

Программная инженерия. Разработка ПО (Python для продвинутых специалистов. Машинное обучение)

Модуль: Предобработка данных и машинное обучение

Лекция 1: Основы машинного обучения. Типы задач машинного обучения. Регрессия и классификация.

Дата: 12.05.2025

- Введение в машинное обучение
- Задачи машинного обучения
- Постановка задачи машинного обучения
- Линейная регрессия

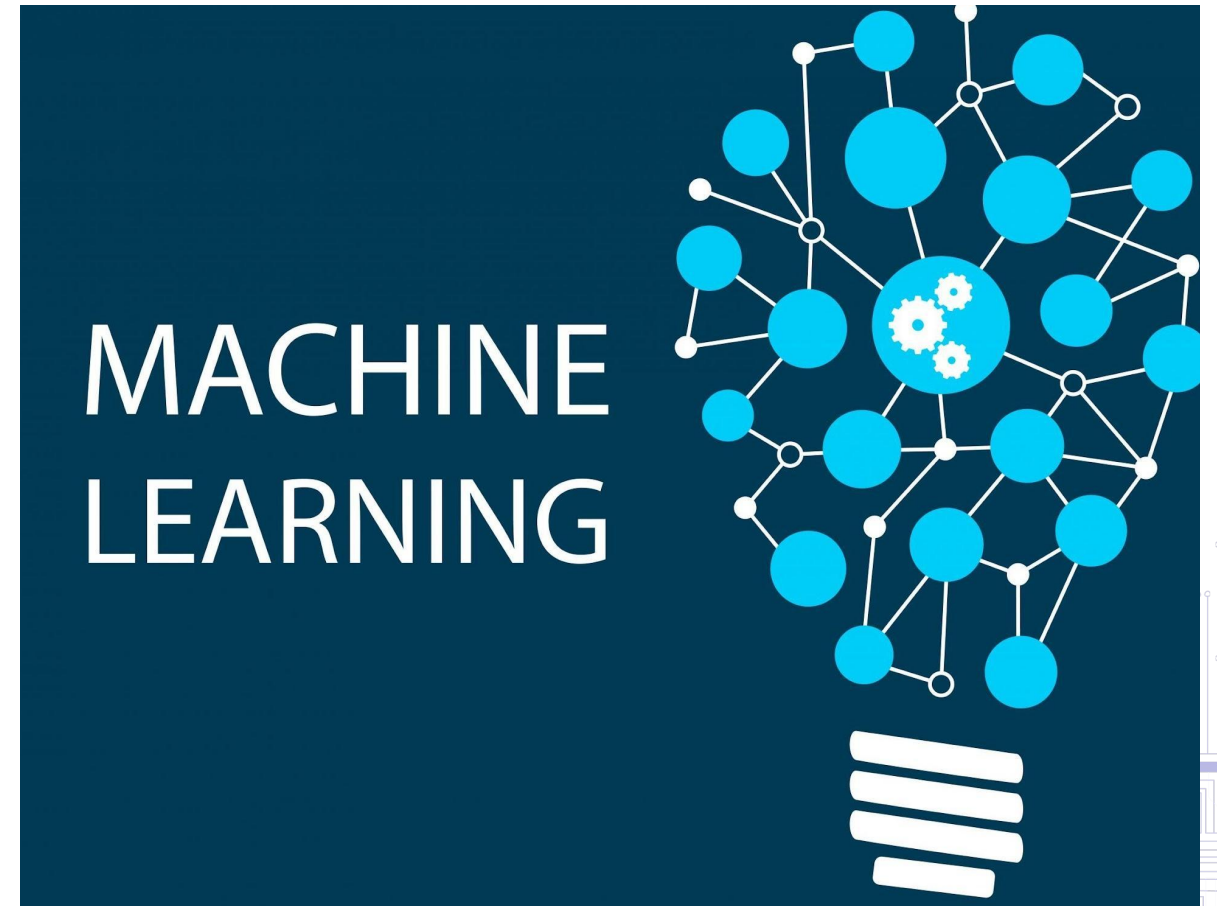
Машинное обучение (ML)

Данные (Data) -> Знания (Knowledge)

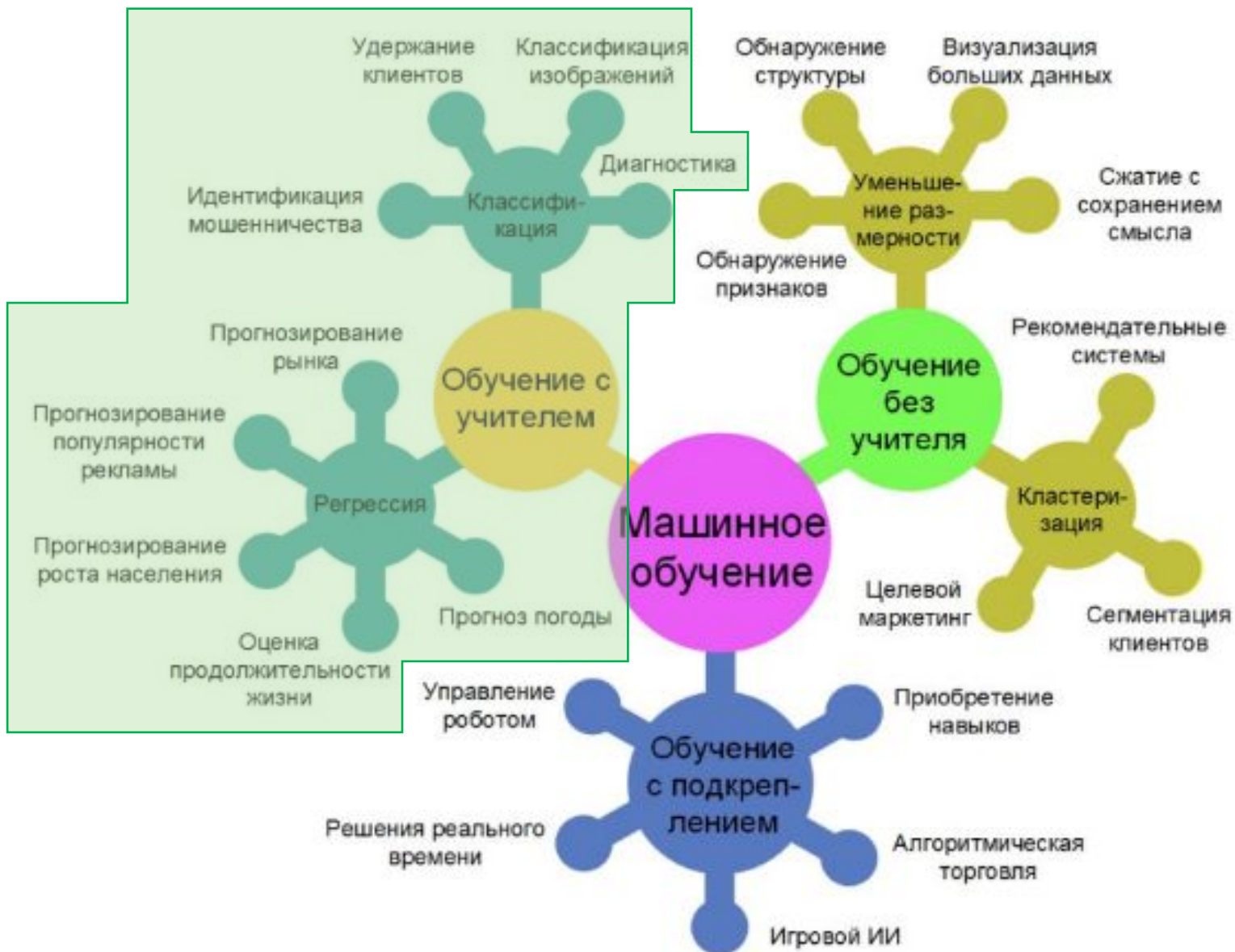
Машинное обучение (machine learning) — это наука, изучающая способы извлечения закономерностей из ограниченного количества примеров.

Особенность – система сама может обучаться на данных и делать выводы/прогнозы

Построенная модель зависит от поставленной задачи и от данных, на которых обучалась



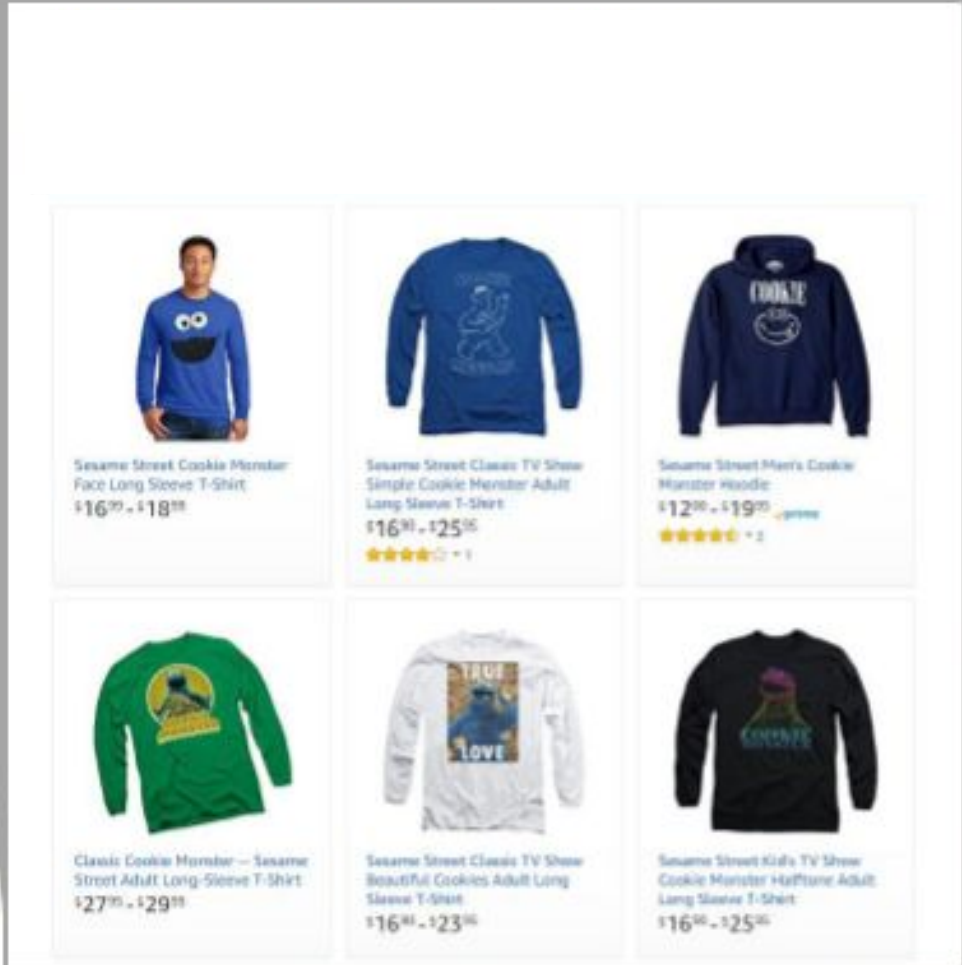
Задачи машинного обучения



Задачи ML



ПЕРЕДОВАЯ
ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Business/ML Problem	Description	Example
Ranking	Помощь пользователям в поиске наиболее релевантной вещи	
Recommendation	Предоставление пользователям то, что они могут быть наиболее заинтересованы в	
Classification	Выяснение того, что что-то такое	
Regression	Прогнозирование численного значения	
Clustering	Сгруппировав похожие объекты вместе	
Anomaly Detection	Поиск необычных вещей	

Задачи ML

Business/ML Problem	Description
Ranking	Помощь пользователям в поиске наиболее релевантной вещи
Recommendation	Предоставление пользователям того, в чем они могут быть наиболее заинтересованы
Classification	Выяснение того, что что-то такое
Regression	Прогнозирование численного значения вещи
Clustering	Сгруппировать похожие объекты вместе
Anomaly Detection	Поиск необычных вещей

Example

Recommendations across the website

Deals recommended for you [See all deals](#)



\$7.00 - \$147.90
Ends in 03:25:54



\$79.99
~~\$189.99~~
Ends in 03:25:54



\$8.99 - \$37.49
Ends in 03:20:55

\$4
Ends in 03:20:55

Amazon's Choice

Amazon's Choice



Panasonic RP-HJE120-PPK In-Ear Stereo Earphones
by Panasonic

\$8.18 | FREE One-Day

Get it by Tomorrow, Apr 24

FREE One-Day Shipping on qualifying orders over \$35

More Buying Choices

\$7.99 (\$7 now offers)

[See newer model of this item](#)

Задачи ML

Business/ML Problem

Description

Example

Ranking

Помощь пользователям в поиске наиболее релевантной вещи

Product classification for our catalog

Recommendation

Предоставление пользователям того, в чем они могут быть наиболее заинтересованы



High-Low Dress



Straight Dress

Classification

Выяснение того, что что-то такое

Regression

Прогнозирование численного значения

Clustering

Сгруппировать похожие объекты вместе



Striped Skirt



Graphic Shirt

Anomaly Detection

Поиск необычных вещей

Задачи ML



ПЕРЕДОВАЯ
ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Business/ML Problem

Description

Example

Ranking

Помощь пользователям в поиске наиболее релевантной вещи

Recommendation

Предоставление пользователям того, в чем они могут быть наиболее заинтересованы

Classification

Выяснение того, что что-то такое

Regression

Прогнозирование численного значения

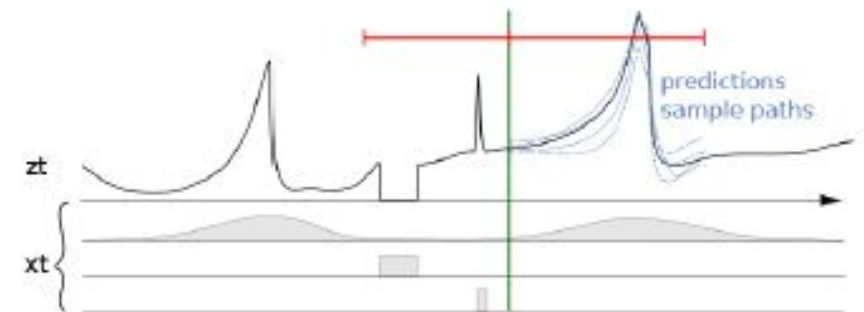
Clustering

Сгруппировать похожие объекты вместе

Anomaly Detection

Поиск необычных вещей

Predicting sales for specific ASINs



Задачи ML



ПЕРЕДОВАЯ
ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Business/ML Problem

Description

Example

Ranking

Помощь пользователям в поиске наиболее релевантной вещи

Recommendation

Предоставление пользователям того, в чем они могут быть наиболее заинтересованы

Classification

Выяснение того, что что-то такое

Regression

Прогнозирование численного значения

Clustering

Сгруппировав похожие объекты вместе

Anomaly Detection

Поиск необычных вещей

Close-matching for near-duplicates



Sheriff Walt Longmire Robert Taylor Trench Coat
by MRASHIONS

\$109.00 - \$160.00



Sheriff Walt Longmire Robert Taylor Trench Coat
by Sparrow

\$154.00

FREE Shipping on eligible orders



Robert Taylor Longmire Sheriff Walt Trench Coat
by MRASHIONS

\$175.00

FREE Shipping on eligible orders

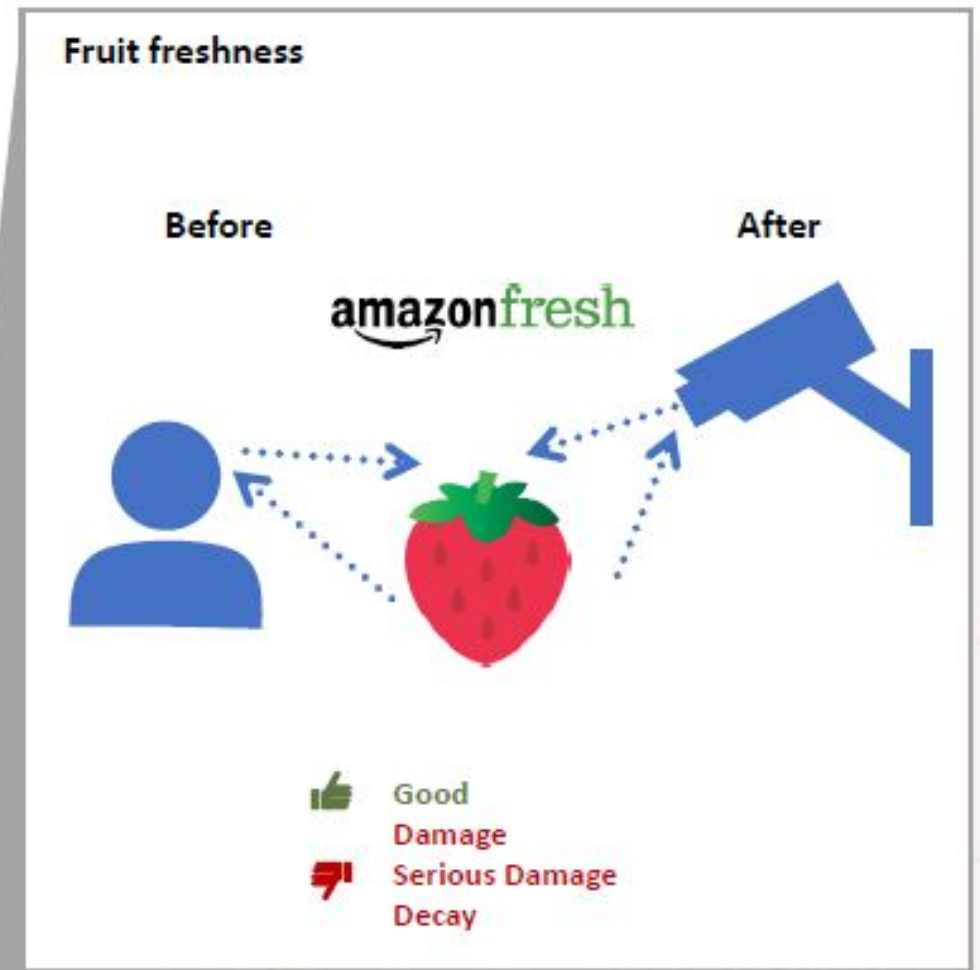
Задачи ML



ПЕРЕДОВАЯ
ИНЖЕНЕРНАЯ ШКОЛА
УНИВЕРСИТЕТА ИННОПОЛИС

Business/ML Problem	Description
Ranking	Помощь пользователям в поиске наиболее релевантной вещи
Recommendation	Предоставление пользователям то, что они могут быть наиболее заинтересованы в
Classification	Выяснение того, что что-то такое
Regression	Прогнозирование численного значения вещи
Clustering	Сгруппировав похожие объекты вместе
Anomaly Detection	Поиск необычных вещей

Example



Примеры ML для классификации

Кредитный скоринг

Объект - кредитная заявка.

Классы - bad (просрочка 90+ в 1-ый год) или good

Примеры признаков: пол, семейное положение, наличие телефона, место проживания, профессия, работодатель, образование, должность, возраст, зарплата, стаж работы, сумма прошлых кредитов, размер просрочки и др.

Особенности задачи: дисбаланс классов (не равное число good, bad)

Постановка диагноза

Объект - пациент в определенный момент времени

Классы - диагноз

Примеры признаков: пол, возраст, головная боль, слабость, пульс, артериальное давление, содержание гемоглобина, другие показатели

Особенности задачи: нужен интерпретируемый алгоритм (*люди могут понять причину, по которой конкретная модель сделала прогноз*), обычно много пропусков

Мониторинг абонентов оператора

Объект - абонент в определенный момент времени.

Классы - уйдет или не уйдет в следующем месяце

Примеры признаков: тарифный план, регион проживания, длительность разговоров, смс, частота оплаты, корпоративный клиент, включение услуг.

Особенности задачи: признаки необходимо вычислять по сырым данным (извлечение информации), большие данные

Анализ изображений и видео

Объект - изображение или видеопоследовательность

Классы - Собаки/Кошки (или решение объехать/не объехать), может быть больше двух

Признаки изображения: цвет, ориентация и размер, содержание, текстура, градиенты

Признаки видео: движение, поворот, зум, световые условия, семантическая и оптическая сегментация

Методы: Используются глубокие нейросетевые архитектуры (Deep Learning)

Этапы построения модели ML



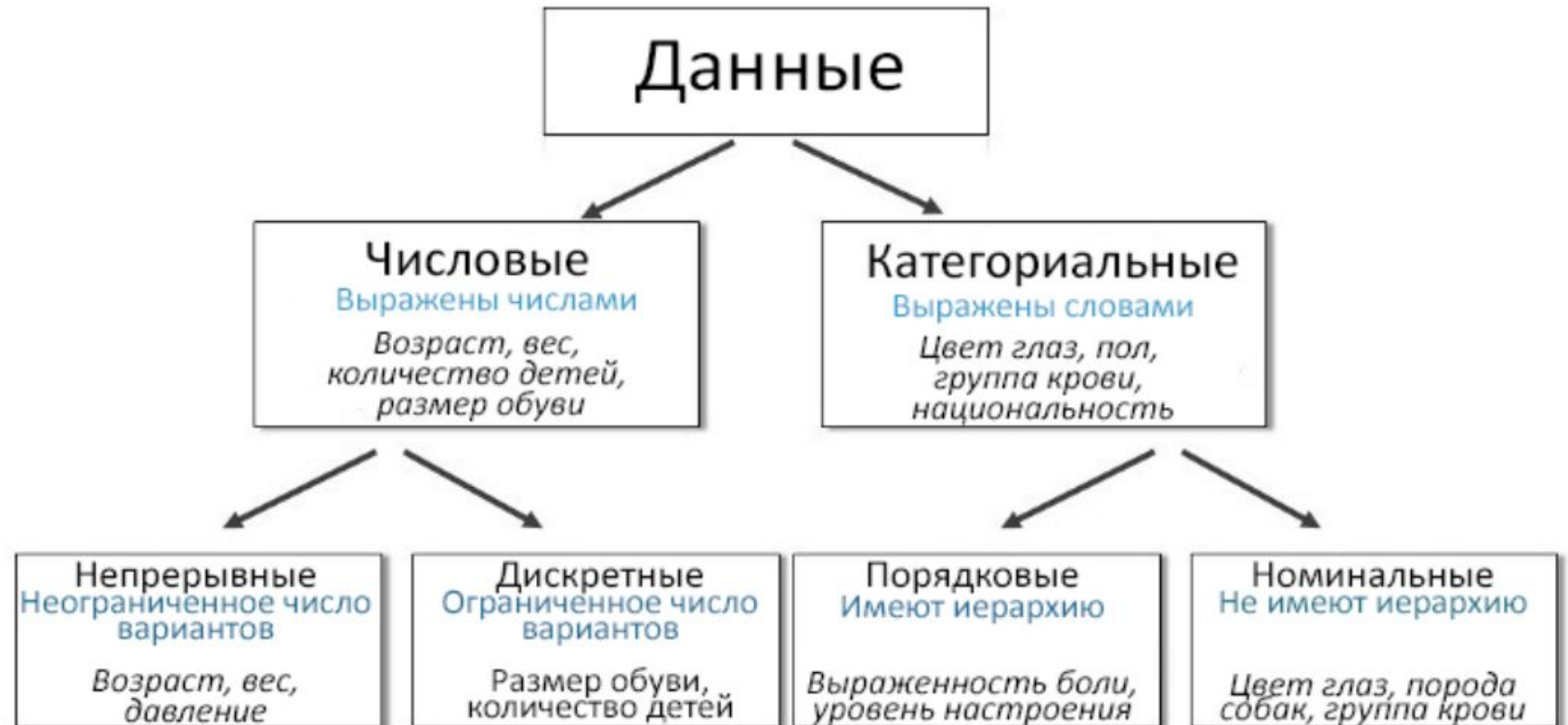
Постановка задачи машинного обучения

Обучение с учителем

X - множество объектов (независимые признаки, предикторы, фичи)

Y - множество ответов (целевой признак, таргет, лейбл)

Данные X и Y - репрезентативная выборка, то есть она должна быть iid (independent and identically distributed)



Регрессия и классификация

id клиента	возраст	образование	...	кол-во детей	купил продукт или не купил	на какую сумму купил?
10001	23	среднее		1	1	8760
10002	45	высшее		0	0	0
10003	56	среднее		2	0	0
10004	32	кандидат наук		3	1	5643
10005	18	среднее		1	1	3421

Если признак целевой категориальный или дискретный - то задача **классификации**

Если целевой признак непрерывный - то задача **регрессии**

Регрессия и классификация

Если целевой категориальный признак принимает только 2 значения, то есть признак бинарный, то и классификация **БИНАРНАЯ**

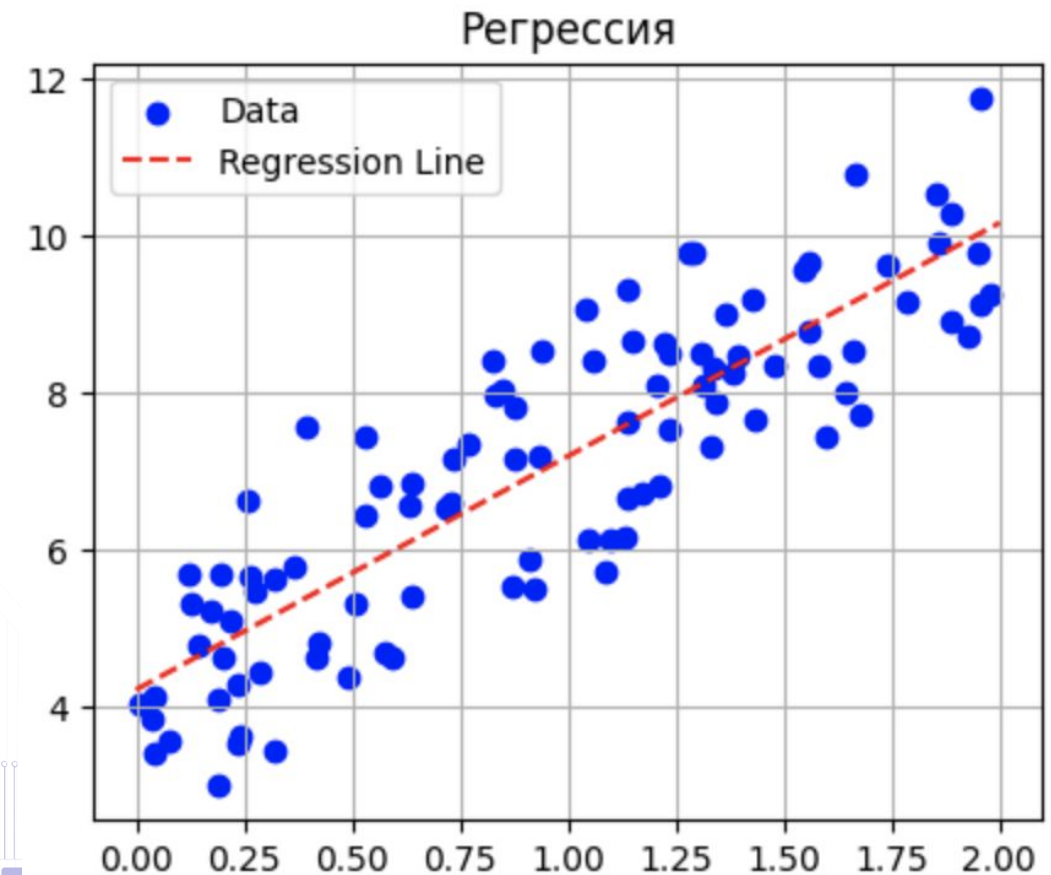
Если классов более 2, то **МНОГОКЛАССОВАЯ**

- купит продукт или не купит
- установит игру/не установит
- вернет кредит/ не вернёт
- клиент купит товары категории: образование, Детские, Для собак или продукты
- оформит кредит или нет



Если признак целевой непрерывный - **РЕГРЕССИЯ**

- на какую сумму клиент совершит покупку
- сколько товаров будет куплено в ближайшие 7 дней
- какую прибыль принесет магазин за 1 год
- какой будет убыток по страховке
- какую долю кредита клиент не вернет



Линейная регрессия

У нас есть данные, мы построили график и увидели следующее распределение

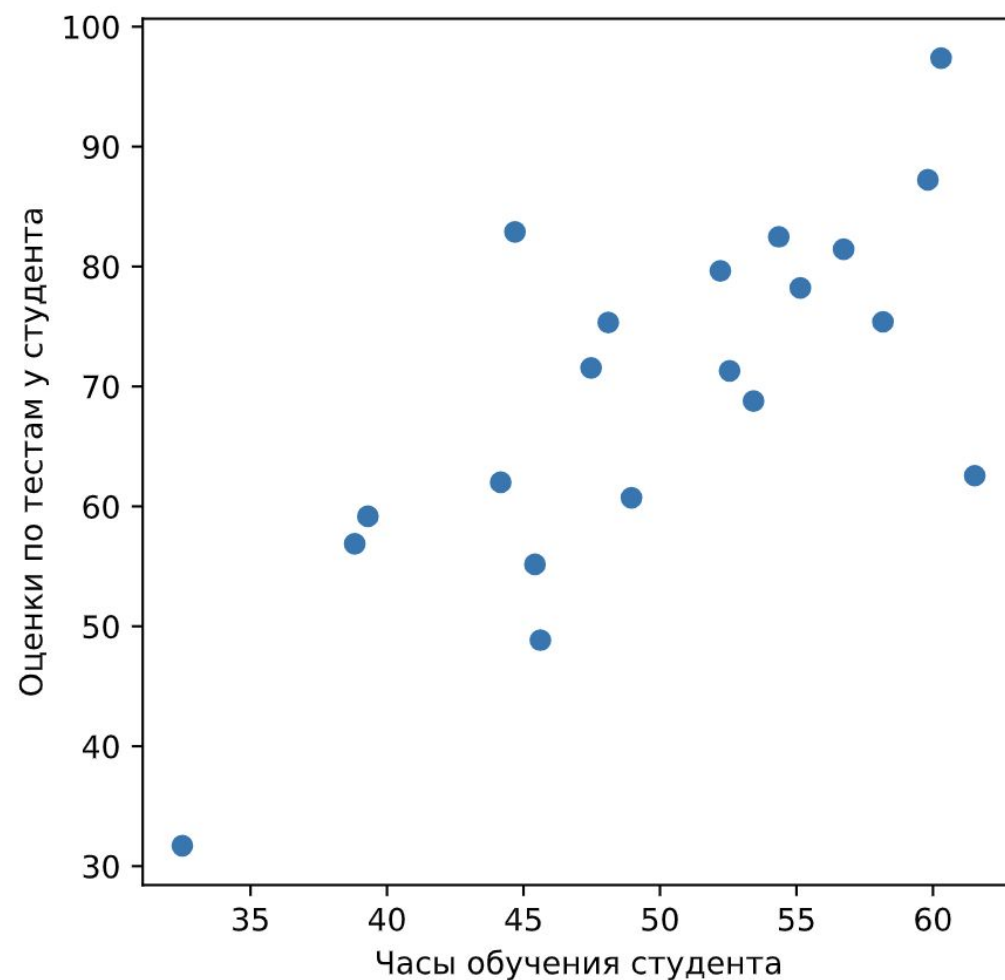
Что делать?

Кажется, что в данных прослеживается закономерность.

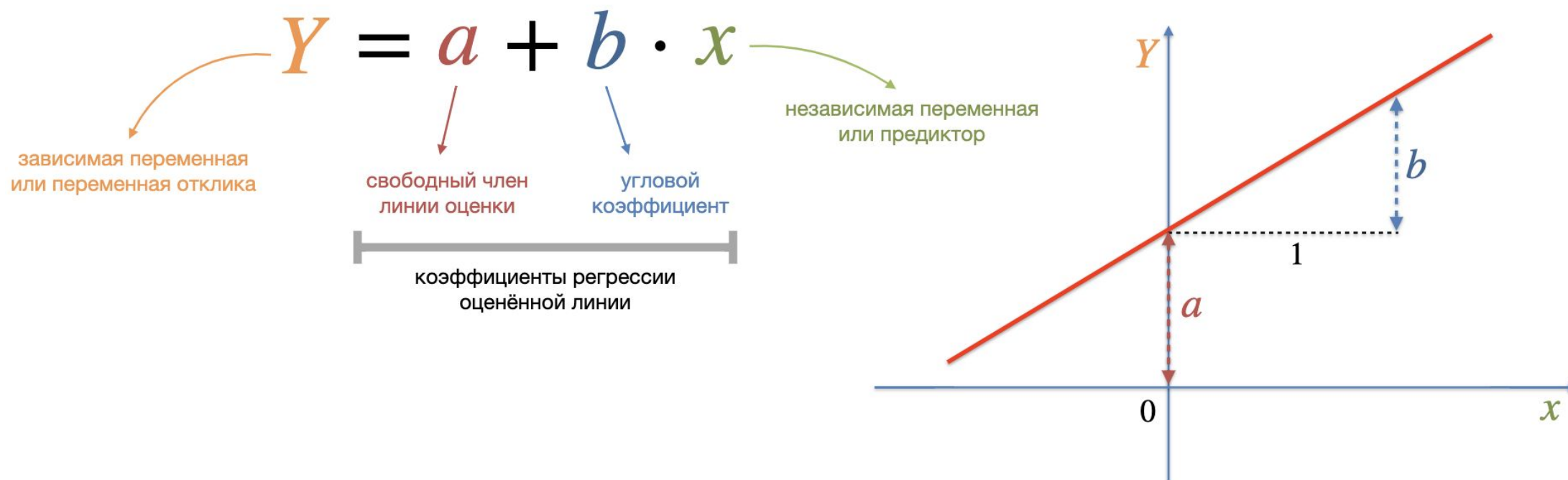
И она похожа на прямую

мы знаем, что уравнение прямой на плоскости

$$y = a + b * x$$

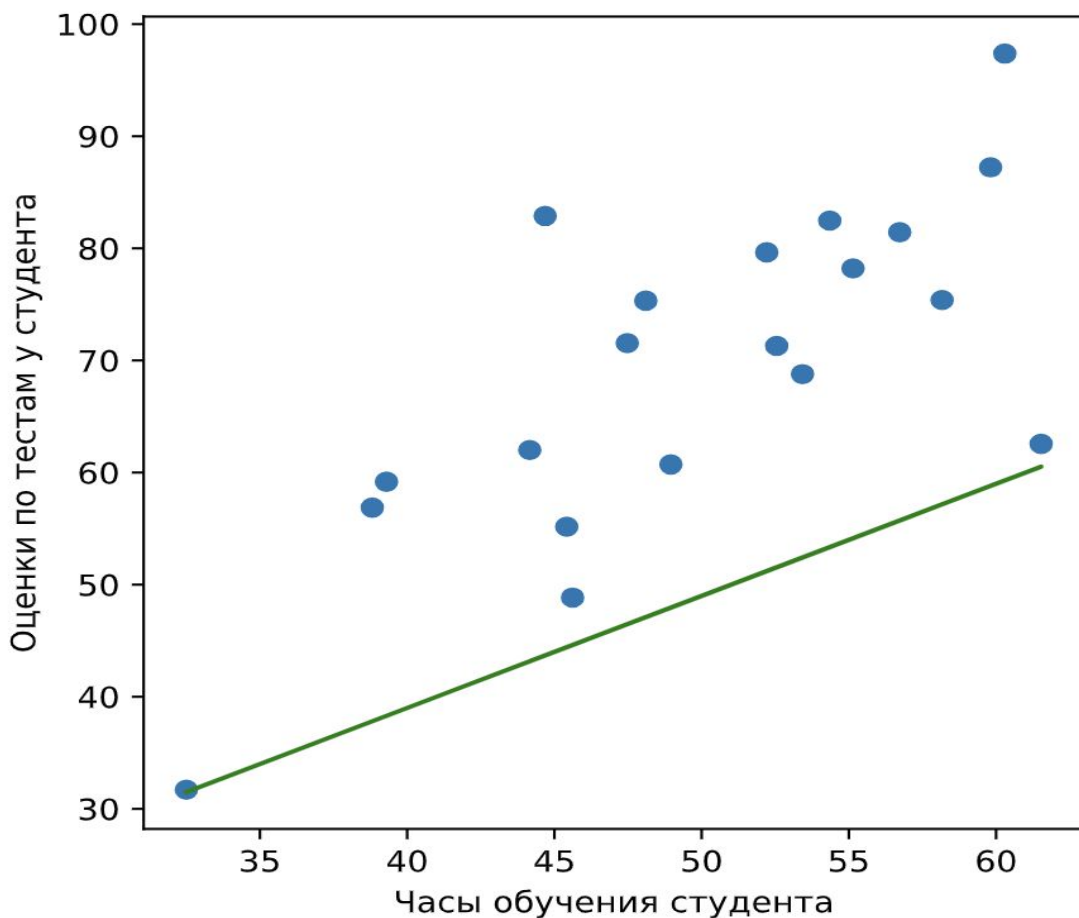


Линейная регрессия



Уравнение прямой - это и есть линейная регрессия, в данном случае парная линейная регрессия

Линейная регрессия

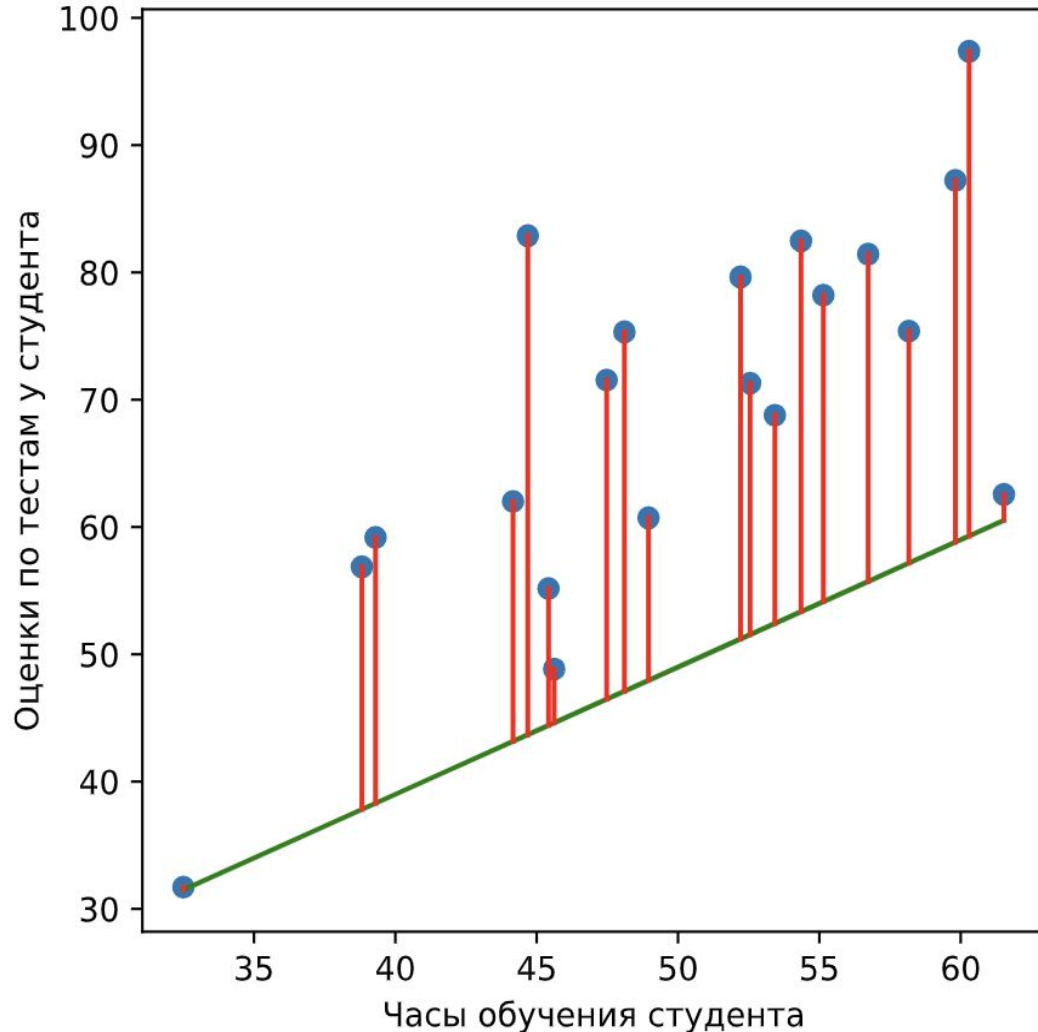


Отлично, построили прямую

Но кажется, что-то не совсем то, что нам хотелось бы

А что нам не нравится?

Линейная регрессия



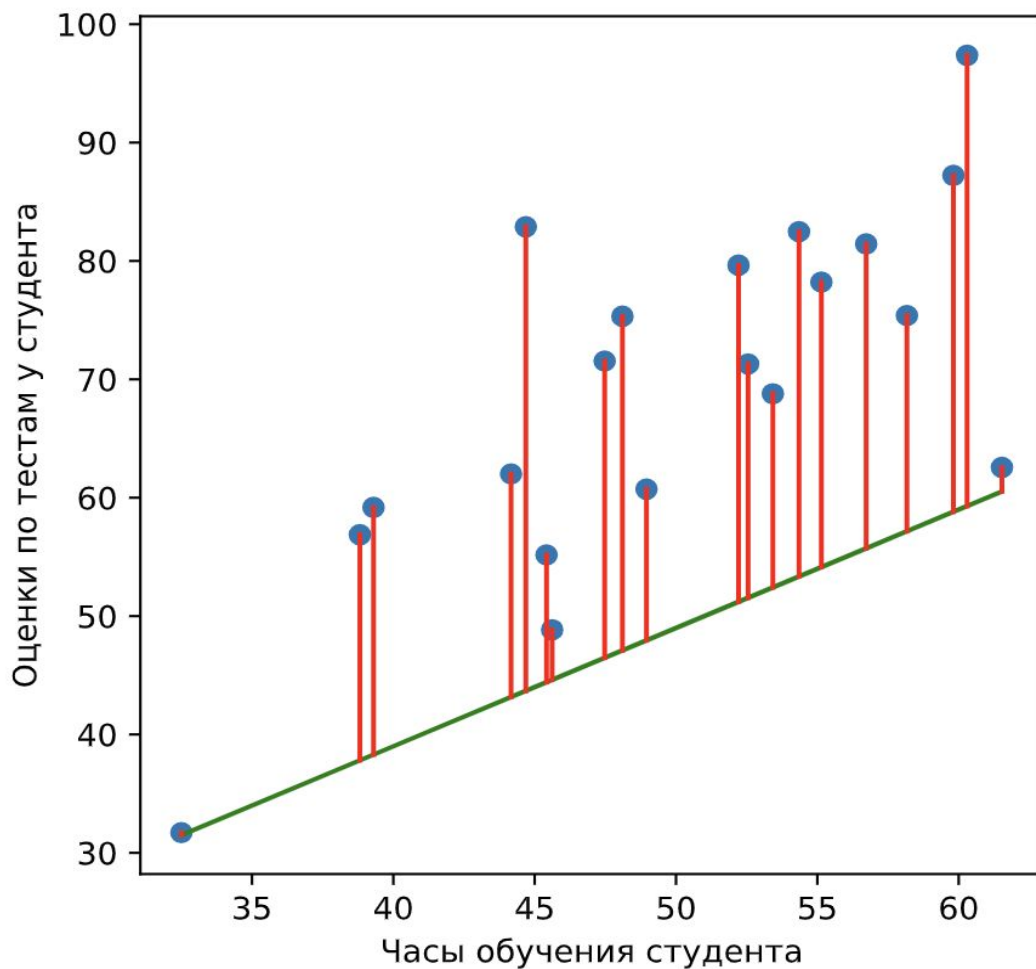
Попробуем оценить, на сколько наша прямая “не попадает” в наши точки

Посчитаем разности между фактическими данными и точками на прямой,

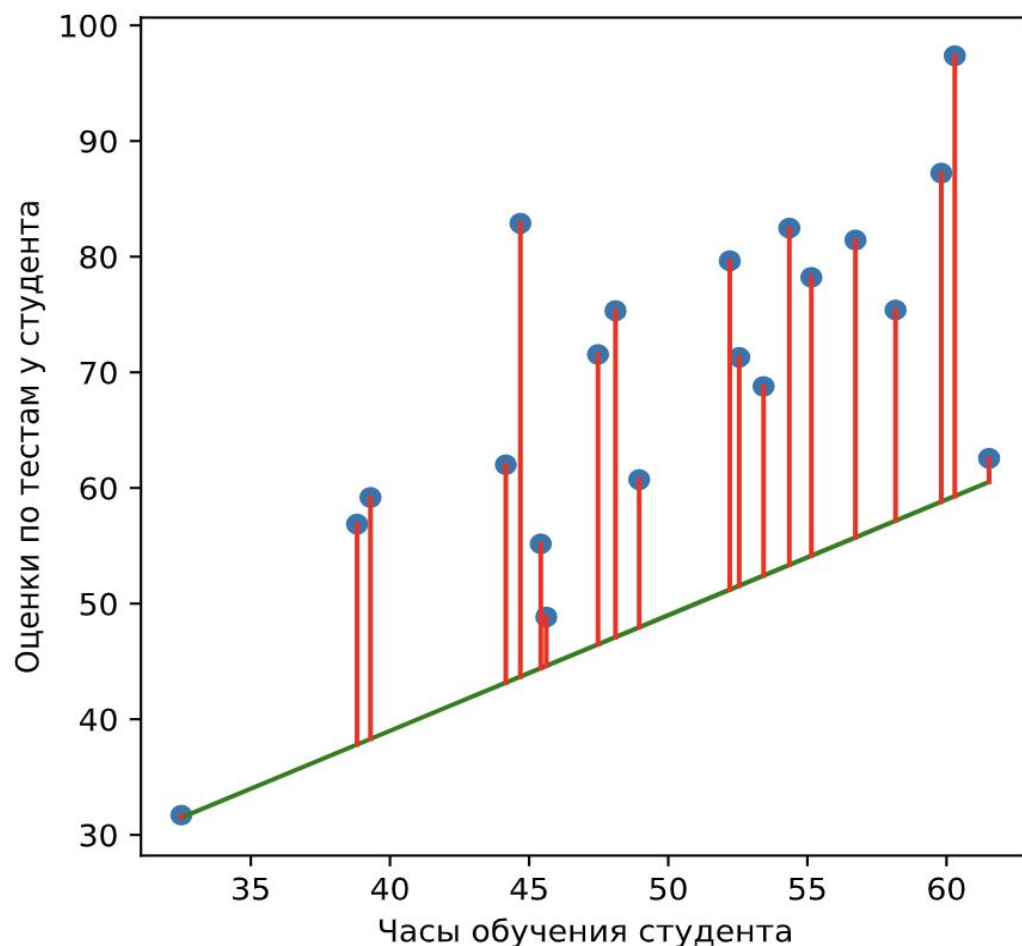
то есть посчитаем абсолютные суммы длин красных отрезков, получим 20.47

Кажется, можно лучше

Линейная регрессия



Или можем подсчитать сумму квадратов отклонения между точками и нашей прямой

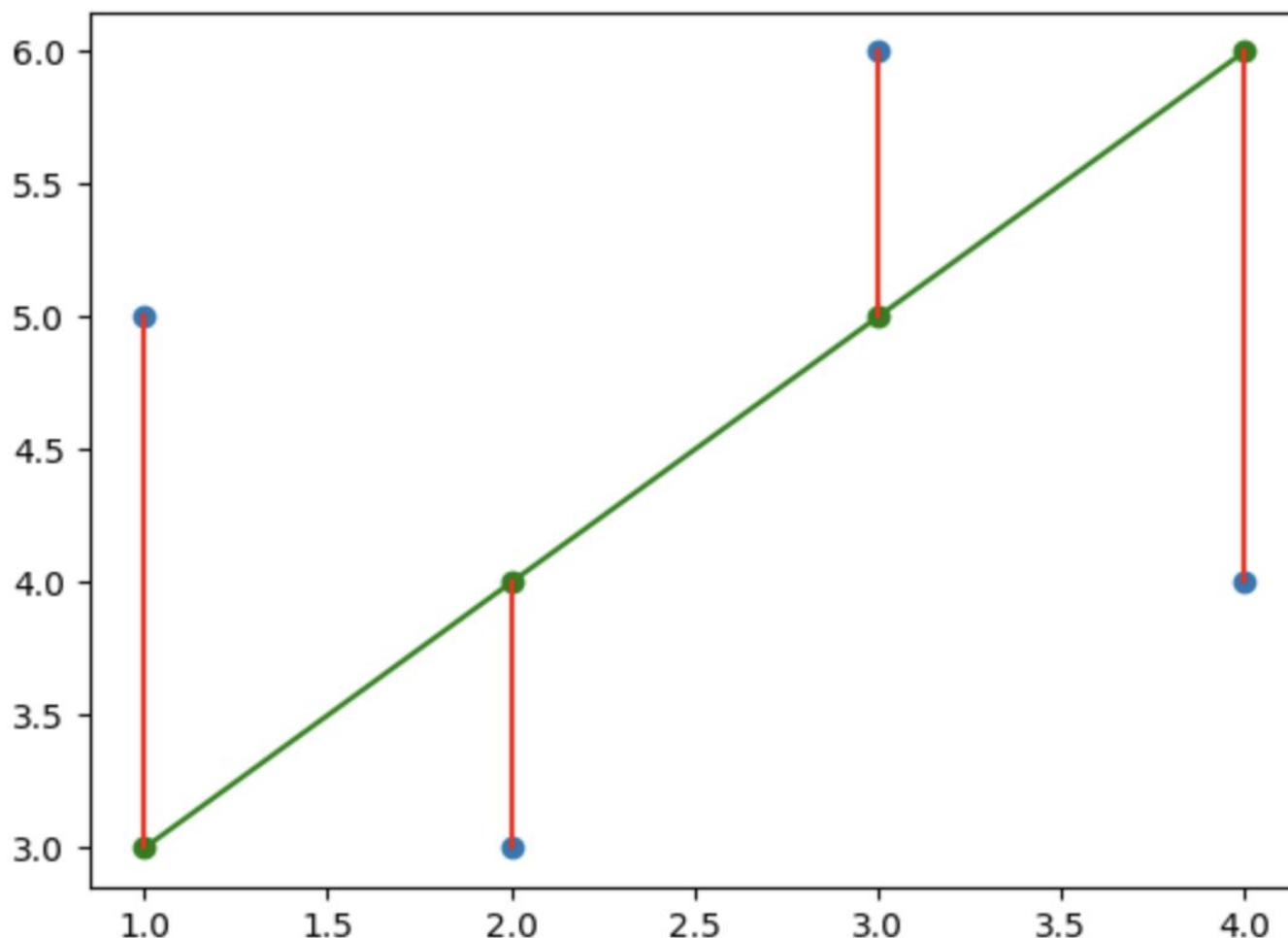


То есть можно считать сумму абсолютных отклонений или сумму квадратов отклонений

Но зачем? можно же просто найти разности?

```
[39] # посчитаем ошибку предсказания MAE и квадратичную ошибку  
      round(mean_absolute_error(y, pred_y),2), round(mean_squared_error(y, pred_y), 2)  
      ↗ (20.47, 528.07)
```

Линейная регрессия



Если будут просто разности, то
сумма ошибок будет равна: $2 - 1 + 1 - 2 = 0$

По такой ошибке мы будто
построили
“идеальный алгоритм”, но это не
так

Линейная регрессия

Необходимо минимизировать сумму квадратов отклонений RSS (Residual Sum of Squares)

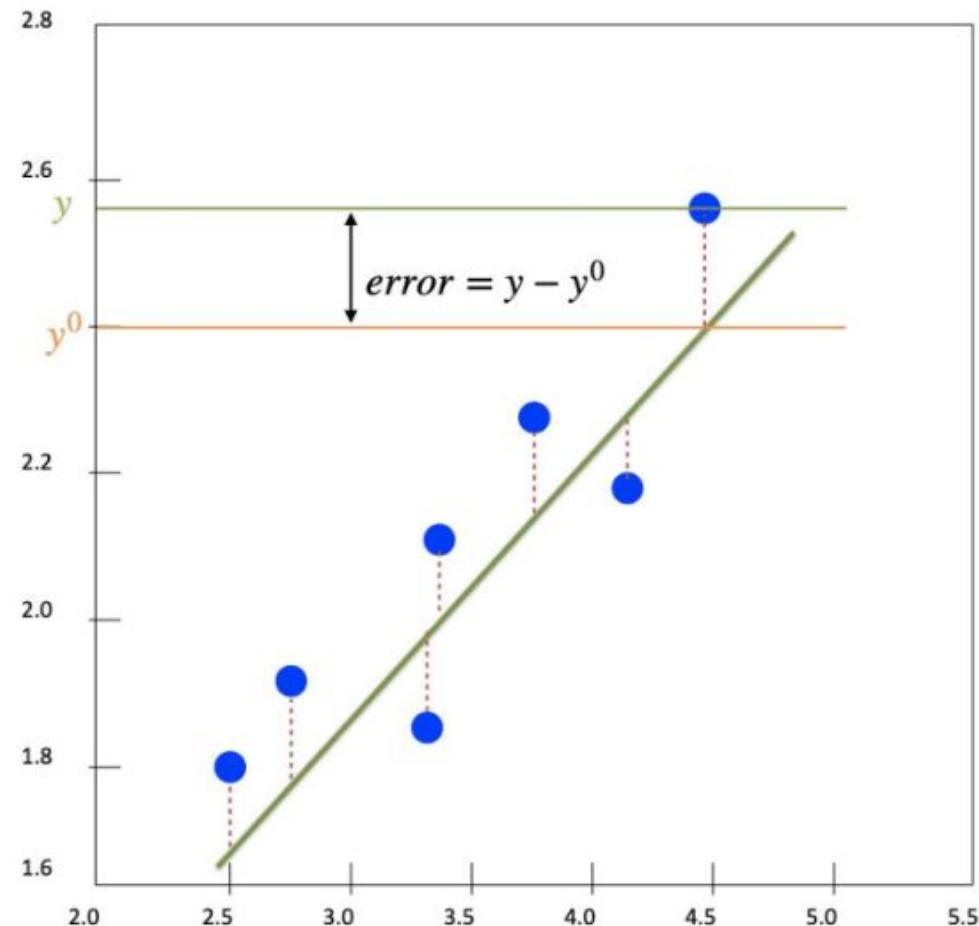
Минимизируемая функция

$$RSS = \sum_i (y_i - (a + bx_i))^2$$

$$\sum_{i=1}^n e_i^2 = RSS \text{ (Residual Sum of Squares)}$$

Для того, чтобы определить прямую, необходимо найти a и b

Нам поможет Метод Наименьших Квадратов МНК (Ordinary Least Squares (OLS))



Линейная регрессия

Метод Наименьших Квадратов (МНК). Аналитическое решение

Минимизируемая функция

$$RSS = \sum_i (y_i - (a + bx_i))^2$$



Результат расчёта

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} \end{cases}$$

Детали расчета (вспоминаем, что такое частные производные):

$$\begin{cases} \frac{\partial RSS}{\partial a} = \sum_i 2(y_i - a - bx_i) = 0 \\ \frac{\partial RSS}{\partial b} = \sum_i 2(y_i - a - bx_i)x_i = 0 \end{cases}$$

$$\begin{cases} \sum_i y_i - na - b \sum_i x_i = 0 \\ \sum_i x_i y_i - a \sum_i x_i - b \sum_i x_i^2 = 0 \end{cases}$$

$$\begin{cases} \bar{y} - a - b\bar{x} = 0 \\ \overline{xy} - a\bar{x} - b\overline{x^2} = 0 \end{cases} \quad \begin{cases} a = \bar{y} - b\bar{x} \\ \overline{xy} - (\bar{y} - b\bar{x})\bar{x} - b\overline{x^2} = 0 \end{cases} \quad \begin{cases} a = \bar{y} - b\bar{x} \\ \overline{xy} - \bar{x}\bar{y} + b[(\bar{x})^2 - \overline{x^2}] = 0 \end{cases}$$

<https://td.chem.msu.ru/uploads/files/courses/special/expmetho ds/statexp/LabLecture03.pdf>

Линейная регрессия

Построим линейную регрессию
сумма всех красных отрезков по модулю
равна 7.88

то есть

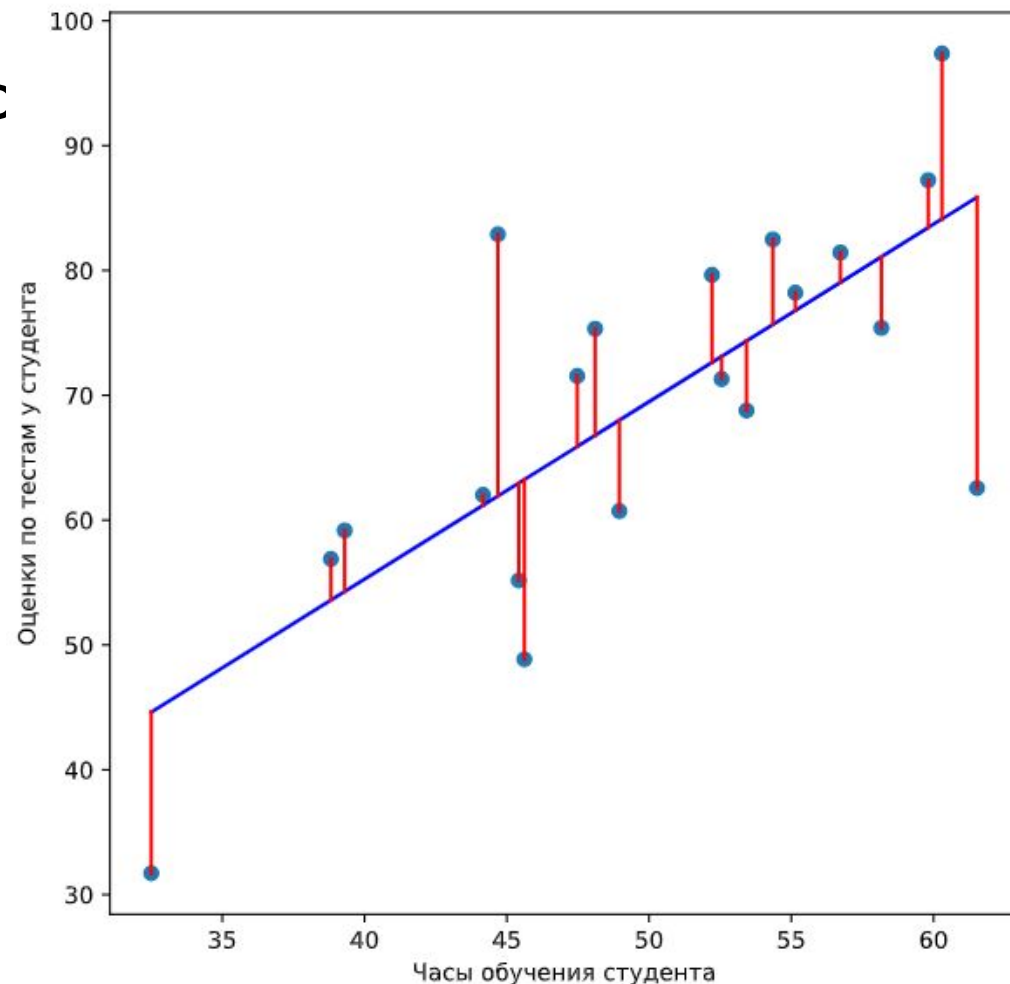
- MAE = 7.88,
- MSE = 98.58,
- RMSE = 9.93

```
round(mean_absolute_error(y, pred_y_m),2), round(mean_squared_error(y, pred_y_m), 2)
```

```
(7.88, 98.58)
```

```
round(mean_squared_error(y, pred_y_m)**(1/2),2)
```

```
9.93
```



Линейная регрессия. Общий случай

Для многомерного случая линейной регрессии, когда у нас есть несколько предикторов (x_1, x_2, \dots, x_p) , модель записывается следующим образом:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

где $(\beta_0, \beta_1, \dots, \beta_p)$ - коэффициенты регрессии, а ε - ошибка.

https://colab.research.google.com/drive/1pOeo9fJr8y6WYSCDi6fwGF_R6d7RtFCc?usp=sharing

Линейная регрессия. Общий случай

Мы хотим найти такие значения $(\beta_0, \beta_1, \dots, \beta_p)$, которые минимизируют сумму квадратов отклонений:

$$S = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}))^2$$

Запишем эту задачу в матричной форме. Обозначим:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Тогда модель можно записать как:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Сумма квадратов отклонений в матричной форме:

$$S = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Для нахождения $\boldsymbol{\beta}$, которое минимизирует S , возьмем производную этой суммы по $\boldsymbol{\beta}$ и приравняем к нулю:

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

Решая это уравнение, получаем:

$$\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}$$

Линейная регрессия. Общий случай

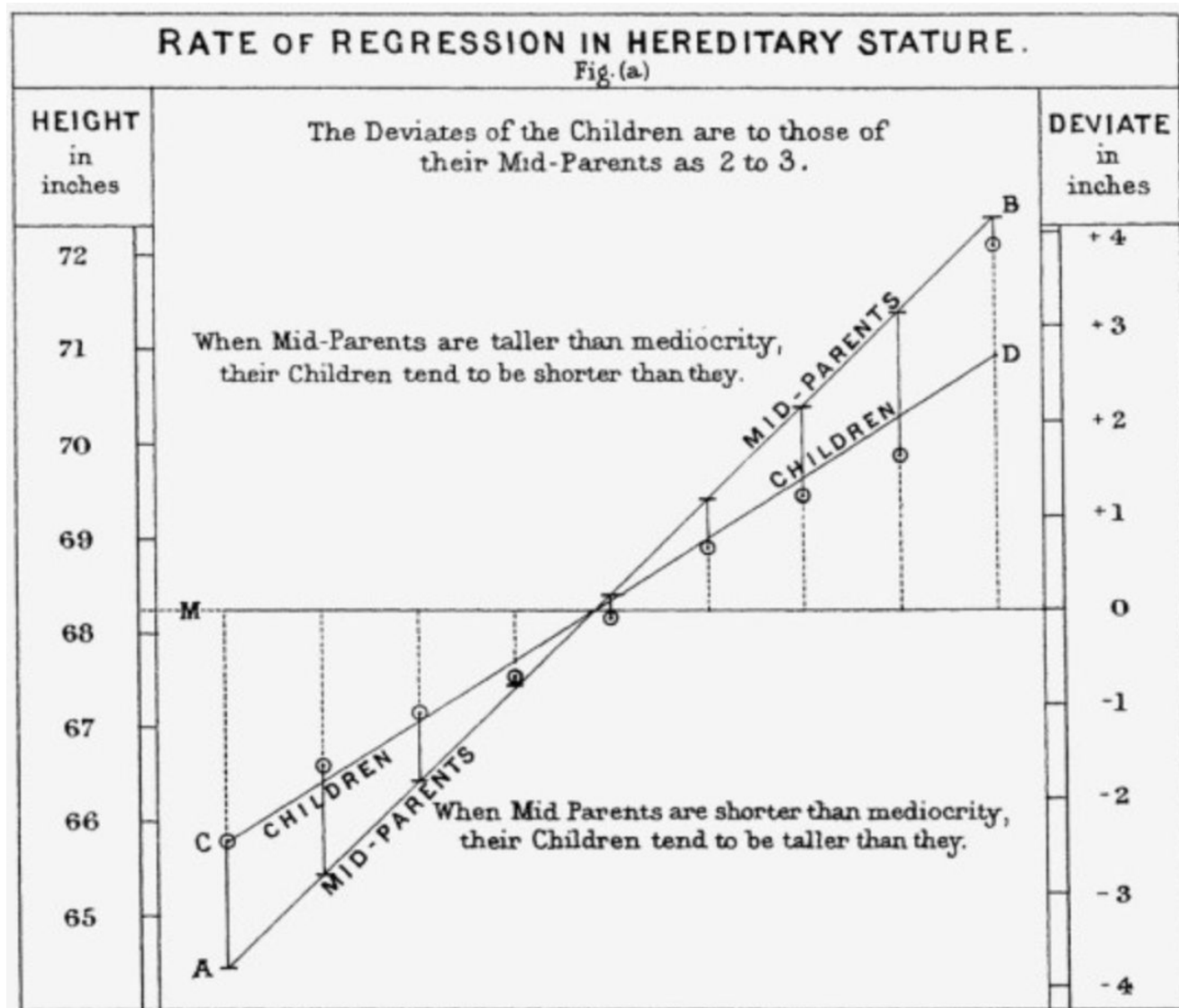
Таким образом, коэффициенты множественной линейной регрессии вычисляются по следующей формуле:

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

где:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Линейная регрессия. История



Почему же регрессия?

В 1886 году Понятие регрессии
ввел сэр Френсис Гальтон,
английский исследователь
широкого профиля.



Передовые
инженерные
школы



МИНОБРНАУКИ
РОССИИ



УНИВЕРСИТЕТ
ИННОПОЛИС



онлайн
университет

Спасибо за внимание