

Введение в машинное обучение

Екатерина Лобачева

научный сотрудник Лаборатории Самсунг-ВШЭ



SAMSUNG
Research



Общая схема машинного обучения

Пример задачи

- Вы хотите купить себе дом
- Имеется несколько вариантов
- Хотим оценить стоимость каждого дома



Обозначения

- x — **объект**, sample — для чего хотим делать предсказания
 - Конкретный дом
- y — ответ, **целевая переменная**, target — что предсказываем
 - Стоимость дома
- Задача: дом → стоимость

Обучающая выборка

- Мы ничего не понимаем в экономике
- Зато имеем много объектов с известными ответами
- $X = (x_i, y_i)_{i=1}^{\ell}$ — обучающая выборка
- ℓ — размер выборки

В нашем случае:

набор домов, проданных в том же городе за последние 2 года

Признаки

- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x^1, \dots, x^d)$ — признаковое описание

Признаки

Какие признаки могут быть использованы в нашем примере?

- Информация о самом доме:
 - Площадь
 - Год постройки
 - ...
- Информация о районе/местоположении:
 - Удаленность от центра
 - Рейтинг безопасности района
 - Уровень экологичности района
 - ...

Алгоритм

- $a(x)$ — алгоритм, модель — функция, предсказывающая ответ для любого объекта
- Линейная модель: $a(x) = w_0 + w_1x_1 + \dots + w_dx_d$
- Например:

$$a(x) = 1.000.000 + 100.000 * (\text{площадь}) \\ - 100.000 * (\text{расстояние до метро})$$

ФУНКЦИЯ ПОТЕРЬ

- Не все алгоритмы полезны
- $a(x) = 0$ — не принесет никакой выгоды
- Функция потерь (или функционал качества) — мера корректности ответа алгоритма
- Предсказали стоимость 15 млн, на самом деле 17 млн — хорошо или плохо?

ФУНКЦИЯ ПОТЕРЬ

- Функция потерь (или функционал качества) — мера корректности ответа алгоритма
- Среднеквадратичная ошибка (Mean Squared Error, MSE):
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$
- Чем меньше, тем лучше
- Должна соответствовать бизнес-требованиям
- Одна из самых важных составляющих анализа данных

Обучение алгоритма

- Есть обучающая выборка и функция потерь
- Семейство алгоритмов \mathcal{A}
 - Из чего выбираем алгоритм
 - Пример: все линейные модели
 - $\mathcal{A} = \{w_0 + w_1x_1 + \dots + w_dx_d \mid w_0, w_1, \dots, w_d \in \mathbb{R}\}$
- Обучение: поиск оптимального алгоритма с точки зрения значения функции потерь на обучающей выборке
- В результате получаем алгоритм, который может делать предсказания для новых объектов.

А вдруг модель просто запомнит примеры?

- Мы хотим, чтобы она нашла зависимости в виде формулы
- А может просто запомнить:
 - Дом именно с такими характеристиками стоит столько-то
 - Не работает на немного других домах
 - Это эффект **переобучения**
- Как проверить?
- **Надо проверять на домах, которые она не видела!**

Обобщающая способность алгоритма

- Перед обучением от данных отделяется **отложенная** выборка:



- Слишком большое обучение — тестовая выборка нерепрезентативна
- Слишком большой тест — модель не сможет обучиться
- Обычно: 70/30, 80/20

Схема работы машинного обучения

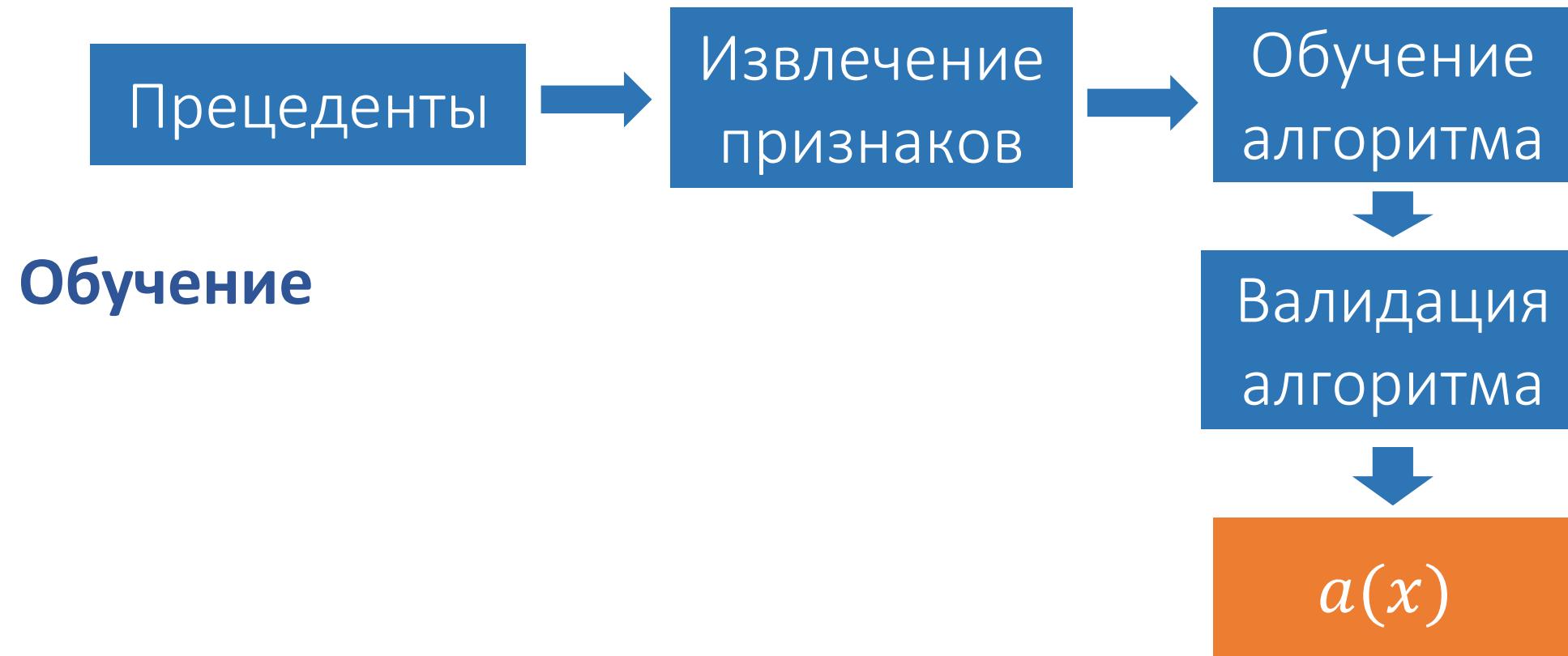
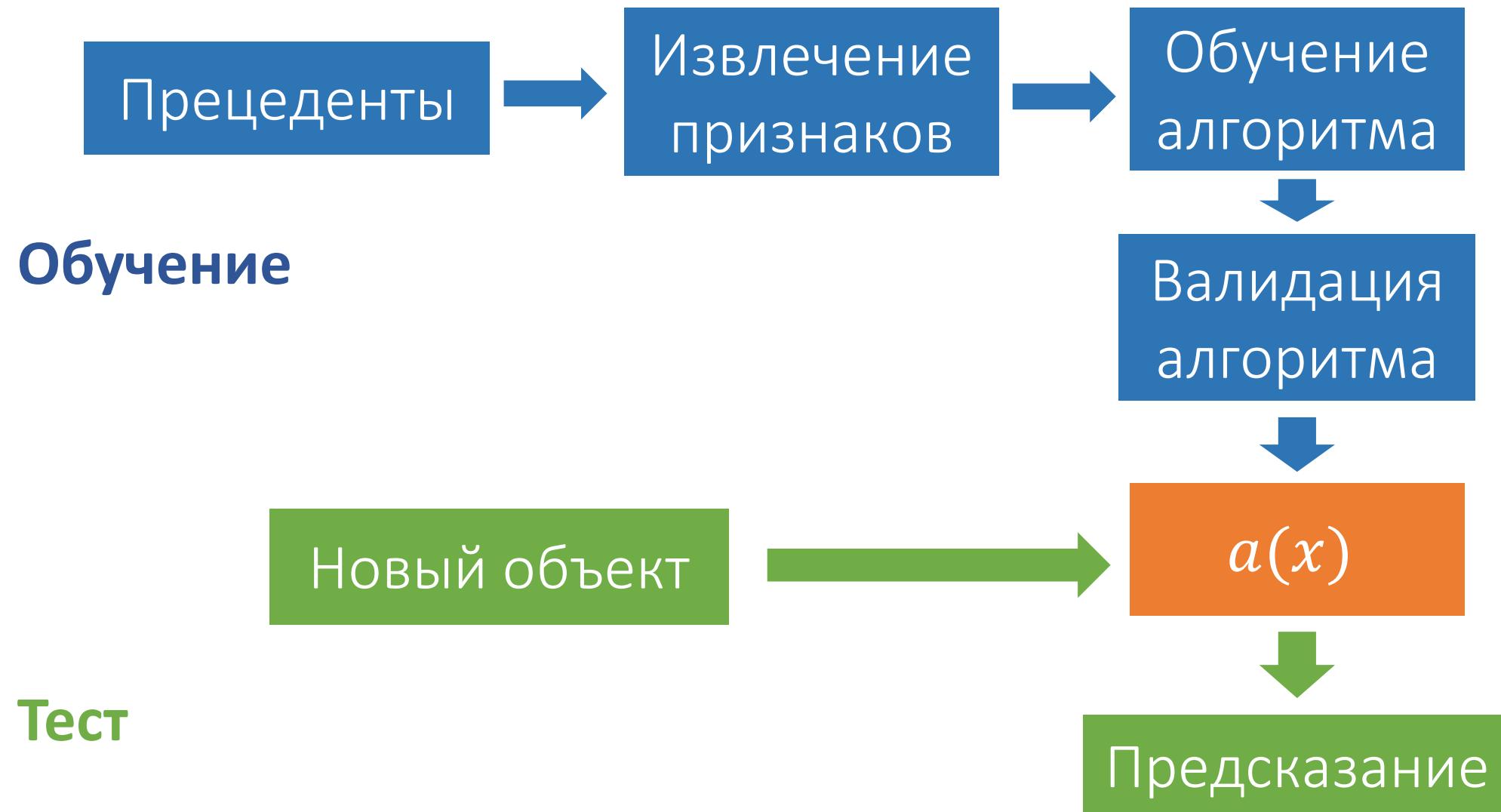


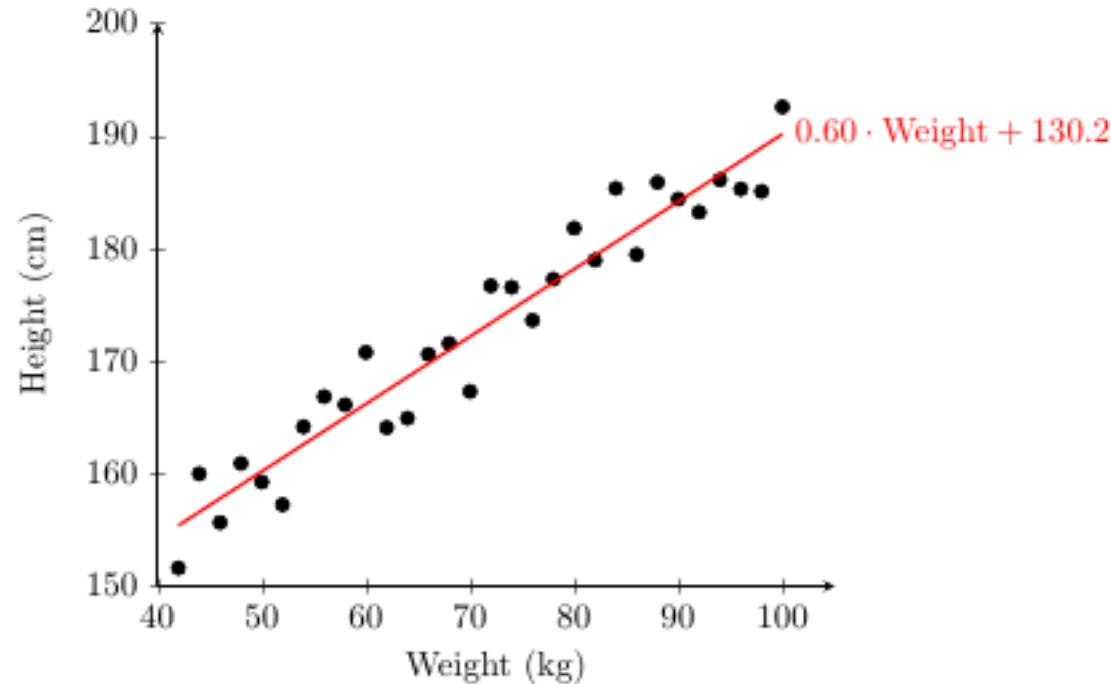
Схема работы машинного обучения



Виды и примеры задач

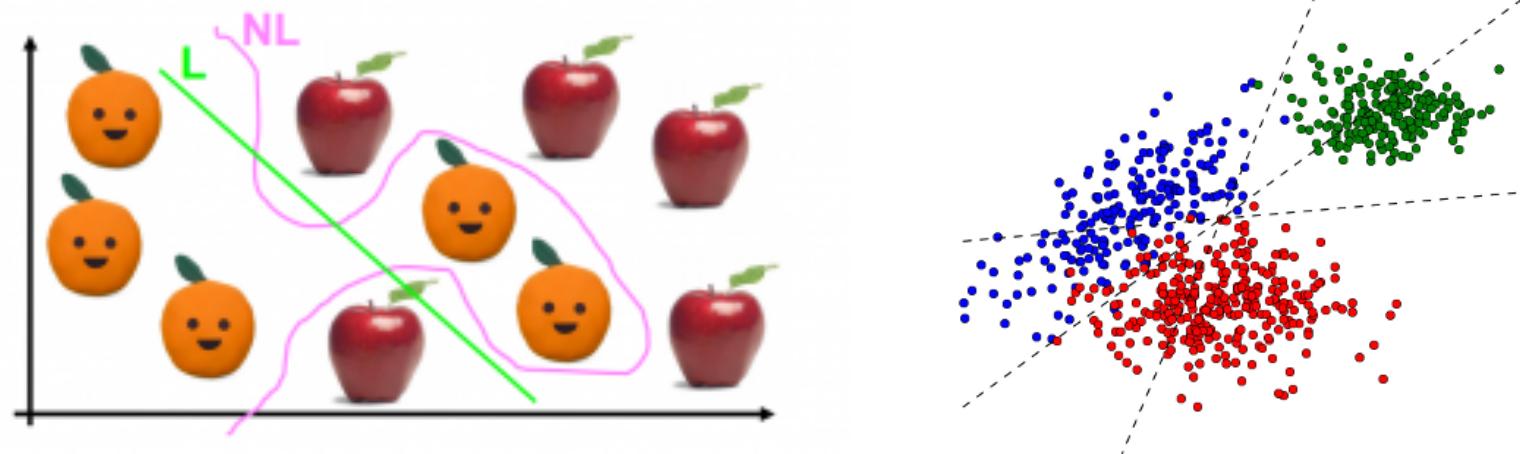
Регрессия

- Вещественные ответы: $\mathbb{Y} = \mathbb{R}$
- (вещественные числа — числа с любой дробной частью)
- Примеры: предсказание зарплат, стоимости недвижимости, объема нефтедобычи...



Классификация

- Конечное число ответов: $|\mathbb{Y}| < \infty$
- Бинарная классификация: $\mathbb{Y} = \{-1, +1\}$
- Многоклассовая классификация: $\mathbb{Y} = \{1, 2, \dots, K\}$
- Примеры: кредитный scoring, предсказание оттока, категоризация писем, определение объекта на фото...



Кластеризация

- \mathbb{Y} — отсутствует
- Нужно найти группы похожих объектов
- Сколько таких групп?
- Как измерить качество?
- Пример: сегментация пользователей мобильного оператора

Ранжирование

- Нужно не предсказать величину, а правильно упорядочить объекты
- Примеры:
 - Ранжирование поисковой выдачи
 - Ранжирование товаров в рекомендательной системе

Яндекс

картинки с котиками — 5 млн ответов

Х Найти

Поиск

Картинки

Видео

Карты

Маркет

Ещё

Картинки с кошками | Fun Cats — Забавные коты
funcats.by > pictures/ ▾
Картинки с кошками. Прикольные коты. 777 изображений. ... 32 изображения. Кошки Стамбула. 41 изображение. Веселые котята.

Уморные котики (57 фото) » Бяки.нет | Картинки
byaki.net > Картинки > 14026-umornye-kotiki-57... ▾
Бяки нет! . NET. Уморные котики (57 фото). 223. Комментариев:9Автор:4ertonok Просмотров:161 395 Картинки28-10-2008, 00:03.

Смешные картинки кошек с надписями | Лолкот.Ру
lolkot.ru ▾
Смешные картинки для новых приколов! Сделать свой прикол очень просто. ... Котик верит в чудеса. Он в носке подарок ищет...

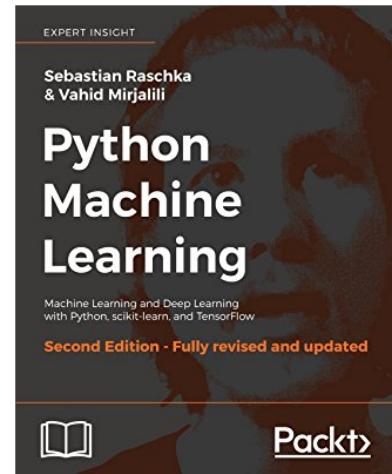
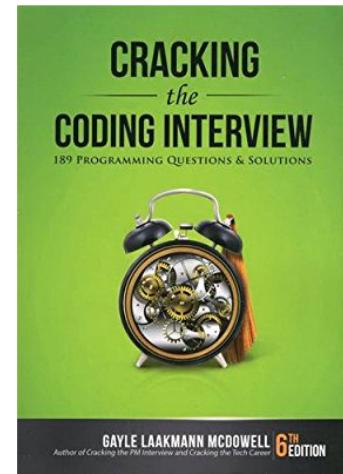
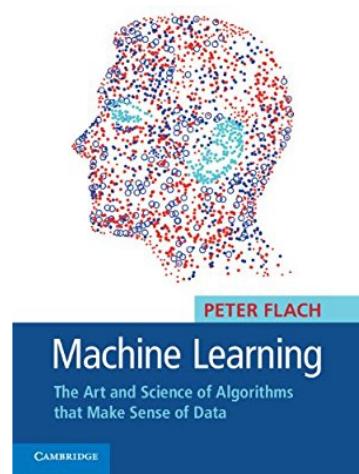
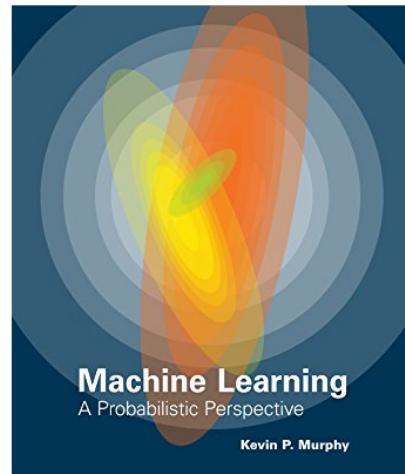
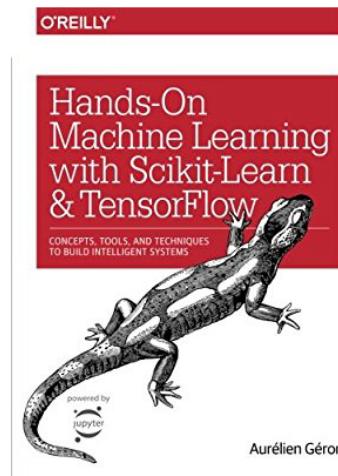
Красивые картинки и фото кошек, котят и котов
foto-zverey.ru > Кошки ▾
Фото и картинки кошек и котят потрясающей красоты и нежности. Здесь мы собрали такие изображения, которые всегда вызывают море положительных эмоций...

Обои для рабочего стола Котята | картинки на стол Котята
7fon.ru > Чёрные обои и картинки > Обои котята ▾
Картинки Котята с 1 по 15. Обои для рабочего стола Котята. ... Скачать Картинки Котята на рабочий стол бесплатно.

Рекомендательные системы

- Полки рекомендаций на Amazon генерируют 35% от всех покупок
- Рекомендации на основе машинного обучения и анализа больших объёмов данных

Recommendations for you in Books

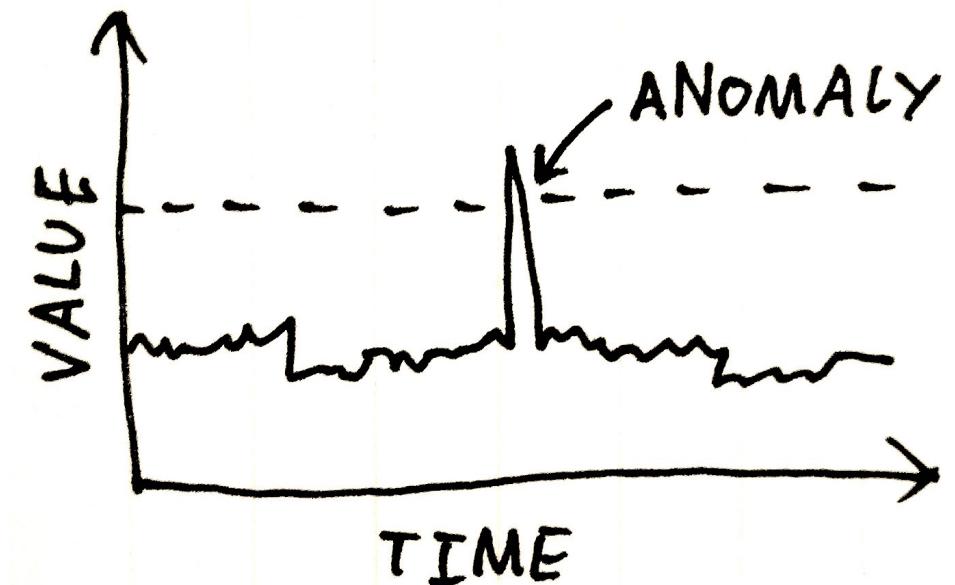


Обнаружение аномалий

- Задача — найти объекты, которые являются нестандартными, выбиваются из общего распределения

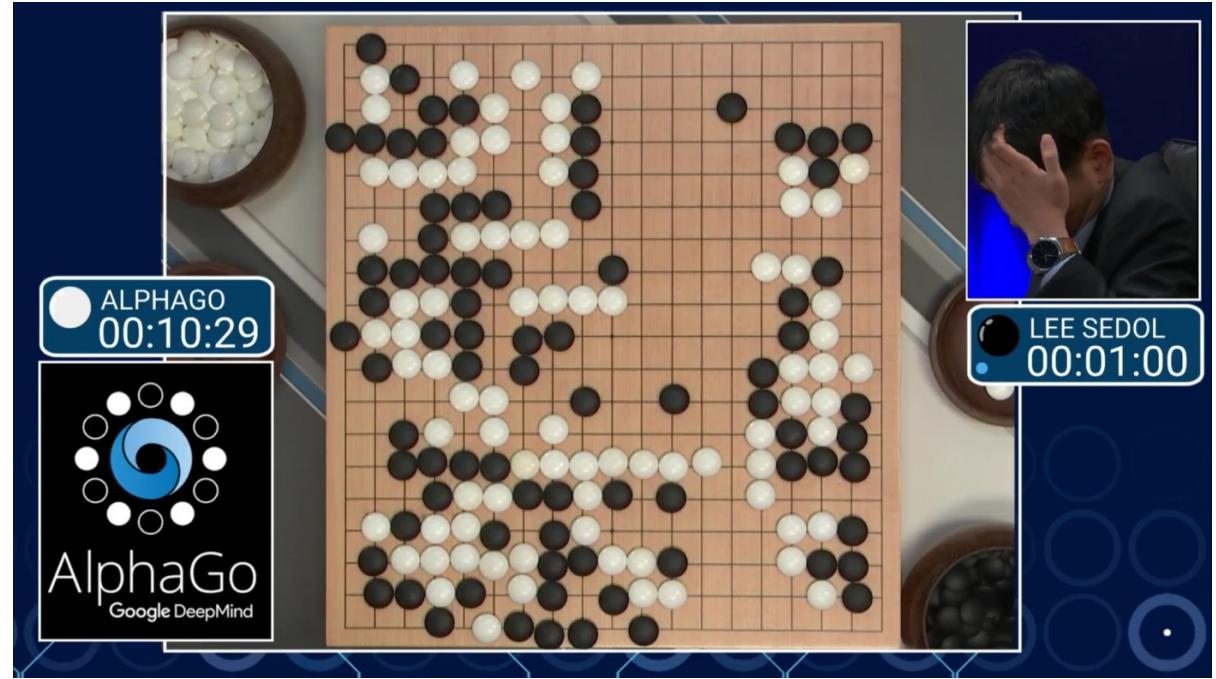
Примеры:

- Обнаружение мошеннических транзакций
- Раннее обнаружение поломок в системах самолёта или автомобиля



AlphaGo

- Модель для игры в Го
- Оценивает успешность хода
- Обучалась путём игры с собой
- Победила чемпиона мира в 2016 году
- Долгое время игра в Го считалась невозможной задачей для компьютера



Перенос стиля

- Задача: картинка, стиль -> картинка со стилем



Картина



Стиль



Результат

Машинный перевод

- Задача: текст на одном языке -> текст на другом языке

The screenshot shows a machine translation interface with two main sections: a source text area and a target text area.

Source Text (Russian):

Машинное обучение (Machine Learning) – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Различают два типа обучения. Обучение по прецедентам, или индуктивное обучение, основано на выявлении общих закономерностей по частным эмпирическим данным.

Mashinnoye obucheniye (Machine Learning) – obshirnyy podrazdel iskusstvennogo intellekta, izuchayushchiy metody postroyeniya algoritmov, sposobnykh obuchat'sya. Razlichayut dva tipa obucheniya. Obucheniye po pretsedentam, ili induktivnoye

[Развернуть](#)

Target Text (English):

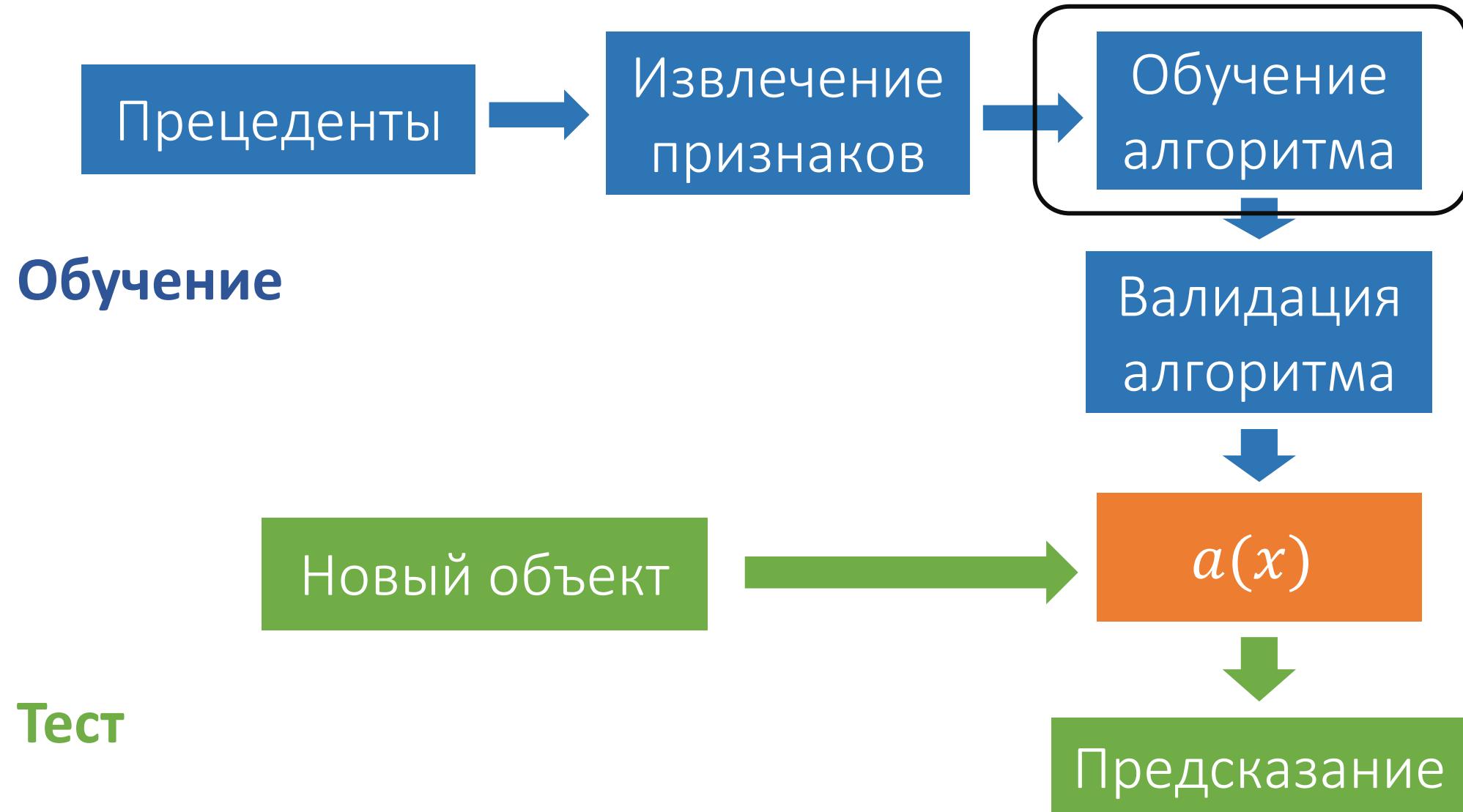
Machine Learning is an extensive subset of artificial intelligence that studies methods for building learning-capable algorithms. There are two types of training. Learning from precedents, or inductive learning, is based on identifying general patterns based on particular empirical data.

Interface Elements:

- Top navigation bar with tabs: ОПРЕДЕЛИТЬ ЯЗЫК, РУССКИЙ, АНГЛИЙСКИЙ, НЕМЕЦКИЙ, ANGЛИЙСКИЙ (highlighted), РУССКИЙ, УКРАИНСКИЙ.
- Central toolbar with icons: microphone, speaker, and document sharing.
- Bottom status bar: 301/5000, Ru, and a dropdown menu.

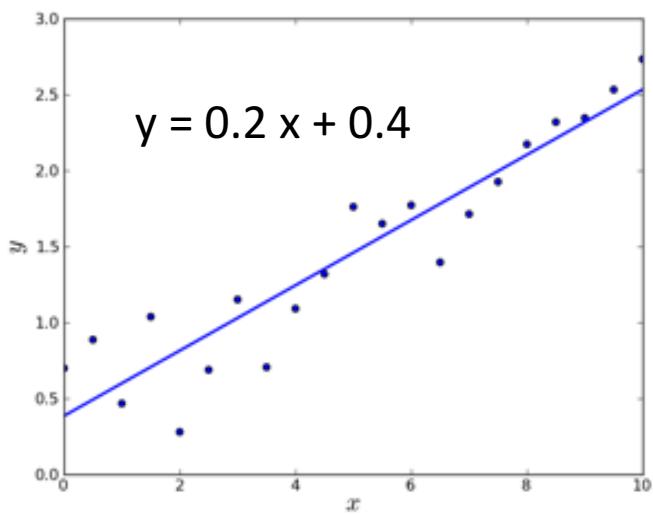
Линейные методы для регрессии и классификации

Схема работы машинного обучения



Основные виды моделей

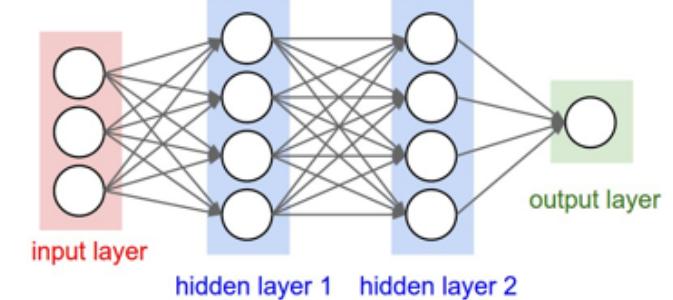
Линейные модели



Решающие деревья и их комбинации

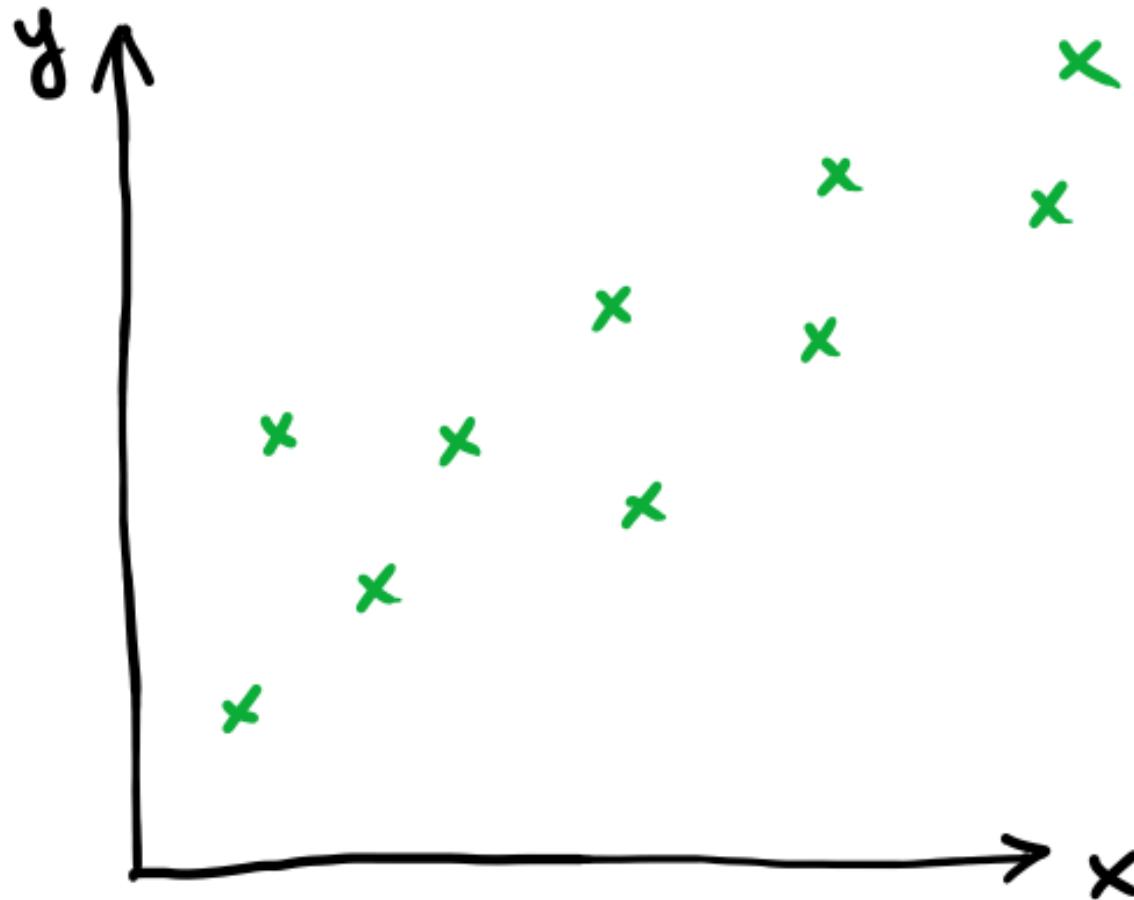


Нейронные сети



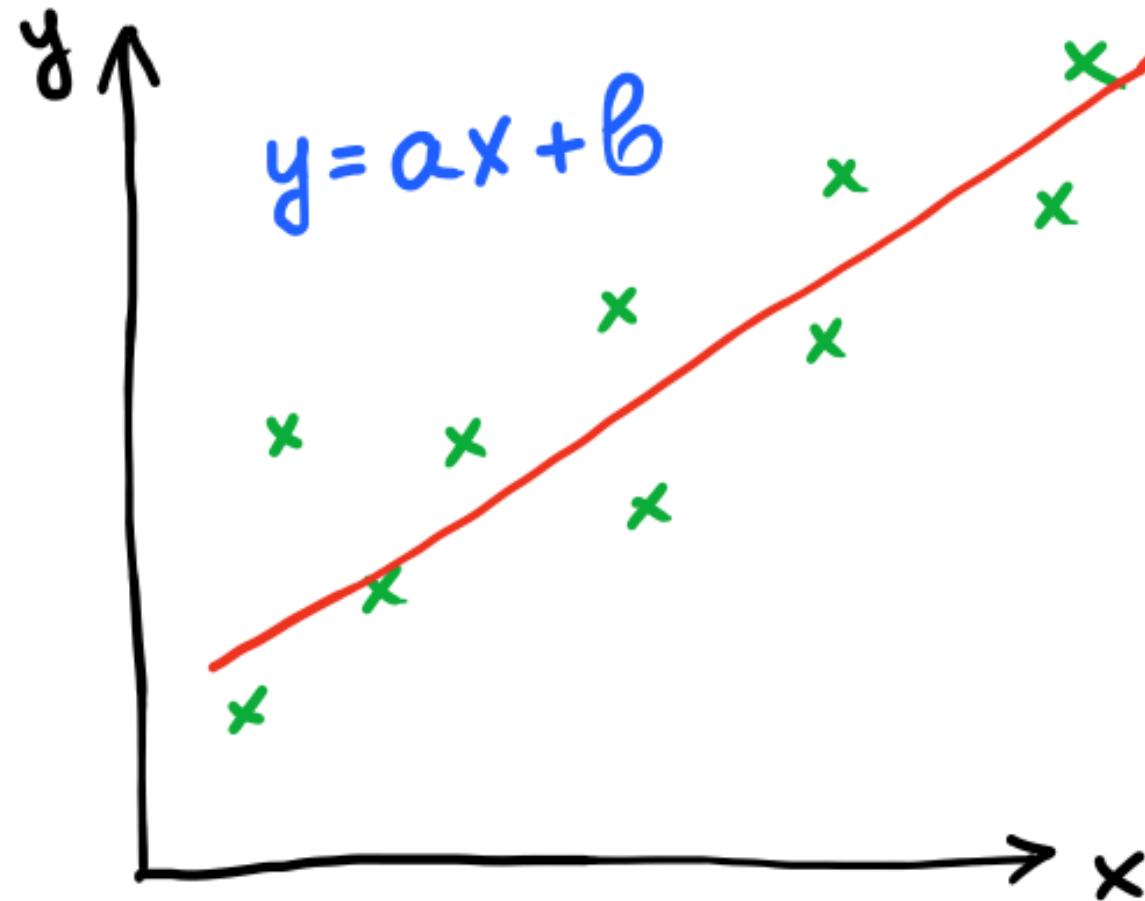
Линейная модель для регрессии

x	y
1	2
3	5
-1	-2
5	?



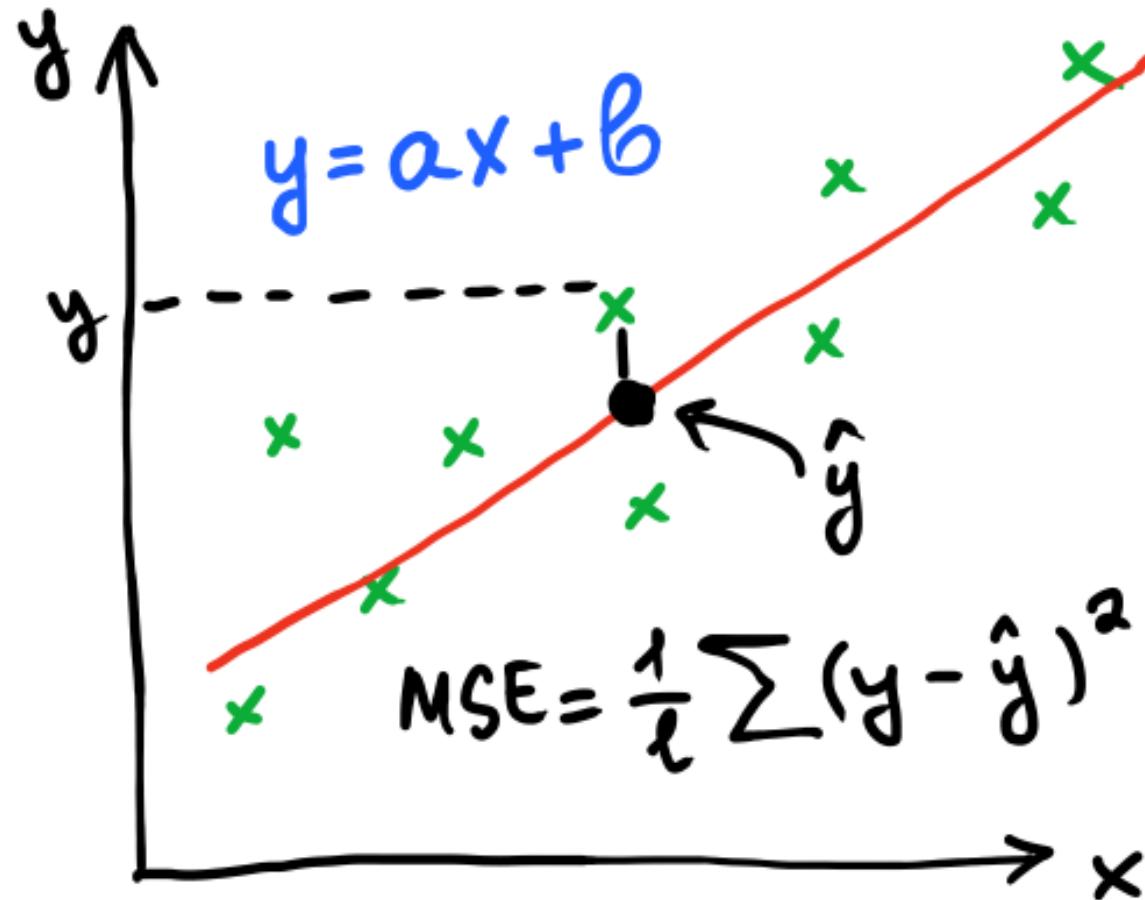
Линейная модель для регрессии

x	y
1	2
3	5
-1	-2
5	?

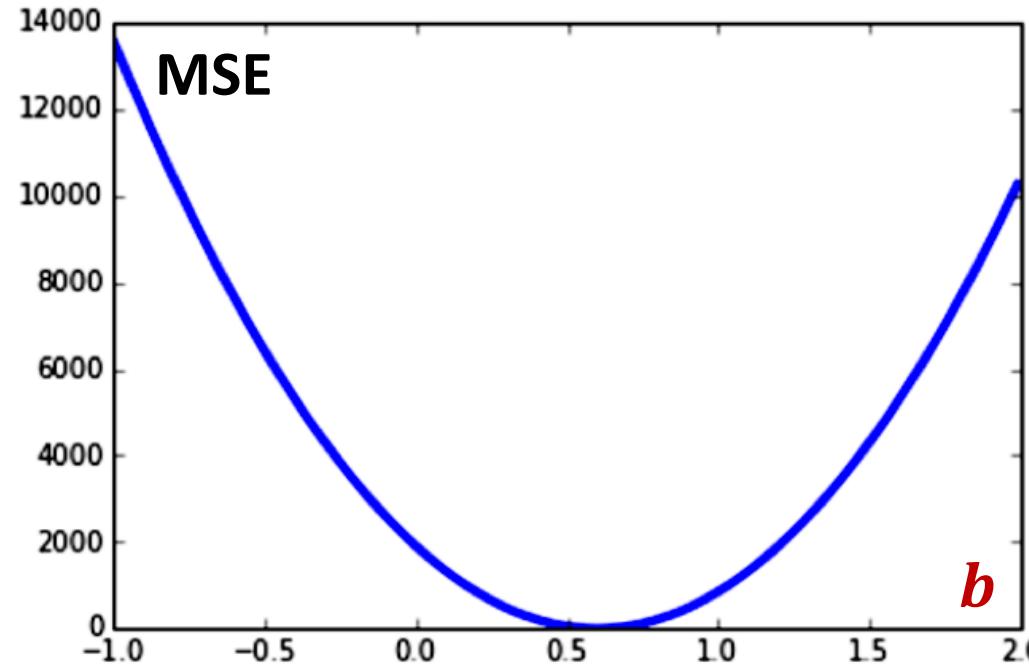


Линейная модель для регрессии

x	y
1	2
3	5
-1	-2
5	?



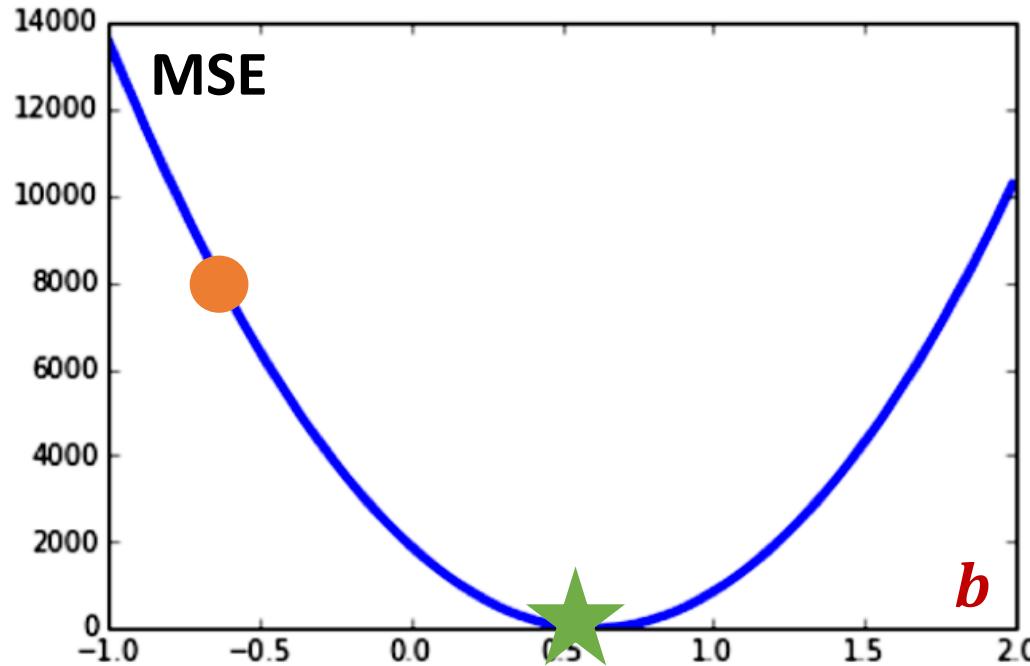
Идея процедуры обучения



MSE при фиксированном a :

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\underbrace{ax_i + b}_{\text{модель}} - y_i)^2$$

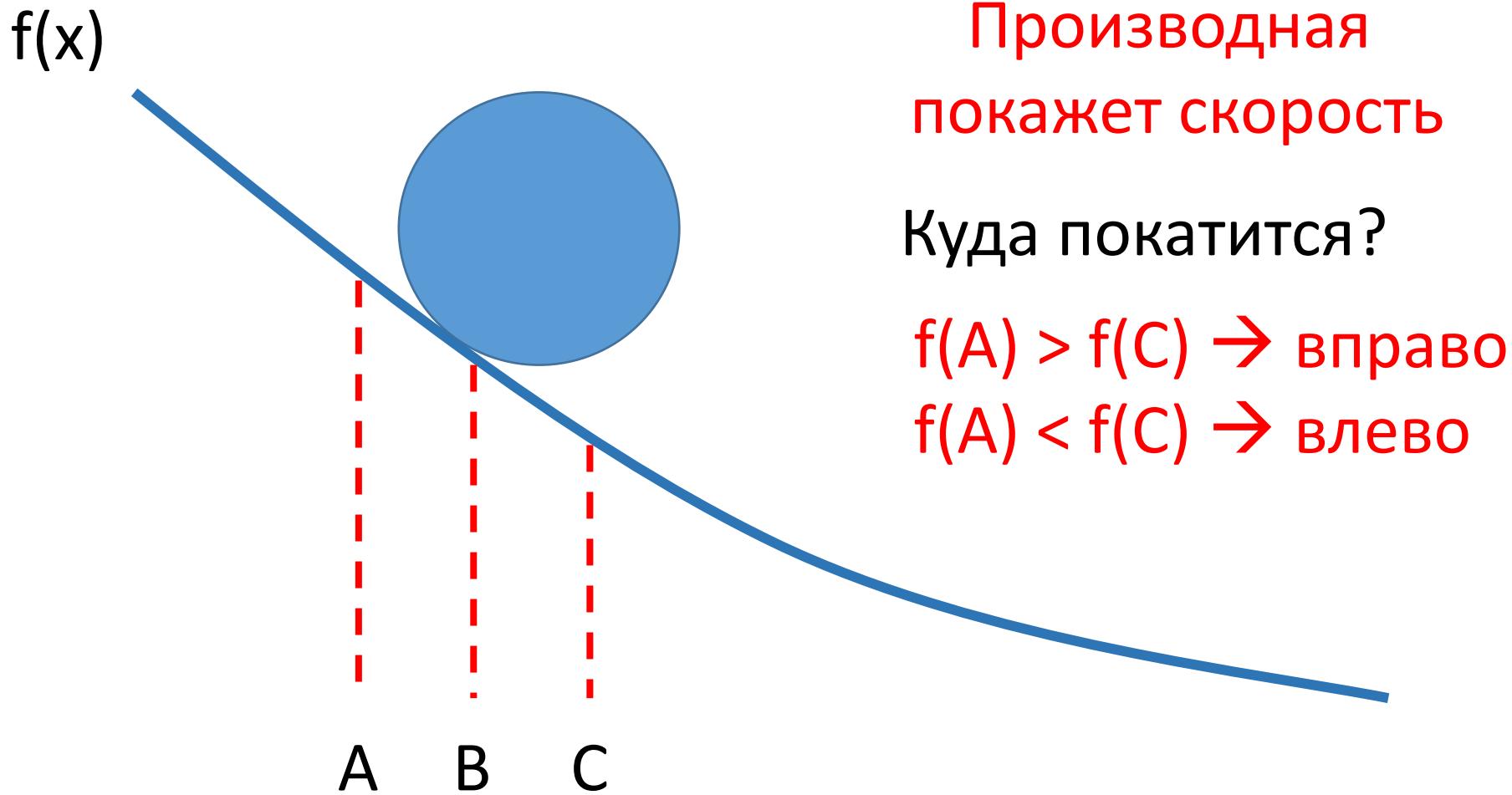
Идея процедуры обучения



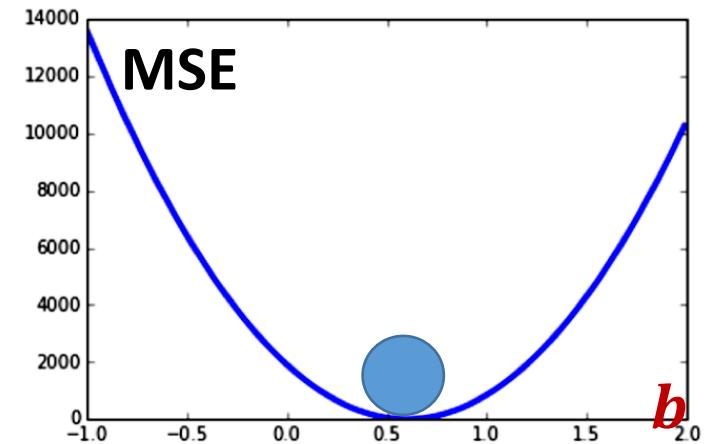
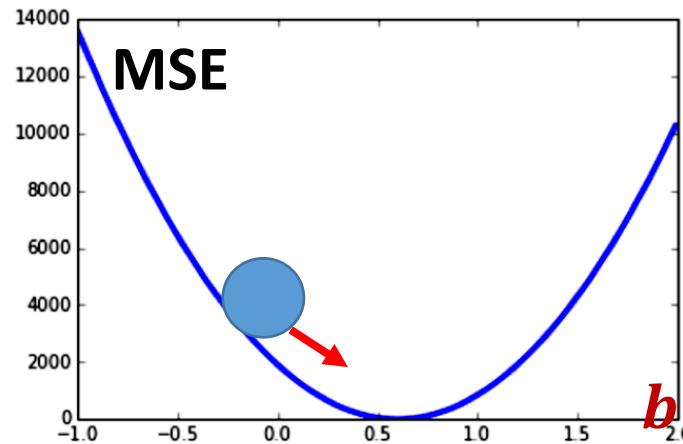
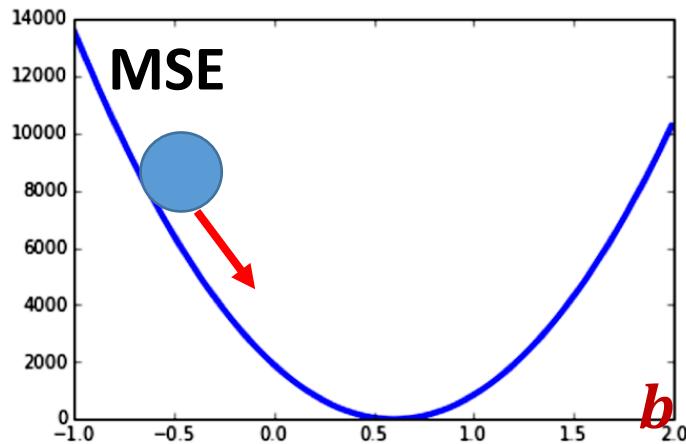
MSE при фиксированном a :

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\underbrace{ax_i + b}_{\text{модель}} - y_i)^2$$

Положим на параболу мячик в точке В



Идея процедуры обучения



MSE при фиксированном a :

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\underbrace{ax_i + b - y_i}_{\text{модель}})^2$$

Обучение линейной регрессии

- Будем считать, что есть признак, всегда равный единице:

$$\begin{aligned}a(x) &= w_1 x_1 + \cdots + w_d x_d = \\&= w_1 * 1 + w_2 x_2 + \cdots + w_d x_d = \\&= \langle w, x \rangle\end{aligned}$$

- Среднеквадратичная ошибка и задача обучения:

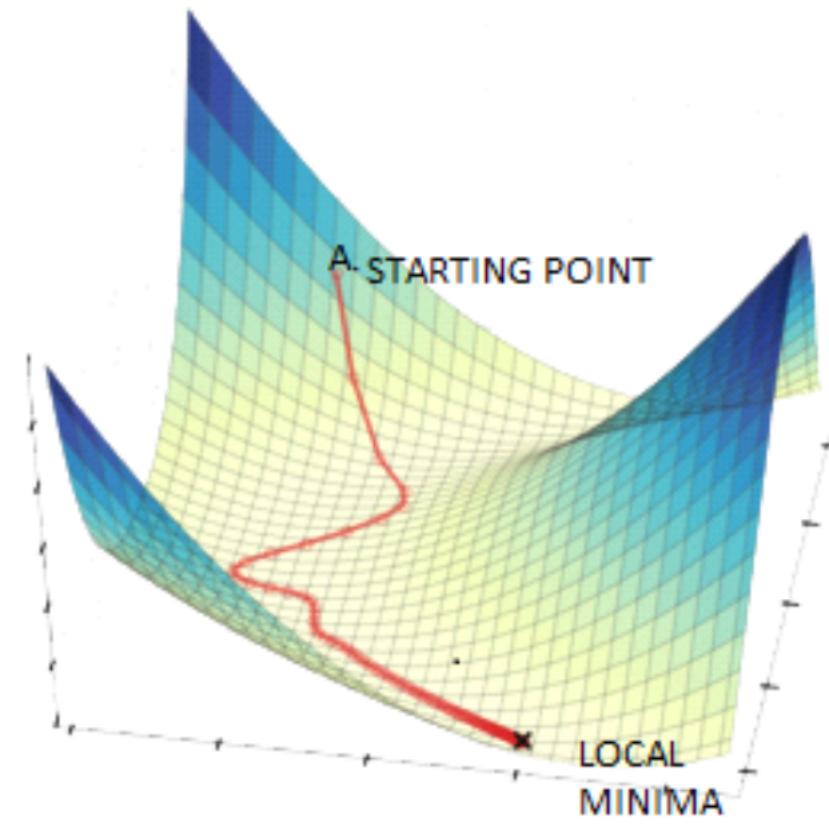
$$Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 = \frac{1}{\ell} \|Xw - y\|^2 \rightarrow \min_w$$

Обучение линейной регрессии

- Градиент — вектор частных производных

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_d} \right)$$

- Вектор градиента указывает в направлении наискорейшего возрастания функции
- Градиентный спуск — итерационных метод минимизации функции, в котором на каждом шаге мы идем в направлении антиградиента



Линейная регрессия

- Среднеквадратичная ошибка и задача обучения:

$$Q(w) = \frac{1}{\ell} \|Xw - y\|^2 \rightarrow \min_w$$

- Подсчет градиента:

$$\nabla Q(w) = \frac{2}{\ell} X^T(Xw - y)$$

Градиентный спуск

1. Начальное приближение: w^0

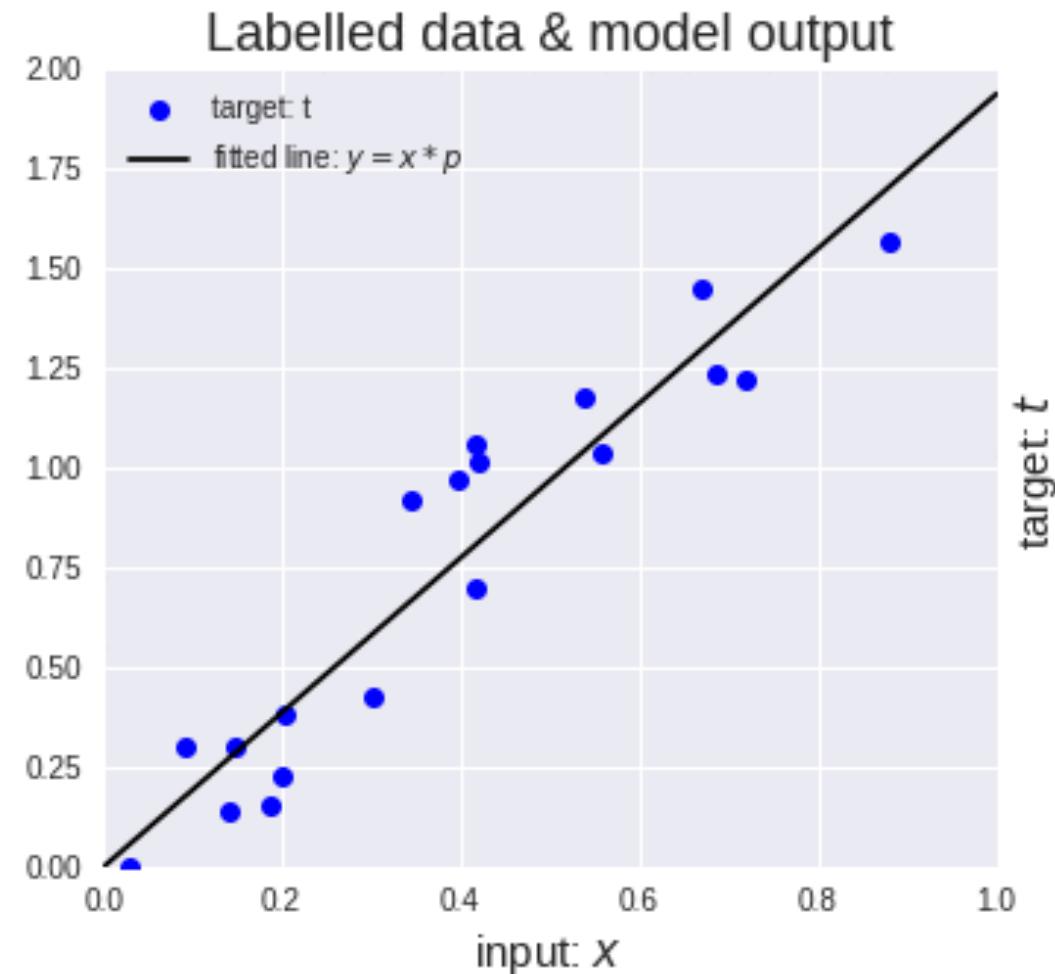
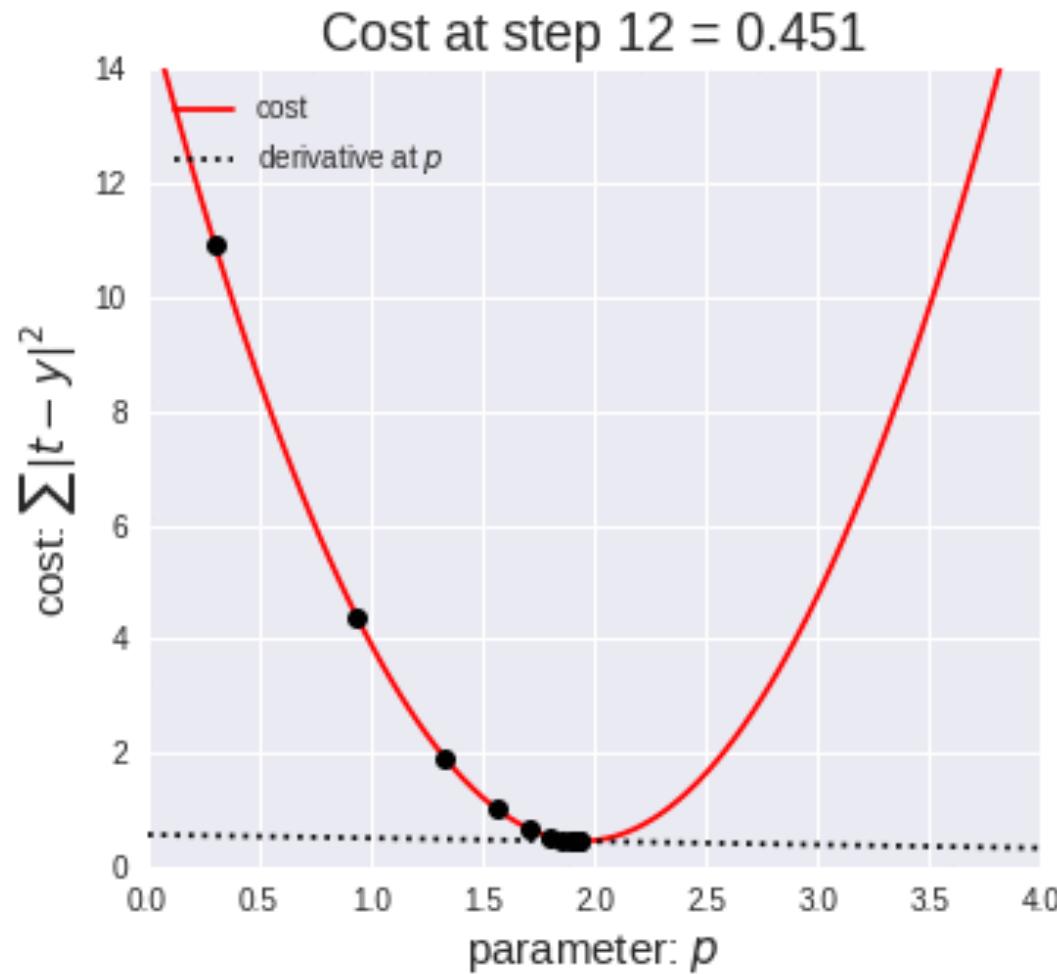
2. Повторять до сходимости:

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

3. Останавливаемся, если выполнено условие сходимости

$$\|w^t - w^{t-1}\| < \varepsilon$$

Демо: линейная регрессия



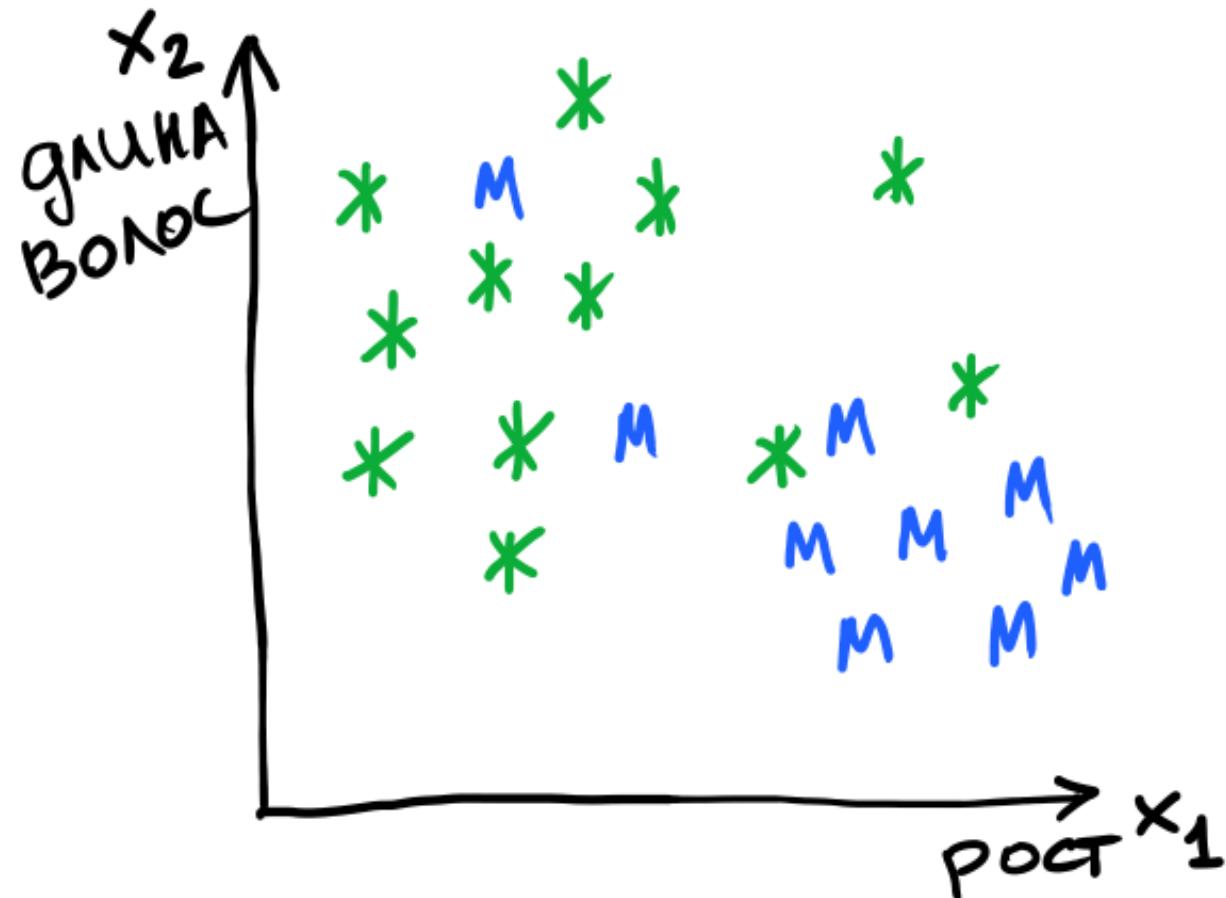
Обучение линейной регрессии

- Задача линейной регрессии выпукла и имеет единственное решение, поэтому нет проблем с локальными оптимумами при градиентном спуске
- Это решение можно найти аналитически, но нужно обращать матрицы

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

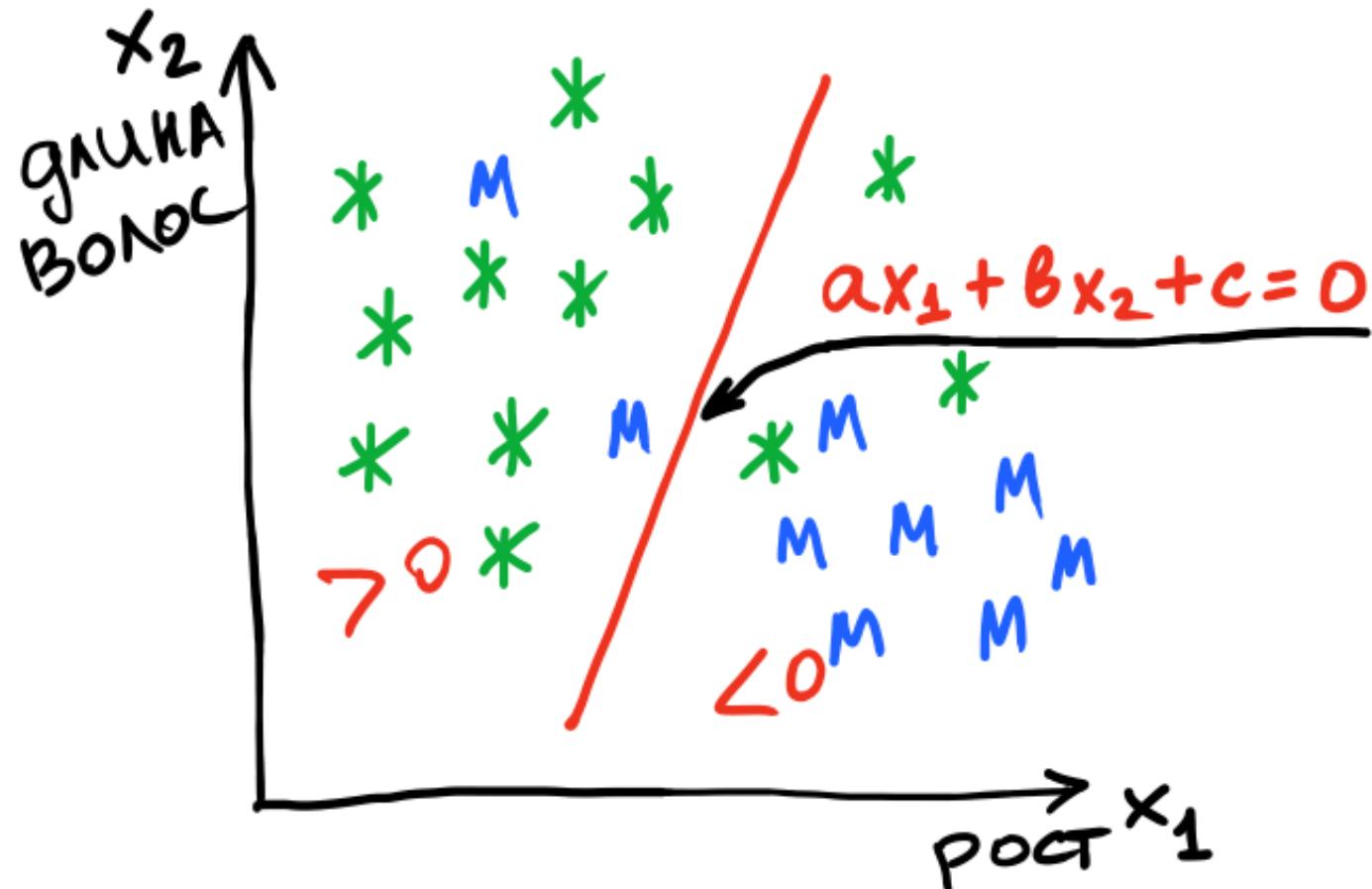
Линейная модель для классификации

x_1	x_2	y
180	5	М
170	20	Ж
160	5	М
190	30	?



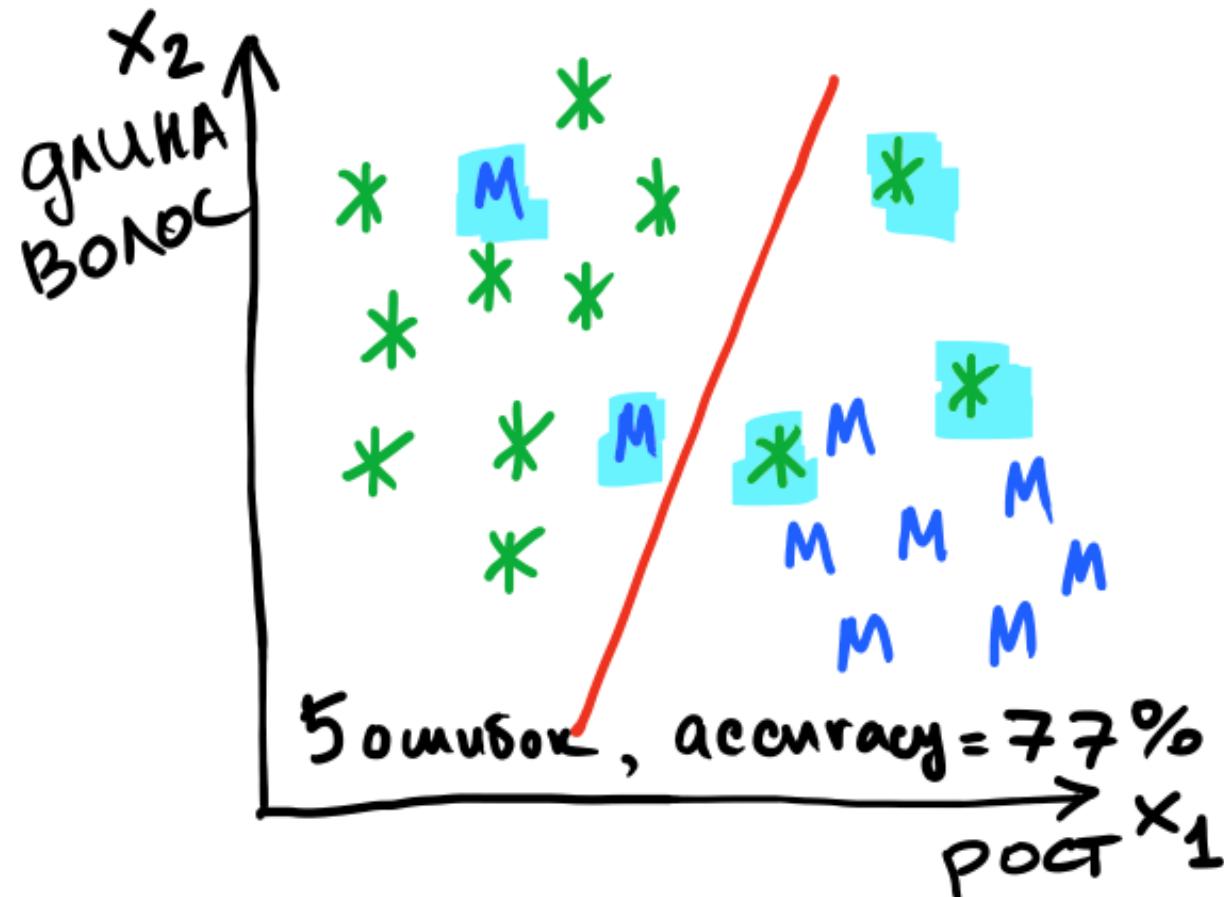
Линейная модель для классификации

x_1	x_2	y
180	5	М
170	20	Ж
160	5	М
190	30	?



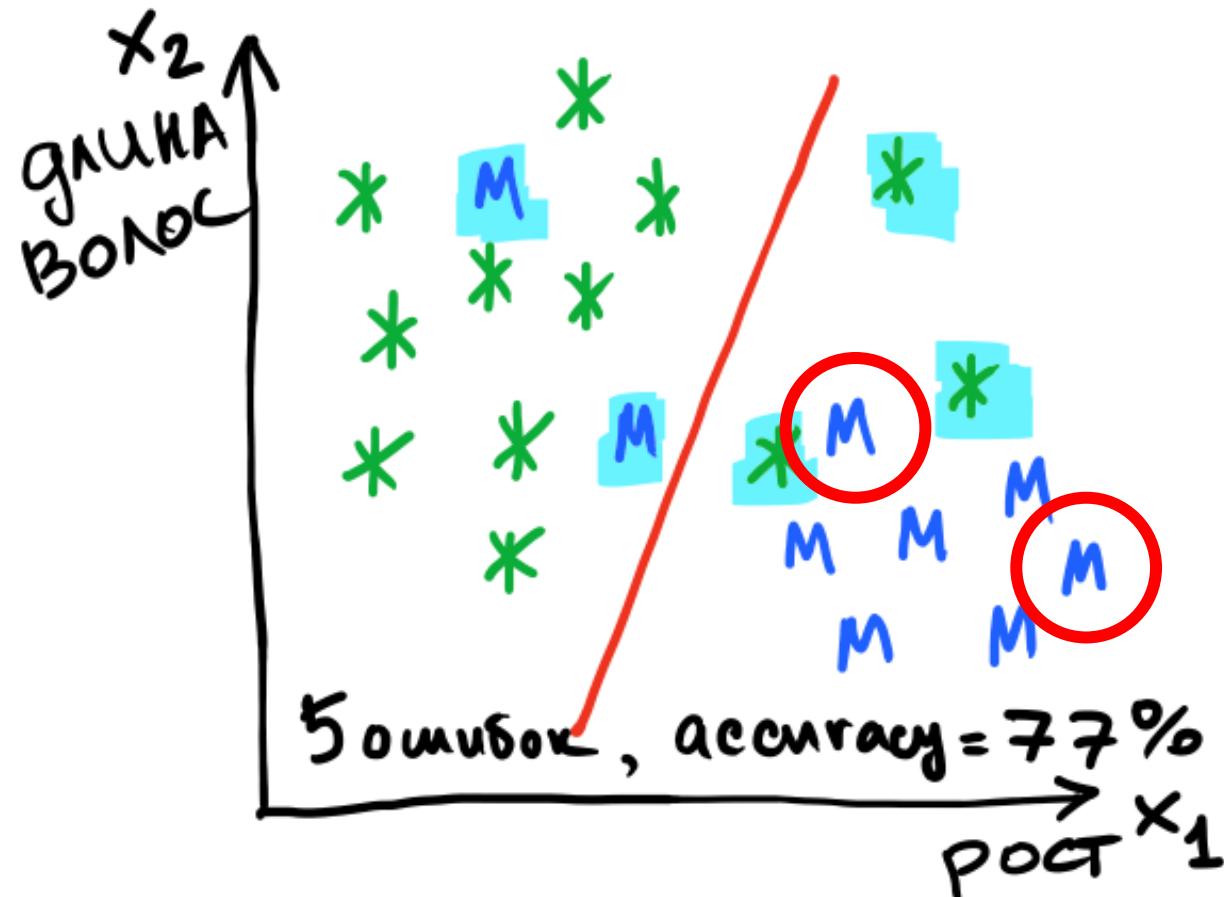
Линейная модель для классификации

x_1	x_2	y
180	5	М
170	20	Ж
160	5	М
190	30	?



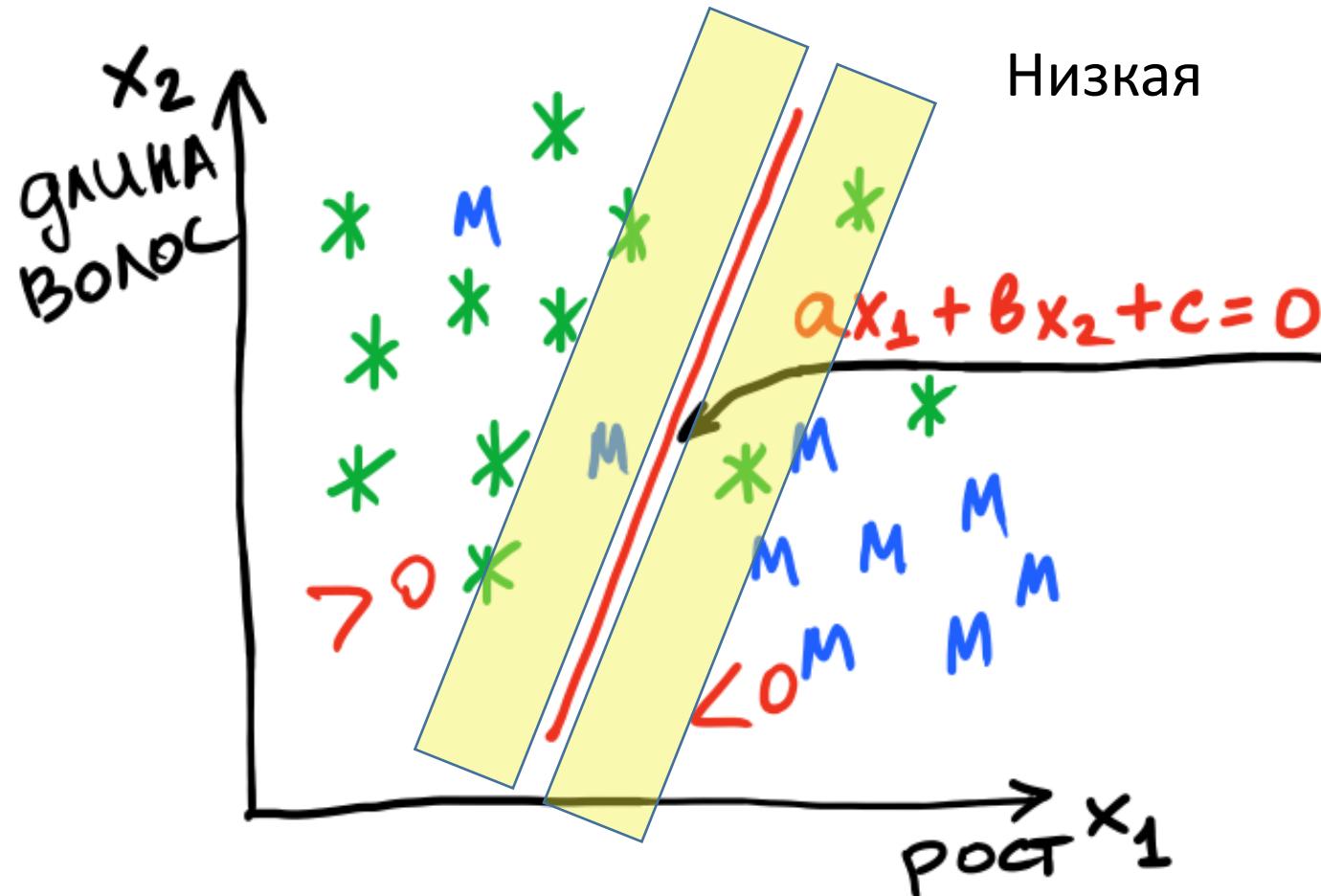
Линейная модель для классификации

x_1	x_2	y
180	5	М
170	20	Ж
160	5	М
190	30	?



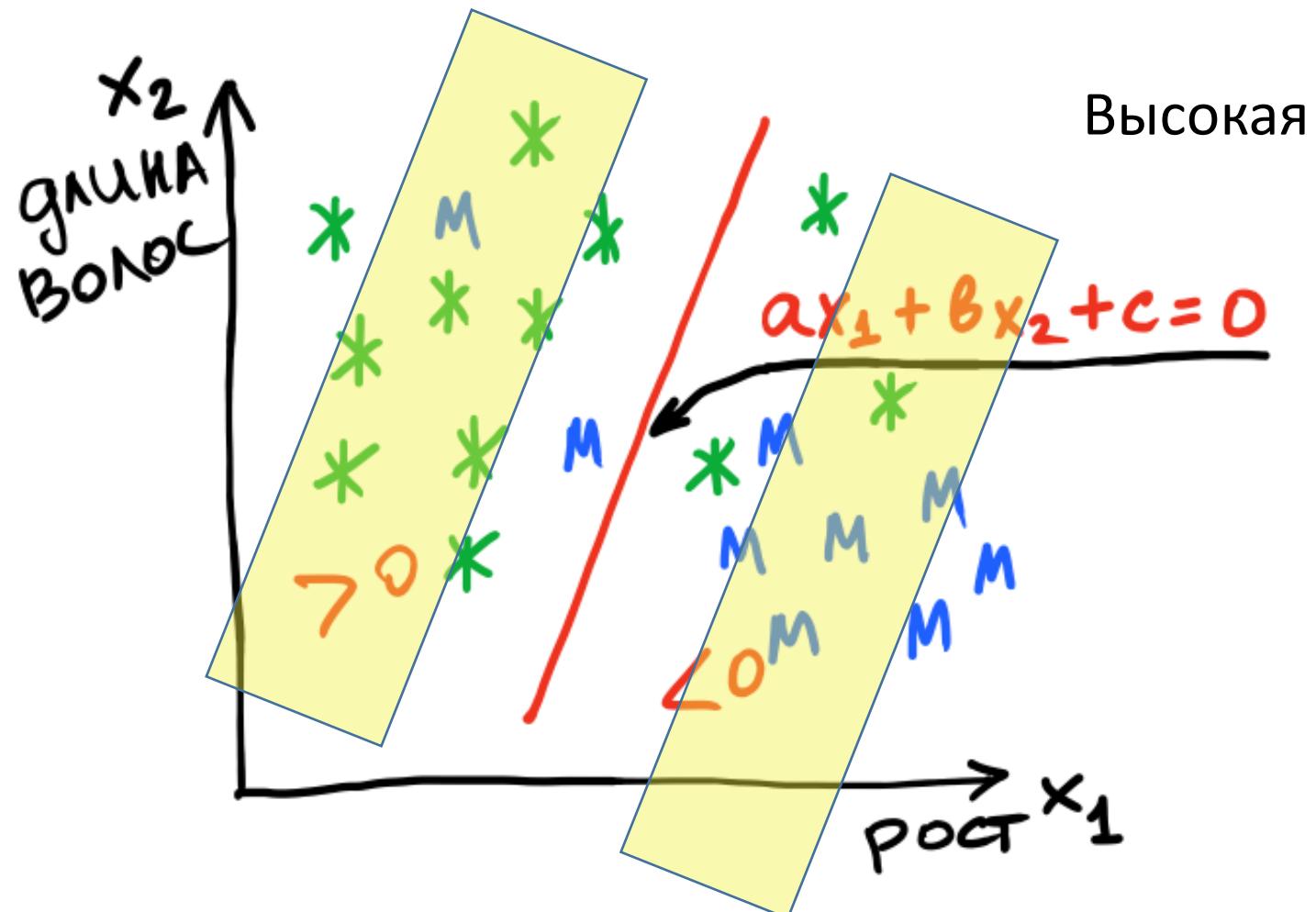
Уверенность в предсказании

x_1	x_2	y
180	5	М
170	20	Ж
160	5	М
190	30	?



Уверенность в предсказании

x_1	x_2	y
180	5	М
170	20	Ж
160	5	М
190	30	?

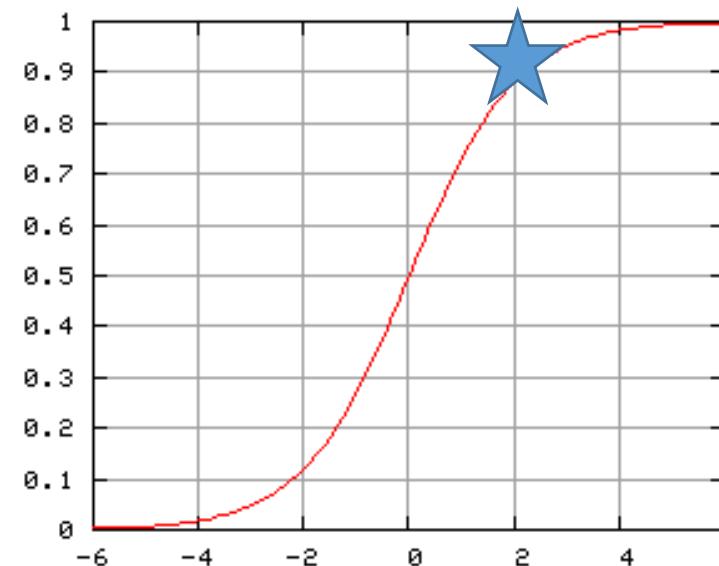


Логистическая регрессия

Линейная модель **классификации**, которая выдает вероятность класса +1.

В случае одного признака: $y(x) = \sigma(ax + b) = \frac{1}{1+e^{-(ax+b)}}$

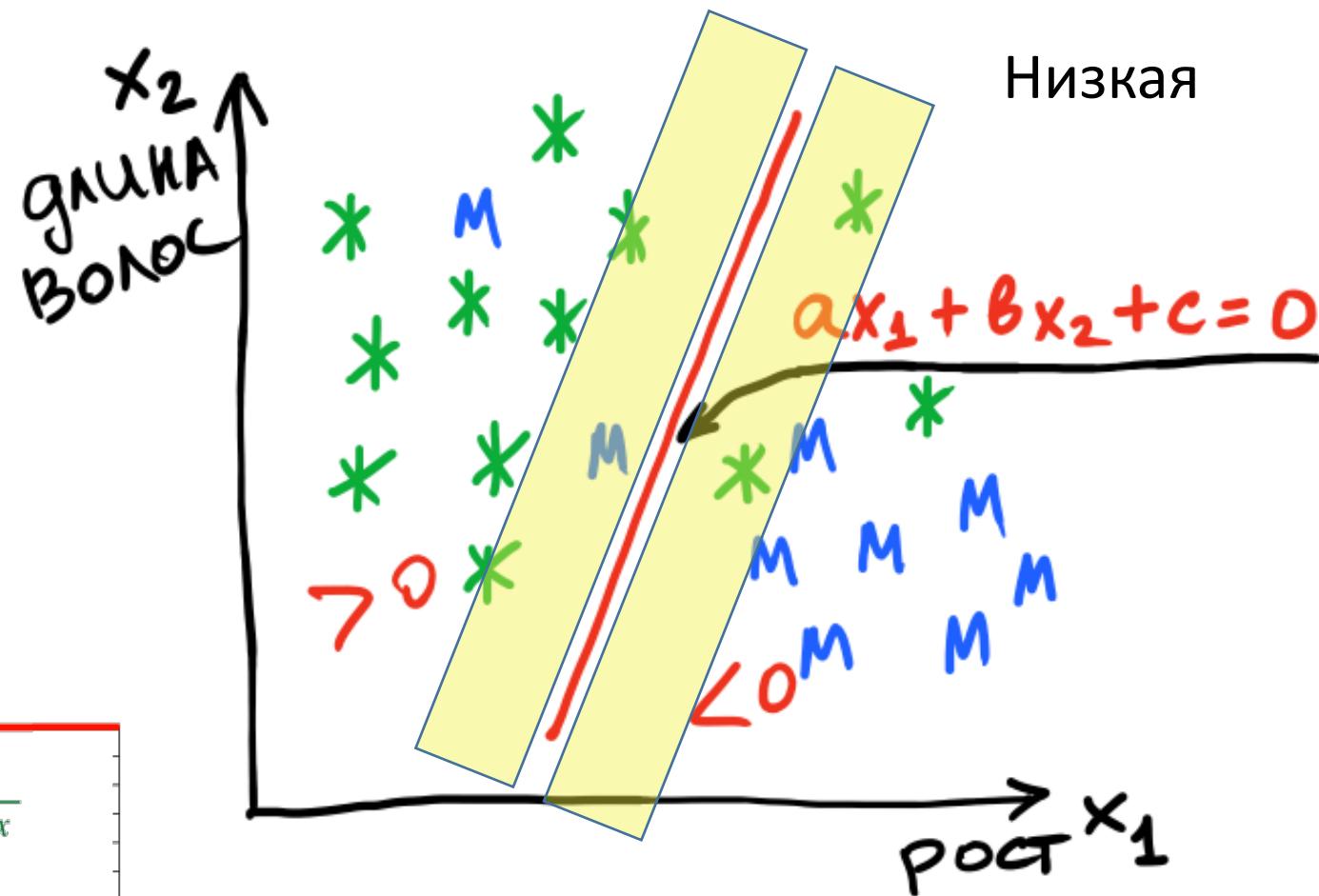
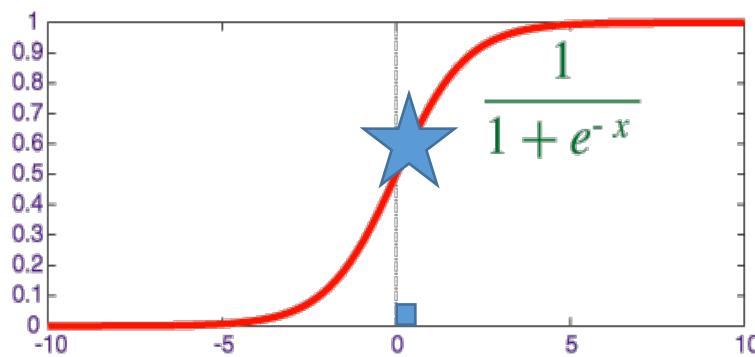
Логистическая функция $\sigma(x)$ взамен знака возвращает уверенность:



Дальше от прямой
– больше
вероятность класса

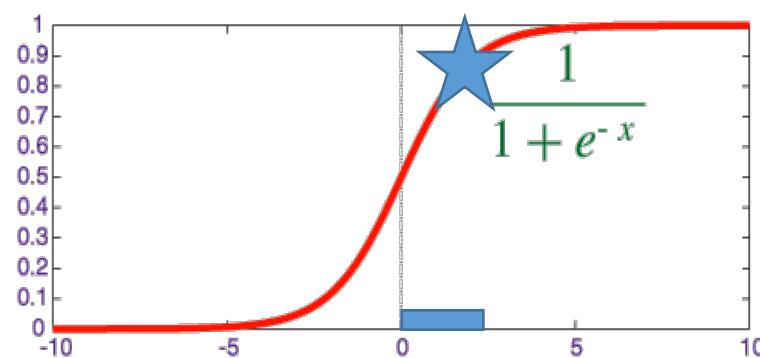
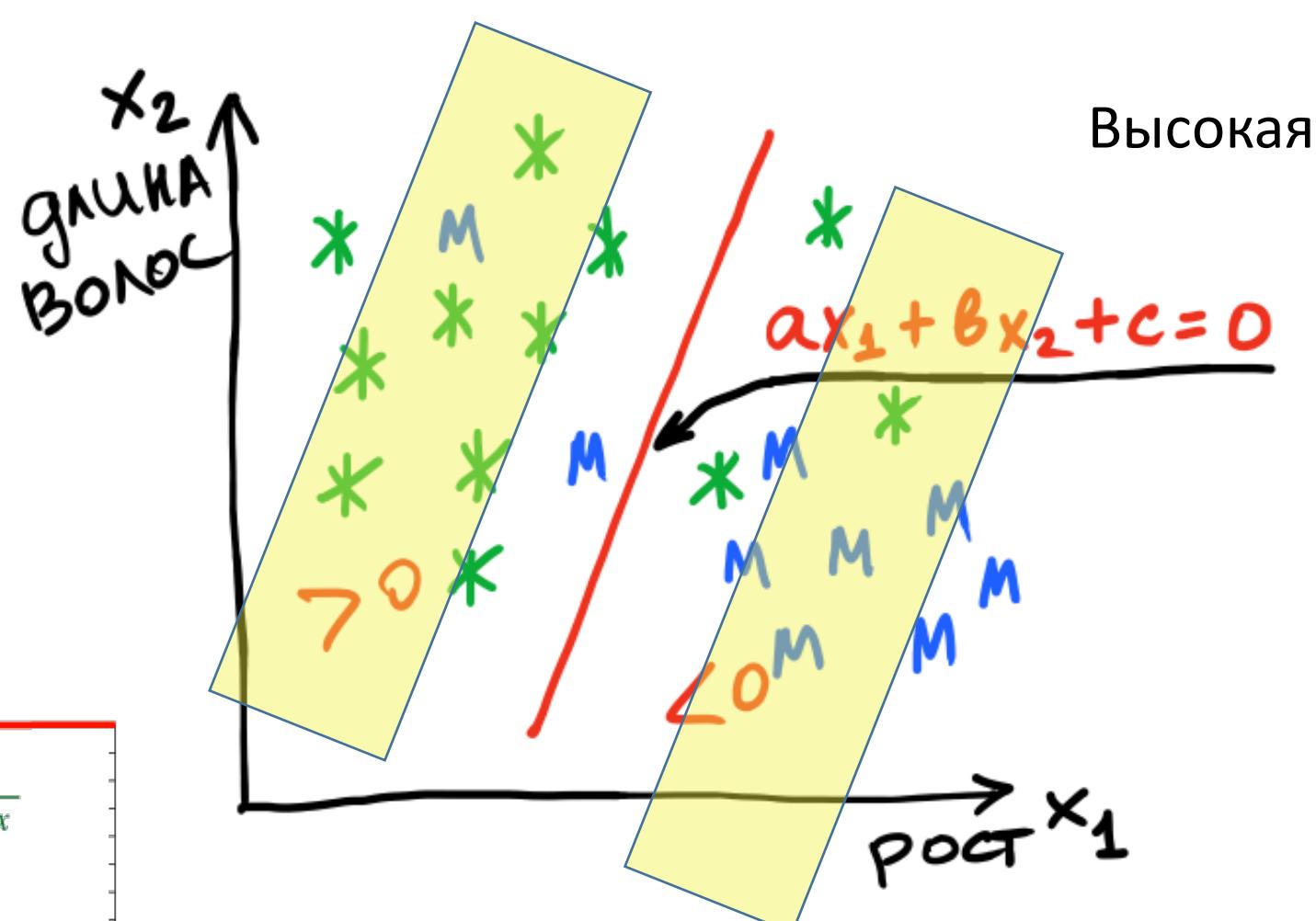
Уверенность в предсказании

X_1	X_2	Y
180	5	M
170	20	X
160	5	M
190	30	?



Уверенность в предсказании

x_1	x_2	y
180	5	М
170	20	Ж
160	5	М
190	30	?



Демо: логистическая регрессия



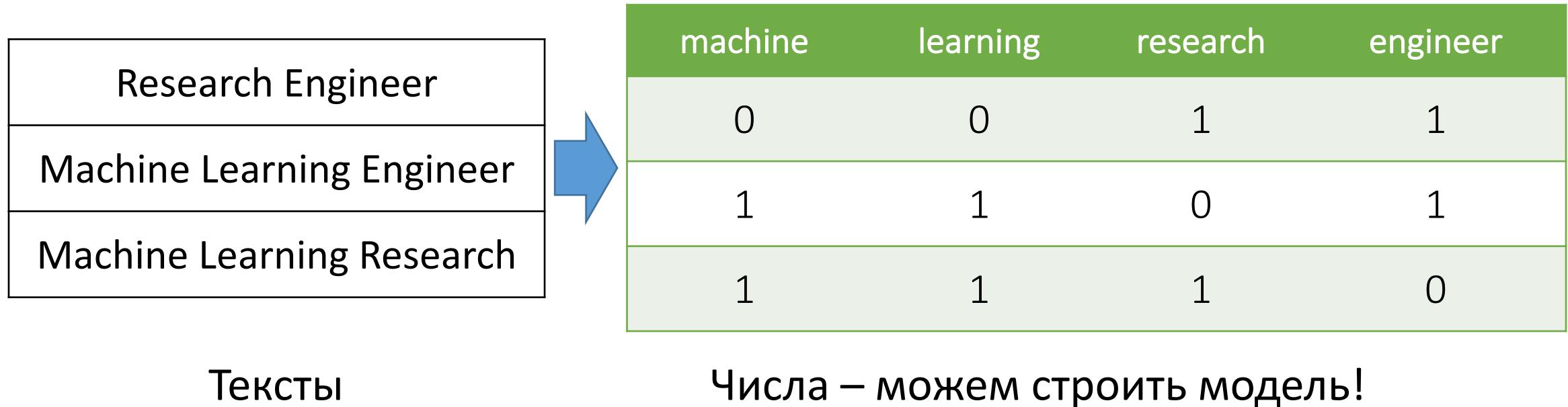
Пример

- Предсказание зарплаты по описанию вакансии

FullDescription	SalaryNormalized
Award winning multi disciplinary consultancy a...	50000
A senior / experienced VB.Net Developer, with ...	37500
Desktop Support Administrator Windows 7, Exch...	28000
RGN Clinical Lead Mid Glamorgan Very competiti...	27527

Пример - извлечение признаков

- Преобразовываем в «мешок слов» или «мешок n-грамм»
- Нормируем признаки - TF-IDF



Пример - предсказания

```
np.exp(clf.predict(tfidf.transform([u"junior window cleaner"])))
```

```
array([ 10213.72923139])
```

```
np.exp(clf.predict(tfidf.transform([u"chief window cleaner"])))
```

```
array([ 20867.5619582])
```

```
np.exp(clf.predict(tfidf.transform([u"chief window cleaner and big data"])))
```

```
array([ 28583.50258539])
```

Пример - веса модели

ngram	weight
director	1.926349
locum	1.519833
senior	1.448557
optometrist	1.445389
london	1.272371
head of	1.247227
dentist	1.233501
contract	1.193814
gmc	1.171238
manager	1.160573

ngram	weight
desired skills	-2.147389
apprenticeship	-1.639893
assistant	-1.556656
apprentice	-1.488979
www justot	-1.479929
justot co	-1.479929
labourer	-1.303409
studentship	-1.288740
graduate	-1.279080
labourers	-1.072269

Линейные модели

Линейная модель в общем виде суммирует значения всех признаков с некоторыми весами

Веса при признаках — параметры, которые необходимо настраивать в процессе обучения

Плюсы:

- Линейные модели способны обучаться на сверхбольших выборках
- Могут работать на данных с большим количеством признаков (например, на текстах)
- Хорошо интерпретируются

Минусы:

- Могут восстанавливать лишь линейные закономерности