

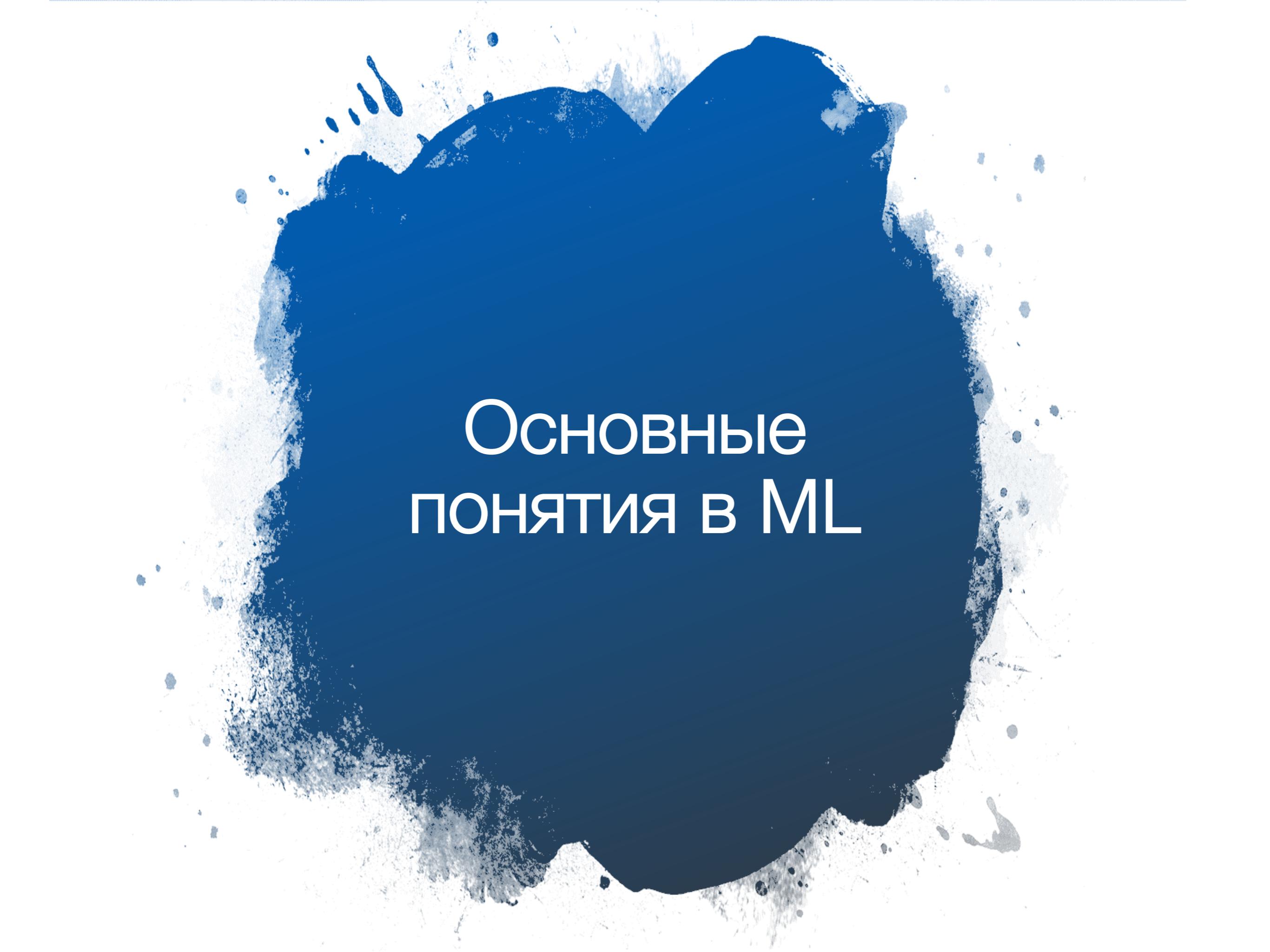


НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

МАШИННОЕ ОБУЧЕНИЕ. ЗАДАЧА РЕГРЕССИИ

Теванян Элен
27.04.2019

Москва 2019



Основные понятия в ML

ПРИМЕР ЗАДАЧИ

- Для улучшения эффективности диспетчерских служб такси важно знать, когда водитель закончит один заказ и будет готов принять следующий.
- Оценка длительности текущей поездки – один из факторов эффективного распределения заказов.
- Как оценить длительность поездки?

ТЕРМИНОЛОГИЯ

- x (**sample**) – объект, для которой хотим делать предсказания
 - Поездки
 - y (**target**) – ответ, целевая переменная, т.е. То, что хотим предсказать
 - Длительность поездки
-
- $(x_i, y_i)_{i=1}^{\ell}$ – обучающая выборка, прецеденты, т.е. все объекты, для которых известны значения целевого признака
 - ℓ – размер выборки.

ПРИЗНАКИ

- Компьютер умеет работать с числовой информацией
- Объекты характеризуются числовой информацией – признаками, факторами, «фичами» (от англ. features)
- m – число признаков
- $x = (x^1, \dots, x^m)$

ПРИЗНАКИ ДЛЯ ЗАДАЧИ

ПРИЗНАКИ ДЛЯ ЗАДАЧИ

- Временные
- Географические
- Погодные
- Маршруты
- Пассажиры

ПРИЗНАКИ ДЛЯ ЗАДАЧИ

- Временные
 - Дата и время посадки
 - Дата и время высадки
- Географические
 - Ширина и долгота места посадки
 - Ширина и долгота места высадки
- Погодные
 - Осадки: дождь, снег, шторм
 - Сила осадков
- Маршруты
 - Наиболее быстрые маршруты
 - Скорость по маршрутам
- Пассажиры
 - Число пассажиров



Наша задача

НАША ЗАДАЧА

- Предсказываем, сколько лайков поставит человек

ДАННЫЕ ДЛЯ ОБУЧЕНИЯ

- Если все совсем просто, то у нас есть таблица

объекты

Person	Age	Nu of Friends	Nu of Likes
Катя Петрова	18	154	4
Лера Жданович	19	263	0
Витя Виктор	18	747	12

признаки

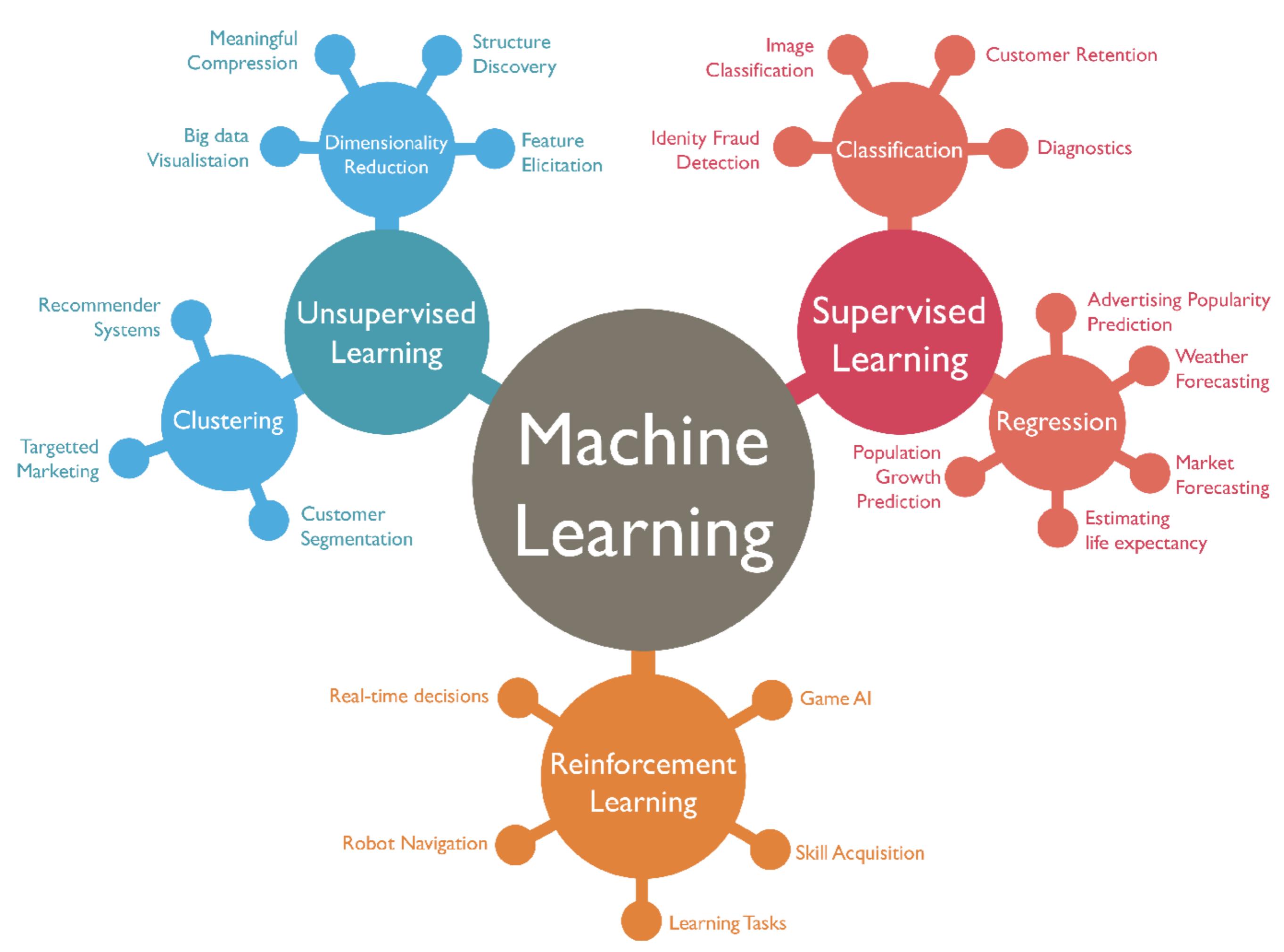
целевая переменная

БУДЕМ УЧИТЬ МОДЕЛЬ

- $a(x)$ – алгоритм/модель/формула
- Это функция, предсказывающая ответ для любого объекта

БУДЕМ УЧИТЬ МОДЕЛЬ

- $a(x)$ – алгоритм/модель/формула
- Это функция, предсказывающая ответ для любого объекта
- Передаем модели возраст и количество друзей человека, а она говорит, сколько лайков поставит



КЛАССЫ ЗАДАЧ

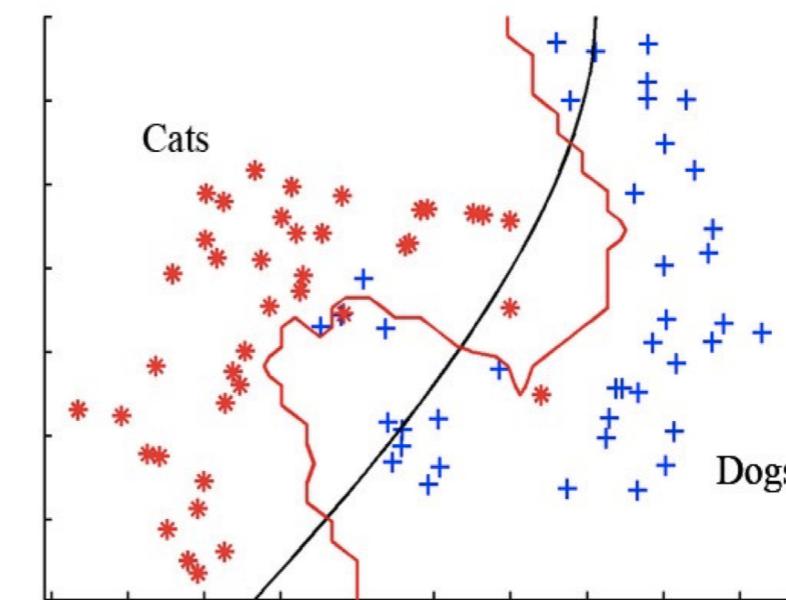
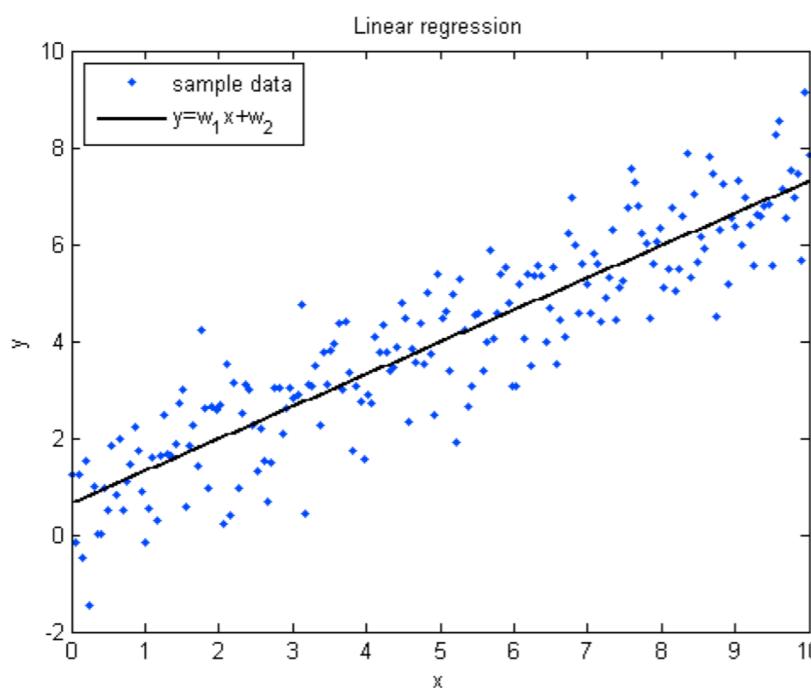
Регрессия

Классификация

Обучение с учителем

Вещественная
целевая переменная

Конечное множество
ответов



РЕГРЕССИЯ. ОПРЕДЕЛЕНИЕ

Регрессия – это способ объяснить зависимость переменной через одну или набор других.

Y – это переменная, которую планируют объяснять. Ее называют зависимой.

X_1, \dots, X_n – это переменные, через которую планируют объяснить Y . Их называют независимыми/регрессорами/предикторами.

РЕГРЕССИЯ. ДЛЯ ЧЕГО?

1. Объяснить разброс/неоднородность зависимой переменной через независимые
2. Предсказать значения зависимой переменной через независимые.
3. Определить вклад каждой из зависимых переменных

Примеры задачи регрессии

СТОИМОСТЬ ДОМА



We use cookies on kaggle to deliver our services, analyze web traffic, and improve your experience on the site. By using kaggle, you agree to our use of cookies.

Got it

Learn more

kaggle

Search



Competitions

Datasets

Kernels

Discussion

Learn

...

Sign In

Featured Prediction Competition

Sberbank Russian Housing Market

Can you predict realty price fluctuations in Russia's volatile economy?

 Sberbank · 3,274 teams · 2 years ago

\$25,000 Prize Money

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Join Competition](#)

Overview	
Description	Housing costs demand a significant investment from both consumers and developers. And when it comes to planning a budget—whether personal or corporate—the last thing anyone needs is uncertainty about one of their biggest expenses. Sberbank , Russia's oldest and largest bank, helps their customers by making predictions about realty prices so renters, developers, and lenders are more confident when they sign a lease or purchase a building.
Evaluation	
Prizes	
Timeline	

ЗП ПО ОПИСАНИЮ ВАКАНСИИ

The screenshot shows the HeadHunter (hh.ru) website. At the top left is the hh logo. To its right is a search bar with the placeholder text "Я ищу...". Further to the right is a dropdown menu labeled "Вакансии". Below the header is a navigation bar with links: "Ищу работу", "Ищу сотрудников", "Помощь", "Компании", and "Проекты".

Маркетолог-аналитик

от 70 000 руб. на руки

АО Телеофис

● Нагатинская, Москва, 1-й Нагатинский проезд, 2с34



Откликнуться



Требуемый опыт работы: 1–3 года

Полная занятость, полный день

TELEOFIS – российская производственная компания, предлагающая широкий ассортимент беспроводного оборудования для построения систем диспетчеризации, контроля и промышленной связи, приглашает аналитика в маркетинг.

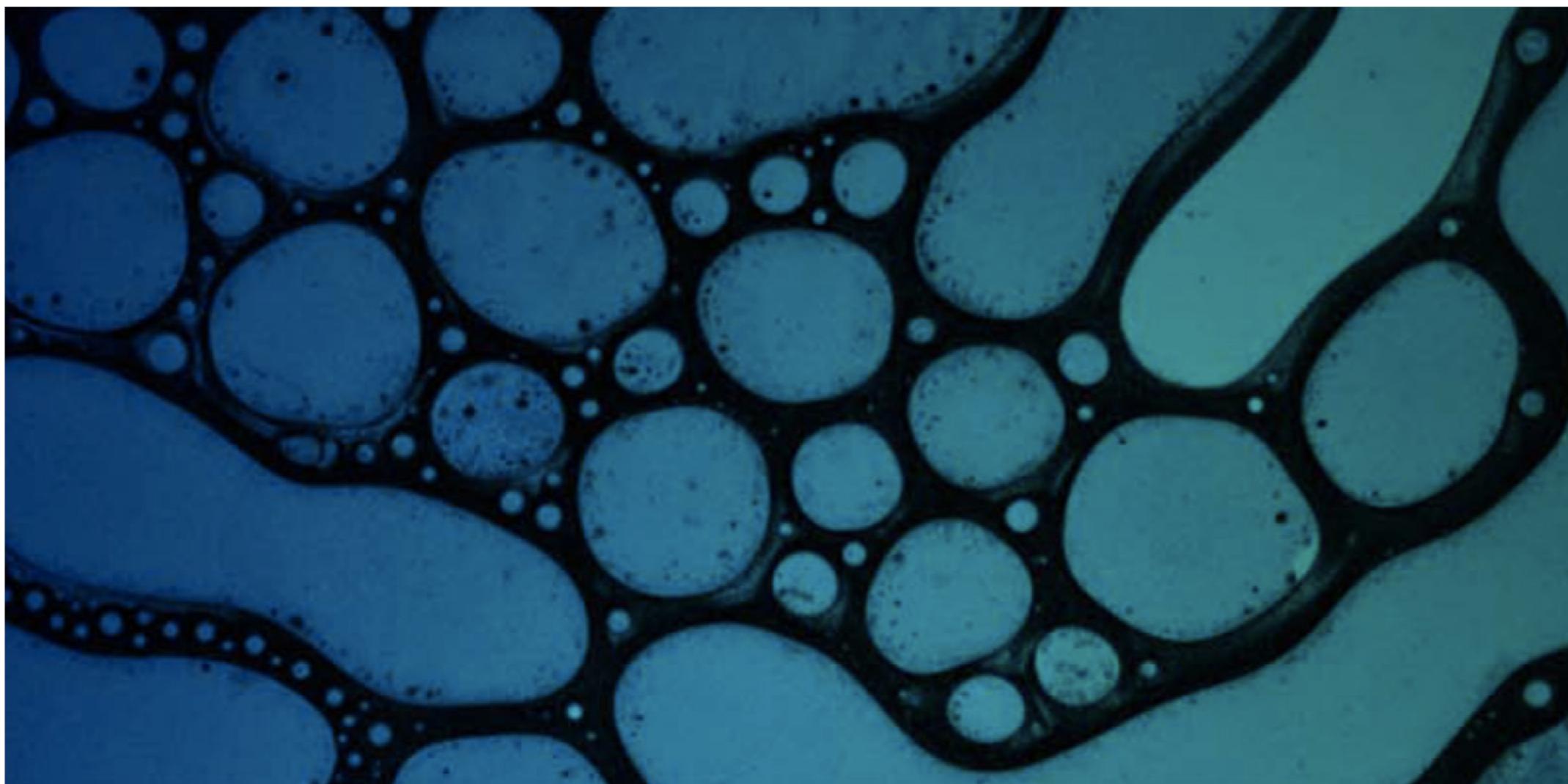
Мы предлагаем:

- Всё по ТК РФ: полностью "белая" заработка плата, оформление, отпуск, больничный;
- ЗП: оклад 70 000 руб. + квартальная премия;
- Оплату обучения, корпоративные тренинги и внешние курсы;
- График 5/2 с 9:00 до 18:00, только офис;

СПРОС НА ТОВАР В БЛИЖАЙШУЮ НЕДЕЛЮ



УРОВЕНЬ ЭКСПРЕССИИ ГЕНОВ



СТОИМОСТЬ СТРАХОВКИ

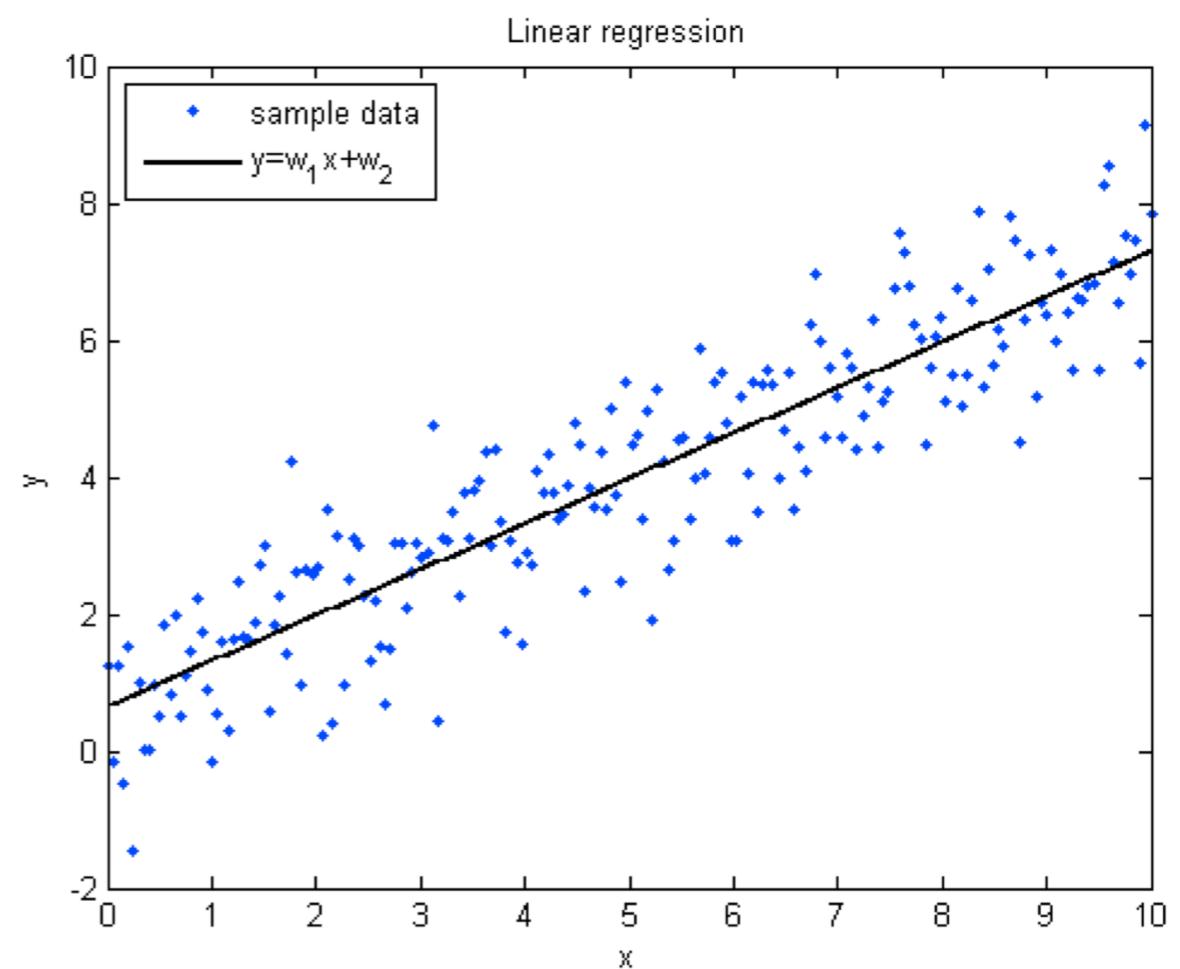


ОБЪЕМ ПОТРЕБЛЕНИЯ ЭЛЕКТРОЭНЕРГИИ



РЕГРЕССИЯ

- Есть обучающая выборка, в которой объекты представлены признаковым описанием и есть значение целевой переменной
- Целевое значение: любое действительное число
- Задача: найти алгоритм, который спрогнозирует для любого объекта его целевое значение



Метрики

MAE

- Средняя абсолютная ошибка

$$MAE = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - \hat{y}_i|$$

MAPE

- Средняя абсолютная процентная ошибка

$$MAPE = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{|y_i - \hat{y}_i|}{y_i}$$

MSE

- Среднеквадратическая ошибка

$$MSE = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - \hat{y}_i)^2$$

RMSE

- Корень из среднеквадратической ошибки

$$RMSE = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - \hat{y}_i)^2}$$

R²

- Доля информации, объясненная моделью, в общем объеме информации выборки

$$R^2 = 1 - \frac{\sum_{i=1}^{\ell} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{\ell} (y_i - \bar{y})^2}$$

$$\bar{y} = \frac{1}{\ell} \sum_{i=1}^{\ell} y_i$$

- Хорошо интерпретируемая величина

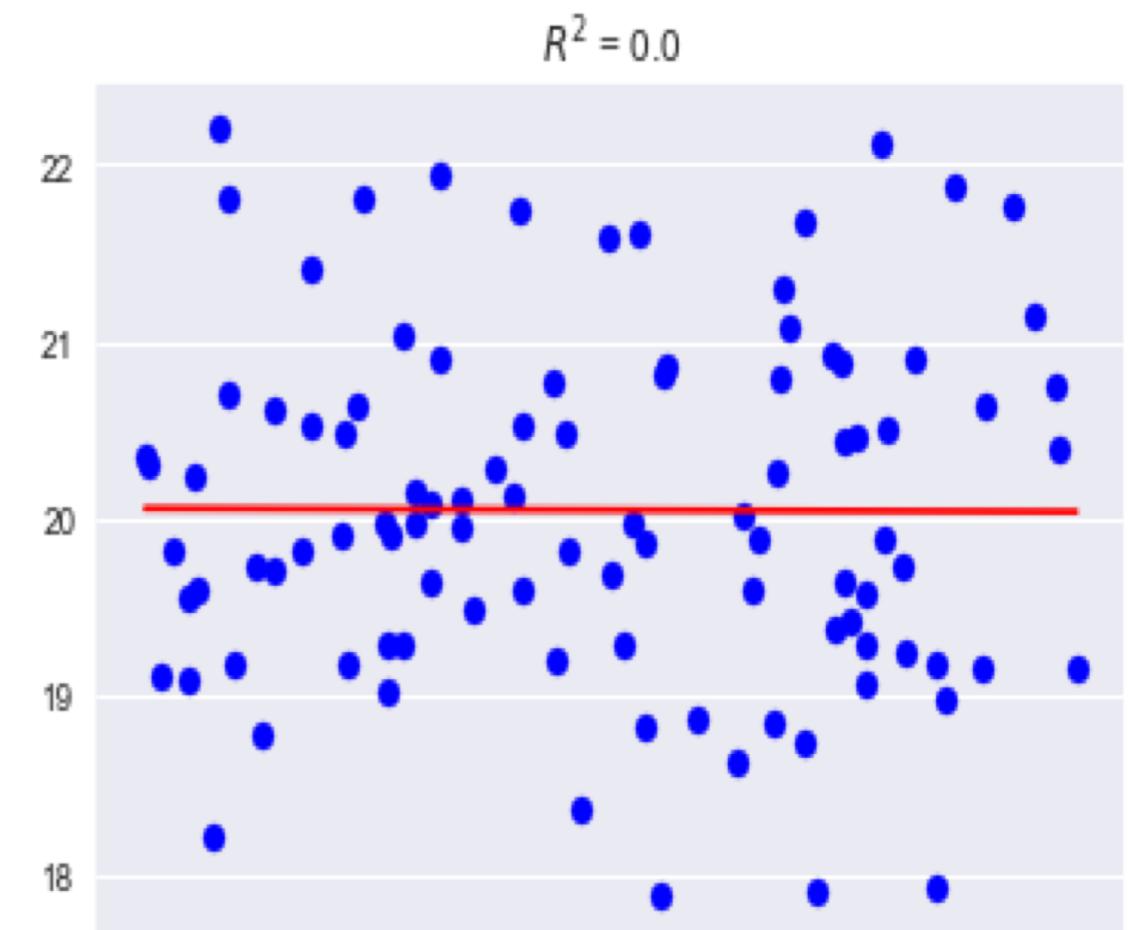
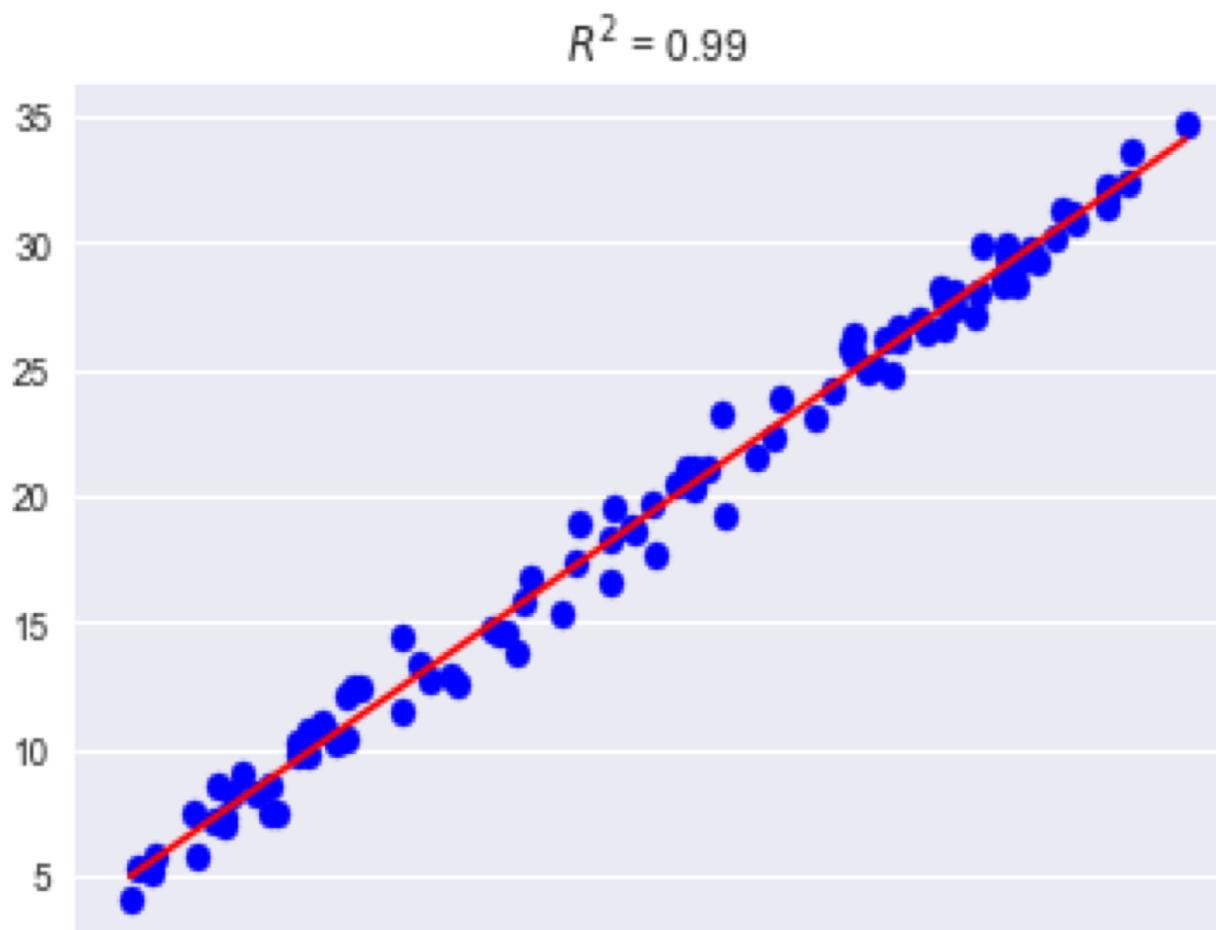
$R^2 < 0$: качество модели плохое, хуже, чем усредненный ответ

$0 \leq R^2 \leq 1$: модель разумна

$R^2 = 0$: модель возвращает средний ответ

$R^2 = 1$: модель идеальна

R²



Линейная регрессия

ЛИНЕЙНАЯ РЕГРЕССИЯ

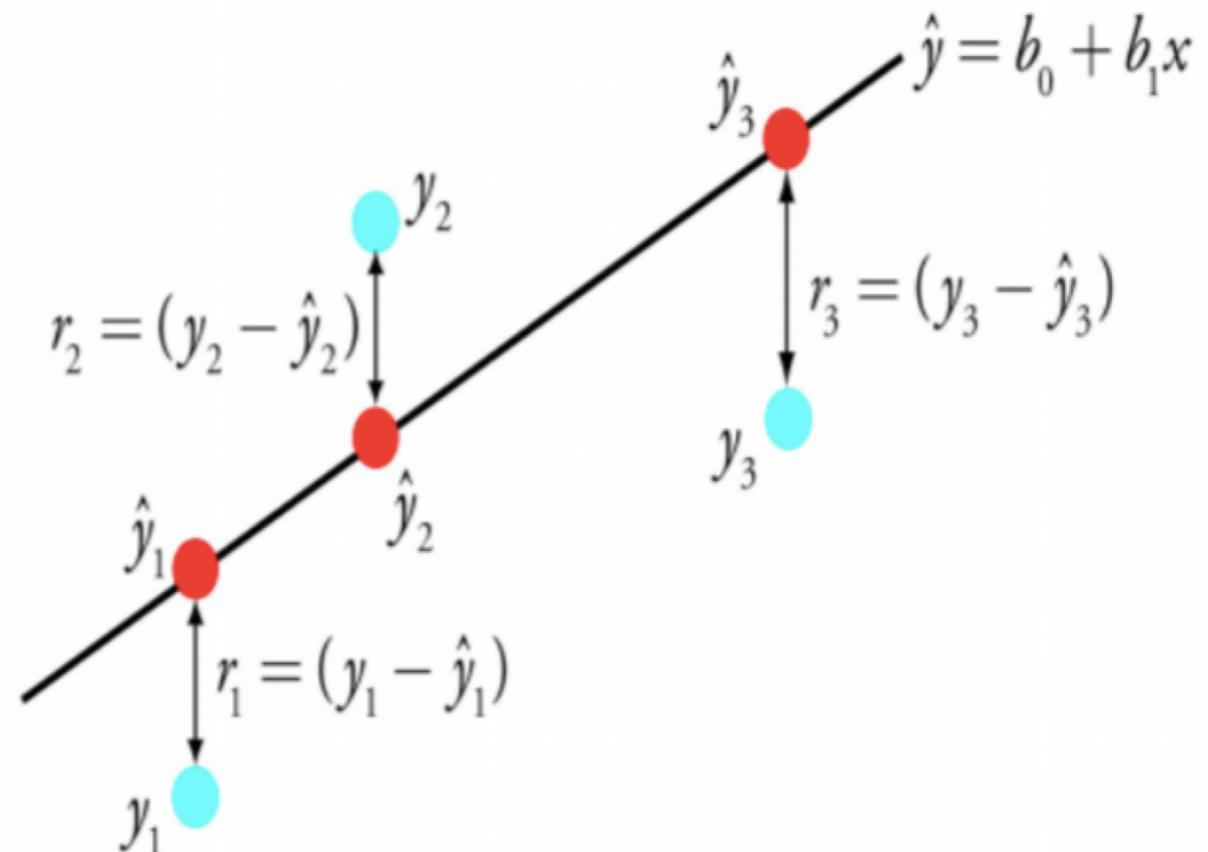
- Одна из самых простых моделей машинного обучения
- В простейшей виде это модель вида (прогнозируем одним признаком):

$$y = w_0 + w_1 x_1$$

- Если прогнозируем несколькими, то:

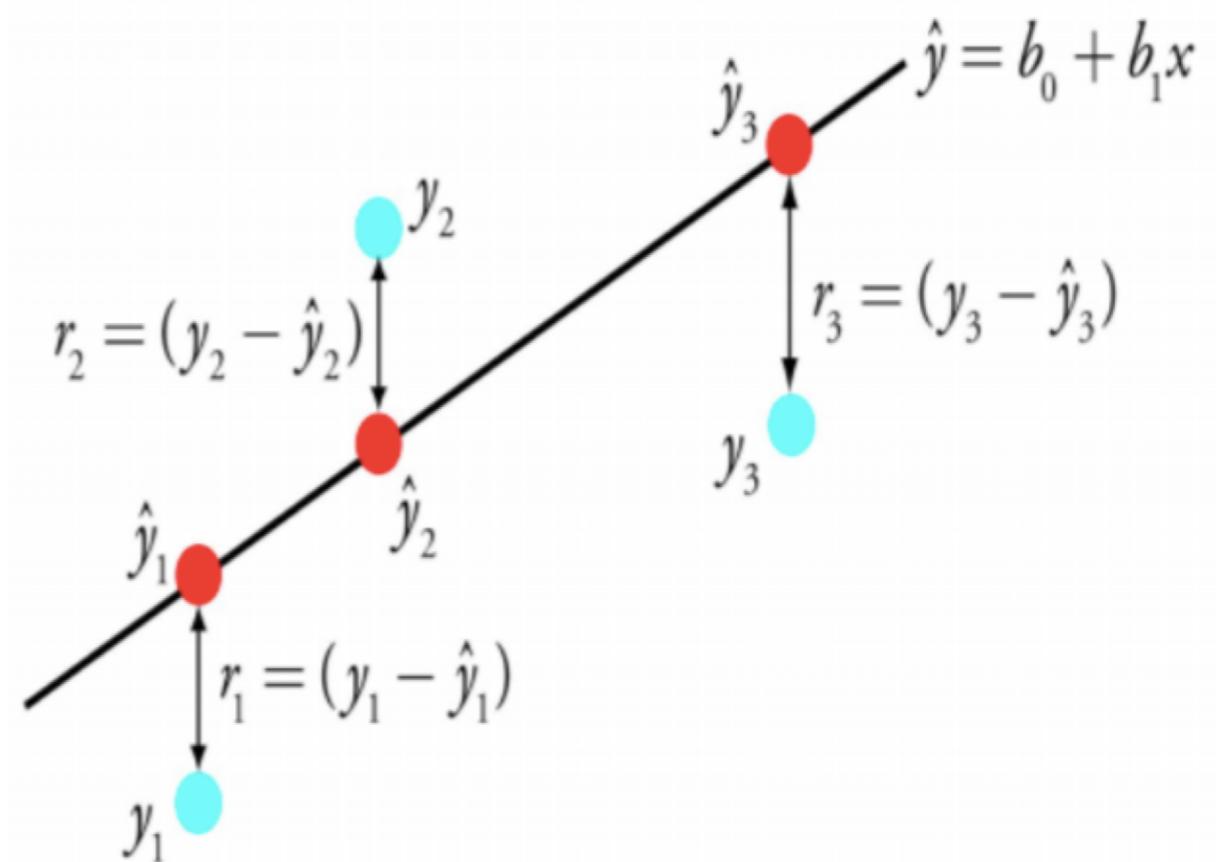
$$y = w_0 + w_1 x_1 + \cdots + w_n x_n$$

- \hat{y} – прогноз



ЛИНЕЙНАЯ РЕГРЕССИЯ

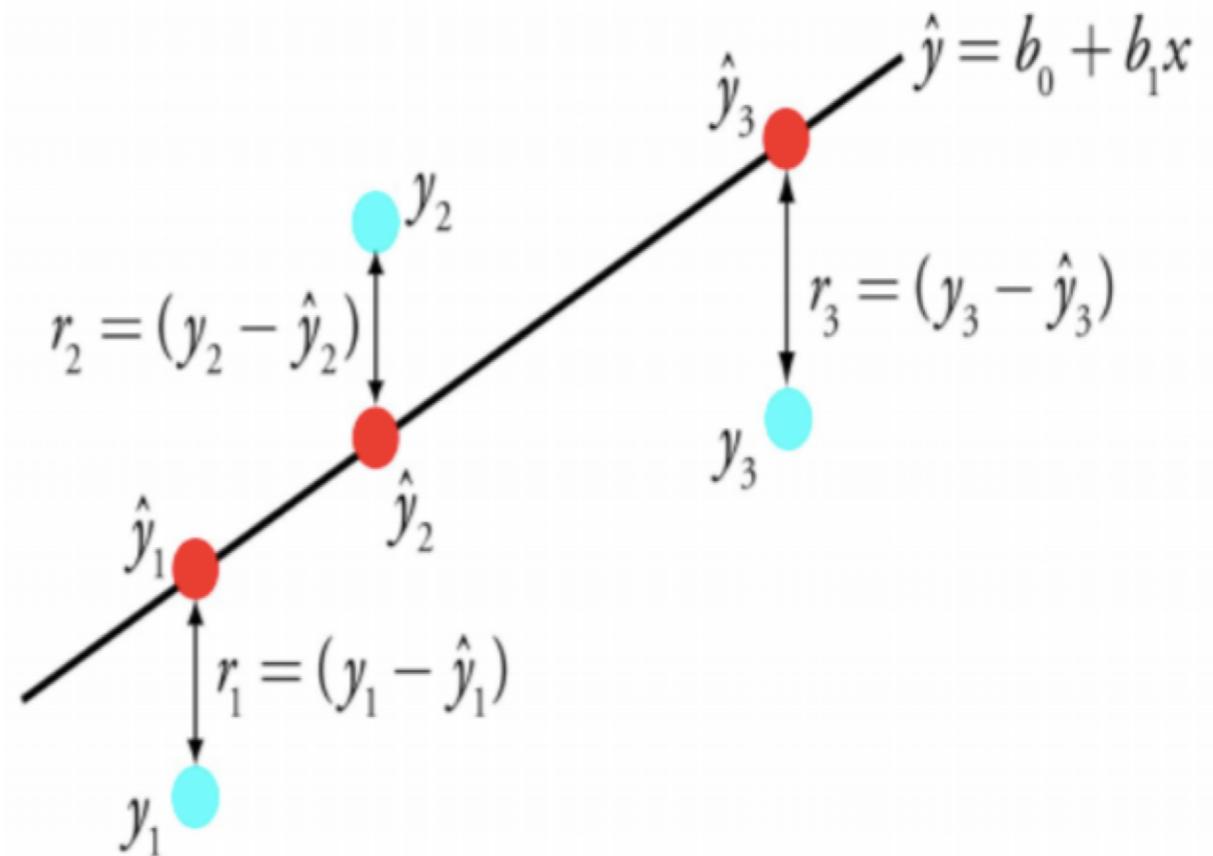
$$\sum_{i=1}^n e^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min$$



ЛИНЕЙНАЯ РЕГРЕССИЯ

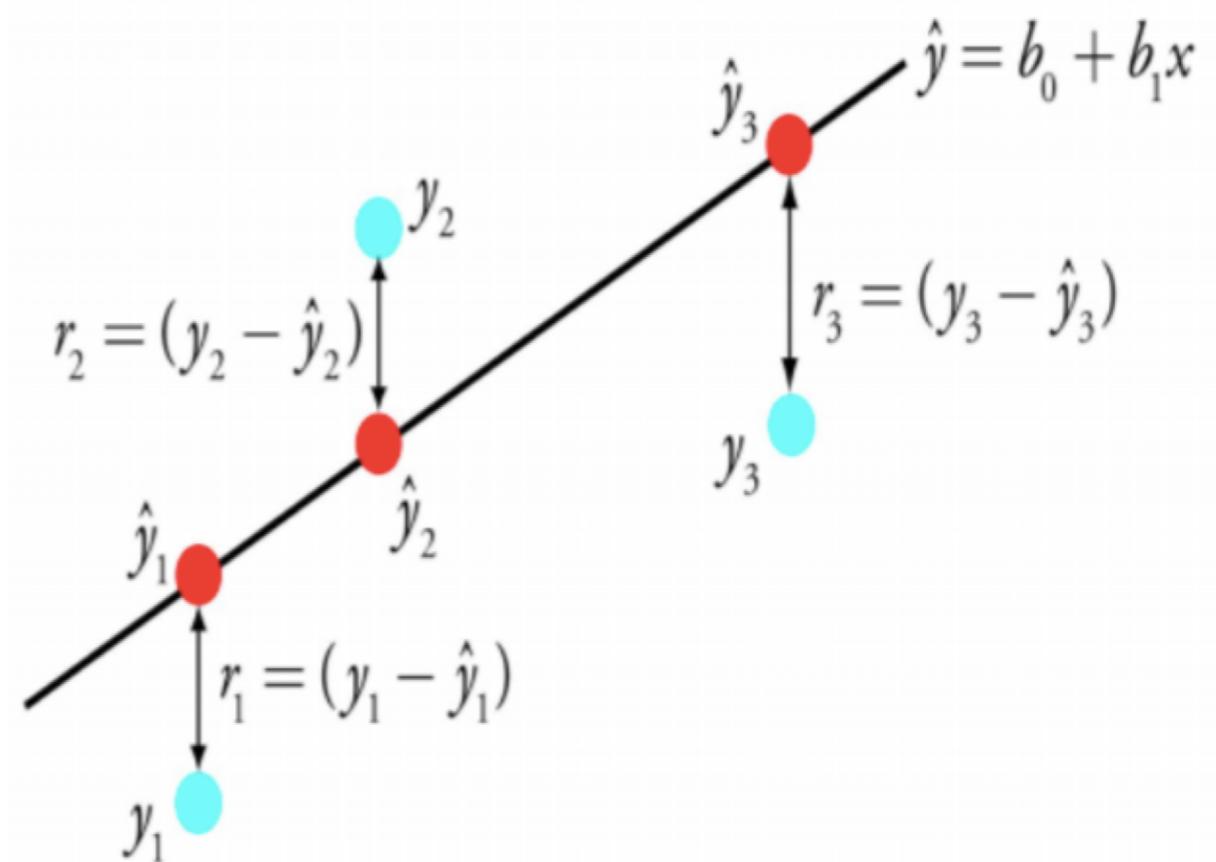
$$\sum_{i=1}^n e^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min$$

$$\hat{y}_i = w_0 + w_1 x_1$$



ЛИНЕЙНАЯ РЕГРЕССИЯ

$$\sum_{i=1}^n e^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - w_0 - w_1 x_i)^2 \rightarrow \min_{w_0, w_1}$$

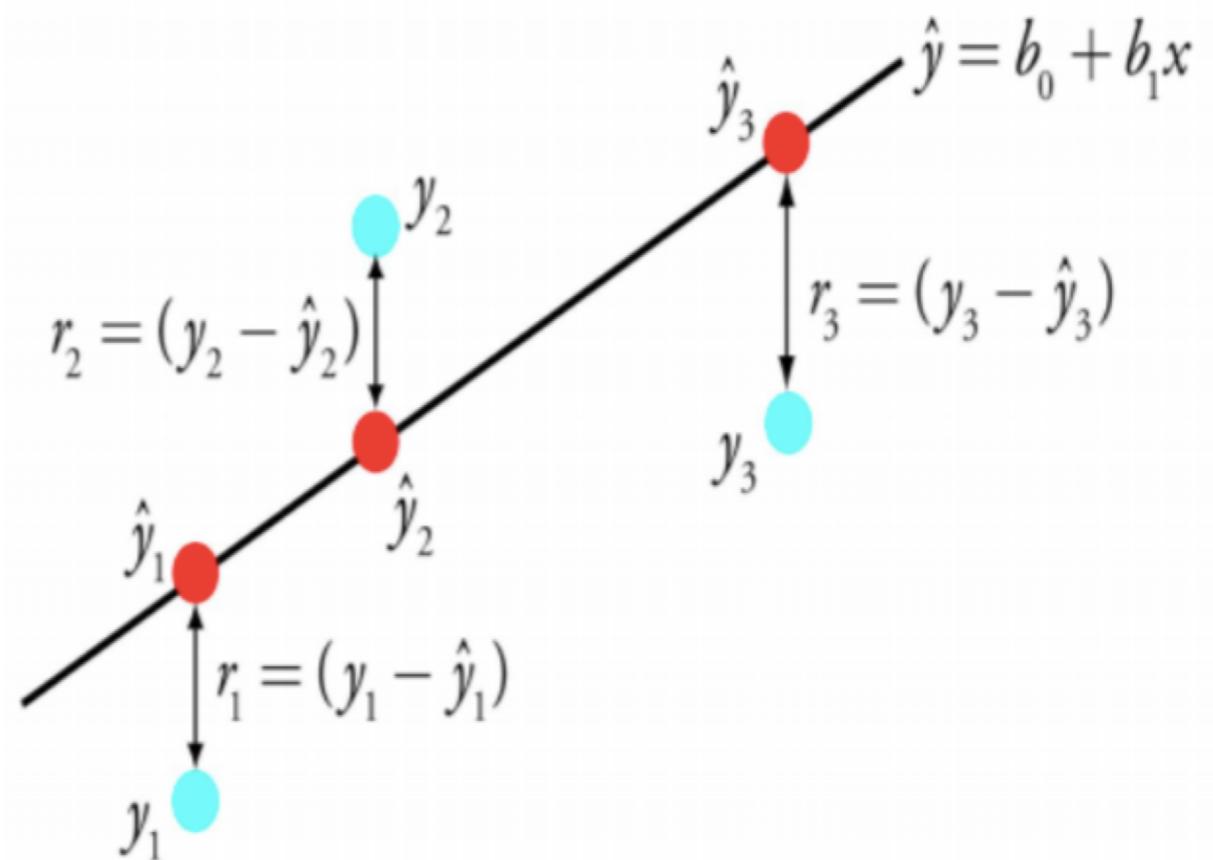


ЛИНЕЙНАЯ РЕГРЕССИЯ

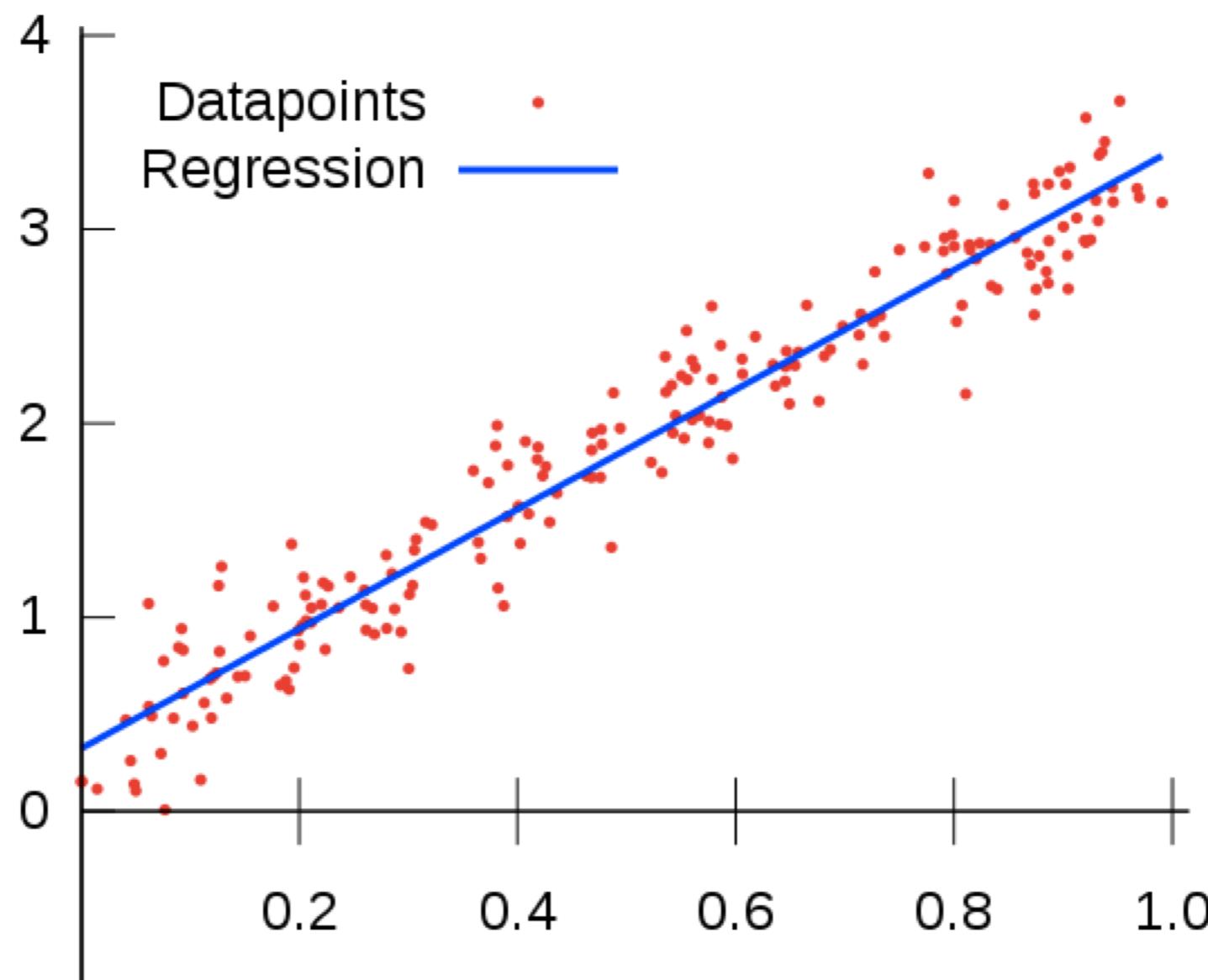
$$\sum_{i=1}^n e^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - w_0 - w_1 x_i)^2 \rightarrow \min_{w_0, w_1}$$

$$w_0 = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2}$$

$$w_1 = \bar{y} - b\bar{x}$$



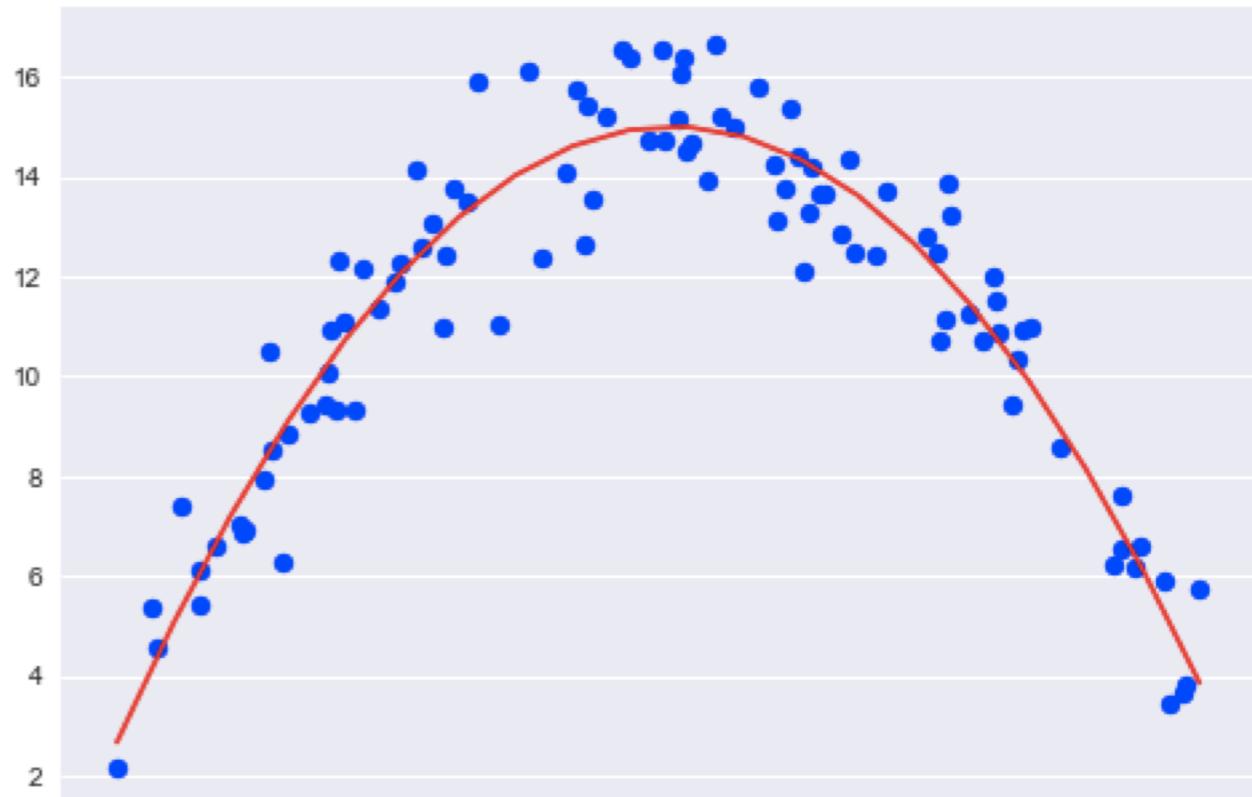
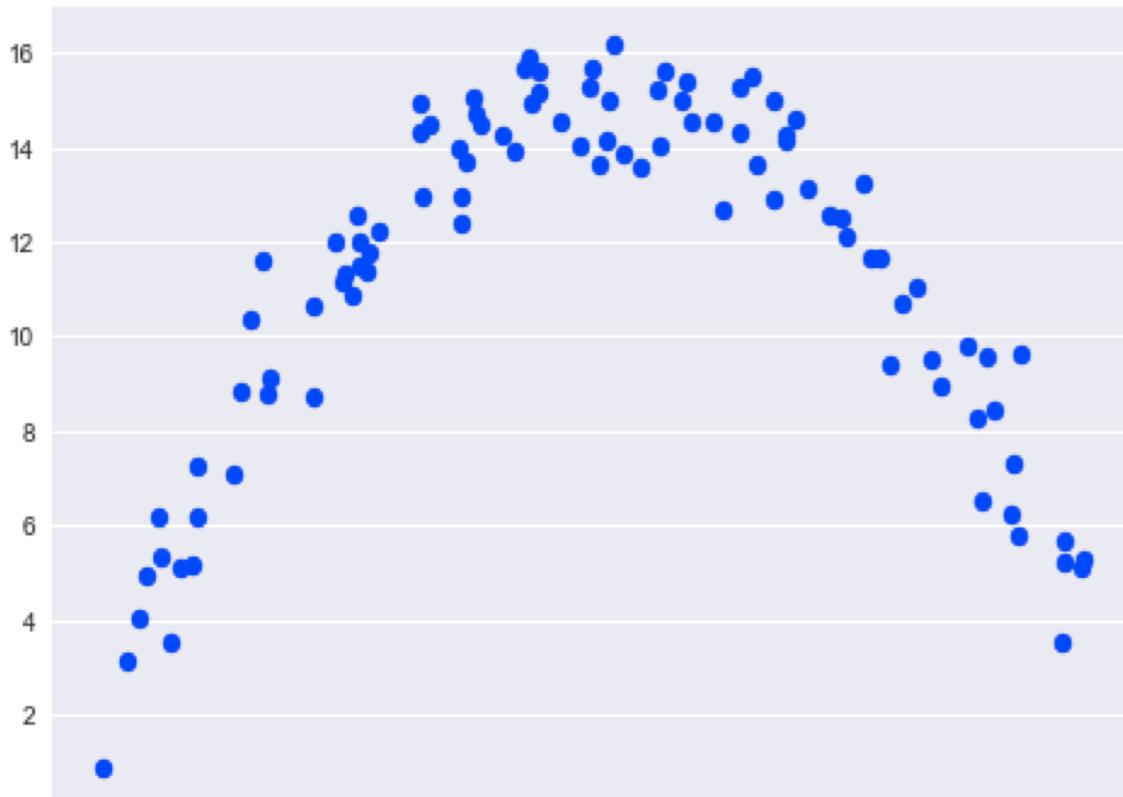
ЛИНЕЙНАЯ РЕГРЕССИЯ



ЛИНЕЙНАЯ РЕГРЕССИЯ

- Легко превращается в нелинейную модель – добавляем имеющийся признак в другом виде:

$$y = w_0 + w_1 x_1 + w_2 x_1^2$$



- Бизнес-задача: определить ключевые драйверы продаж



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ