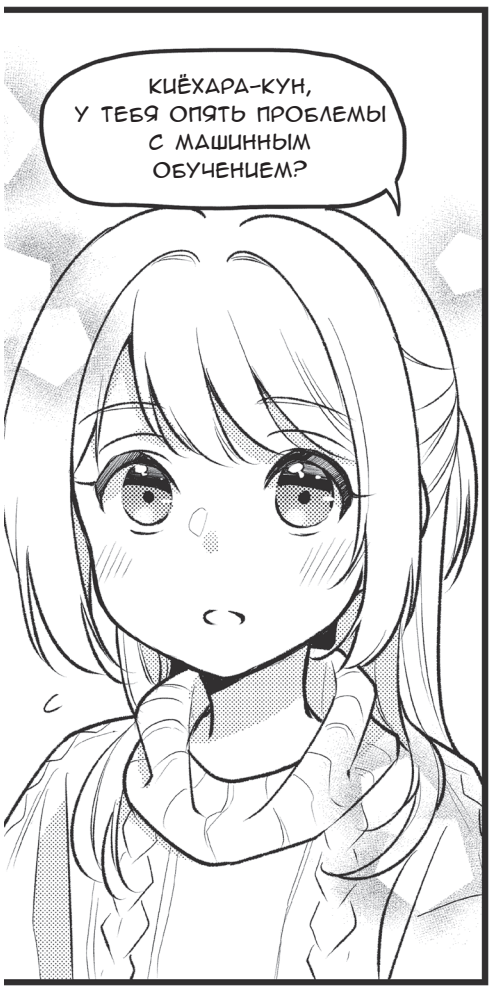

КАЖДЫЙ МЕСЯЦ НАШ ДЕПАРТАМЕНТ УСТРАИВАЕТ РАССЫЛКУ ДЛЯ ОДИНОКИХ ПОЖИЛЫХ ЛЮДЕЙ.
В ЭТОЙ РАССЫЛКЕ Я СОВЕТОВАЛ ИМ ОТДЕЛЬНЫЕ МЕРОПРИЯТИЯ, ОСНОВЫВАЯСЬ НА ИХ ПРЕДПОЧТЕНИЯХ.
ЛЮБИТ ЦВЕТЫ
Весенняя распродажа растений такого-то числа в «Зеленом саду»
ХОТЕЛОСЬ БЫ, ЧТОБЫ РАССЫЛКА ПРОДОЛЖАЛАСЬ И ПОСЛЕ ТОГО, КАК Я УЙДУ НА ПЕНСИЮ.
ДО ЭТОГО Я ОТПРАВЛЯЛ ВСЕ САМОСТОЯТЕЛЬНО, НО ХОТЕЛОСЬ БЫ КАК-ТО СИСТЕМАТИЗИРОВАТЬ ЭТО ДЕЛО...
ВОТ ОНО ЧТО.
ВЫ ЖЕ НЕ СМОЖЕТЕ ВСЕ ЭТО ПЕРЕДАТЬ ПРОСТО ТАК СВОЕМУ ПРЕЕМНИКУ?
ДА, И Я ПОДУМАЛ, МОЖЕТ ВЫ, КИЁХАРА-КУН, СМОЖЕТЕ ЧТО-НИБУДЬ СДЕЛАТЬ.
НУ КАК? ИЛИ СЛОЖНО?
НУ...
ДАННЫХ ДЛЯ ОБУЧЕНИЯ НЕТ, ПОЭТОМУ ЭТО ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ.
Я ПОПРОБУЮ, НО МНЕ НУЖНО ИЗУЧИТЬ ВОПРОС.
КОНЕЧНО! СПАСИБО, КИЁХАРА-КУН!


ТАК...
ХРУМ
ХРУМ
ЧТО-ТО ТЫ НЕДАВНО СТАЛ УЧИТЬСЯ ВО ВРЕМЯ ОБЕДА...
УГУ.
ДИРЕКТОР СКАЗАЛ, ЧТО ОТЗЫВЫ НА САЙТ С ПРОГНОЗАМИ ДИАБЕТА ХОРОШИЕ, И ПОПРОСИЛ СДЕЛАТЬ КОЕ-ЧТО, ДЛЯ ЧЕГО НАДО ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ...
НУ... МОЖЕТ, ТЫ К СЭМПАЮ СВОЕМУ ОБРАТИШЬСЯ?
ЧТО?
РАССКАЖИ МНЕ, МОЛ, КАК НАДО СДЕЛАТЬ ТО И ЭТО.
ЧАВК
ЧАВК
ЧАВК
ЧТО?
КСТАТИ, ЧТО-ТО ПИСЬМА ОТ НЕЕ КАКИЕ-ТО ФОРМАЛЬНЫЕ!
МОЖЕТ, Я ЧТО-ТО НЕ ТО СКАЗАЛ?
НЕРВНИЧАЕТ
.....


НУ, Я МОГУ СПРОСИТЬ У ТОВАРИЩА ПО ИГРЕ, КОТОРЫЙ ПОМОГ МНЕ С ЖУРНАЛОМ ПРО МАШИННОЕ ОБУЧЕНИЕ.
ДА? СПАСИБО!
Неделю спустя
ТАК...
ЧТОБЫ ПРОАНАЛИЗИРОВАТЬ МОДЕЛИ ПОВЕДЕНИЯ ПОЖИЛЫХ, НУЖНА КЛАСТЕРИЗАЦИЯ, А ЧТОБЫ ВЫБРАТЬ НУЖНУЮ ИНФОРМАЦИЮ – РАЗЛОЖЕНИЕ МАТРИЦЫ.
НО Я СЛИШКОМ МНОГОГО НЕ ЗНАЮ...
ДЗИНЬ
М-М-М?
СООБЩЕНИЕ.
ТЫК
ТЫК
О, ЭТО ОТ САЯКА-СЭМПАЙ!
Саяка-сэмпай
Извини, что отвлекаю
Не хочешь встретиться в кафе после работы?
ДЗЫНЬ!
ДОБРО ПОЖАЛОВАТЬ!
CAFE


СЭМПАЙ!
ИЗВИНИТЕ, ЧТО ОПОЗДАЛ!
ТОП
ТОП
НИЧЕГО СТРАШНОГО!


ЧТО-ТО СЛУЧИЛОСЬ?
И СЕГОДНЯ ОНА ТОЖЕ МИЛАЯ!
НУ...


КИЁХАРА-КУН,
У ТЕБЯ ОПЯТЬ ПРОБЛЕМЫ С МАШИННЫМ ОБУЧЕНИЕМ?

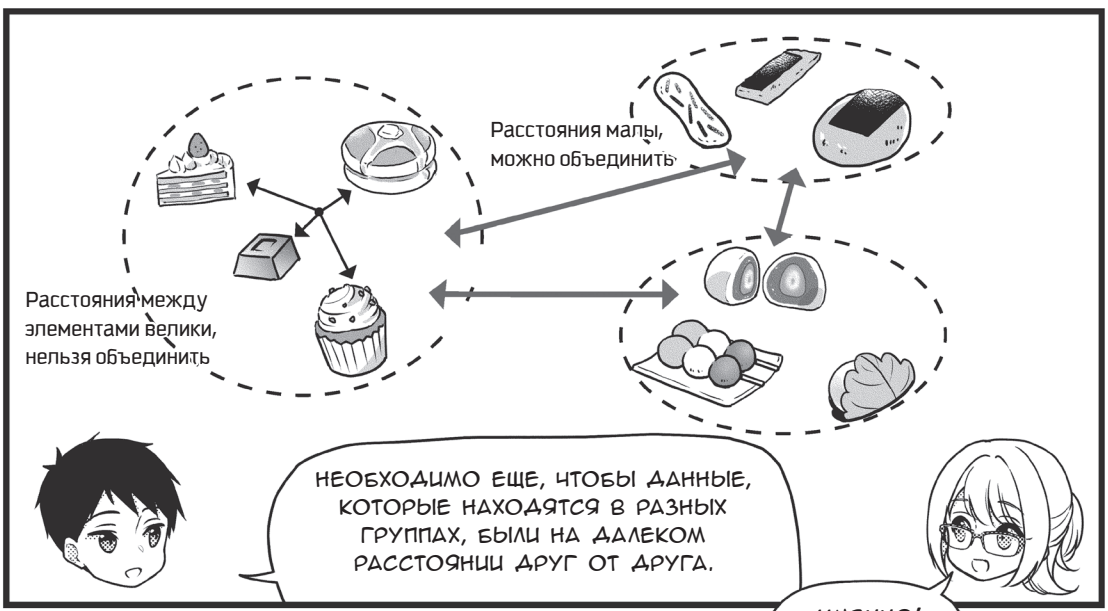

ТЫ ЭКСТРАСЕНС, ЧТО ЛИ?
А?


НЕТ, НИЧЕГО, ИЗВИНИТЕ, ЗАБУДЬТЕ, ПОЖАЛУЙСТА!
Я ПОНЯЛА, ЧТО ТЫ ОЧЕНЬ УСТАЛ.


НУ...
НА САМОМ ДЕЛЕ...
РАССКАЖИТЕ МНЕ ПРО ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ.
АГА!


Я попытался сам найти информацию про кластеризацию и разложение матрицы.
Направление верное!
Хорошо!
Сэмпай, вы мне расскажете про обучение без учителя, правда?
Конечно!
Улыбается
Я же рада видеть, что ты чему-то учишься!
Ага!
Радостно

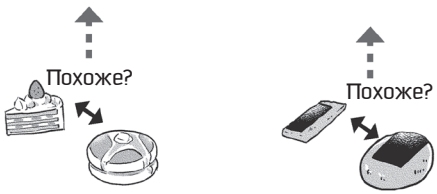
6.1. КЛАСТЕРИЗАЦИЯ



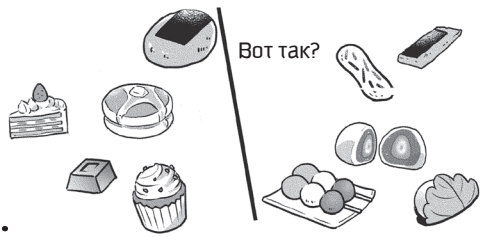
КЛАСТЕРИЗАЦИЯ ДЕЛИТСЯ НА **ИЕРАРХИЧЕСКУЮ**, КОГДА НОВЫЕ КЛАСТЕРЫ СОЗДАЮТСЯ ПУТЕМ ОБЪЕДИНЕНИЯ БОЛЕЕ МЕЛКИХ КЛАСТЕРОВ И ДЕРЕВО СОЗДАЕТСЯ ОТ ЛИСТЬЕВ К СТВОЛУ; И НА **РАЗДЕЛЯЮЩУЮ (ДИВИЗИВНУЮ)**...

...КОГДА НОВЫЕ КЛАСТЕРЫ СОЗДАЮТСЯ ПУТЕМ ДЕЛЕНИЯ БОЛЕЕ КРУПНЫХ КЛАСТЕРОВ НА БОЛЕЕ МЕЛКИЕ И ДЕРЕВО СОЗДАЕТСЯ ОТ СТВОЛА К ЛИСТЬЯМ.

Иерархическая кластеризация



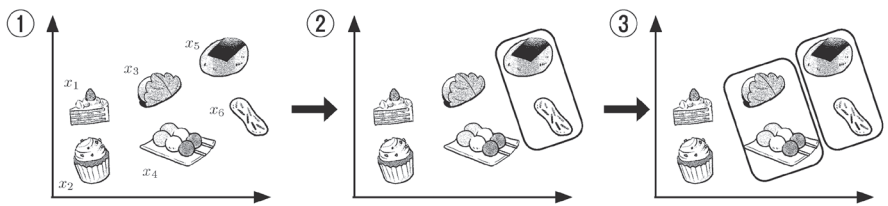
Разделяющая кластеризация



6.1.1. Иерархическая кластеризация

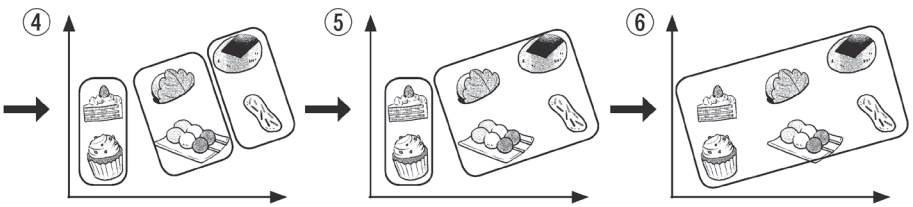
ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ – ЭТО ПРОЦЕСС, ПРИ КОТОРОМ СОЗДАЮТСЯ НОВЫЕ КЛАСТЕРЫ ПУТЕМ ОБЪЕДИНЕНИЯ БОЛЕЕ МЕЛКИХ КЛАСТЕРОВ. КЛАСТЕРЫ ПОСТЕПЕННО РАСТУТ.

АЛГОРИТМ ВЫГЛЯДИТ ТАК:



① Данные объединяются в кластеры

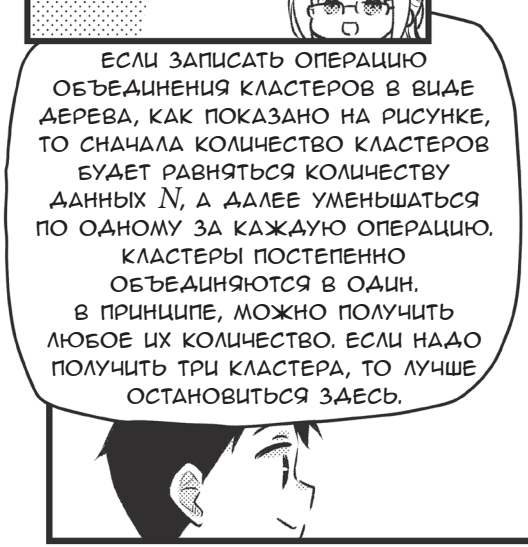
②–⑤ Кластеры, которые находятся рядом, объединяются – так получаются новые кластеры

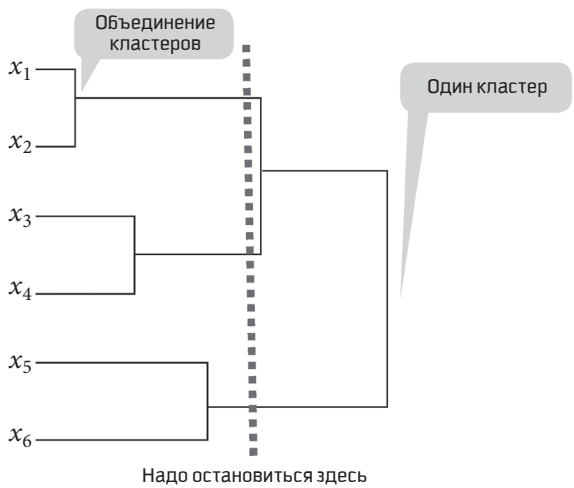


⑥ В конце концов данные объединяются в один кластер

ПОНЯТНО, КОГДА ДАННЫЕ НАХОДЯТСЯ РЯДОМ, НО КАК ВЫЧИСЛИТЬ РАССТОЯНИЕ МЕЖДУ ДАННЫМИ И КЛАСТЕРОМ, А ТАКЖЕ МЕЖДУ ДВУМЯ КЛАСТЕРАМИ?

Метод одиночной связи	Метод полной связи	Центроидный метод	Метод Уорда
Сходство определяется величиной расстояния между самыми близкими элементами	Сходство определяется величиной расстояния между самыми дальними элементами	Сходство определяется расстоянием между центроидами кластеров	После объединения кластеров вычисляется квадрат среднего расстояния между данными и центром кластера, из него вычитается эта величина до объединения
Есть тенденция к удлинению кластеров в одном направлении	Есть тенденция к избеганию удлинения кластеров в одном направлении	Тенденция к удлинению кластеров находится между одиночной и полной связью	Часто получаются сравнительно хорошие кластеры





6.1.2. Разделяющая кластеризация

А что за метод – разделяющая кластеризация?

Разделяющая кластеризация – метод, который определяет функцию оценки качества разделения данных и выдает результат лучшего значения этой функции.

А чем он отличается от иерархической кластеризации?

В иерархической кластеризации деление идет снизу вверх, поэтому с точки зрения целого могут создаваться искаженные кластеры. С этой точки зрения кластеры, которые создаются при разделяющей кластеризации, более оптимальны.

Так. Получается, что разделяющая кластеризация всегда дает лучшие результаты, чем иерархическая?..

Ну...

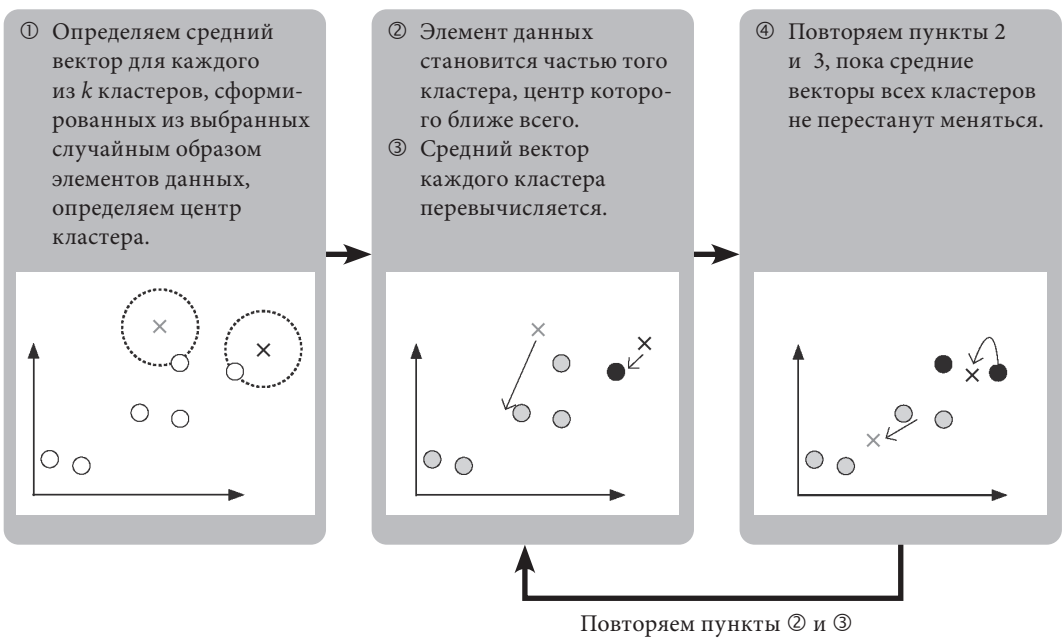
Например, какие вычисления надо провести, если надо перебрать и найти лучший результат функции оценки при разделении N элементов данных на два кластера?

Один элемент данных может быть разделен по кластерам одним из двух способов. Два – из четырех. А три, получается, из восьми.

Получается 2 в степени *N*. Если *N* – большое число, то на практике невозможно оценить все значения функции.

В таком случае в качестве обычного метода используют поиск оптимального разделения. Основной метод разделяющей кластеризации – метод *k*-средних, можешь рассказать о нем?

Да. Выглядит он так. У нас изначально должно быть задано количество кластеров *k*.



А какая у этого метода функция оценки?

Ну, сумма расстояний между каждым элементом данных и средним вектором кластера, к которому он принадлежит.

А почему при этом величина функции оценки уменьшается?

На втором этапе изменение кластера означает, что найден средний вектор, который находится еще ближе, поэтому величина функции оценки уменьшится.
Далее, на третьем этапе, снова рассчитывается средний вектор, и так повторяется, пока он не перестанет меняться; это будет означать, что найдена наименьшая сумма расстояний данных в кластере, поэтому величина функции оценки становится меньше.

Да. Однако при использовании этого метода мы находим локальное оптимальное решение. Я специально говорю «локальное», потому что этот метод не позволяет сказать, является ли оно общим для целого.

Значит, можно найти ответ и получше, но метод останавливается на полпути.

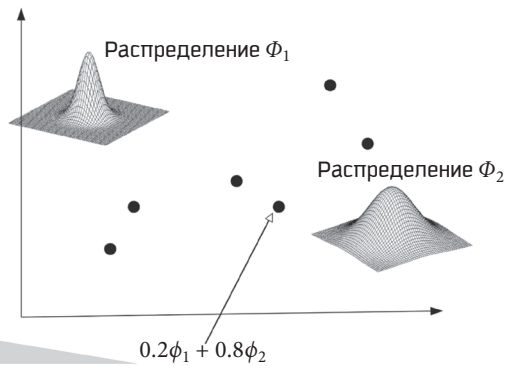
Да. Однако если провести несколько итераций с разными входными данными, то кластеризация методом *k*-средних даст хороший результат.

КСТАТИ, СУЩЕСТВУЕТ ТАК НАЗЫВАЕМЫЙ **EM-АЛГОРИТМ**, ОСНОВАННЫЙ НА ФУНКЦИИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ, КОТОРЫЙ ПОМОГАЕТ НЕ ТОЛЬКО РАЗДЕЛИТЬ ДАННЫЕ ПО КЛАСТЕРАМ, НО И ГЕНЕРИРОВАТЬ НОВЫЕ ДАННЫЕ В КАЖДОМ КЛАССЕ.

ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ? СЛОЖНОВАТО.

В ОБЩЕМ, ЕЕ МОЖЕТ ЛЕГКО ОБЪЯСНИТЬ ГИПОТЕЗА РАСПРЕДЕЛЕНИЯ ПРАВИЛЬНЫХ ОТВЕТОВ.

ПОРЯДОК ТАКОЙ ЖЕ, КАК И В МЕТОДЕ k-СРЕДНИХ: СНАЧАЛА МЫ СТРОИМ СРЕДНИЕ ВЕКТОРЫ И НАХОДИМ МАТРИЦУ РАСПРЕДЕЛЕНИЯ. ЭТО ЭКВИВАЛЕНТНО ПОМЕЩЕНИЮ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ В СООТВЕТСТВУЮЩЕМ МЕСТЕ В ПРОСТРАНСТВЕ ПРИЗНАКОВ.



ЗАТЕМ С ПОМОЩЬЮ МЕТОДА k-СРЕДНИХ МЫ ОПРЕДЕЛЯЕМ, К КАКОМУ КЛАСТЕРУ ПРИНАДЛЕЖИТ КАЖДЫЙ ЭЛЕМЕНТ ДАННЫХ.

С ПОМОЩЬЮ EM-АЛГОРИТМА МОЖНО НАЙТИ РАСПРЕДЕЛЕНИЕ: НАПРИМЕР, В ОДНОМ КЛАСТЕРЕ 20 %, А В ДРУГОМ 80 %.

БАМ! БАМ!

20%

80%

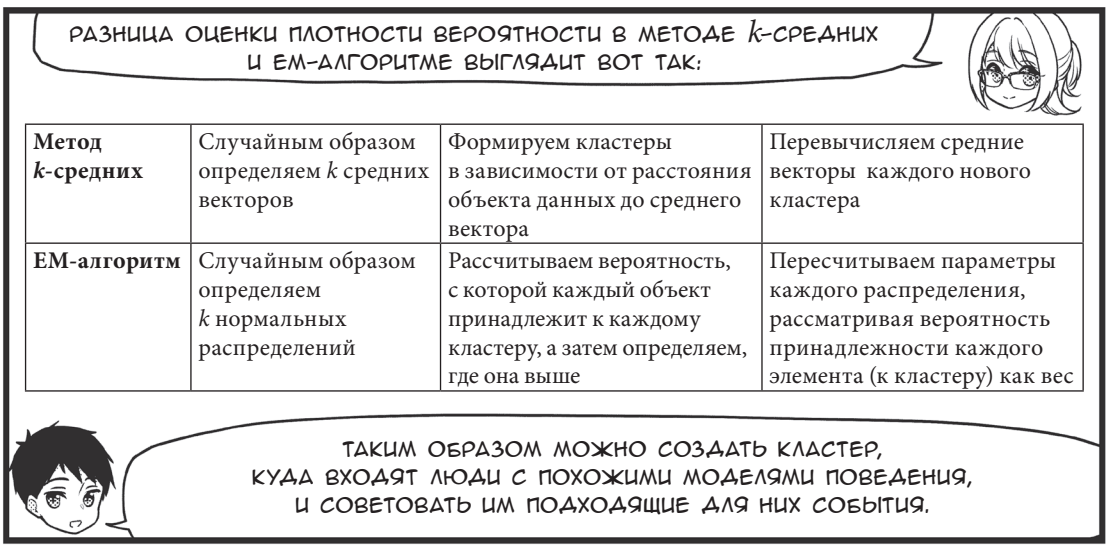
И КАК УДАЛОСЬ НАЙТИ ЭТО РАСПРЕДЕЛЕНИЕ?

ЕГО МОЖНО РАССЧИТАТЬ НА ОСНОВАНИИ ПОДХОДЯЩЕГО РАСПРЕДЕЛЕНИЯ ВСЕХ ДАННЫХ.

ЭТО КАК-ТО СЛУЧАЙНО!

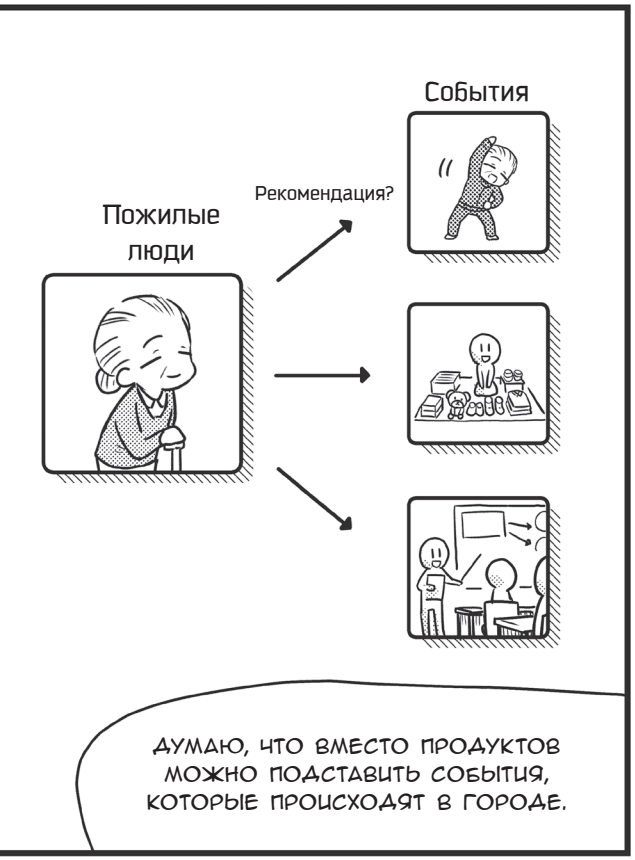
НО СНАЧАЛА, ИСПОЛЬЗУЯ МЕТОД k-СРЕДНИХ, МЫ СЛУЧАЙНЫМ ОБРАЗОМ ОПРЕДЕЛЯЛИ СРЕДНИЙ ВЕКТОР И НА ОСНОВАНИИ ЕГО СТРОИЛИ КЛАСТЕРЫ, ПРАВДА?

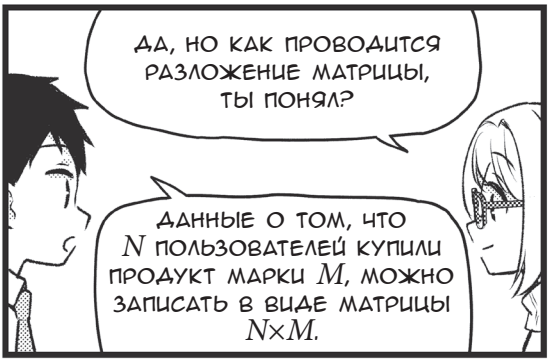
НУ... ДА.



Метод *k*-средних	Случайным образом определяем *k* средних векторов	Формируем кластеры в зависимости от расстояния объекта данных до среднего вектора	Перевычисляем средние векторы каждого нового кластера
EM-алгоритм	Случайным образом определяем *k* нормальных распределений	Рассчитываем вероятность, с которой каждый объект принадлежит к каждому кластеру, а затем определяем, где она выше	Пересчитываем параметры каждого распределения, рассматривая вероятность принадлежности каждого элемента (к кластеру) как вес

6.2. Разложение матрицы

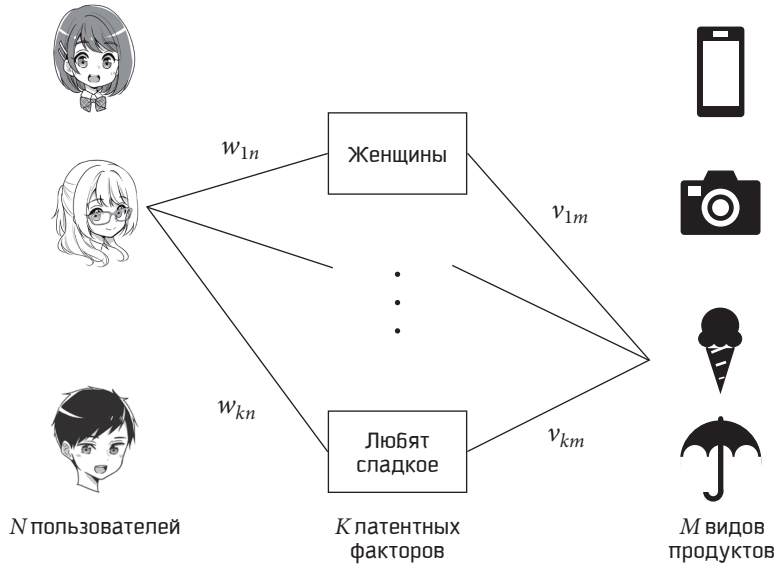

Пожилые люди
Рекомендация?
События
Думаю, что вместо продуктов можно подставить события, которые происходят в городе.


Да, но как проводится разложение матрицы, ты понял?
Данные о том, что N пользователей купили продукт марки M, можно записать в виде матрицы $N \times M$.


Однако если в такой матрице и N, и M – большие числа, то матрица получается немного бессмысленной.
Да, и более того, товары, которые покупает пользователь, составляют лишь небольшую часть всех товаров на сайте.


Есть гипотеза о существовании нескольких латентных факторов, на основании которых пользователи покупают вещи. Когда пользователей группируют на основании этих факторов, то в фокусе оказываются признаки вроде "женщина", "местная", "любит сладкое".
Поскольку эти данные добываются без учителя, их не всегда можно интерпретировать.
Женщины
Местные
Любят сладости


"Женщина", "местная", "любит сладкое"...
Ну-ну...



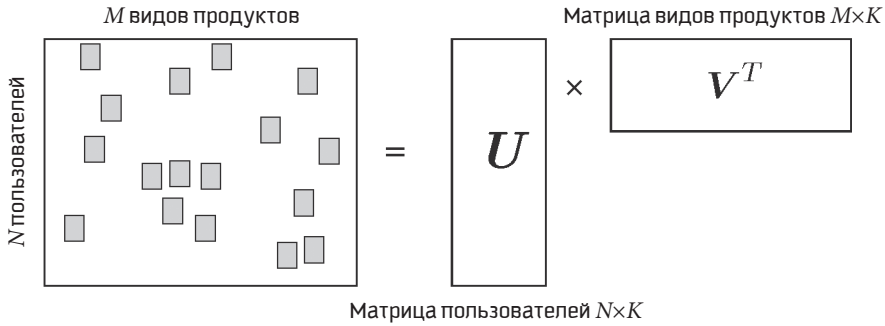
У нас есть K латентных факторов, и это число K меньше как количества пользователей N, так и товаров M.

Если мы будем следовать данной гипотезе, то величина элементов изначальной матрицы будет определяться по формуле ниже:

$x_{nm} = w_1 v_{1m} + w_2 v_{2m} + \cdots + w_K v_{Km}$.

А как с ее помощью разложить матрицу?

Разложение матрицы при помощи этой формулы – на следующей странице.



Если матрица $N \times M$ большая, то информация о пользователях собирается в матрицу $\boldsymbol{U}$, в которой N столбцов и K строк, а информация о товарах собирается в матрицу $\boldsymbol{V}$, в которой M столбцов и K строк.

А зачем в целом $\boldsymbol{U}$ и $\boldsymbol{V}$ нужны?

Ну, если изначальная матрица – $\boldsymbol{X}$, то можно сравнить $\boldsymbol{X}$ с $\boldsymbol{UV}^T$ и, обозначив ошибку за $\boldsymbol{E}$, можно решить задачу оптимизации $\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{UV}^T$. Учитывая совместимость с методом градиента, который используется в качестве процедуры оптимизации, мы рассмотрим квадрат ошибки.

Для вычисления ошибки матрицы мы вычисляем квадратный корень из суммы квадратов каждого элемента матрицы. С использованием нормы Фробениуса оптимальный вариант выглядит так:

$$\min_{\boldsymbol{U},\boldsymbol{V}} \frac{1}{2} \|\boldsymbol{E}\|_{Fro}^2 = \min_{\boldsymbol{U},\boldsymbol{V}} \frac{1}{2} \|\boldsymbol{X} - \boldsymbol{UV}^T\|_{Fro}^2$$

Это можно решить с использованием сингулярного разложения, которое изучается в линейной алгебре, но для этого надо заменить $\boldsymbol{X}$ на 0, если значение $\boldsymbol{X}$ не указано. Несмотря на то что никакой информации поначалу нет, далее она появится, и она будет отличаться от изначально введенных данных.

И что мы будем делать?

Так. Поскольку величина $\boldsymbol{X}$ существует, надо максимально уменьшить квадрат ошибки. Этим задача близка к регрессии. И, как и в задаче регрессии, тут появляются те же проблемы.

Переобучение. Тогда используем регуляризацию.

Да. Мы проводим оптимизацию с помощью алгоритма Alternating Least Squares.

$$\min_{\boldsymbol{U},\boldsymbol{V}} \sum_{(i,j)\in\Omega} (x_{ij} - \boldsymbol{u}_i^T \boldsymbol{v}_j)^2 + \lambda_1 ||\boldsymbol{U}||_{Fro}^2 + \lambda_2 ||\boldsymbol{V}||_{Fro}^2$$

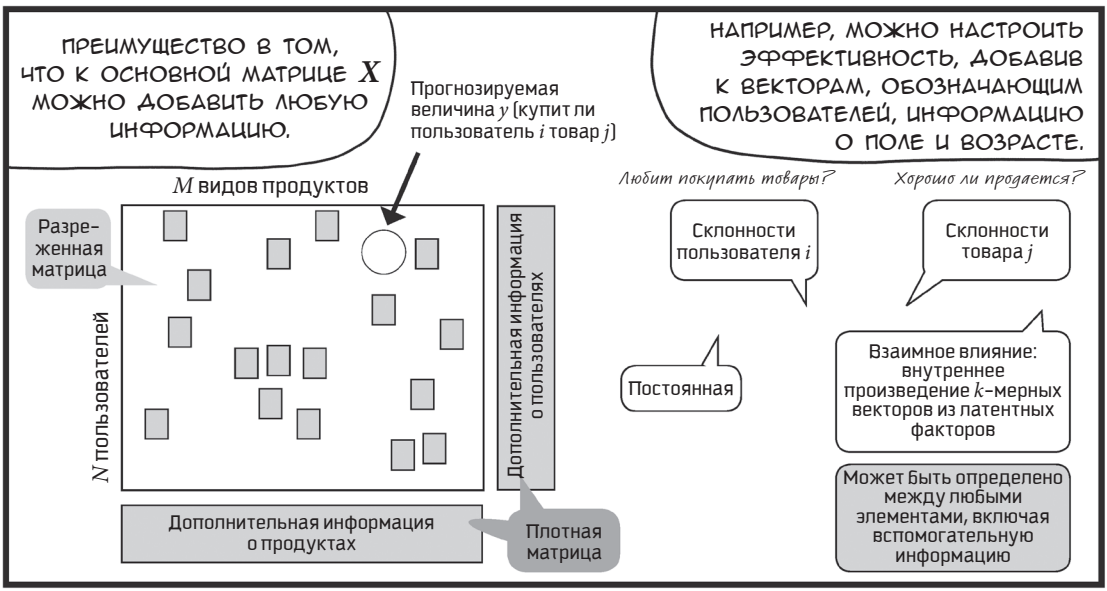
Здесь Ω – это индекс элемента матрицы $\boldsymbol{X}$, $\boldsymbol{u}_i$ – k-мерный вектор в i-й строке матрицы $\boldsymbol{U}$, то же самое верно для $\boldsymbol{v}_j$ – k-мерного вектора в j-й строке матрицы $\boldsymbol{V}$.

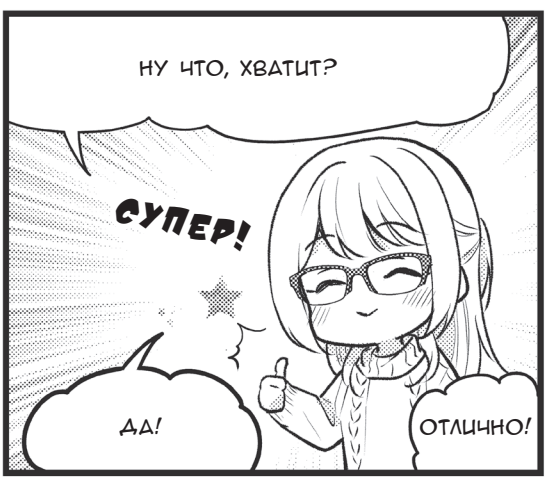
Ты говорил, что будешь использовать метод градиентов, но можно ли его использовать, если у нас два объекта оптимизации, $\boldsymbol{U}$ и $\boldsymbol{V}$?

$\boldsymbol{U}$ и $\boldsymbol{V}$ могут взаимно оптимизироваться, по всей видимости.

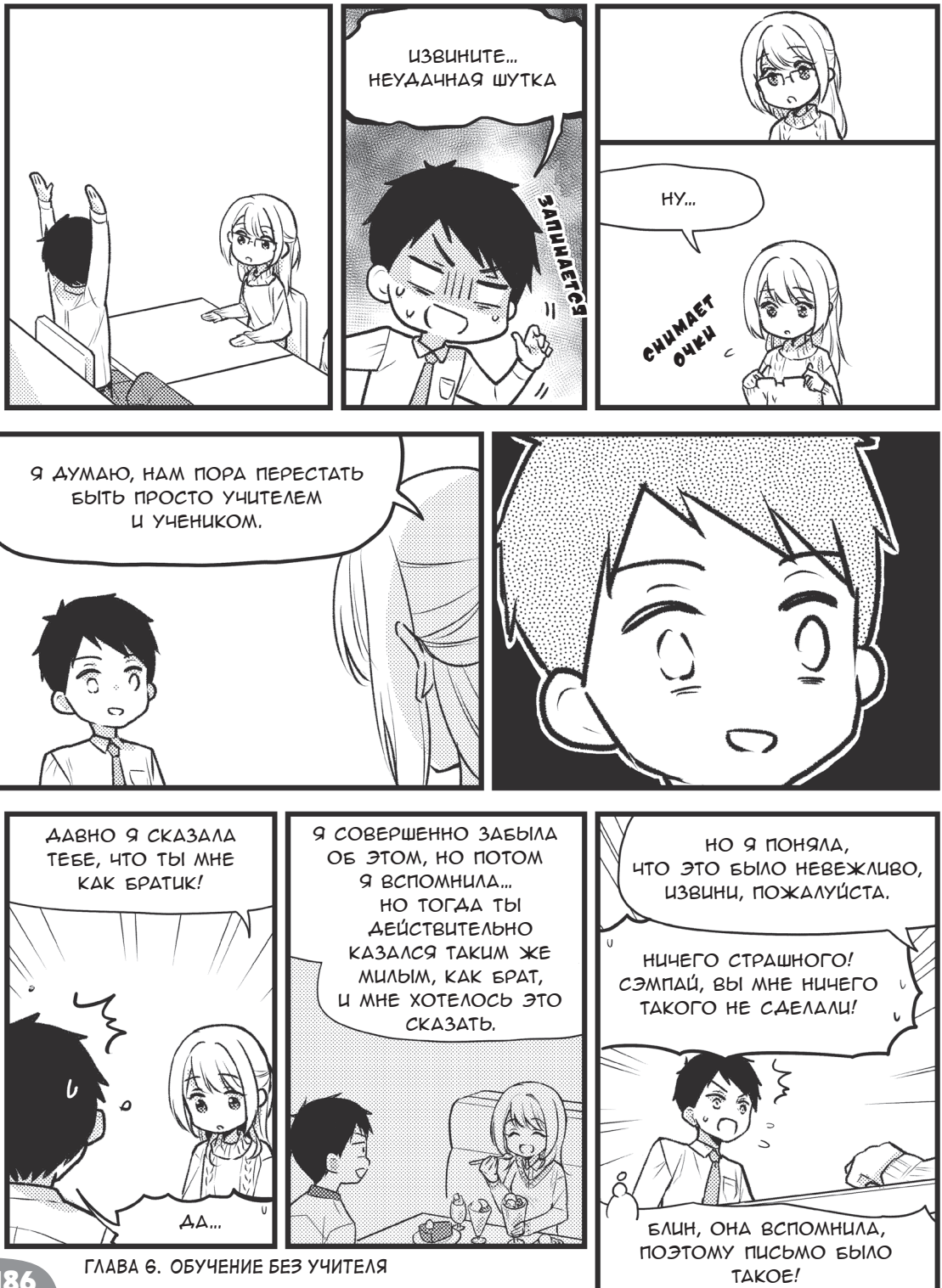

МОЛОДЕЦ.
А ТЫ СЛЫШАЛ О НЕОТРИЦАТЕЛЬНОМ МАТРИЧНОМ РАЗЛОЖЕНИИ?
ДА, ОНО ВСТРЕЧАЛОСЬ В РАЗНЫХ МЕСТАХ, НО Я НЕ МОГ ПОНЯТЬ, ЧТО ТАКОЕ НЕОТРИЦАТЕЛЬНОЕ.
СУТЬ В ТОМ, ЧТО БОЛЬШИНСТВО ЭЛЕМЕНТОВ МАТРИЦЫ X НЕОТРИЦАТЕЛЬНЫЕ.
АГА, КТО СКОЛЬКО ТОВАРОВ КУПИЛ ИЛИ ОЦЕНИЛ КИНО НА СТОЛЬКО-ТО БАЛЛОВ.
ДОПУСТИМ, X – КОЛИЧЕСТВО ТОВАРОВ В МАГАЗИНЕ, И ПОКУПАТЕЛЬ КУПИЛ В МАГАЗИНЕ ОДИН ТОВАР. И ЭТО ОДНО ЧИСЛО ОБОЗНАЧАЕТ ПОКУПАТЕЛЯ, КОТОРЫЙ КУПИЛ ОДИН ТОВАР.
1 – ЭТО ВЕРОЯТНОСТЬ ТОГО, ЧТО ЭТОТ ПОКУПАТЕЛЬ:
...ЖЕНЩИНА.
...СИДИТ НА ДИЕТЕ...
...СЛАДКОЕЖКА...
МОЖНО ОБЪЯСНИТЬ, ЧТО ТОВАР – ЭТО СУММА ВСЕХ ВЫЧИСЛЕНИЙ ВЕРОЯТНОСТЕЙ, СВЯЗАННЫХ С ЭТИМИ СКРЫТЫМИ ФАКТОРАМИ.
ДЕЙСТВИТЕЛЬНО, МОЖНО ЖЕ ОБЪЯСНИТЬ ЕГО КАК ВЕРОЯТНОСТЬ...
А ВЕРОЯТНОСТЬ НЕ МОЖЕТ БЫТЬ ОТРИЦАТЕЛЬНОЙ!
ЕСЛИ U И V МОГУТ ПРИНИМАТЬ ПРОИЗВОЛЬНЫЕ ЗНАЧЕНИЯ, ТО ДЛЯ УМЕНЬШЕНИЯ ОШИБКИ МОГУТ ПОТРЕБОВАТЬСЯ КРАЙНИЕ ЗНАЧЕНИЯ.
А ТУТ ОНИ ОГРАНИЧЕНЫ ДО НЕОТРИЦАТЕЛЬНЫХ.
ЕСЛИ МЫ ИХ ОГРАНИЧИЛИ ДО НЕОТРИЦАТЕЛЬНЫХ, ТО ЭТО ОЧЕНЬ ПОХОЖЕ НА РЕГУЛЯРИЗАЦИЮ.


ДА,
ЭТА МОДЕЛЬ ВЫЧИСЛЯЕТ ЗНАЧЕНИЕ, КОТОРОЕ МОЖНО ОЦЕНИТЬ ПО ВНУТРЕННЕМУ ПРОИЗВЕДЕНИЮ U И V В КАЖДОМ ИЗМЕРЕНИИ И СУММЕ ВЕСОВ ЭЛЕМЕНТОВ ПРОИЗВОЛЬНОГО ИЗМЕРЕНИЯ.
А ПРО FACTORIZATION MACHINE ТЫ СЛЫШАЛ?


ПРЕИМУЩЕСТВО В ТОМ, ЧТО К ОСНОВНОЙ МАТРИЦЕ X МОЖНО ДОБАВИТЬ ЛЮБУЮ ИНФОРМАЦИЮ.
Прогнозируемая величина y (купит ли пользователь i товар j)
НАПРИМЕР, МОЖНО НАСТРОИТЬ ЭФФЕКТИВНОСТЬ, ДОБАВИВ К ВЕКТОРАМ, ОБОЗНАЧАЮЩИМ ПОЛЬЗОВАТЕЛЕЙ, ИНФОРМАЦИЮ О ПОЛЕ И ВОЗРАСТЕ.
M видов продуктов
Любит покупать товары?
Хорошо ли продается?
Разре-женная матрица
N пользователей
Дополнительная информация о пользователях
Склонности пользователя i
Склонности товара j
Постоянная
Взаимное влияние: внутреннее произведение k-мерных векторов из латентных факторов
Может быть определено между любыми элементами, включая вспомогательную информацию
Дополнительная информация о продуктах
Плотная матрица

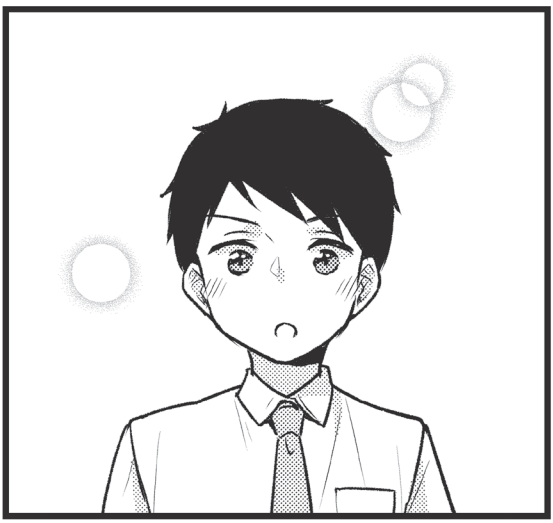

НУ ЧТО, ХВАТИТ?
СУПЕР!
ДА!
ОТЛИЧНО!


УРА! СПАСИБО!
ТЕПЕРЬ Я МОГУ ПРОВОДИТЬ ОБУЧЕНИЕ МОДЕЛИ БЕЗ УЧИТЕЛЯ, БЕЗ СВОЕГО УЧИТЕЛЯ, – БЕЗ ВАС, СЭМПАЙ!
ВООДУШЕВЛЕННО


ИЗВИНИТЕ... НЕУДАЧНАЯ ШУТКА
ЗАПИНАЕТСЯ
НУ...
СНИМАЕТ ОЧКИ
Я ДУМАЮ, НАМ ПОРА ПЕРЕСТАТЬ БЫТЬ ПРОСТО УЧИТЕЛЕМ И УЧЕНИКОМ.
ДАВНО Я СКАЗАЛА ТЕБЕ, ЧТО ТЫ МНЕ КАК БРАТИК!
ДА...
Я СОВЕРШЕННО ЗАБЫЛА ОБ ЭТОМ, НО ПОТОМ Я ВСПОМНИЛА... НО ТОГДА ТЫ ДЕЙСТВИТЕЛЬНО КАЗАЛСЯ ТАКИМ ЖЕ МИЛЫМ, КАК БРАТ, И МНЕ ХОТЕЛОСЬ ЭТО СКАЗАТЬ.
НО Я ПОНЯЛА, ЧТО ЭТО БЫЛО НЕВЕЖЛИВО, ИЗВИНИ, ПОЖАЛУЙСТА.
НИЧЕГО СТРАШНОГО! СЭМПАЙ, ВЫ МНЕ НИЧЕГО ТАКОГО НЕ СДЕЛАЛИ!
БЛИН, ОНА ВСПОМНИЛА, ПОЭТОМУ ПИСЬМО БЫЛО ТАКОЕ!


МЛАДШИЙ...
УЧЕНИК...
БРАТИК...
ВСЕ ЭТИ КАТЕГОРИИ НЕСУТ
В СЕБЕ УНИЗИТЕЛЬНОЕ ЗНАЧЕНИЕ.


ТЕПЕРЬ Я БУДУ ОБРАЩАТЬСЯ
К ТЕБЕ КИЁХАРА-КУН!




ДА!


ИЗВИНИ
ЗА "ЖЕНЩИНУ",
"МЕСТНУЮ",
"ЛЮБИТ СЛАДОСТИ"
В МОЕМ ПРИМЕРЕ.
ДА, ЭТО ВСЕ
ПРО МЕНЯ.


СЕГОДНЯ ПЛАТЫ ЗА ОБУЧЕНИЕ НЕ БУДЕТ, МЫ РАЗДЕЛИМ СЧЕТ ПОПОЛАМ!
ДА!
Хотя вы же меня многому научили сегодня
Спасибо
АЗЫНЬ!
АЗЫНЬ!

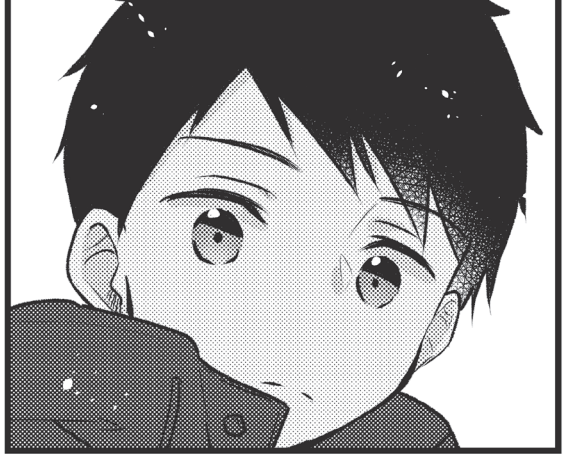

КСТАТИ, СЭМПАЙ, У ВАС ЖЕ БЫЛИ ДРУГИЕ ОЧКИ.
ЭТИ МНЕ ПОДАРИЛИ ДРУЗЬЯ В ЧЕСТЬ НОВОЙ РАБОТЫ.


ЕЙ ОЧЕНЬ ИДЕТ
В-А-ДАМ
О-О-Ч-
ЕЙ ОЧЕНЬ ИДЕТ
ЕЙ ОЧЕНЬ ИДЕТ
ОЧЕ-ЕНЬ
НА САМОМ ДЕЛЕ Я НЕ НАСТОЛЬКО БЛИЗОРУКАЯ.... КОГДА Я НАДЕВАЮ ОЧКИ, ТО БУДТО БЫ ПЕРЕКЛЮЧАЮСЬ НА РАБОТУ.


ХОРОШО, КОГДА ЕСТЬ ТАКОЙ ПЕРЕКЛЮЧАТЕЛЬ.
ЕЙ ОЧЕНЬ ИДЕТ!


КИЁХАРА-КУН, У ТЕБЯ ТОЖЕ ЕСТЬ ТАКОЙ ПЕРЕКЛЮЧАТЕЛЬ?
ЧТО?


КОГДА ТЫ ХОЧЕШЬ СДЕЛАТЬ ЧТО-ТО РАДИ КОГО-ТО ЕЩЕ!




ПРАВДА, ЧТО ЛИ?
ДА.


ТЫ ВЕДЬ СТАРАЛСЯ НА УРОКАХ ПО МАШИННОМУ ОБУЧЕНИЮ, ПОТОМУ ЧТО ЭТИ ПРОЕКТЫ БЫЛИ НУЖНЫ КОМУ-ТО ЕЩЕ.
Возможен ли у вас диабет?
ДА
НЕТ
Выберите нужный ответ
НАЧАТЬ


ПОМНИШЬ, КАК НА ФЕСТИВАЛЕ КУЛЬТУРЫ ТЫ С ДРУЗЬЯМИ ЖАРИЛ ЯКИСОБУ? ВЫ ОШИБЛИСЬ В ПРОПОРЦИЯХ, И ЕЕ БЫЛО НЕМНОГО БОЛЬШЕ, ЧЕМ НУЖНО.
КОНЕЧНО!


ТЫ ОБЫЧНО НЕ ШЕВЕЛИШЬСЯ, НО ТОГДА ТЫ НАЧАЛ ДЕЛАТЬ ОБЪЯВЛЕНИЯ: "ЛАПША ЗА ПОЛЦЕНЫ!", "ПРИБЫЛЬ БУДЕТ ПОЖЕРТВОВАНА!" ОБЪЯВЛЕНИЯ БЫЛИ И В СОЦСЕТЯХ, ПОЭТОМУ ВЫ СМОГЛИ ВСЕ ПРОДАТЬ.
Лапша за полцены!
Вся прибыль будет пожертвована
ДА... ПОМНЮ...


Я ИНОГДА ВОЛНУЮСЬ, МОЖЕТ БЫТЬ, ТЫ НЕ ХОЧЕШЬ ЧТО-ТО ДЕЛАТЬ ДЛЯ СЕБЯ?
НУ...
И ЭТО, МОЖЕТ БЫТЬ, НЕ СОВСЕМ ТО, ЧТО НАДО, НО МНЕ КАЖЕТСЯ, В ЭТОМ И ЕСТЬ ТВОЯ СИЛЬНАЯ СТОРОНА.
Я ТАК ДУМАЮ ИНОГДА.


СПАСИБО.
Ого... драгоценные слова


И это награда за то, что я старался весь год
НУ, ХВАТИТ, ПОЖАЛУЙ...
Я, НАВЕРНОЕ, НЕ СМОГУ ПРОВОДИТЬ ВАС, СЭМПАЙ.
КАК РАЗ ЖЕ КОНЕЦ ГОДА... СМОТРИ, НЕ ПЕРЕРАБОТАЙ!
АГА!
НО ЕСЛИ ЧТО-ТО СЛУЧИТСЯ, ПИШИ. ПУСТЬ ТЫ МНЕ БОЛЬШЕ И НЕ УЧЕНИК, И НИКАКОЙ ПЛАТЫ ЗА ОБУЧЕНИЕ ТОЖЕ НЕ НАДО!
КОНЕЧНО.


НУ ЧТО, ПОКА!
СЭМПАЙ!


ПРИЯТНОЙ ПОЕЗДКИ!


СПАСИБО!
ДО СВИДАНИЯ!

В кабинете у Саяка (7)

Математическое повторение (6)

В сегодняшней беседе меня заботит норма Фробениуса в матрице. Норма – это же величина вектора?

При правильном определении для d-мерного вектора x по отношению к p, если $1 \leq p < \infty$,

$$\sqrt[p]{|x_1|^p + \cdots + |x_d|^p},$$

то это называется L_p-нормой x. Если $p = 2$, то норма L_2 обычно имеет смысл величины вектора.

Трудно. А зачем тогда нужно p?

Представь, что $p = 1$. Разве ты не видела нигде L_1-норму?

Если $p = 1$, то L_1-норма – это сумма абсолютных значений всех элементов… Ага, если мы заменим $\boldsymbol{x}$ весом $\boldsymbol{w}$, то получим дополнительный член лассо-регрессии!

Да. Изменение стандарта размера также меняет эффект, который влияет на размер. Ридж-регрессия с нормой L_2 в качестве члена регуляризации и лассо-регрессия с нормой L_1 в качестве члена регуляризации будут по-разному влиять на коэффициенты.

Если помнить это, то можно связать многие вещи.

Таким образом, L_2-норма вектора – это квадратный корень из суммы квадратов всех его элементов. Точно так же норма Фробениуса матрицы – это квадратный корень из суммы квадратов всех элементов матрицы.

Я могу еще представить величину вектора, а величину матрицы – нет.

Нет нужды ее представлять. Надо просто понимать, что это все нужно только для того, чтобы минимизировать ошибку E до нулевой матрицы.

Я знала, что можно минимизировать ошибку способом, похожим на тот, что описан в главе 1, но не думала, что найду такой ответ.

Кстати, допустим, на сайте есть 100 пользователей и 50 товаров, и матрица будет размером 100×50, то есть в ней будет 5000 элементов. А если мы добавим 10 латентных факторов, то количество элементов будет:

$100 \times 10 + 50 \times 10 = 1500$.

Но 5000 единиц информации не могут быть представлены 1500 единицами!

Ты говоришь правильно, но идея разложения матрицы не в том, чтобы можно было восстановить предыдущую матрицу, а сделать приближение более низкого ранга.

В случае если значение каждого элемента исходной матрицы полностью независимо от других элементов, этот метод не дает результата разложения, близкого к исходной матрице.

Но в данном случае предполагается, что пользователи с похожими характеристиками демонстрируют схожее покупательское поведение, или же пользователи, приобретающие продукты с аналогичными характеристиками, имеют сходные тенденции покупок. Говоря техническими терминами, предполагается, что в данных, у которых много измерений, есть структуры с более низкими измерениями.

Я, наверное, лучше промолчу...

Начнем с разложения по собственным числам. Для матрицы $d \times d$, т. е. для квадратной матрицы d-порядка $\boldsymbol{M}$, рассмотрим пару из действительного числа λ и d-мерного вектора $\boldsymbol{x}$, удовлетворяющих следующим условиям:

$\boldsymbol{Ax} = \lambda \boldsymbol{x}, \quad \boldsymbol{x} \neq 0.$

Тогда формула преобразуется в $(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = 0$.

$\boldsymbol{I}$ – единичная матрица, если есть обратная матрица $\boldsymbol{A} - \lambda \boldsymbol{I}$, то $\boldsymbol{x} = 0$, что противоречит условиям. Значит, обратной матрицы $\boldsymbol{A} - \lambda \boldsymbol{I}$ нет. В этом случае определитель равен 0, или же

$\det(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = 0.$

Если у нас квадратная матрица второго порядка $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, то ее определитель равен $ad - bc$. В первой главе мы узнали, как можно найти обратную матрицу – $1/(ad - bc)$, но если знаменатель равен 0, то найти обратную матрицу не получится!

Да. В случае квадратной матрицы d-порядка $\det(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = 0$ представляет собой многочлен и будет иметь d решений. λ – собственное число, а соответствующий вектор $\boldsymbol{x}$ называется собственным вектором.

Угу.

Тогда, используя собственные числа и собственные векторы, матрица $\boldsymbol{M}$ будет записана как произведение матриц по формуле ниже:

$\boldsymbol{M} = \boldsymbol{U}diag(\lambda_1, ..., \lambda_d)\boldsymbol{U}^{-1}$,

где $\boldsymbol{U}$ – вектор-строка с d собственными векторами, которые находятся рядом, а $diag$ – диагональная матрица, в которой упорядоченные числа расположены в диагональных элементах.

Мне пока понятно. Если в $\boldsymbol{M}$ d строк и d столбцов, а в $\boldsymbol{U}$ и $\boldsymbol{U}^{-1}$ – тоже d строк и d столбцов, то если выбрать правильные значения, можно провести преобразование.

А разве то, что $\boldsymbol{M}$ должна быть квадратной матрицей, – это не необычное условие? Разве, когда на сайте количество пользователей и товаров одинаковое, это не странно?

Конечно. Поэтому это разложение по собственным числам превращается в сингулярное разложение.

$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$.

Здесь $\boldsymbol{M}$ – матрица $n \times m$, $\boldsymbol{U}$ – матрица $n \times n$, $\boldsymbol{V}$ – матрица $m \times m$, $\boldsymbol{\Sigma}$ – матрица $n \times m$. Таким образом, при перемножении $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ получится матрица $n \times m$.

$\boldsymbol{\Sigma}$ – не квадратная матрица, и диагональной матрицы из нее не получится.

$\boldsymbol{\Sigma}$ – матрица с r сингулярными числами, лежащими на главной диагонали (r меньше m и n), а оставшиеся элементы дополняются нулями до n строк и m столбцов.

Сингулярные числа рассчитываются по собственному значению, и здесь мы можем думать о них как о числах, полученных в результате разложения.

$$\boldsymbol{\Sigma} = \begin{array}{|c|c|} \hline \sigma_1 \quad \ddots \quad \sigma_r & 0 \\ \hline 0 & 0 \\ \hline \end{array}$$

Здесь важно знать, что сингулярные значения от σ_1 до σ_r расположены по возрастанию.

Каждое сингулярное значение умножается на элементы ***U*** и ***V***, в результате чего получается элемент ***M***, и, конечно, большая величина сингулярного значения оказывает большое влияние на определение значения ***M***. Это сразу станет ясно, если представить, что первое сингулярное число настолько велико, что будто возвышается над остальными.

А если сингулярные значения со второго и ниже маленькие, то ***M*** будет меняться незначительно. Тогда они будут играть какую-то роль?

Да. Можно выбрать несколько больших сингулярных значений и затем смотреть, как меняется значение ***M***.

Если сумма нескольких выбранных сингулярных значений составляет большую часть от суммы всех сингулярных значений, то получится матрица, которая не сильно отличается от исходной ***M***.

Ага. Если мы выберем k элементов из больших сингулярных чисел, чтобы сделать $\mathbf{\Sigma}$, то ***U*** будет матрицей $n \times k$, ***V*** – $m \times k$, а $\mathbf{\Sigma}$ – $k \times k$. Если k меньше, чем n или m, то произведением небольших матриц мы сможем получить большую.

Да. Как я и говорила, в данных, у которых много измерений, есть структуры с более низкими измерениями.

Теперь я поняла. Даже когда ты будешь в Токио, Саяка, то учи меня математике! (Хотя не представляю, как можно спокойно работать в Токио…)

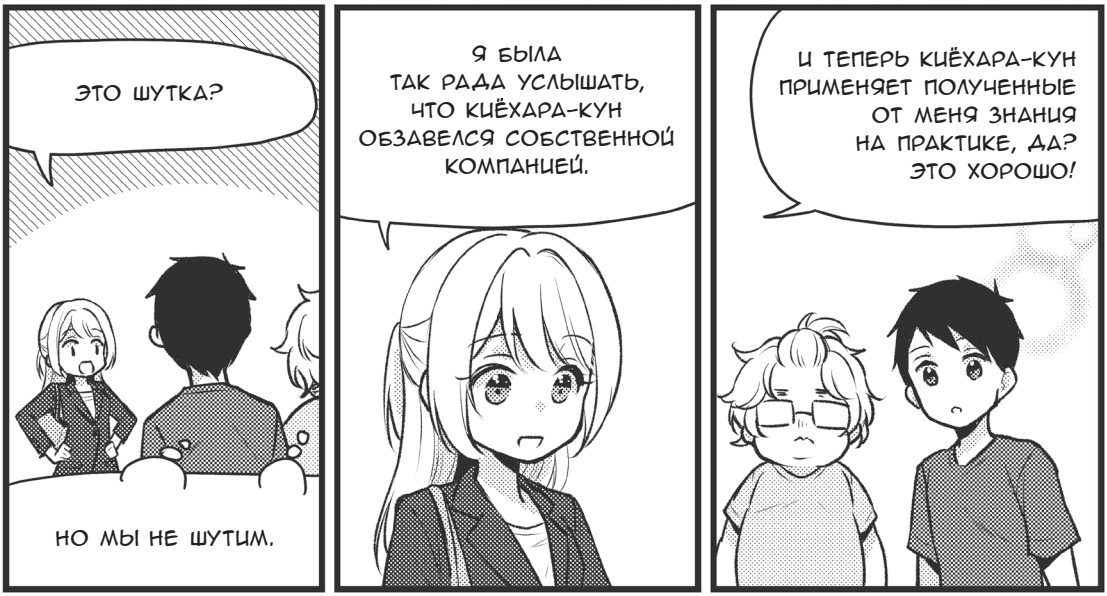
ЭПИЛОГ

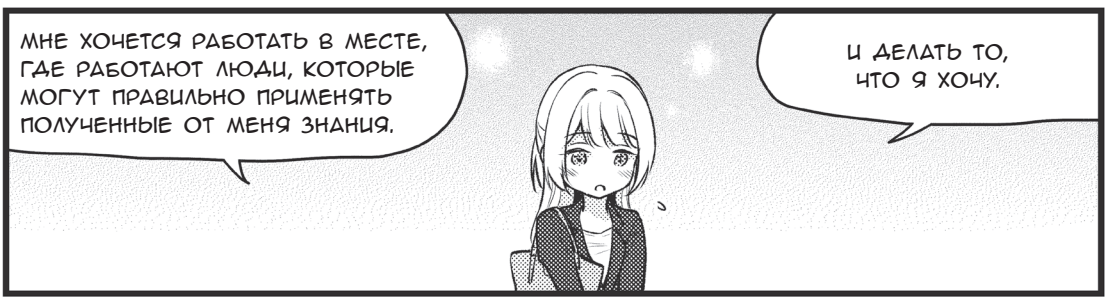
Полгода спустя...

Количество проектов, связанных с машинным обучением, увеличилось, поэтому Киёхара больше не мог концентрироваться на работе в администрации и создал свою компанию.


ТОП-ТОП-ТОП
ТОП-ТОП-ТОП
НАДО НЕ ТОЛЬКО МЕНЯ БЛАГОДАРИТЬ, НО И ВСЕХ В МЭРИИ, ЧТО НАШЛИ ОФИС ПОДЕШЕВЛЕ И ДАЛИ СТАРУЮ МЕБЕЛЬ.
СОГЛАСЕН.
ТОП-ТОП-ТОП
ТОП-ТОП-ТОП
ПОЙДУ-КА Я ПОРАБОТАЮ.
КОНЕЧНО!
ТОП-ТОП-ТОП
НО Я НЕ ДУМАЛ, КУДЗЁ-САН, ЧТО ВЫ ПОЙДЕТЕ СО МНОЙ.
Я ПОДУМАЛ, С КИЁХАРА СКУЧНО НЕ БУДЕТ.
СПАСИБО БОЛЬШОЕ ТЕБЕ!
КИЁХАРА-КУН?
БУХ!


ДА ЧТО Ж...
...ТАКОЕ!
САЯКА-СЭМПАЙ,
ЧТО ВЫ ТУТ ДЕЛАЕТЕ?
ВСЕМ
ПРИВЕТ!
НАДО ЖЕ?
ЭТО ТВОЯ СЭМПАЙ?
А? ДА.
СЭМПАЙ, А ЧТО
СЛУЧИЛОСЬ С ВАШЕЙ
РАБОТОЙ?
Я УВОЛИЛАСЬ.
РАДОСТНО
ЧТО?
НУ... НА ВСТУПИТЕЛЬНОЙ ЦЕРЕМОНИИ
ОКАЗАЛОСЬ, ЧТО У ПРЕЗИДЕНТА
КОМПАНИИ КАКИЕ-ТО СВОИ ОЖИДАНИЯ
ОТ НИ. Я ПОДУМАЛА, ЧТО ПОЛГОДА
ПОРАБОТАЮ И ВЕРНУСЬ ОБРАТНО,
А КОГДА УЗНАЛА СИТУАЦИЮ
ВНУТРИ КОМПАНИИ, ПОНЯЛА,
ЧТО ОН НА САМОМ ДЕЛЕ ТАК ДУМАЕТ,
И ПОЭТОМУ УВОЛИЛАСЬ!
Мудрое
решение
НАДО ЖЕ.


Это шутка?
Но мы не шутим.
Я была так рада услышать, что Киёхара-кун обзавелся собственной компанией.
И теперь Киёхара-кун применяет полученные от меня знания на практике, да? Это хорошо!


Мне хочется работать в месте, где работают люди, которые могут правильно применять полученные от меня знания.
И делать то, что я хочу.

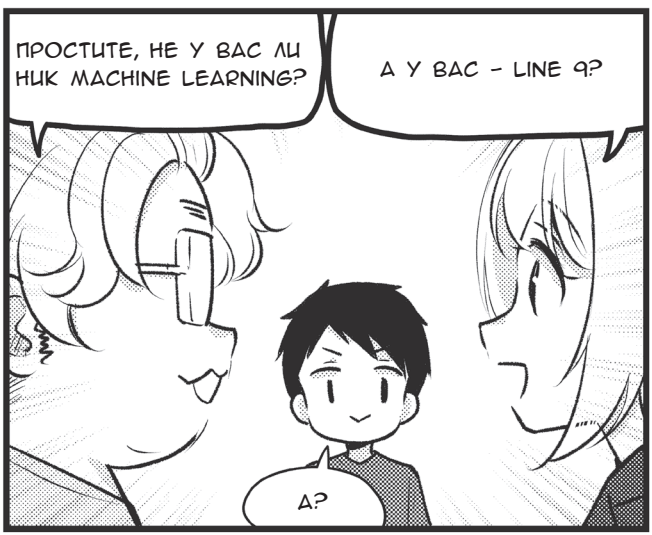

Вы возьмете меня на работу?


ДА! КОНЕЧНО, ВОЗЬМЕМ!

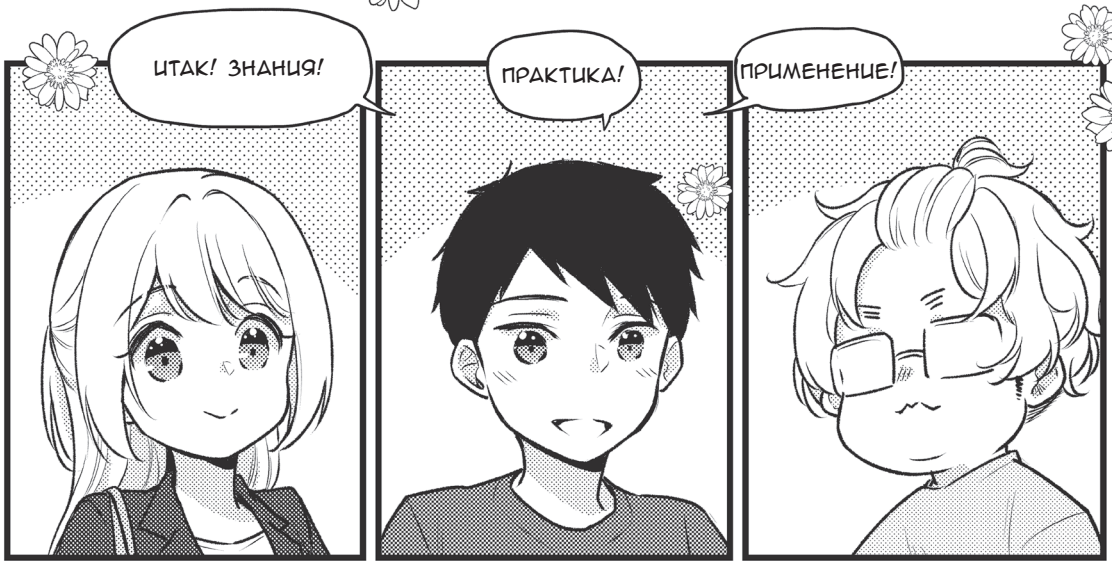

ПРАВДА? ЗДОРОВО! СПАСИБО!
ВЗАИМНО! УРА! С ТОБОЙ МЫ СПРАВИМСЯ СО ВСЕМ!


А Я ПРОГРАММИСТ, КУДЗЁ.
КУДЗЁ-САН? Я О ВАС СЛЫШАЛА. Я МИЯНО!


.....


ПРОСТИТЕ, НЕ У ВАС ЛИ НИК MACHINE LEARNING?
А У ВАС – LINE 9?
А?


ДА!
ЧТО? ЧТО?
ВЫ ЗНАКОМЫ?
ДА! ПО ОДНОЙ ИГРЕ!
Я Ж ГОВОРИЛ, ЧТО У МЕНЯ ЕСТЬ ЗНАКОМАЯ, КОТОРАЯ ИГРАЕТ В ИГРЫ.
А?
Я ПОБЛАГОДАРИЛ КУДЗЁ-САН ЗА ЭТУ СТАТЬЮ.
НА САМОМ ДЕЛЕ Я ИГРАЛ ТУТ С КАКИМ-ТО СПЕЦОМ ПО МАШИННОМУ ОБУЧЕНИЮ, ПОГОВОРИЛ С НИМ, И ОН ПОСОВЕТОВАЛ МНЕ ЭТУ КНИГУ.
ТОЧНО!
БИНГО! Я ТАК И ПОДУМАЛ, ЧТО ЭТО ВЫ.
УРА!
Я ТОЖЕ ПОДУМАЛА!
УРА!
.....!
А КОГДА РЕЧЬ ШЛА ОБ ОБУЧЕНИИ БЕЗ УЧИТЕЛЯ, ПРИШЛА СЭМПАЙ.
НУ, Я МОГУ СПРОСИТЬ У ТОВАРИЩА ПО ИГРЕ, КОТОРЫЙ ПОМОГ МНЕ С ЖУРНАЛОМ ПРО МАШИННОЕ ОБУЧЕНИЕ.
ДА? СПАСИБО!
КЮЁХАРА-КУН, У ТЕБЯ ОПЯТЬ ПРОБЛЕМЫ С МАШИННЫМ ОБУЧЕНИЕМ?
ДА, МЫ ЗНАЛИ, ЧТО ТЕБЕ БЫЛО ПЛОХО, И РЕШИЛИ ПОМОЧЬ...
?
ХЕ-ХЕ...
КУДЗЁ-САН!

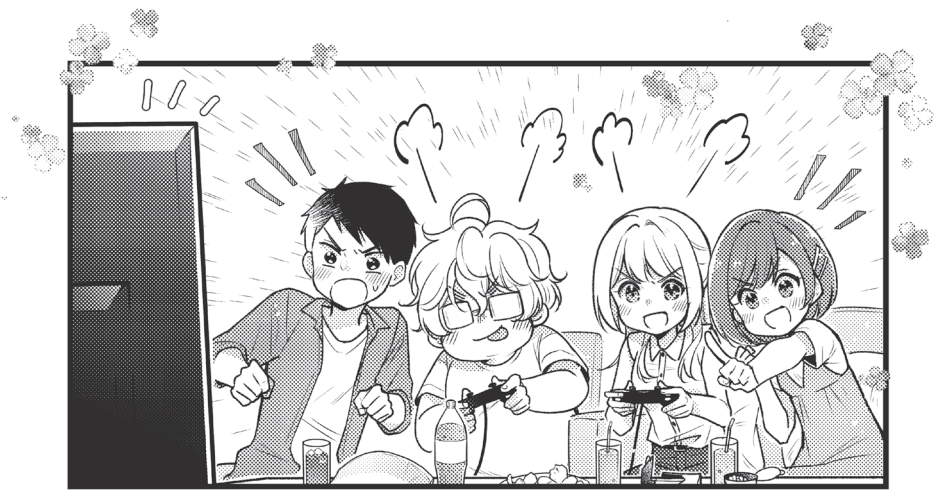

Итак! Знания!
Практика!
Применение!


Постараемся им следовать!
Да!
Ага!


Для начала сделаем приличный офис из этой комнаты!
Да! Я пойду переоденусь в спортивную одежду.
Вперед!
Отдохну-ка я!
Кудзё-сан!
Так, Киёхара, который нашел работу в компании своих хороших друзей, стал злиться...
Не надо тут этих разговоров!
Вот и я

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Книги издательства «ДМК ПРЕСС» можно купить оптом и в розницу
в книготорговой компании «Галактика» (представляет интересы издательств
«ДМК ПРЕСС», «СОЛОН ПРЕСС», «КТК Галактика»).
Адрес: г. Москва, пр. Андропова, 38;
тел.: **(499) 782-38-89**, электронная почта: **books@alians-kniga.ru**.
При оформлении заказа следует указать адрес (полностью),
по которому должны быть высланы книги; фамилию, имя и отчество получателя.
Желательно также указать свой телефон и электронный адрес.
Эти книги вы можете заказать и в интернет-магазине: **www.a-planeta.ru**.



Араки Масахиро (автор), Ватари Макана (художник)

Занимательное программирование

Машинное обучение

Манга

Главный редактор *Д. А. Мовчан*
dmkpress@gmail.com
Научный редактор *М. Е. Петровичева*
Переводчик *А. С. Слащева*
Корректор *Г. И. Синяева*
Верстальщик *А. А. Чаннова*

Формат 70×100 1/16.
Гарнитура Anime Ace. Печать офсетная.
Усл. п. л. 17,39. Тираж 500 экз.

Веб-сайт издательства www.dmkpress.com