



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

РЕГРЕССИЯ И ЕЕ ТОНКОСТИ

Теванян Элен

12.11.2019

Москва 2019



FLASHBACK

ПРОИЗВОДНАЯ

- Пусть дана функция $f(x)$
- Производной в точке называется:

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x_0 + \Delta x) - f(x_0)}{x_0 + \Delta x - x_0}$$

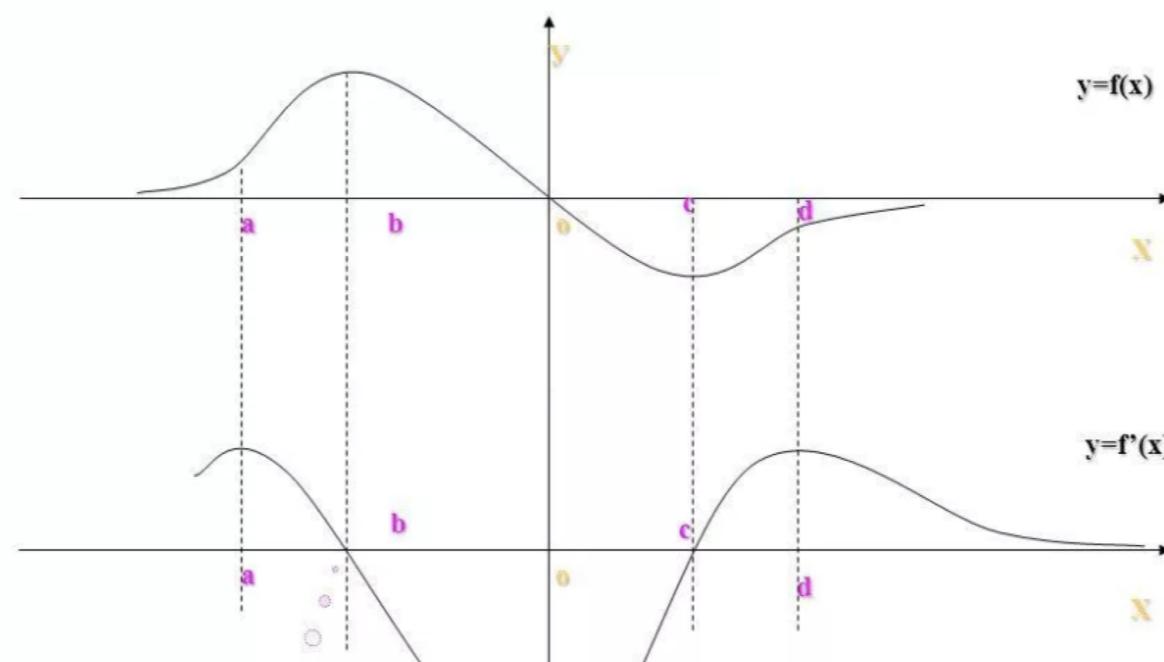


ТАБЛИЦА ПРОИЗВОДНЫХ

Таблица производных	
1. $(u^\alpha)' = \alpha u^{\alpha-1} u'$, $\alpha = const$	
2. $\left(\frac{1}{u}\right)' = -\frac{1}{u^2} u'$	3. $(\sqrt{u})' = \frac{1}{2\sqrt{u}} u'$
4. $(e^u)' = e^u u'$	5. $(a^u)' = a^u \ln a u'$
6. $(\ln u)' = \frac{1}{u} u'$	7. $(\log_a u)' = \frac{1}{u \ln a} u'$
8. $(\sin u)' = \cos u \cdot u'$	9. $(\arcsin u)' = \frac{1}{\sqrt{1-u^2}} u'$

КЛАССЫ ЗАДАЧ

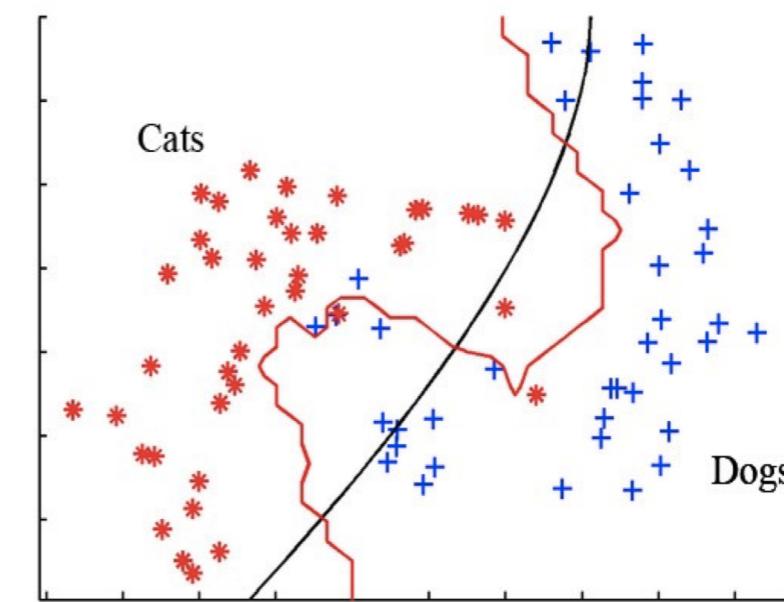
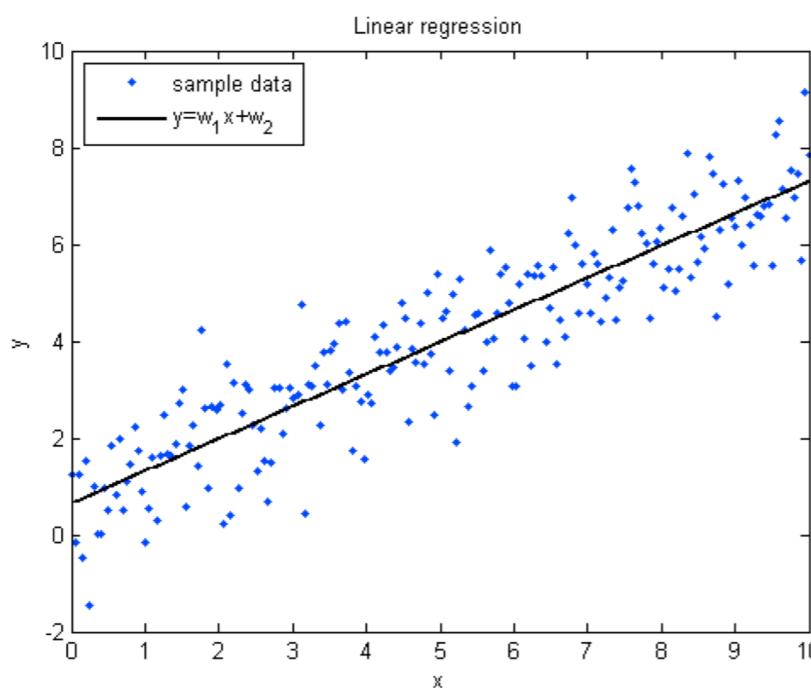
Регрессия

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	Although the housing market is relatively stable in Russia, the country's volatile economy makes

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.

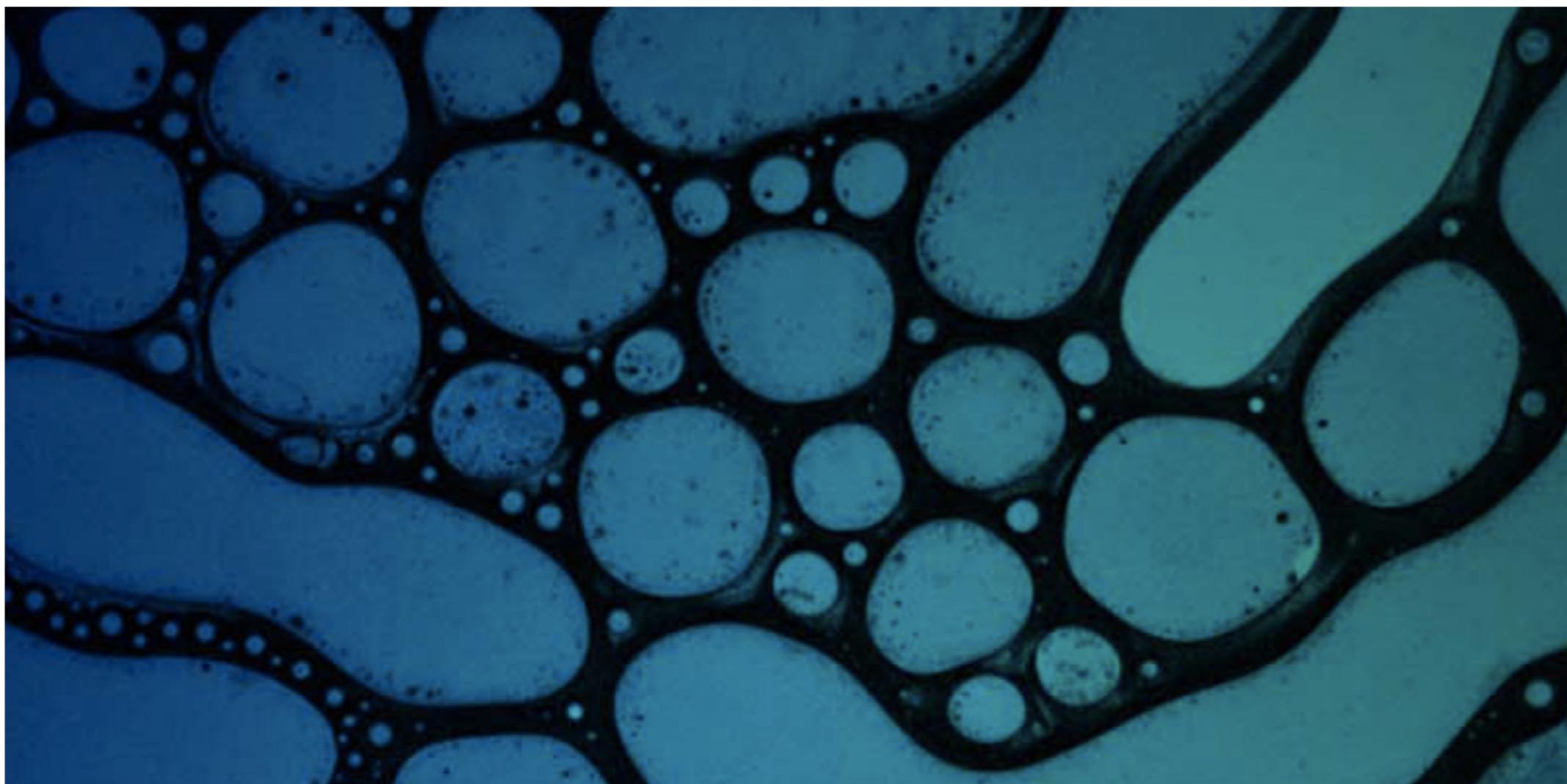
Мы предлагаем:

- 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

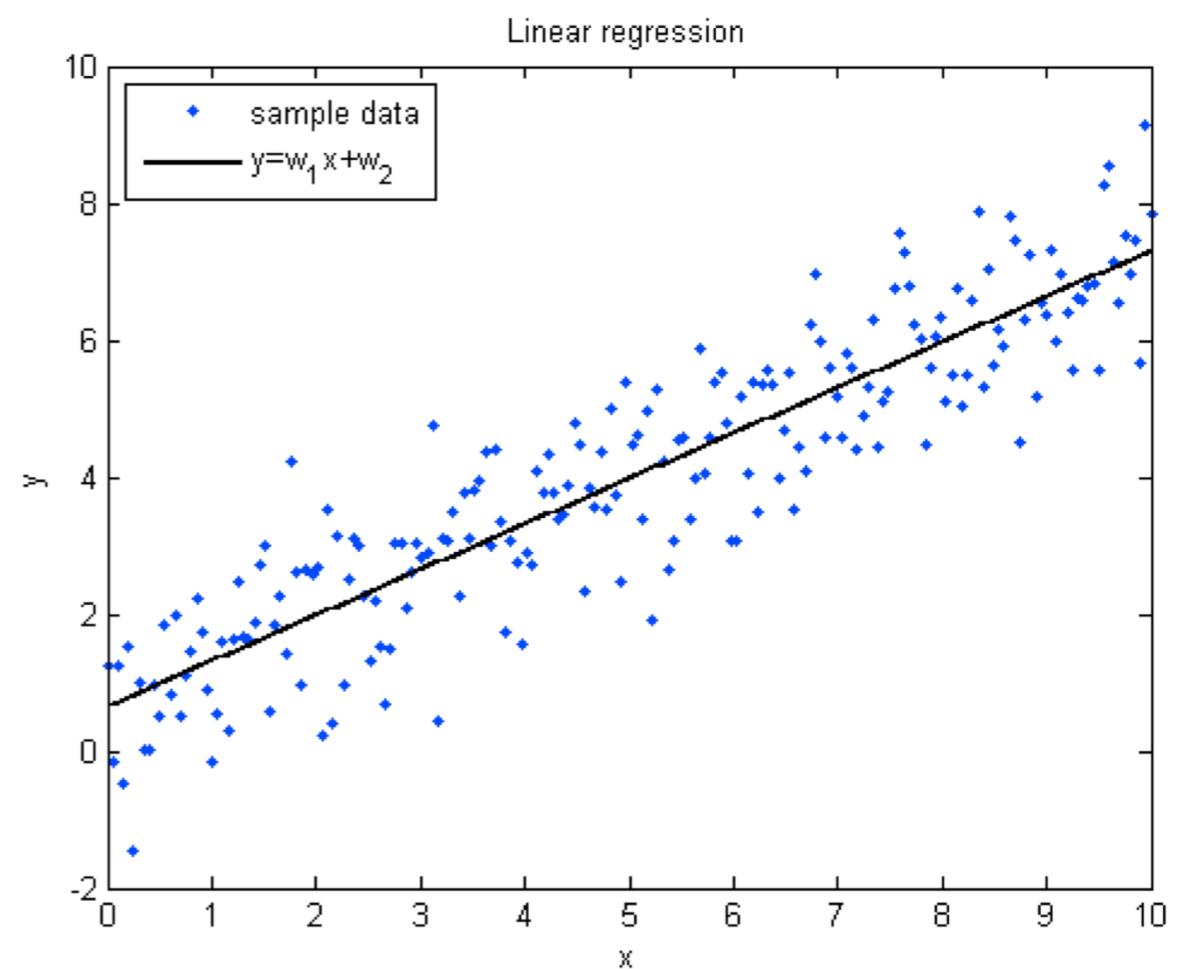


ОБЪЕМ ПОТРЕБЛЕНИЯ ЭЛЕКТРОЭНЕРГИИ



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

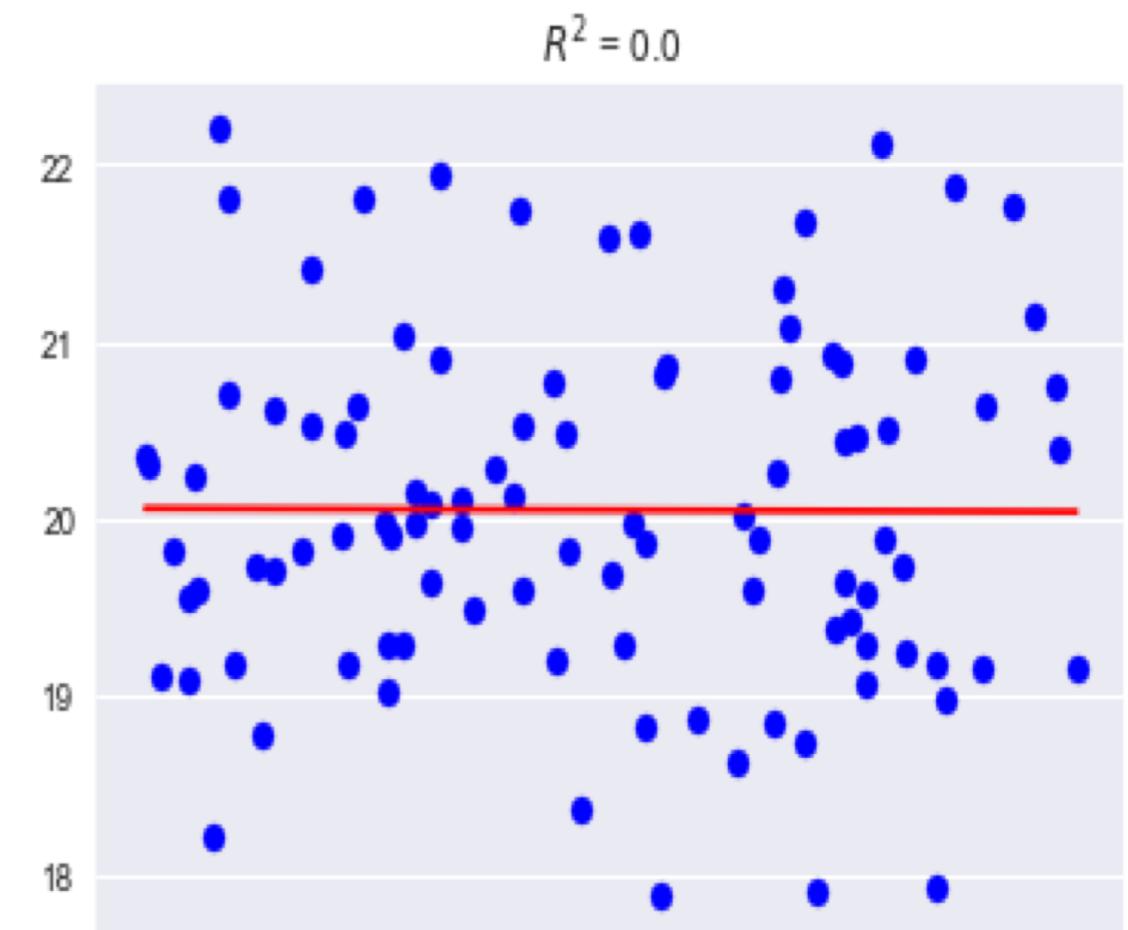
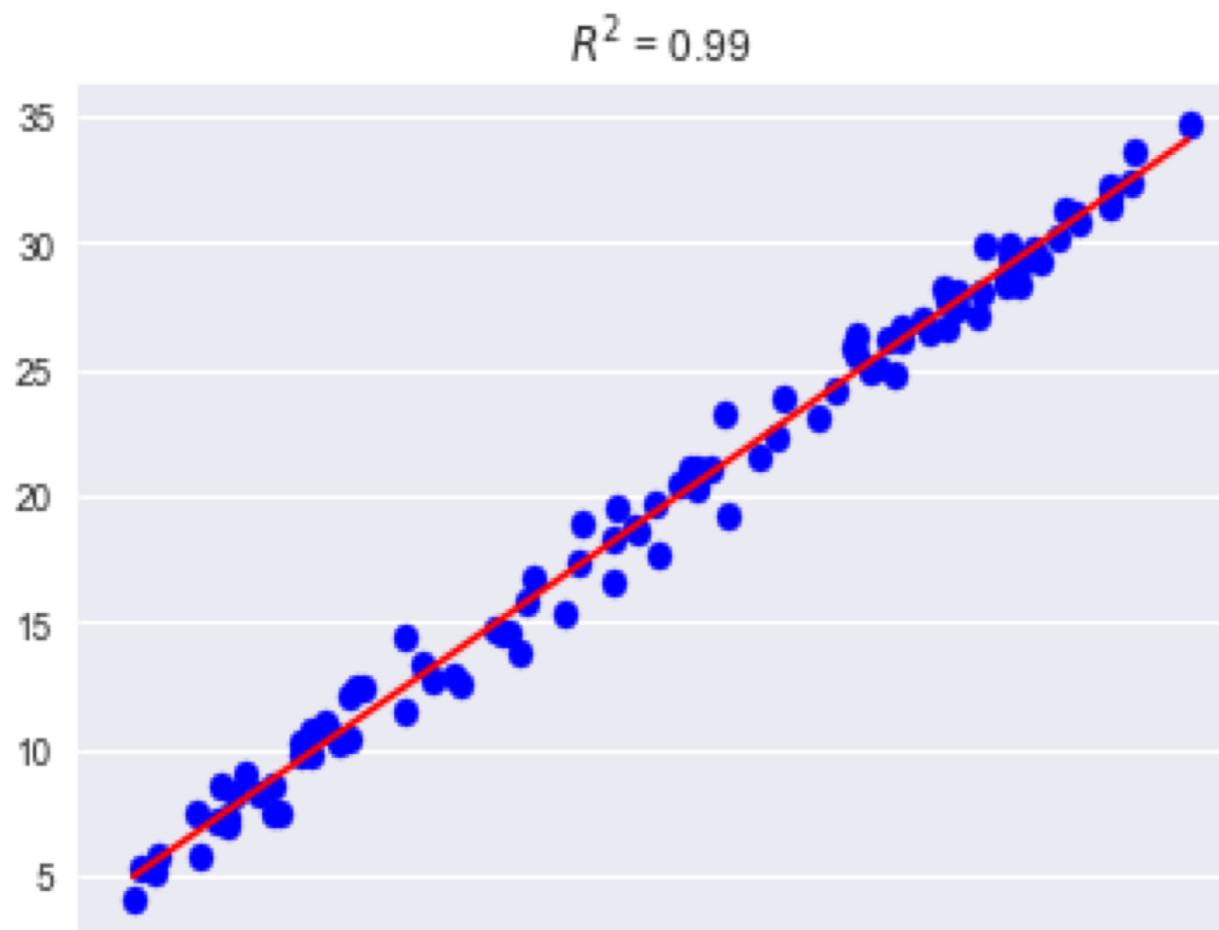
$R^2 < 0$: качество модели плохое, хуже, чем усредненный ответ

$0 \leq R^2 \leq 1$: модель разумна

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

$R^2 = 1$: модель идеальна

R²



Математика регрессии

ЛИНЕЙНАЯ РЕГРЕССИЯ

- $x^i = (f_1^i, \dots, f_n^i)$ – каждый объект описан признаками
- $\hat{y} = w_0 + w_1 f_1 + \dots + w_n f_n$ – модель линейной регрессии.
- Как найти w_0, \dots, w_n ?
- Давайте начнем с простого. Пусть мы хотим строить прогноз только по одной характеристике.
- $\hat{y} = w_0 + w_1 f_1$
- Помните про $\sum_{i=1}^l e_i^2 \rightarrow \min$?

ЛИНЕЙНАЯ РЕГРЕССИЯ

- $x^i = (f_1^i, \dots, f_n^i)$ – каждый объект описан признаками
- $\hat{y} = w_0 + w_1 f_1 + \dots + w_n f_n$ – модель линейной регрессии.
- Как найти w_0, \dots, w_n ?
- Давайте начнем с простого. Пусть мы хотим строить прогноз только по одной характеристике.
- $\hat{y} = w_0 + w_1 f_1$
- Помните про $\sum_{i=1}^l e_i^2 \rightarrow \min$?

ОДНОМЕРНАЯ РЕГРЕССИЯ

- $\hat{y} = w_0 + w_1 f_1$
- $\sum_{i=1}^l e_i^2 = \sum_{i=1}^l (y_i - \hat{y}_i)^2 =$
- $= \sum_{i=1}^l (y_i - w_0 - w_1 f_1)^2 \rightarrow \min$ -

прекрасная оптимизационная задача,

которую мы можем решить

ОДНОМЕРНАЯ РЕГРЕССИЯ

- $\hat{y} = w_0 + w_1 f_1$
- $\sum_{i=1}^l e_i^2 = \sum_{i=1}^l (y_i - \hat{y}_i)^2 =$
- $= \sum_{i=1}^l (y_i - w_0 - w_1 f_1)^2 \rightarrow \min$ -

прекрасная оптимизационная задача,

которую мы можем решить

ОДНОМЕРНАЯ РЕГРЕССИЯ

- $\hat{y} = w_0 + w_1 f_1$
- $\sum_{i=1}^l e_i^2 = \sum_{i=1}^l (y_i - \hat{y}_i)^2 =$
- $= \sum_{i=1}^l (y_i - w_0 - w_1 f_1)^2 \rightarrow \min$

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

$$w_1 = \bar{y} - w_0 \bar{x}$$

АНАЛИТИЧЕСКОЕ РЕШЕНИЕ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

- X – матрица с признаками
- y - целевая переменная
- w – веса модели

$$w = (X^T X)^{-1} X^T y$$

ПРОБЛЕМЫ

$$w = (X^T X)^{-1} X^T y$$

- Обращение матрицы
- $X^T X$ может быть вырожденной или плохо обусловленной

ГРАДИЕНТ

- Градиент – это вектор частных производных функции многих переменных

$$\nabla f(x_1, \dots, x_d) = \left(\frac{\partial f}{\partial x_j} \right)_{j=1}^d$$

- Градиент – направление наискорейшего роста функции
- Антиградиент – направление наискорейшего убывания

$$-\nabla f$$

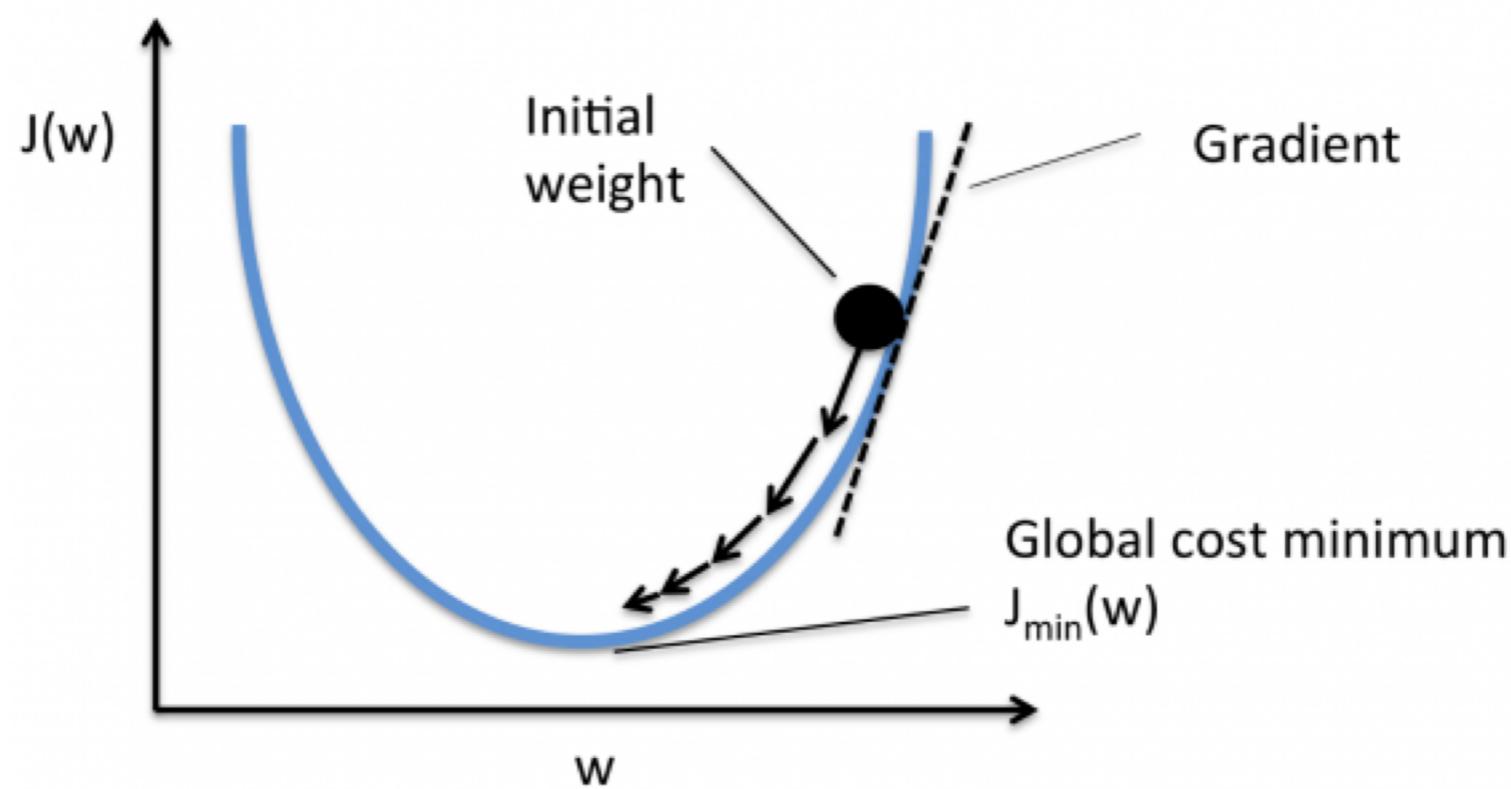
ГРАДИЕНТНЫЙ СПУСК

- 1) Стартуем в точке
- 2) Двигаемся в сторону антиградиента

$$-\nabla f$$

- 3) Пересчитываем антиградиент
- 4) Возвращаемся к п.2

ГРАДИЕНТНЫЙ СПУСК



Регуляризация регрессии

ВИДЫ РЕГУЛЯРИЗАЦИИ

$$R(w) = \|w\|_2 = \sum_{i=1}^d w_i^2,$$

$$R(w) = \|w\|_1 = \sum_{i=1}^d |w_i|.$$

ВИДЫ РЕГУЛЯРИЗАЦИИ

$$R(w) = \|w\|_2 = \sum_{i=1}^d w_i^2, \quad L_2, \text{ Ридж}$$

$$R(w) = \|w\|_1 = \sum_{i=1}^d |w_i|. \quad L_1, \text{ Лассо}$$

АНАЛИТИЧЕСКОЕ РЕШЕНИЕ ДЛЯ L₂

- Градиент – это вектор частных производных функции многих переменных

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



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ