Системы искусственного интеллекта

Лекция 2

Линейная регрессия

Запорожцев Иван Федорович
zaporozhtsev.if.work@gmail.com

Линейная регрессия

Гипотеза о линейной зависимости целевой переменной, ищем решение в виде:

$$a(X_1,\ldots,X_n) = w_0 + w_1X_1 + \ldots + w_nX_n$$

Практика:

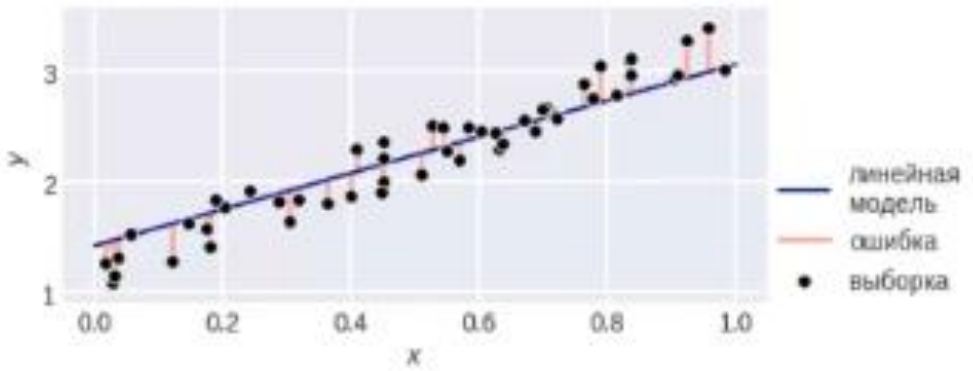
- часто неплохо работает и при монотонных зависимостях
- хорошо работает, когда есть много «однородных» признаков

Цель – число продаж на следующей неделе

- признак 1 – число заходов на страницу продукта
- признак 2 – число добавлений в корзину
- признак 3 – число появлений продукта в поисковой выдачи
- ...

Линейная регрессия от одной переменной

$$a(X_1) = w_0 + w_1 X_1$$



Обучение:

$$\{(x_1, y_1), \ldots, (x_m, y_m)\}, \; x_i \in \mathbb{R}$$

Хотели бы...

$$\begin{cases} w_0 + w_1 x_1 = y_1 \\ \cdots \\ w_0 + w_1 x_m = y_m \end{cases}$$

Невязки / отклонения (residuals):

$$e_1 = y_1 - w_0 - w_1 x_1$$
$$\cdots$$
$$e_m = y_m - w_0 - w_1 x_m$$

Обучение:

$$\{(x_1, y_1), \ldots, (x_m, y_m)\}, \; x_i \in \mathbb{R}$$

Хотели бы...

$$\begin{cases} w_0 + w_1 x_1 = y_1 \\ \cdots \\ w_0 + w_1 x_m = y_m \end{cases}$$

Невязки / отклонения (residuals):

$$e_1 = y_1 - w_0 - w_1 x_1$$
$$\cdots$$
$$e_m = y_m - w_0 - w_1 x_m$$

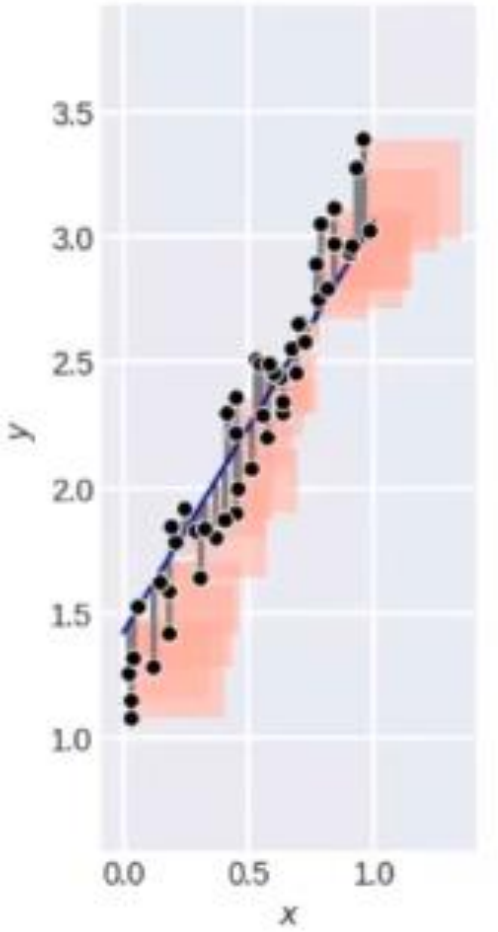
Задача минимизации суммы квадратов отклонений (residual sum of squares)

$$\text{RSS} = e_1^2 + \ldots + e_m^2 \to \min$$

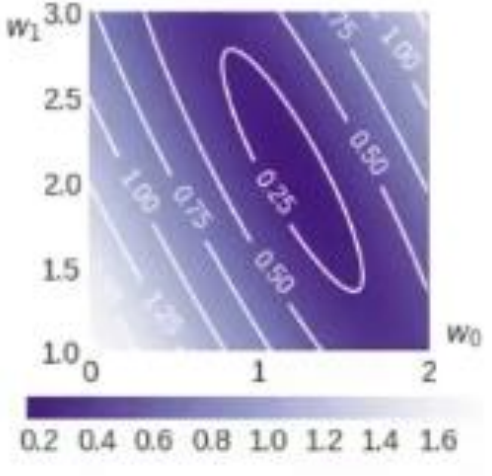
$$L(w) = \sum_{i=1}^{m} (y_i - a_w(x_i))^2 = \sum_{i=1}^{m} (y_i - (w_0 + w_1 x_i))^2$$

Линейная регрессия от одной переменной

Геометрический смысл ошибки



$$a(X_1) = w_0 + w_1 X_1$$



$$\sum_{i=1}^{m} (y_i - w_0 - w_1 x_i)^2$$

Отличается от суммы расстояний до поверхности!

Линейная регрессия от одной переменной

Нетрудно показать:

$$w_1 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} = \frac{\mathrm{cov}(\{x_i\},\{y_i\})}{\mathrm{var}(\{x_i\})},$$

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

где $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$, $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$.

Полученное уравнение прямой (проходит через «центр масс»):

$$(y - \bar{y}) = \frac{\mathrm{cov}(\{x_i\},\{y_i\})}{\mathrm{var}(\{x_i\})}(x - \bar{x})$$

Общий случай

Многих переменных

$$a(X_1,\ldots,X_n) = w_0 + w_1X_1 + \cdots + w_nX_n = x^{\mathrm{T}}w$$

веса (параметры) – $w = (w_0, w_1, \ldots, w_n)^{\mathrm{T}}$

объект – $x = (X_0, X_1, \ldots, X_n)^{\mathrm{T}}$

$X_0 \equiv 1$ **– фиктивный признак для удобства**

обучение: $\{(x_1, y_1), \ldots, (x_m, y_m)\} \quad x_i \in \mathbf{R}^{n+1}$

опять хотим решить $Xw = y$

$$\begin{cases} x_1^{\mathrm{T}}w = y_1 \\ \cdots \\ x_m^{\mathrm{T}}w = y_m \end{cases}$$

Как решать?

Общий случай

многих переменных:
в матричной форме

$$Xw = y$$

В матрице X по строкам записаны описания объектов, в векторе y значения их целевого признака (здесь есть коллизия в обозначении y)

Будем решать так:

$$\| Xw - y \|_2^2 = \sum_{i=1}^{m} (x_i^T w - y_i)^2 \to \min_w$$

Общий случай

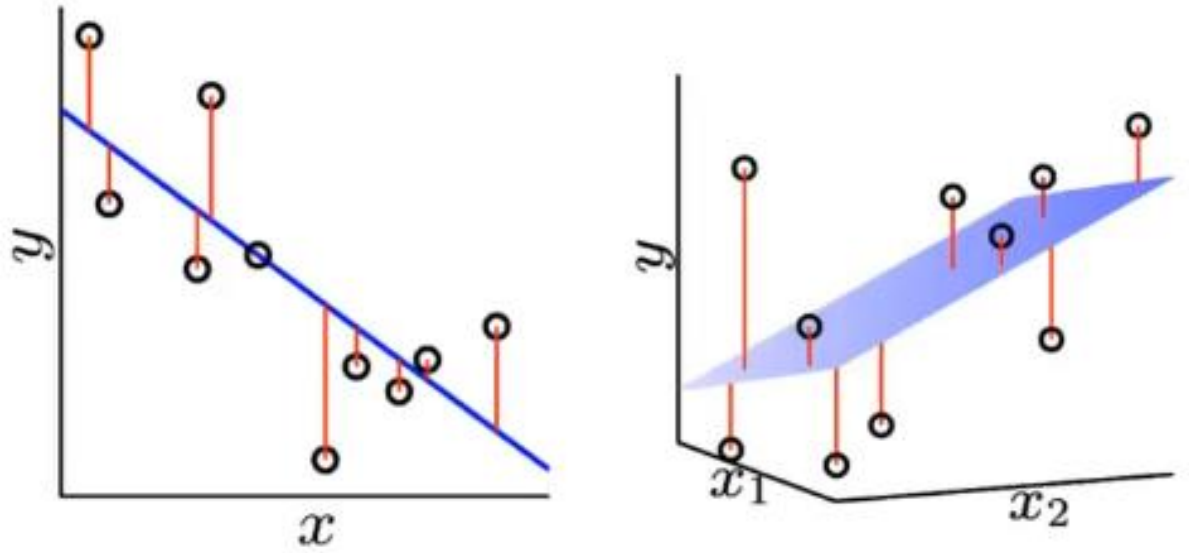
многих переменных:

в матричной форме

$$Xw = y$$

$$\| Xw - y \|_2^2 = \sum_{i=1}^{m} (x_i^{\mathrm{T}} w - y_i)^2 \to \min_w$$

Будет единственный минимум – локальный = глобальный



Решение задачи минимизации: прямой метод

$$\| Xw - y \|_2^2 \to \min_w$$

$$\| Xw - y \|_2^2 = (Xw - y)^{\mathrm{T}}(Xw - y) = w^{\mathrm{T}}X^{\mathrm{T}}Xw - w^{\mathrm{T}}X^{\mathrm{T}}y - y^{\mathrm{T}}Xw + y^{\mathrm{T}}y$$

$$\nabla \| Xw - y \|_2^2 = 2X^{\mathrm{T}}Xw - 2X^{\mathrm{T}}y = 0$$

$X^{\mathrm{T}}Xw = X^{\mathrm{T}}y$ **решение существует, если столбцы л/н**

$w = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ **помним, что** $\mathrm{rg}(X^{\mathrm{T}}X) = \mathrm{rg}(X)$

хотим решить $Xw = y$

$(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$

псевдообратная матрица Мура-Пенроуза
обобщение обратной на неквадратные матрицы

Обобщённая линейная регрессия: вместо X – что угодно

Выражаем целевое значение через л/к базисных функций (они фиксированы)

$$a(X_1,\ldots,X_n) = w_0 + w_1\varphi_1(X_1,\ldots,X_n) + \cdots + w_k\varphi_k(X_1,\ldots,X_n)$$

$$C_{Ni}\left(I_{Ni}, I_{Cu}, I_{Fe}\right) = w_0 + w_1 I_{Ni} + w_2 \frac{1}{I_{Cu}} + w_3 \frac{I_{Cu}}{I_{Fe}}$$

FEATURE GENERATION

data iku

PATIENT ID	PATIENT AGE	NUMBER OF DIAGNOSES
55629189	15	9
86057875	25	6
82442376	35	7
42519267	45	5
82637451	55	9
114882984	65	7
48330782	75	8

AGE X NUMBER DIAGNOSES
135
150
245
225
495
455
600

Проблема вырожденности матрицы

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

Проблемы, когда матрица $X^{\mathrm{T}} X$ плохо обусловлена...

$$\mu(X^{\mathrm{T}} X) = \| X^{\mathrm{T}} X \| \cdot \| (X^{\mathrm{T}} X)^{-1} \| = \frac{\lambda_{\max}(X^{\mathrm{T}} X)}{\lambda_{\min}(X^{\mathrm{T}} X)}$$

Решения:

1. Регуляризация – здесь и в «сложности»
2. Селекция (отбор) признаков – «селекция»
3. Уменьшение размерности (в том числе, PCA) – USL
4. Увеличение выборки

Регуляризация:

Упрощённое объяснение смысла

$$a(X_1,\ldots,X_n) = w_0 + w_1X_1 + \ldots + w_nX_n$$

Если есть два похожих объекта,
то должны быть похожи метки, пусть
отличаются в j-м признаке, тогда
ответы модели отличаются на $\varepsilon_j w_j$

Поэтому не должно быть очень больших
по модулю весов (у признаков,
по которым могут отличаться
похожие объекты)

Поэтому вместе с $\| Xw - y \|_2^2 \to \min$

хотим $\| w \|_2^2 \to \min$

Не на все коэффициенты нужна регуляризация!

Почему?

Регуляризация

Иванова

$$\begin{cases} \| Xw - y \|_2^2 \to \min \\ \| w \|_2^2 \le \lambda \end{cases}$$

Тихонова

$$\| Xw - y \|_2^2 + \lambda \| w \|_2^2 \to \min$$

Удобнее: безусловная оптимизация

$$\| w \|_2^2 = w_1^2 + w_2^2 + \ldots + w_n^2 \text{ – нет } w_0^2$$

Эти две формы эквивалентны: решение одного можно получить как решение другого

На самом деле, регуляризация упрощает модель

Регуляризация и гребневая регрессия

$$\arg\min_{w} \| Xw - y \|_2^2 + \lambda \| w \|_2^2 = (X^T X + \lambda I)^{-1} X^T y$$

$$\lambda \geq 0$$

Доказать!

Такая регрессия называется **гребневой регрессией** (Ridge Regression)

Виден другой смысл регрессии: складываем две матрицы Грама

Неотрицательно определённая + положительно определённая

Боремся с вырожденностью матрицы

Коэффициент регуляризации (shrinkage penalty)

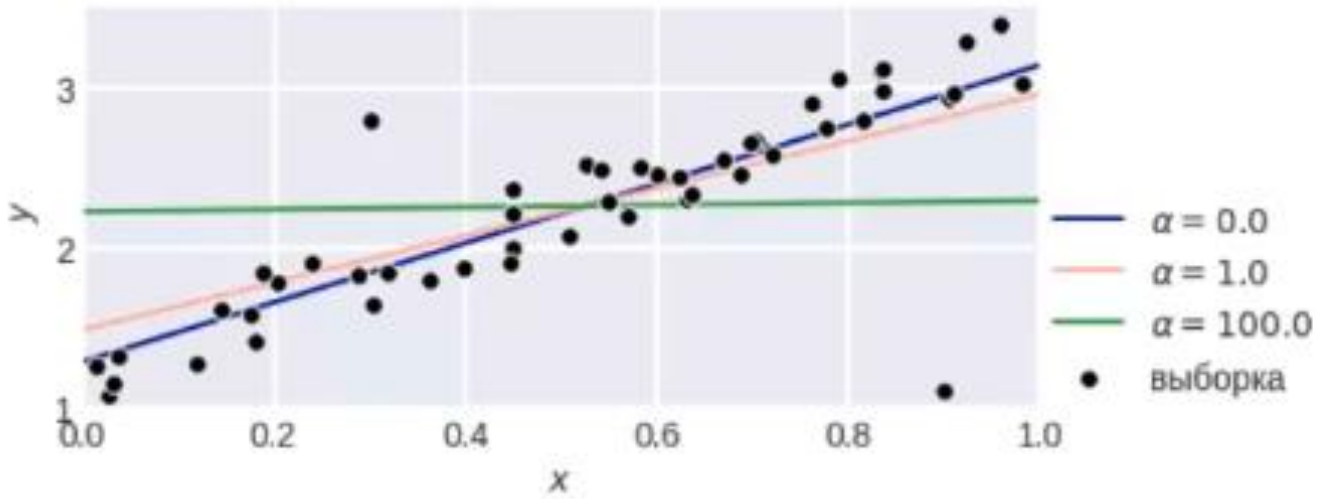
$\lambda = 0$ – получаем классическое решение

$\lambda \to +\infty$ – меньше «затачиваемся на данные» и больше регуляризуем

Значение параметра регуляризации можно выбрать на скользящем контроле

Минутка кода

Регуляризация и гребневая регрессия




```
from sklearn.linear_model import Ridge
model = Ridge(alpha=0.0) # гребневая регрессия
# обучение
model.fit(x_train[:, np.newaxis], y_train)
# обратите внимание: np.newaxis
# контроль
a_train = model.predict(x_train[:, np.newaxis])
a_test = model.predict(x_test[:, np.newaxis])
```

Регуляризация и гребневая регрессия

Для ridge-регрессии нужна правильная нормировка признаков!

Нет инвариантности (в отличие от линейной) от умножения признаков на скаляры

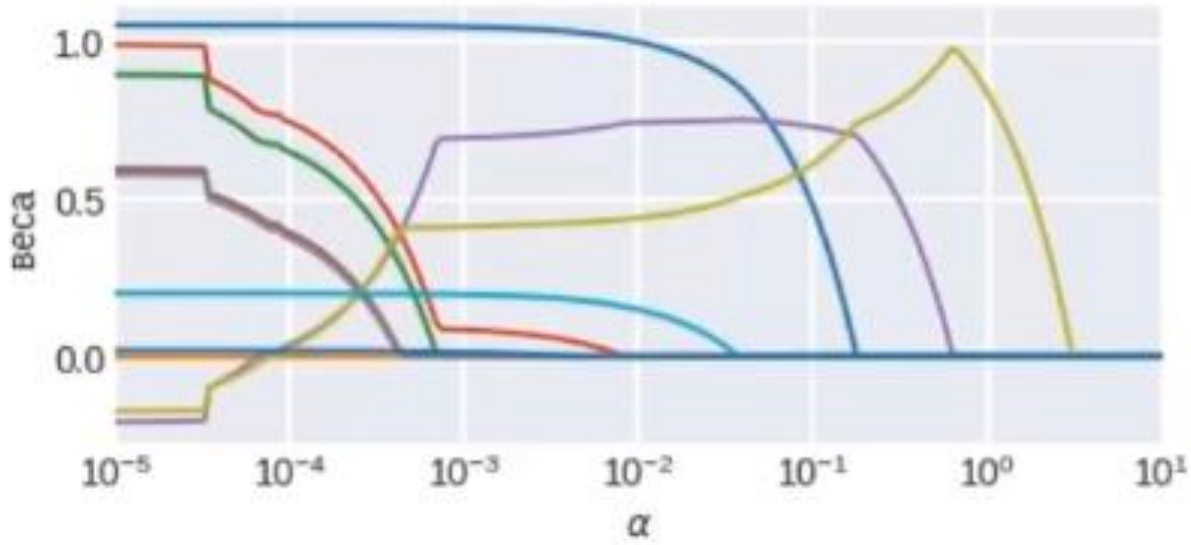
Перед регуляризацией – стандартизация!!!

LASSO

Least Absolute Shrinkage and Selection Operator

$$\sum_{i=1}^{m}\left(y_i - a(x_i)\right)^2 + \lambda\sum_{j=1}^{n}|\, w_j \,| \to \min$$

$$\lambda \geq 0$$

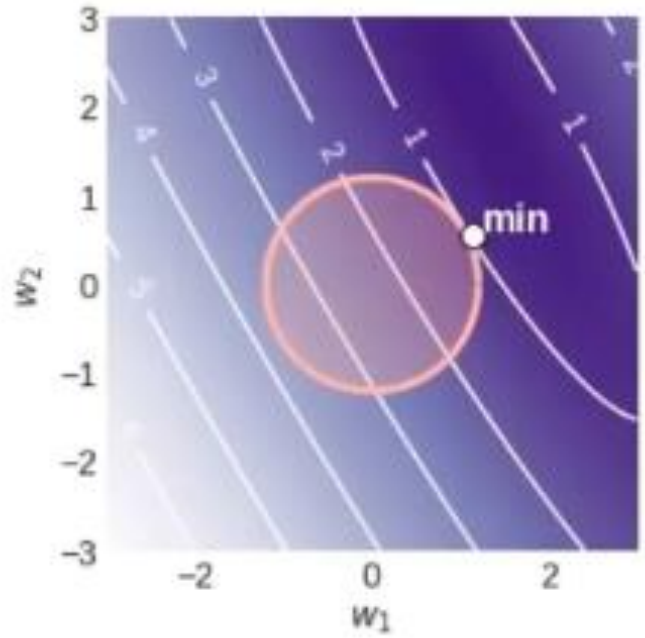


Здесь значения коэффициентов существенно меньше (т.к. при $\Sigma|\cdot|$, а не $\Sigma(\cdot)^2$)

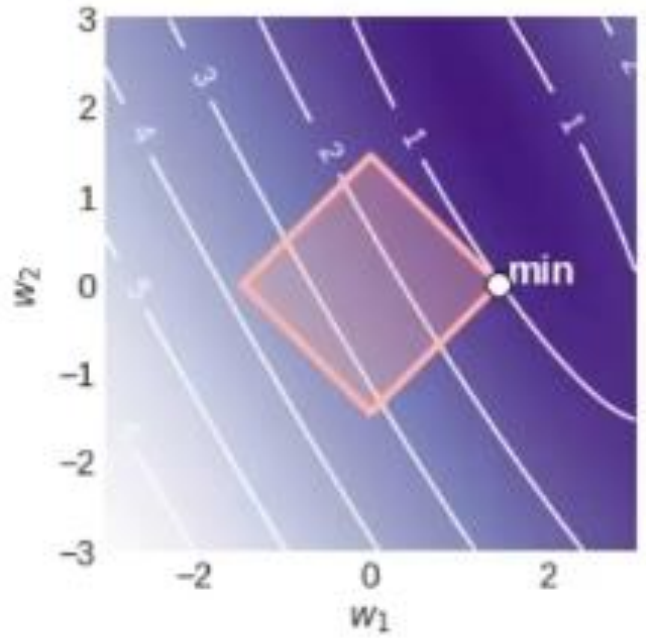
Здесь коэффициенты интенсивнее зануляются при увеличении $\lambda \geq 0$

Геометрический смысл Ridge и LASSO

$$\sum_{i=1}^{m}\left(y_i - w_0 - \sum_{j=1}^{n} w_j x_{ij}\right)^2 \to \min_w, \quad \sum_{j=1}^{n} w_j^2 \le s$$



$$\sum_{i=1}^{m}\left(y_i - w_0 - \sum_{j=1}^{n} w_j x_{ij}\right)^2 \to \min_w, \quad \sum_{j=1}^{n} |w_j| \le s$$



Семейство регуляризированных линейных методов

Ridge $\| y - Xw \|_2^2 + \lambda \| w \|_2^2 \to \min_w$

LASSO
Least Absolute Shrinkage and Selection Operator

$\| y - Xw \|_2^2 + \lambda \| w \|_1 \to \min_w$

Elastic Net = LASSO + Ridge $\| y - Xw \|_2^2 + \lambda_1 \| w \|_1 + \lambda_2 \| w \|_2^2 \to \min_w$

Проблема вырожденности / плохой обусловленности матрицы

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

Решения:

1. Регуляризация
2. **Селекция (отбор) признаков**
3. Уменьшение размерности (в том числе, PCA)
4. Увеличение выборки

Какие признаки включить в модель

$$a(X_1,\ldots,X_n) = w_0 + w_1X_1 + \ldots + w_nX_n$$

Маленький обзор стратегий:

1 стратегия – **перебор** – умный перебор подмножества признаков

2 стратегий – **оценка** – оценка качества признаков (фильтры)

3 стратегия – **автомат** – встроенные методы (ex: LASSO)

Проблема вырожденности / плохой обусловленности матрицы

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

Решения:

1. Регуляризация
2. Селекция (отбор) признаков
3. **Уменьшение размерности (в том числе, PCA)**
4. Увеличение выборки

	x1	x2	x3	y
0	0.44	0.62	0.51	−0.25
1	0.03	0.53	0.07	−0.51
2	0.55	0.13	0.43	0.41
3	0.44	0.51	0.10	0.04
4	0.42	0.18	0.13	0.12
5	0.33	0.79	0.60	−0.45

→

	x1−x2	y
0	−0.18	−0.25
1	−0.50	−0.51
2	0.42	0.41
3	−0.07	0.04
4	0.24	0.12
5	−0.46	−0.45

Обоснование необходимости аналогично селекции

Проблема вырожденности / плохой обусловленности матрицы

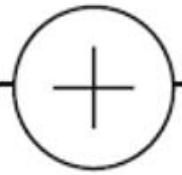
$$w = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$$

Решения:

1. Регуляризация
2. Селекция (отбор) признаков
3. Уменьшение размерности (в том числе, PCA)
4. **Увеличение выборки**

На модельном примере:

$$m \le n \quad \Rightarrow \quad \mathrm{rg}(X^{\mathrm{T}}X) = n$$



При увеличении выборки могут исчезнуть линейные зависимости между столбцами

Оптимизация: градиентный спуск: проблема выбора темпа

$$\frac{1}{2}\sum_{i=1}^{m}(a(x_i \mid w) - y_i)^2 \to \min \qquad a(x \mid w) = w^{\mathrm{T}}x$$

$\nabla f(w_0)$ – направление наискорейшего возрастания функции

$$f(w) = f(w_0) + (w - w_0)^{\mathrm{T}}\nabla f(w_0) + o(\| w - w_0 \|)$$

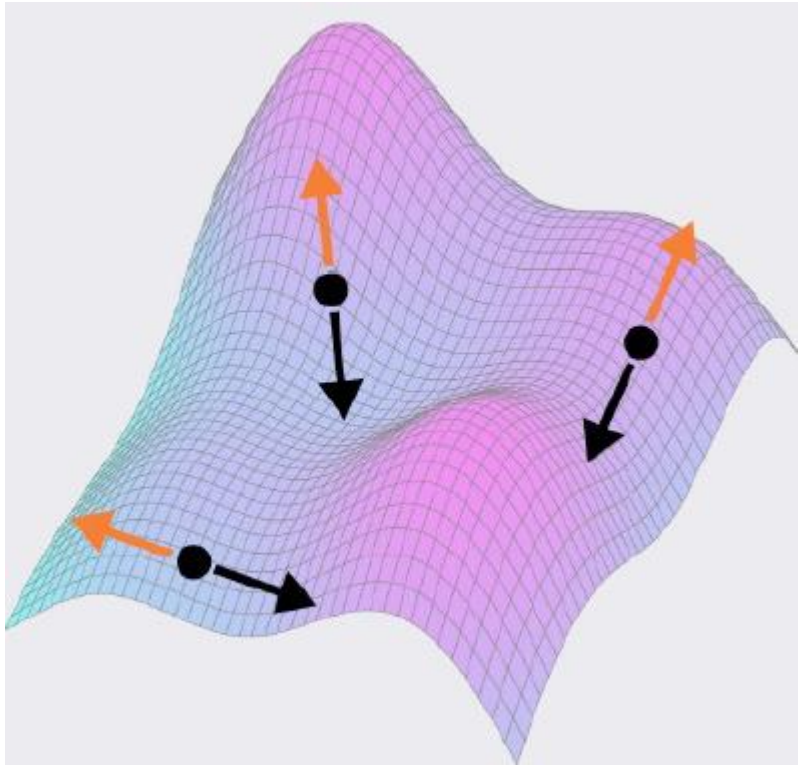
$$f(w) - f(w_0) \approx (w - w_0)^{\mathrm{T}}\nabla f(w_0)$$

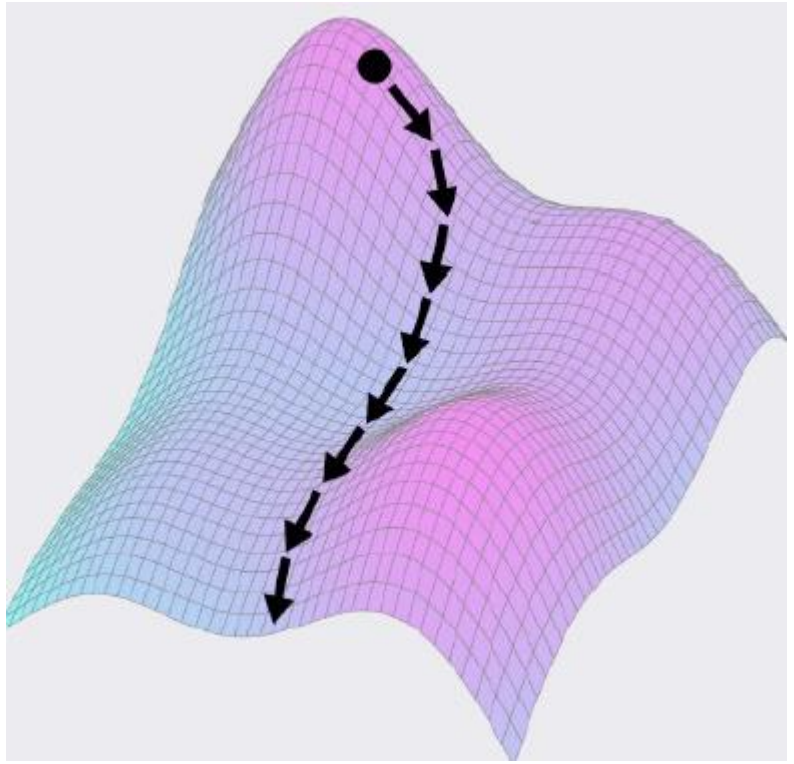
Если выбирать из всех векторов $w - w_0$ единичной нормы, то по неравенству

$$|(w - w_0)^{\mathrm{T}}\nabla f(w_0)| \le 1 \| \nabla f(w_0) \| = \frac{\nabla f(w_0)^{\mathrm{T}}}{\| \nabla f(w_0) \|}\nabla f(w_0)$$

Антиградиент $(-\nabla f(w_0))$ – направление наискорейшего убывания функции

Оптимизация: градиентный спуск: проблема выбора темпа



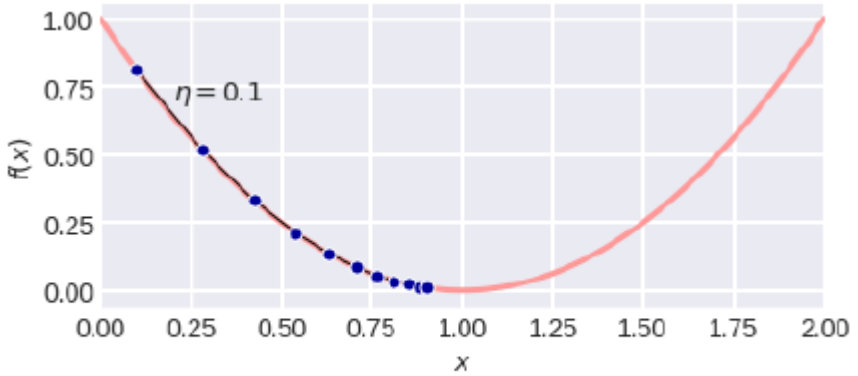


Оптимизация: градиентный спуск: проблема выбора темпа

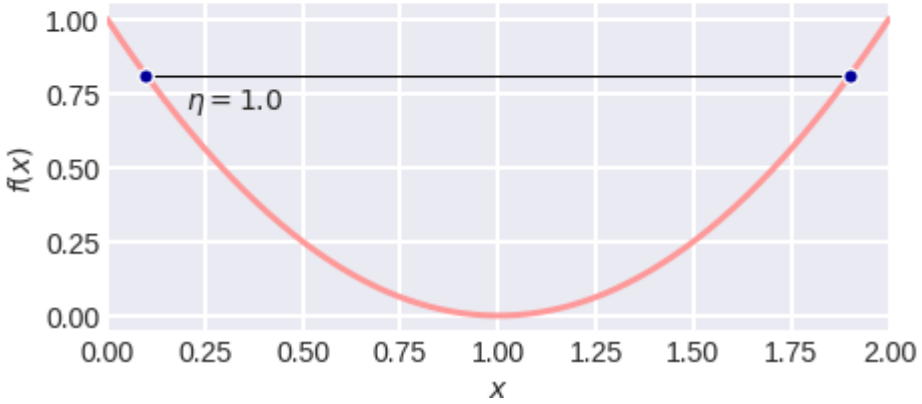
$$w^{(t+1)} = w^{(t)} - \eta \nabla L(w^{(t)})$$

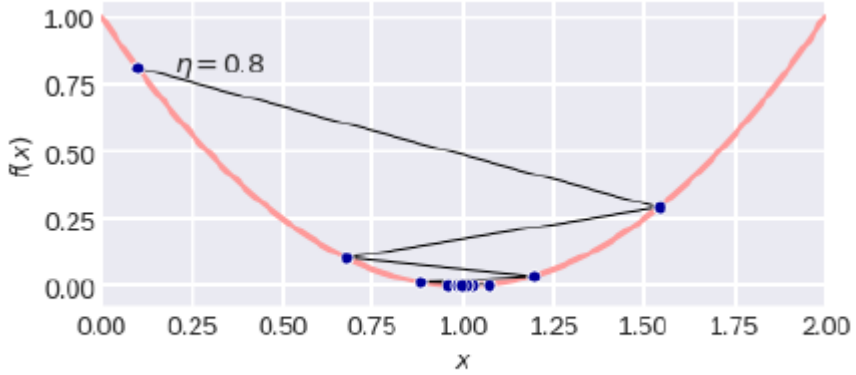
$\eta > 0$ – **шаг / темп обучения**
(step size / learning rate)

Хотим $\lim_{t \to \infty} w^{(t)} = \arg\min_{w} L(w)$



Темп, возможно, маленький

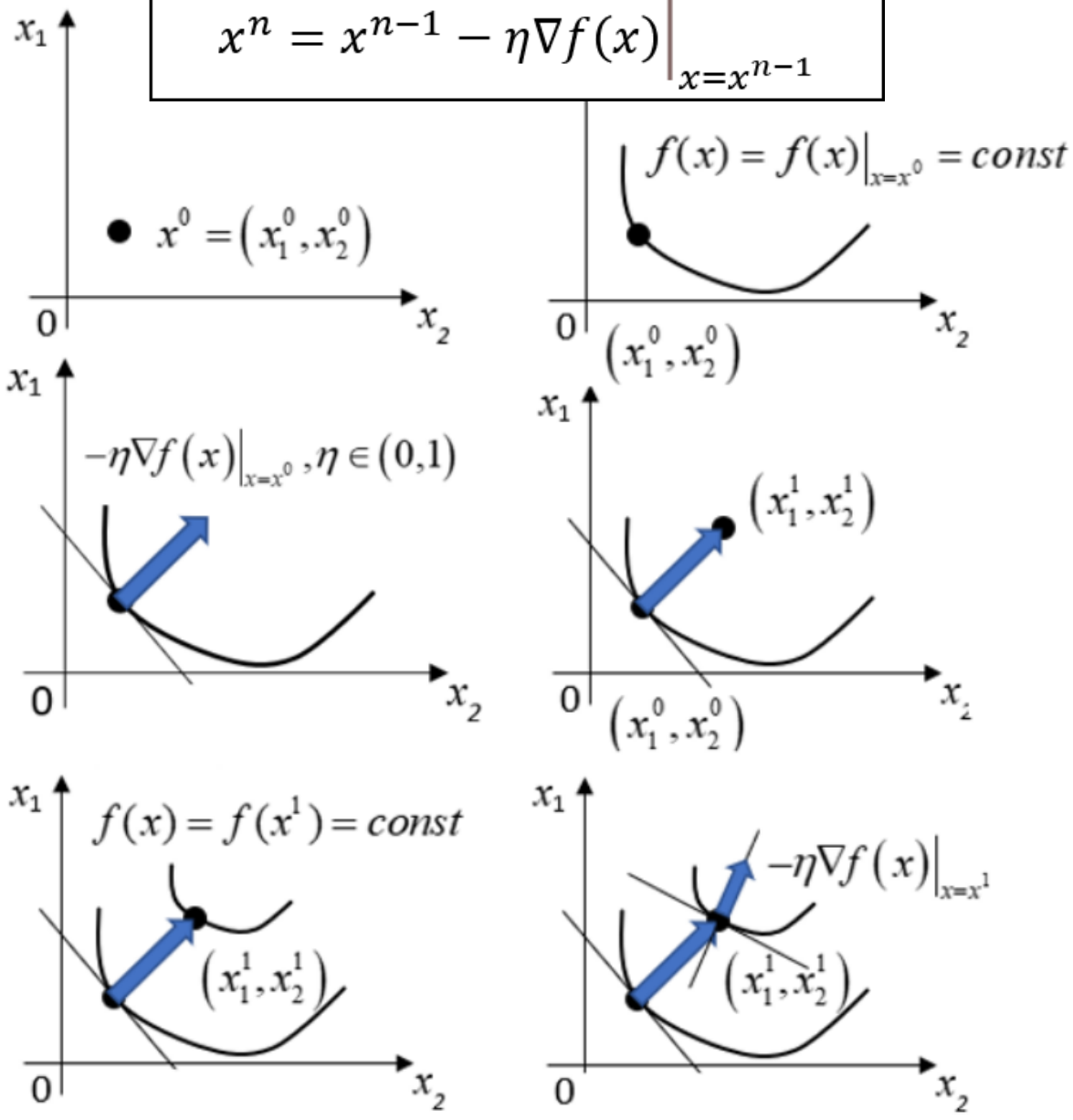


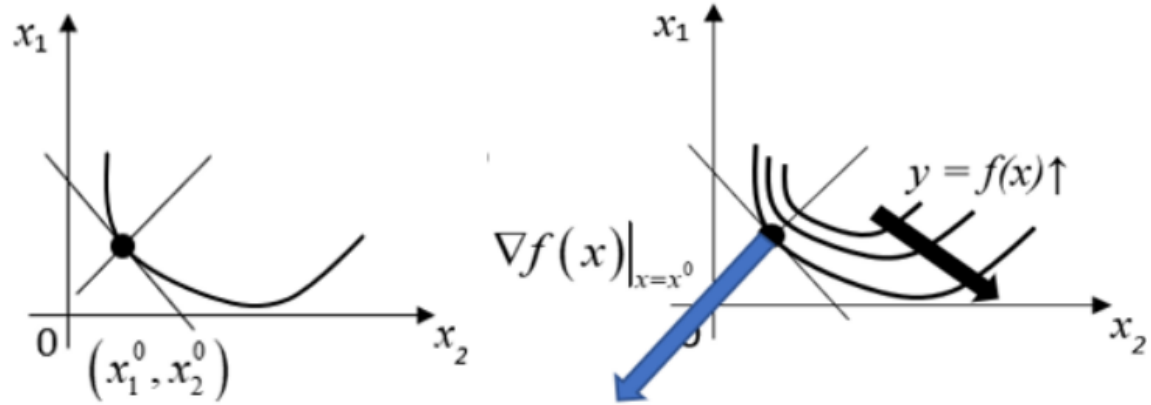


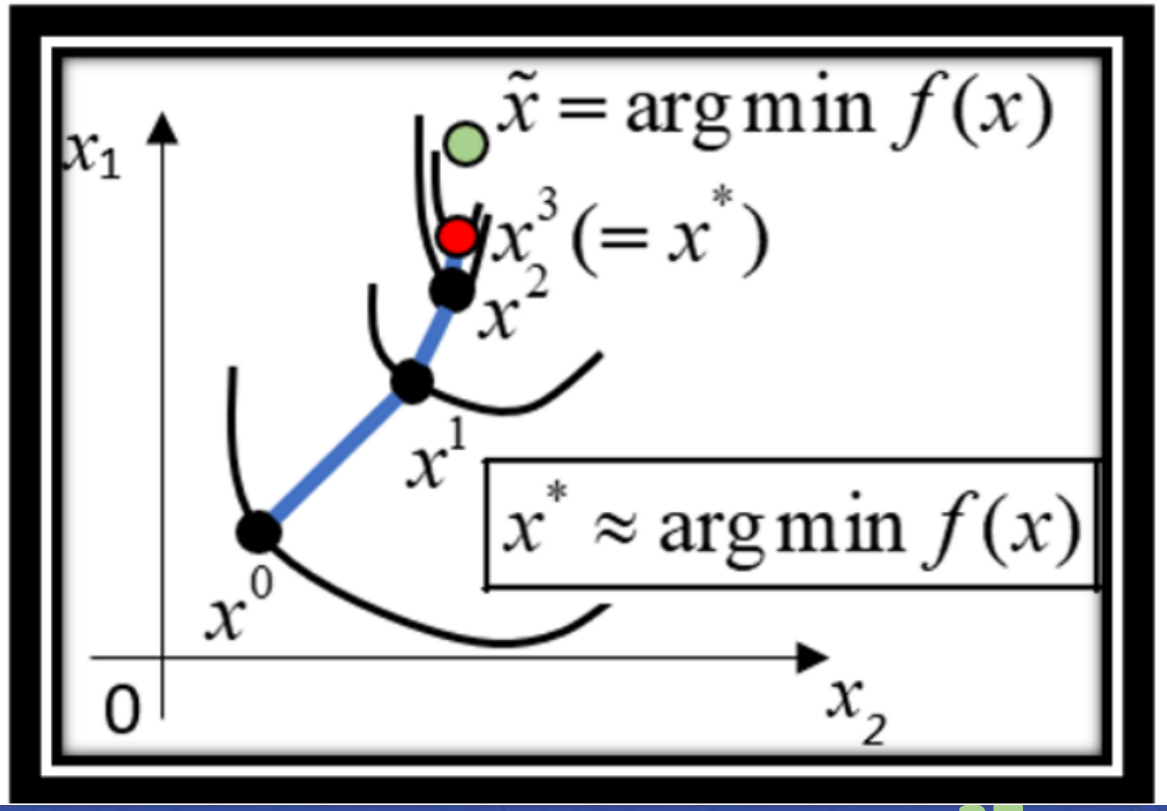
Темп, возможно, большой

Диаграмма последовательных приближений и линий уровня

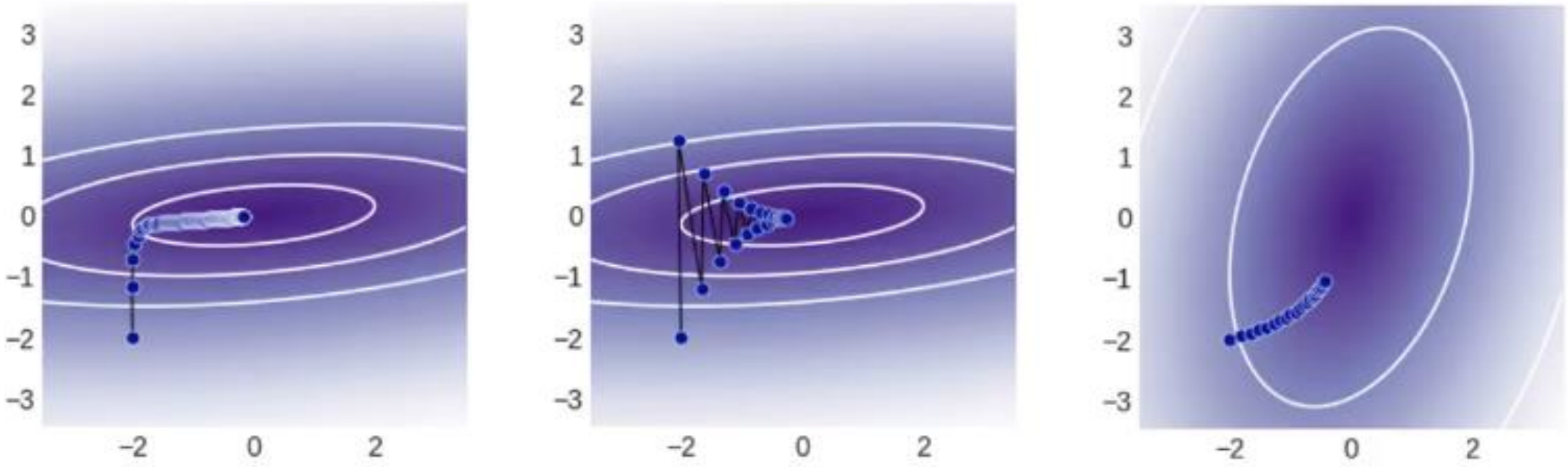
$$x^n = x^{n-1} - \eta \nabla f(x)\Big|_{x=x^{n-1}}$$







Оптимизация: проблема масштаба признаков



Вот для чего нормируют признаки

Линейная регрессия: градиентный метод обучения

$$L(w) = \frac{1}{2}\sum_{i=1}^{m}(a(x_i \mid w) - y_i)^2 \to \min$$

$$a(x \mid w) = w^{\mathrm{T}}x$$

$$L(w) = \sum_{t=1}^{m} L_t(w)$$

$$\nabla L(w) = \sum_{t=1}^{m} \nabla L_t(w)$$

$$w^{(t+1)} = w^{(t)} - \eta \nabla L_i(w^{(t)})$$

$$w^{(t+1)} = w^{(t)} - \eta\sum_{i=1}^{m}(a(x_i \mid w^{(t)}) - y_i)\frac{\partial a(x_i \mid w^{(t)})}{\partial w}$$

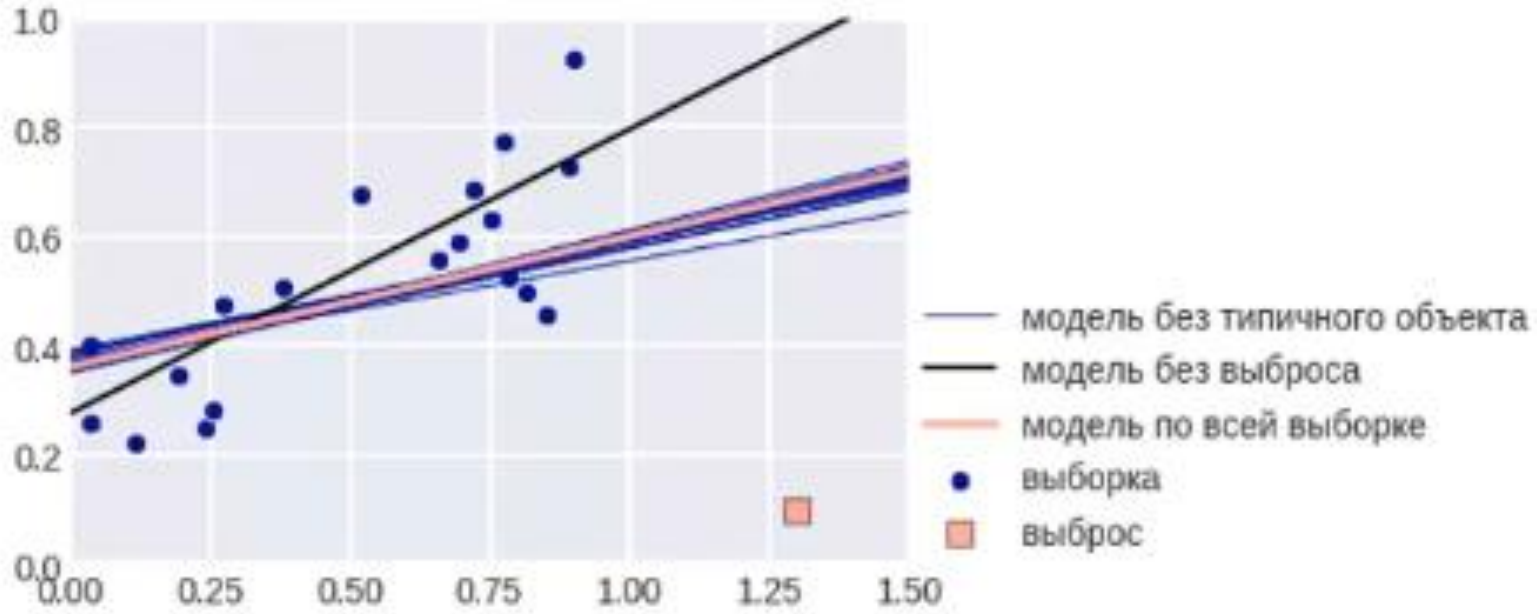
$$w^{(t+1)} = w^{(t)} - \eta\sum_{i=1}^{m}(a(x_i \mid w^{(t)}) - y_i)x_i$$

Gradient Descent

$$w^{(t+1)} = w^{(t)} - \eta_t(a(x_i \mid w^{(t)}) - y_i)x_i$$

Stochastic Gradient Descent

Неустойчивость к выбросам

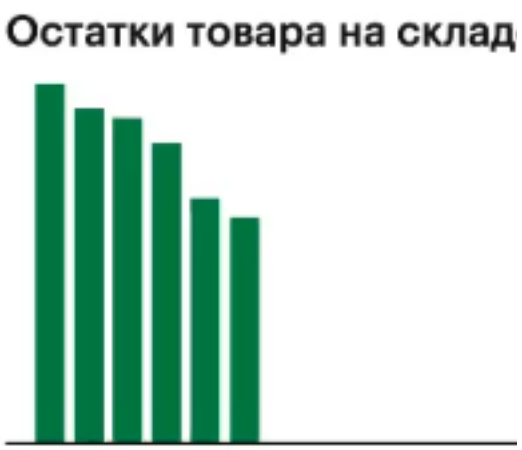


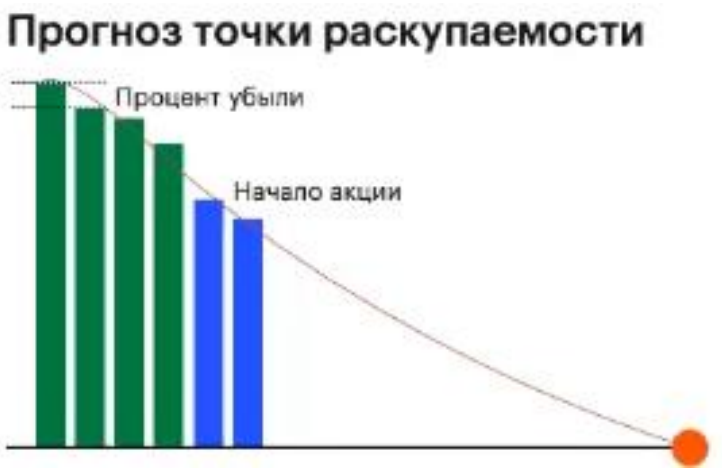
- удаление выбросов
- устойчивая регрессия (ошибки с весами)

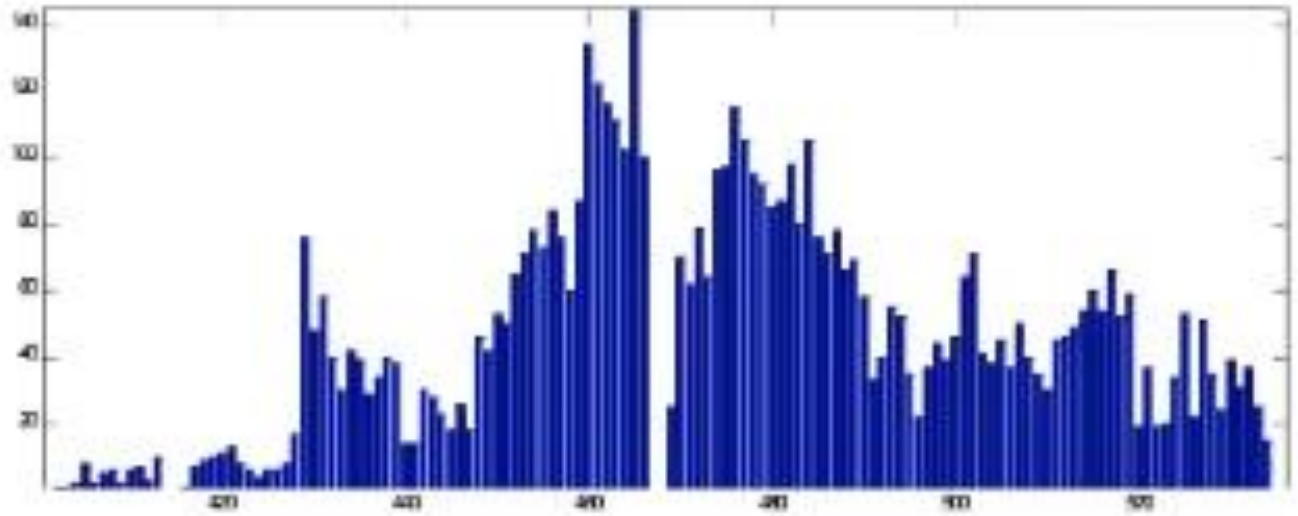
Прогнозирование спроса, точки раскупаемости

- # покупок за k дней
- # просмотров за k дней
- # корзин за k дней
- # дней без покупок
- изменение цены за последние k дней
- есть ли маркетинговая акция

$$Y = \max\left[\sum_t w_t X_t, 0\right]$$

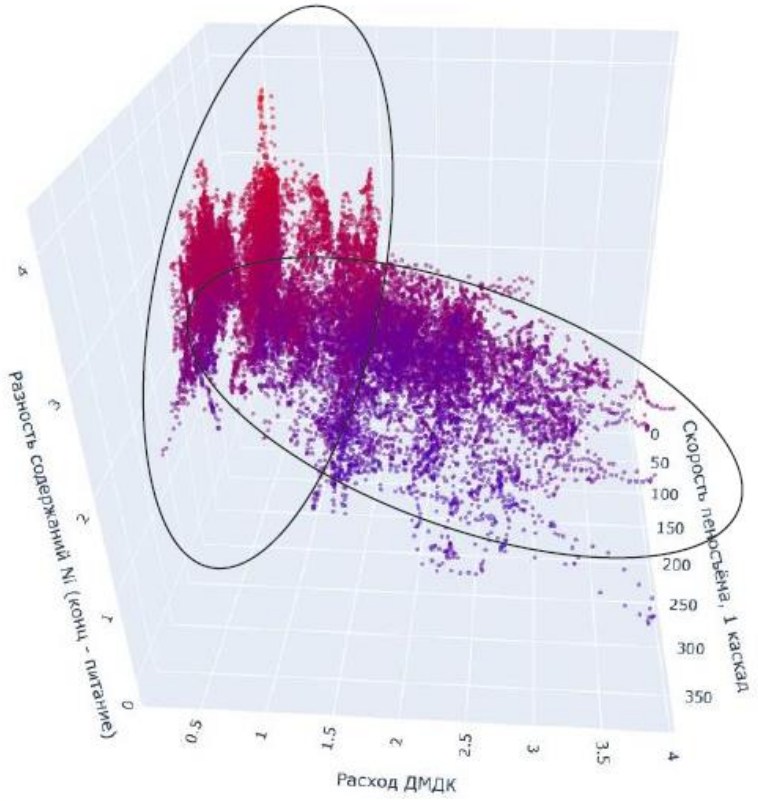


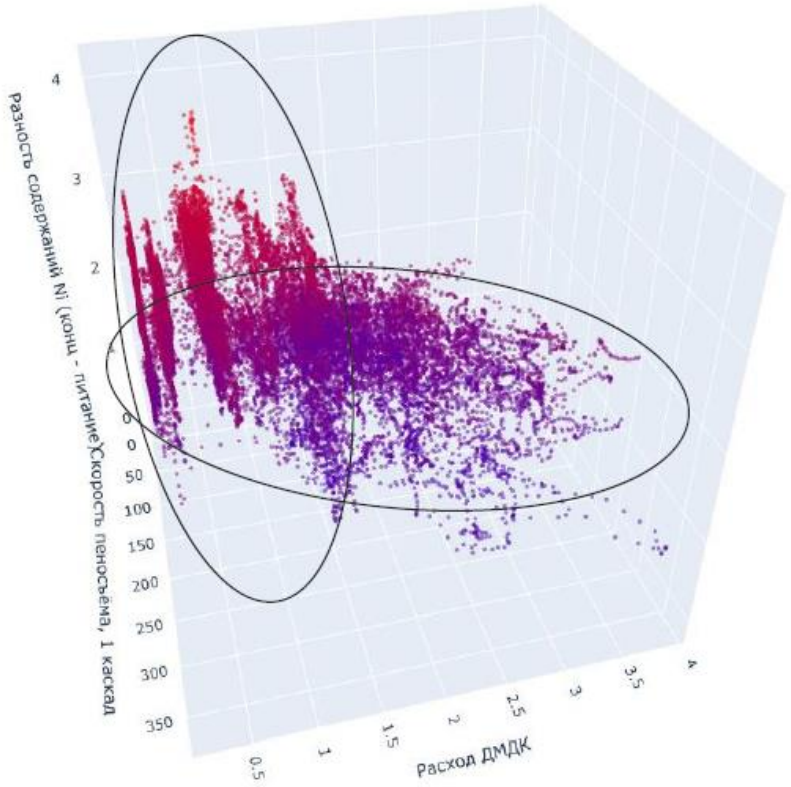


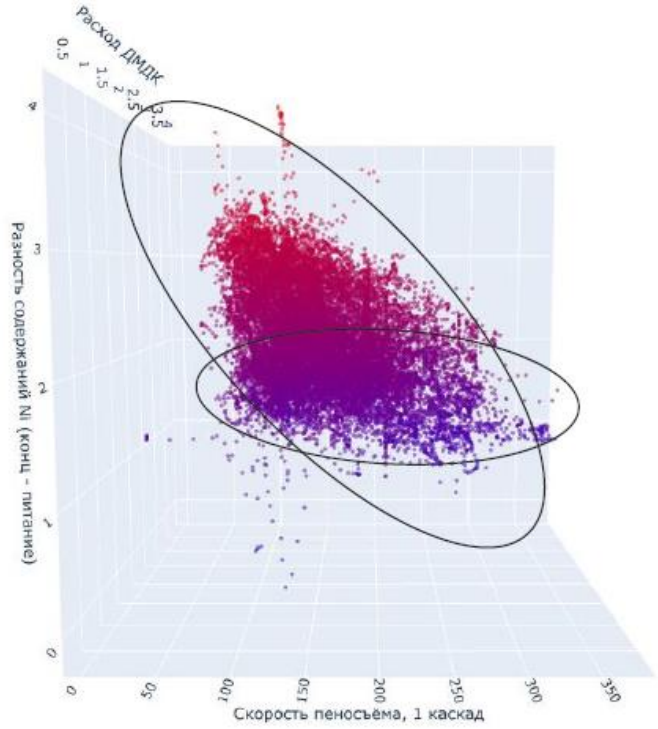


Кол-во купленных единиц по дням. Падает спрос?

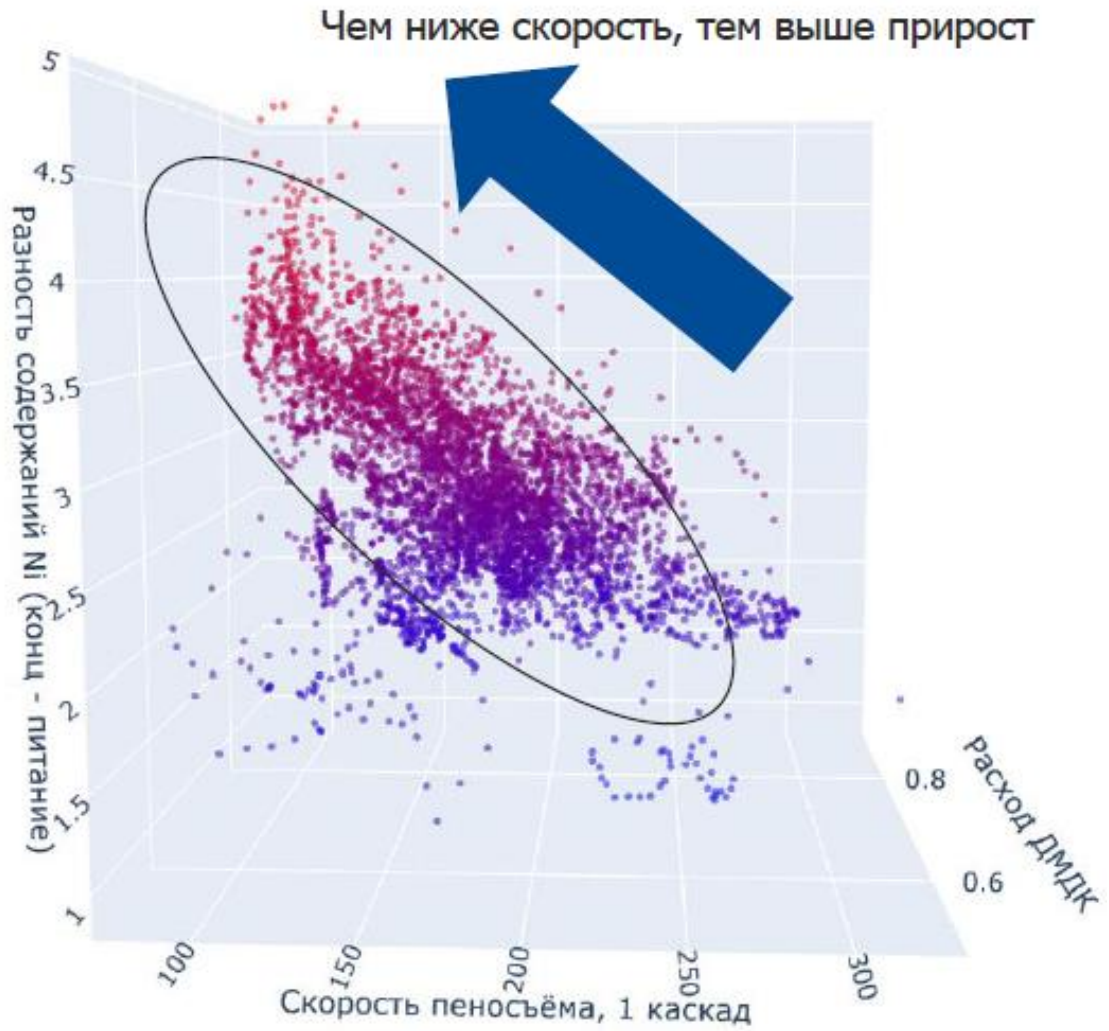
«Эффективные» диапазоны. Пример флотации







«Эффективные» диапазоны. Пример флотации

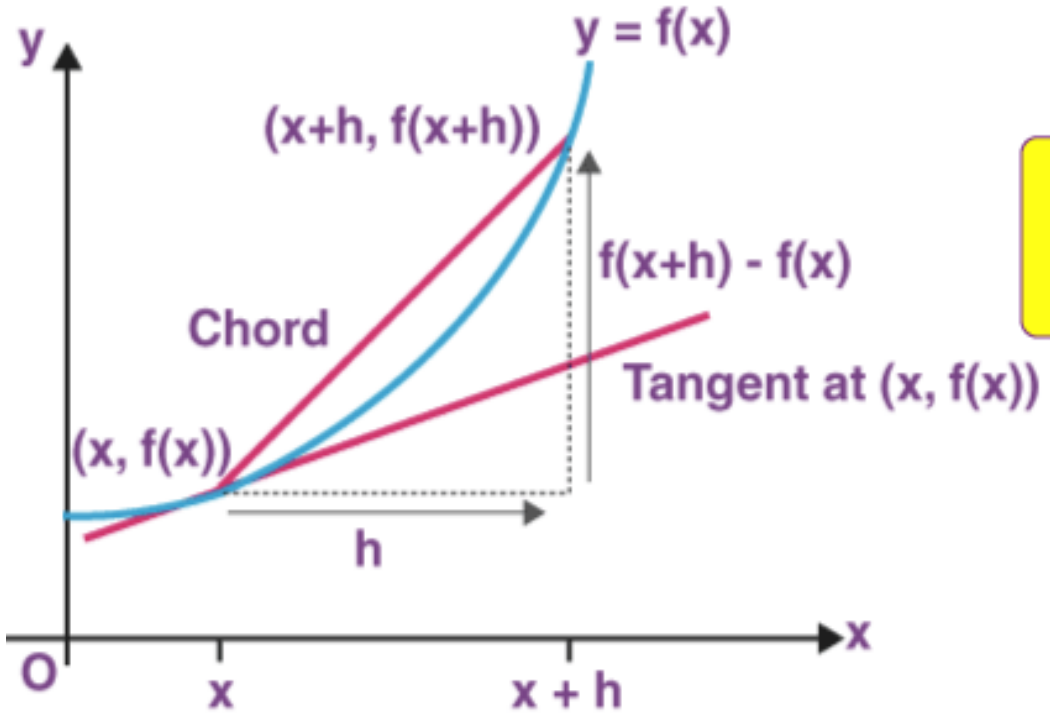


Плюсы и минусы линейных алгоритмов

+	простой, надёжный, быстрый, популярный метод
+	интерпретируемость (нахождение закономерностей)
+	интерполяция и экстраполяция
+	может быть добавлена нелинейность, с помощью генерации новых признаков
+	хорош для теоретических исследований (в Ridge есть явная формула)
+	коэффициенты асимптотически нормальны (можно тестировать гипотезы о влиянии признаков)
+	глобальный минимум в оптимизируемом функционале

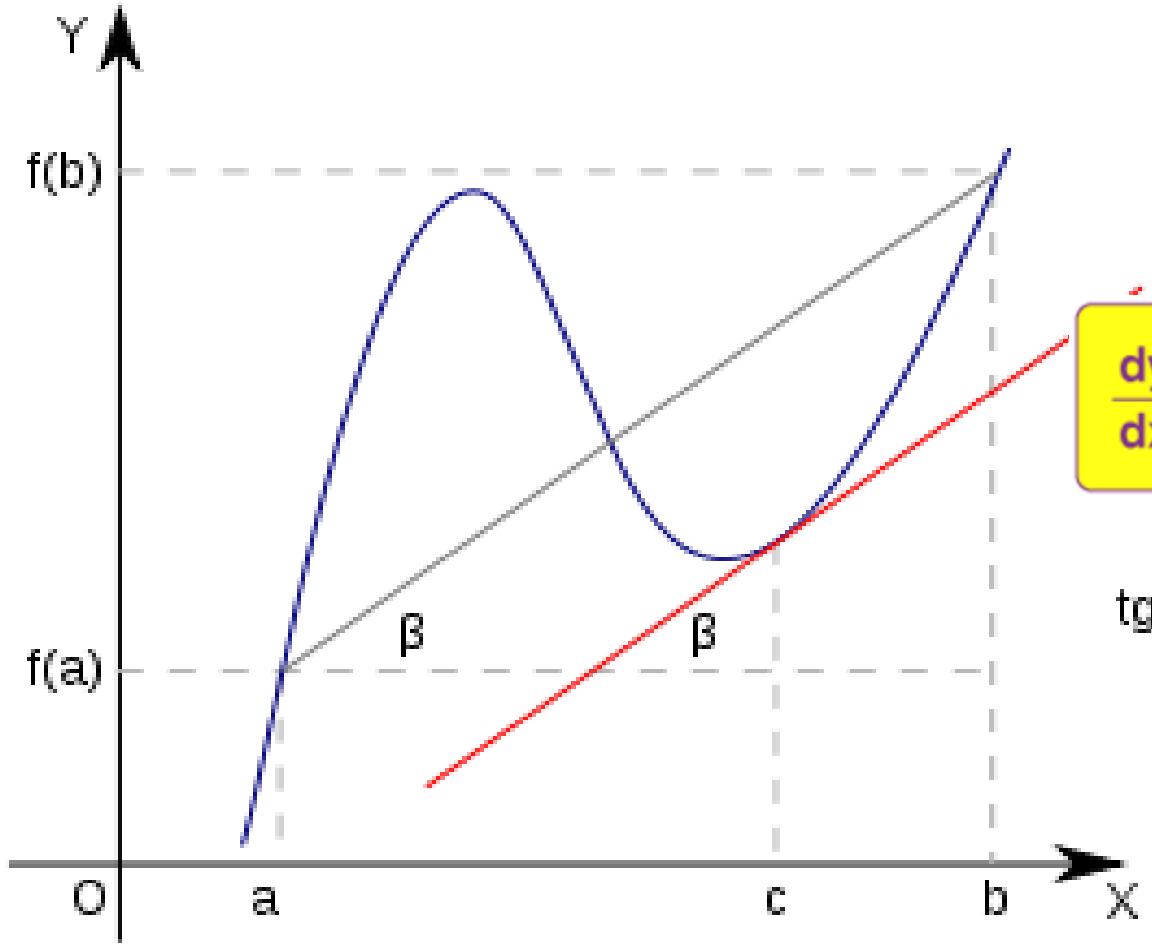
–	линейная гипотеза вряд ли верна
–	в теоретическом обосновании ещё предполагается нормальность ошибок (зависит от функции ошибок)
–	«страдает» из-за выбросов
–	признаки в одной шкале и однородные (см. успешные примеры)
–	проблема коррелированных признаков → необходимость регуляризации, селекции, PCA, data↑

Производная – это разность
Производная – это тангенс угла
Производная – это скорость



$$\frac{dy}{dx} = \lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$$

Вычисляем производную без предела (теорема Лагранжа)

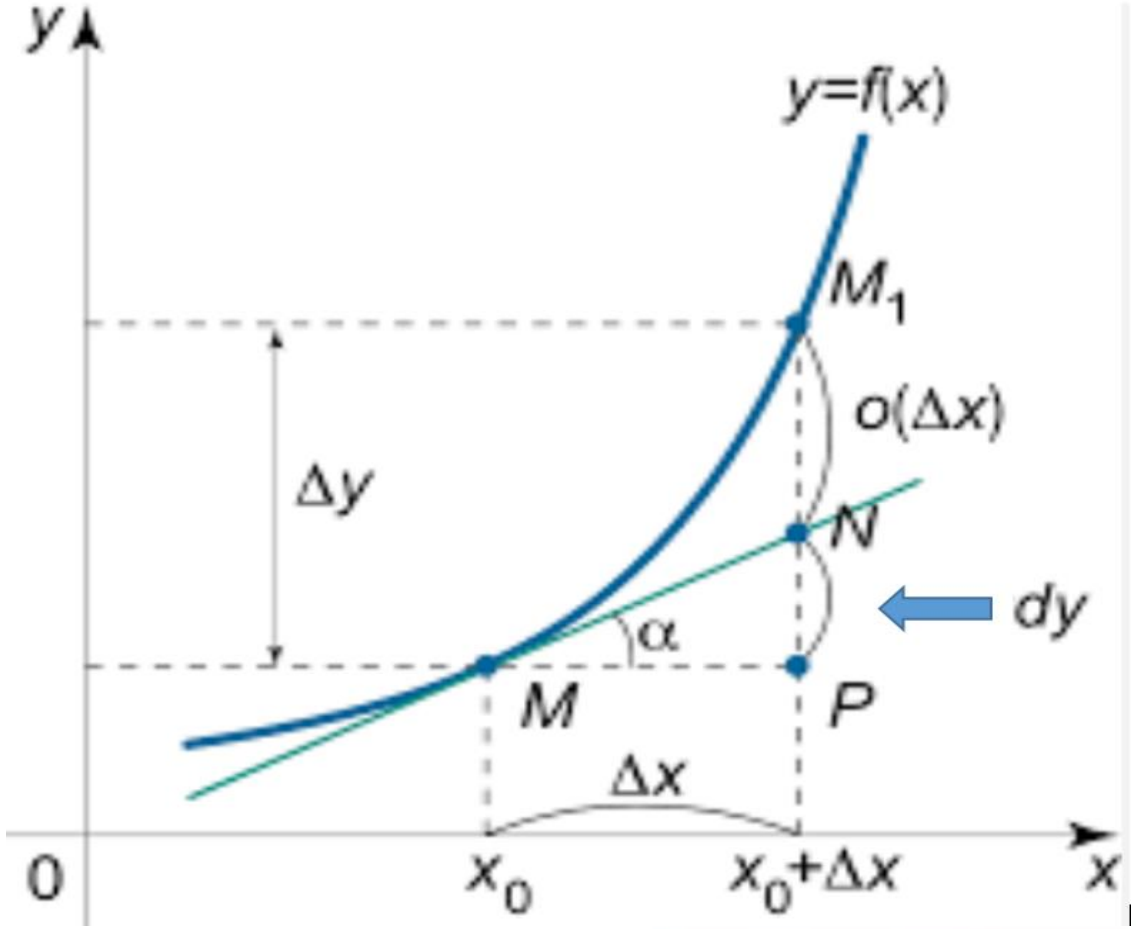


$$\frac{dy}{dx} = \lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$$

$$\operatorname{tg} \beta = \frac{f(b)-f(a)}{b-a} = f'(c)$$

В малом все процессы линейные

The rate of change of function is called a **derivative** and **differential** is the actual change of linear model



Спасибо за внимание!

Запорожцев Иван Федорович
zaporozhtsev.if.work@gmail.com