Пример задачи

- Сеть ресторанов
- Хотим открыть еще один
- Несколько вариантов размещения
- Какой из вариантов принесет максимальную прибыль?

* см. kaggle.com, TFI Restaurant Revenue Prediction

Обозначения

- x — объект, sample — для чего хотим делать предсказания
  - Конкретное расположение ресторана
- $\mathbb{X}$ — пространство всех возможных объектов
  - Все возможные расположения ресторанов
- y — ответ, целевая переменная, target — что предсказываем
  - Прибыль в течение первого года работы
- $\mathbb{Y}$ — пространство ответов — все возможные значения ответа
  - Все вещественные числа

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

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

- Про демографию:
  - Средний возраст жителей ближайших кварталов
  - Динамика количества жителей
- Про недвижимость:
  - Средняя стоимость квадратного метра жилья поблизости
  - Количество школ, банков, магазинов, заправок
  - Расстояние до ближайшего конкурента
- Про дороги:
  - Среднее количество машин, проезжающих мимо за день

Алгоритм

- $a(x)$ — алгоритм, модель — функция, предсказывающая ответ для любого объекта
- Отображает $\mathbb{X}$ в $\mathbb{Y}$
- Линейная модель: $a(x) = w_1 x^1 + \cdots + w_d x^d$

Функция потерь

- Не все алгоритмы полезны
- $a(x) = 0$ — не принесет никакой выгоды
- Функция потерь — мера корректности ответа алгоритма
- Предсказали \$10000 прибыли, на самом деле \$5000 — хорошо или плохо?
- Квадратичное отклонение: $(a(x) - y)^2$

Функционал качества

- Функционал качества, метрика качества — мера качества работы алгоритма на выборке
- Среднеквадратичная ошибка (Mean Squared Error, MSE):

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i) - y_i)^2$$

- Чем меньше, тем лучше

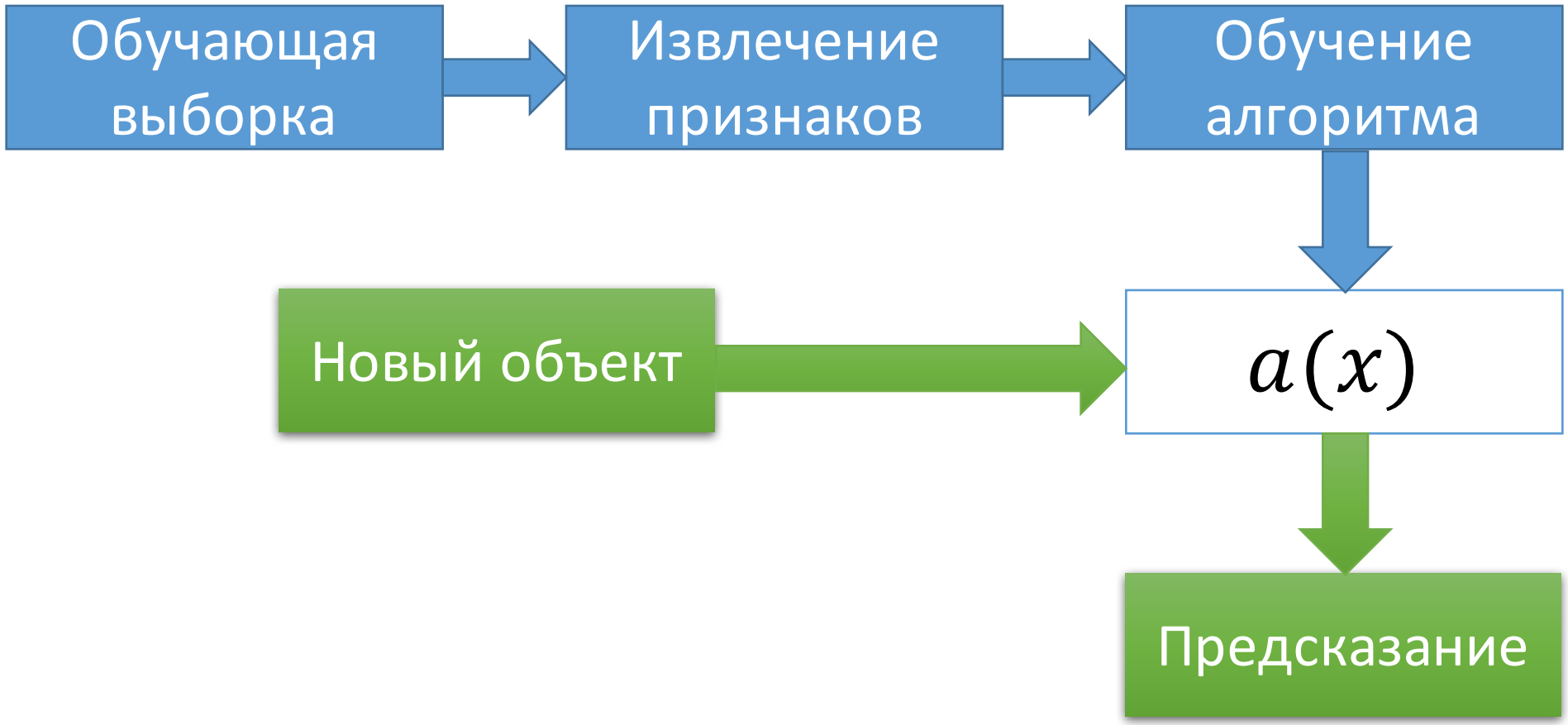
Функционал качества

- Должен соответствовать бизнес-требованиям
- Одна из самых важных составляющих анализа данных

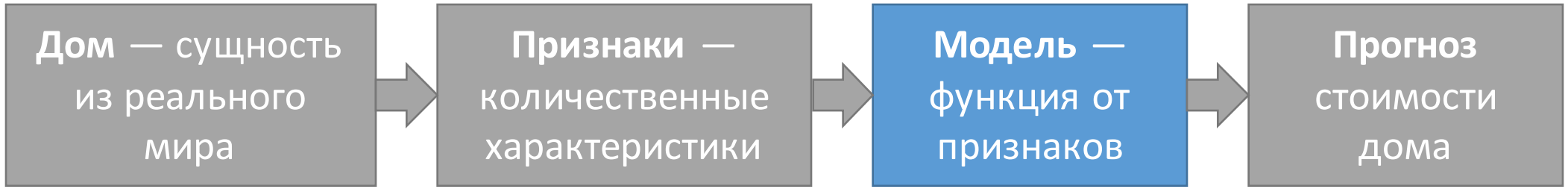
Обучение алгоритма

- Есть обучающая выборка и функционал качества
- Семейство алгоритмов $\mathcal{A}$
  - Из чего выбираем алгоритм
  - Пример: все линейные модели
  - $\mathcal{A} = \{w_1 x^1 + \cdots + w_d x^d \mid w_1, \ldots, w_d \in \mathbb{R}\}$
- Обучение: поиск оптимального алгоритма с точки зрения функционала качества

Машинное обучение



Предсказание стоимости дома



Предсказание стоимости дома

Обучающая выборка:

Площадь	Цена
50	250
60	340
10	20
90	800

Возможные признаки:

- площадь
- площадь^2
- площадь^3
- $\sin(\text{площадь})$
- $\sqrt{\text{площадь}}$
- и так далее

Возможные модели:

- $w_1 * \text{площадь}$
- $w_1 * \text{площадь}^2$
- $w_1 * \text{площадь} + w_2 * \text{площадь}^2$
- и так далее

Вид модели — работа эксперта либо полный перебор.

Выбор весов w_1, w_2 — автоматический процесс (на основе данных)

Предсказание стоимости дома

Модель $a(x) = 5 * \text{площадь}$

Площадь	Прогноз	Цена	$(a - y)^2$
50	250	250	0
60	300	340	1600
10	50	20	900
90	450	800	122500

MSE: 31 250

RMSE: 176,78

Модель $a(x) = 0.1 * \text{площадь}^2$

Площадь	Прогноз	Цена	$(a - y)^2$
50	250	250	0
60	360	340	400
10	10	20	100
90	810	800	100

MSE: 150

RMSE: 12,25

Предсказание стоимости дома

Признаков может быть больше:

- Площадь
- Год постройки
- Наличие бассейна
- Число комнат
- Удалённость от центра
- Рейтинг полицейского участка
- И так далее

Возможные модели:

- Линейная: $w_1 * \text{площадь} + w_2 * \text{год} + w_3 * \text{бассейн} + w_4 * \text{комнаты} + w_5 * \text{удалённость} + w_6 * \text{полиция}$
- Решающие деревья
- Нейронные сети
- Метод k ближайших соседей
- И так далее

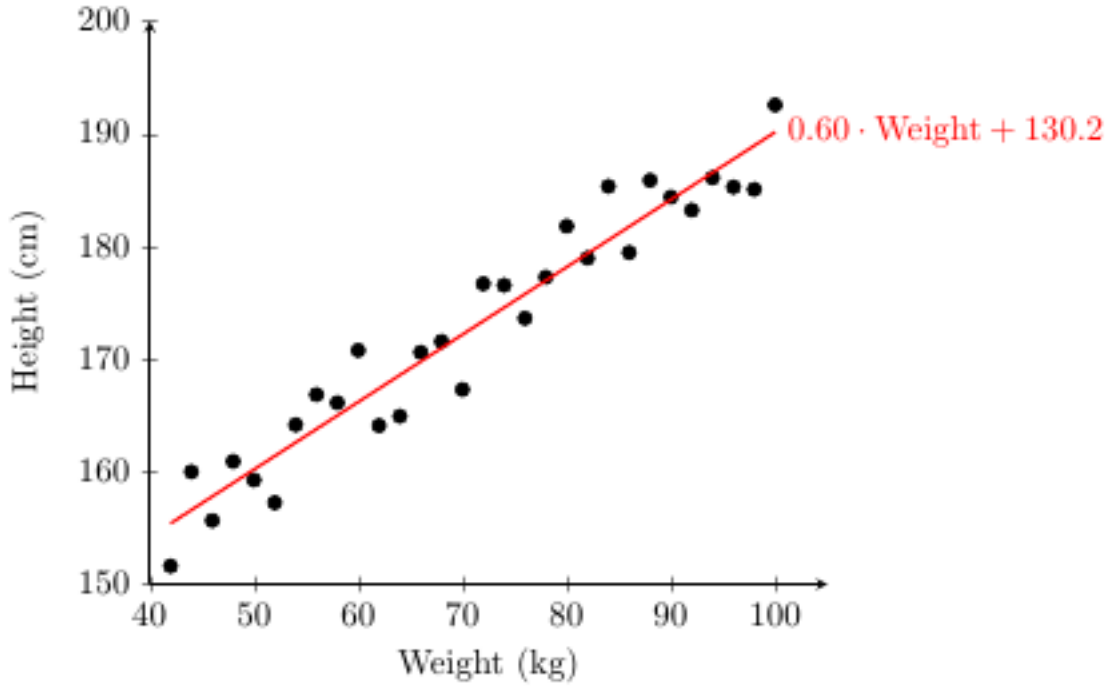
Что нужно знать

1. Как сформулировать задачу?
2. Какие признаки использовать?
3. Откуда взять обучающую выборку?
4. Как выбрать метрику качества?
5. Как обучить алгоритм?
6. Как оценить качество алгоритма?

Типы ответов

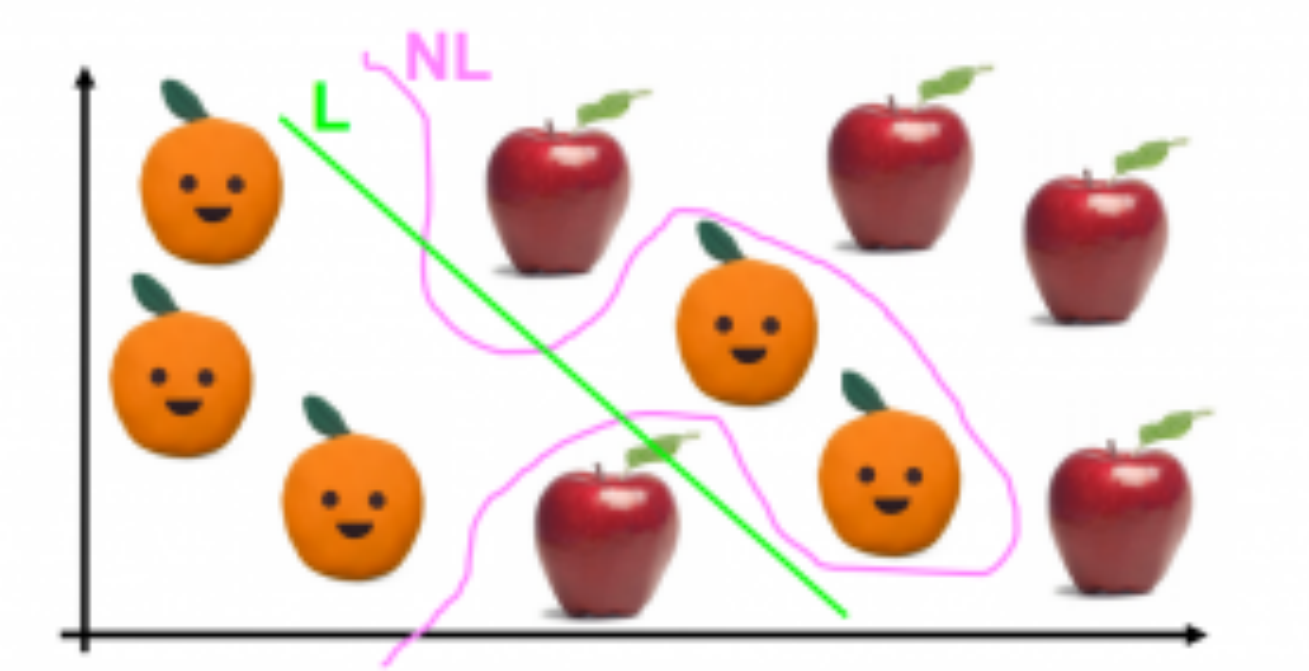
Регрессия

- Вещественные ответы: $\mathbb{Y} = \mathbb{R}$
- (вещественные числа — числа с любой дробной частью)
- Пример: предсказание роста по весу



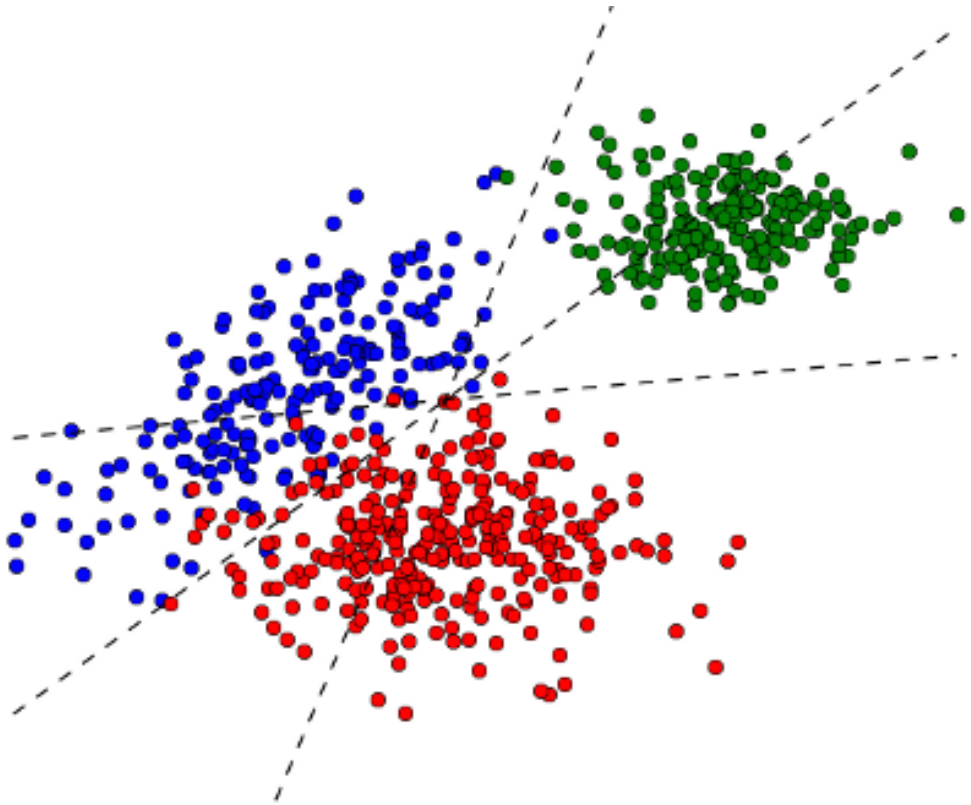
Классификация

- Конечное число ответов: $|\mathbb{Y}| < \infty$
- Бинарная классификация: $\mathbb{Y} = \{-1, +1\}$



Классификация

- Многоклассовая классификация: $\mathbb{Y} = \{1, 2, \dots, K\}$



Классификация

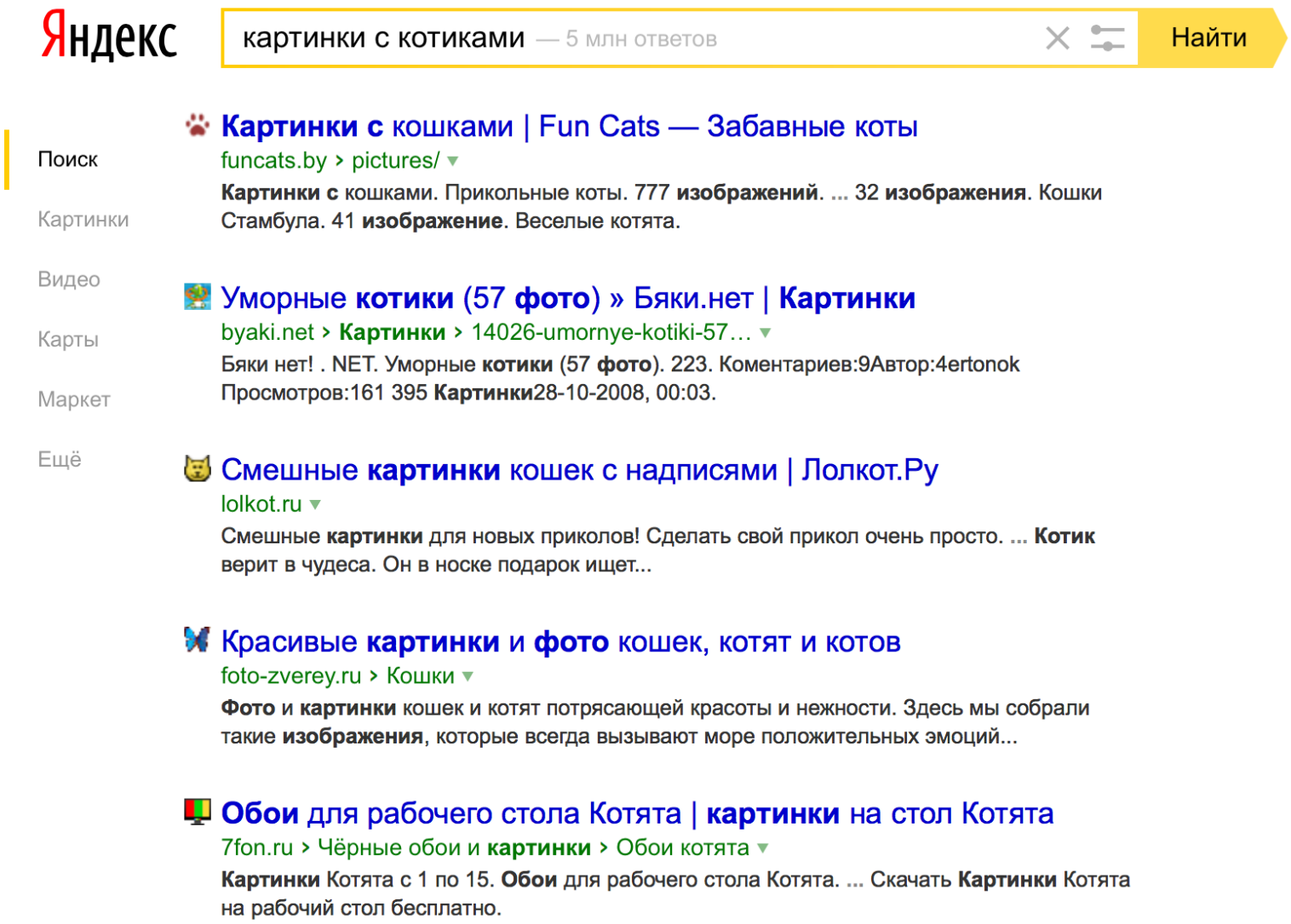
- Классификация с пересекающимися классами: $\mathbb{Y} = \{0, 1\}^K$
  - (multi-label classification)
- Ответ — набор из К нулей и единиц
- i-й элемент ответа — принадлежит ли объект i-му классу

- Какие темы присутствуют в статье?
- (математика, биология, экономика)

Ранжирование

- Набор документов $d_1, \ldots, d_n$
- Запрос q
- Задача: отсортировать документы по *релевантности* запросу
- $a(q, d)$ — оценка релевантности

Ранжирование

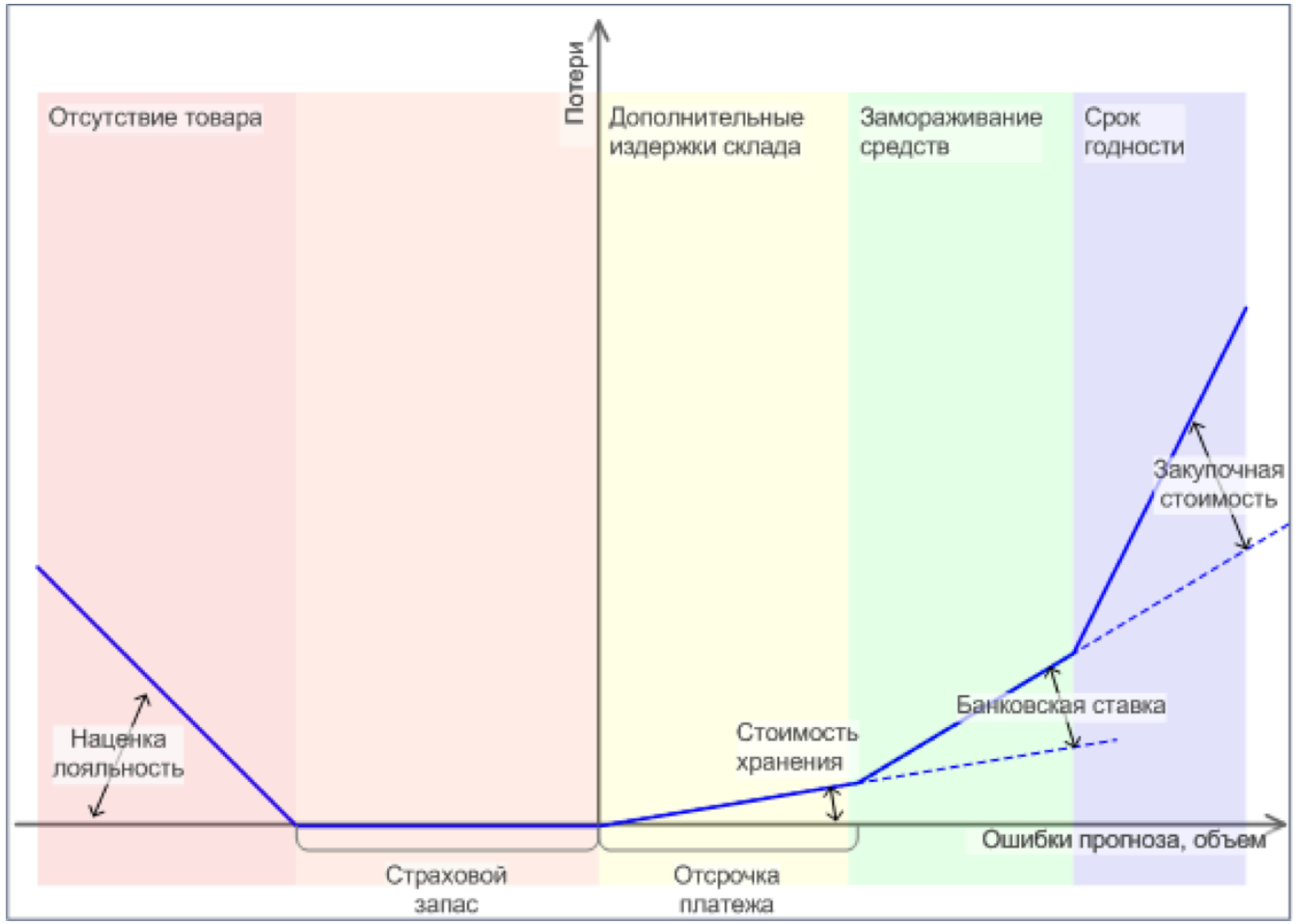


Кластеризация

- $\mathbb{Y}$ — отсутствует
- Нужно найти группы похожих объектов
- Сколько таких групп?
- Как измерить качество?

- Пример: сегментация пользователей мобильного оператора

Пример специализированной функции потерь



Типы признаков

Типы признаков

- f_j — j-й признак
- D_j — множество значений признака

Бинарные признаки

- $D_j = \{0, 1\}$
- Доход клиента выше среднего по городу?
- Цвет фрукта — зеленый?

Вещественные признаки

- $D_j = \mathbb{R}$
- Возраст
- Площадь квартиры
- Количество звонков в колл-центр

Категориальные признаки

- D_j — неупорядоченное множество
- Цвет глаз
- Город
- Образование (может быть упорядоченным)

- Очень трудны в обращении

Порядковые признаки

- D_j — упорядоченное множество
- Воинское звание
- Роль в фильме (первого плана, второго плана, массовка)
- Тип населенного пункта

Множествозначные признаки

- (set-valued)
- D_j — множество всех подмножеств некоторого множества
- Какие фильмы посмотрел пользователь?
- Какие слова входят в текст?

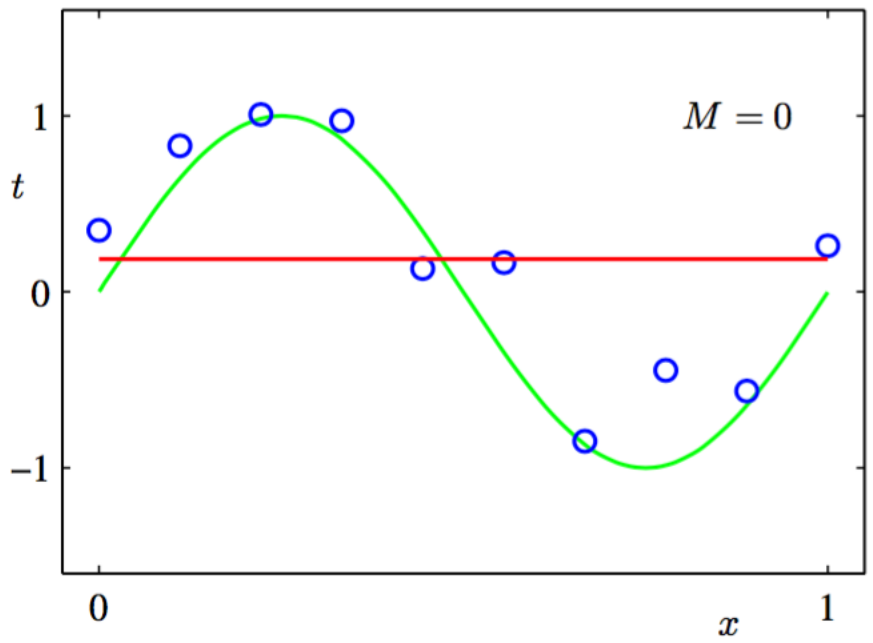
Обобщающая способность

Обобщающая способность

- Выбираем алгоритм с лучшим качеством на обучающей выборке
- Как он будет вести себя на новых данных?
- Смог ли он выразить y через x?

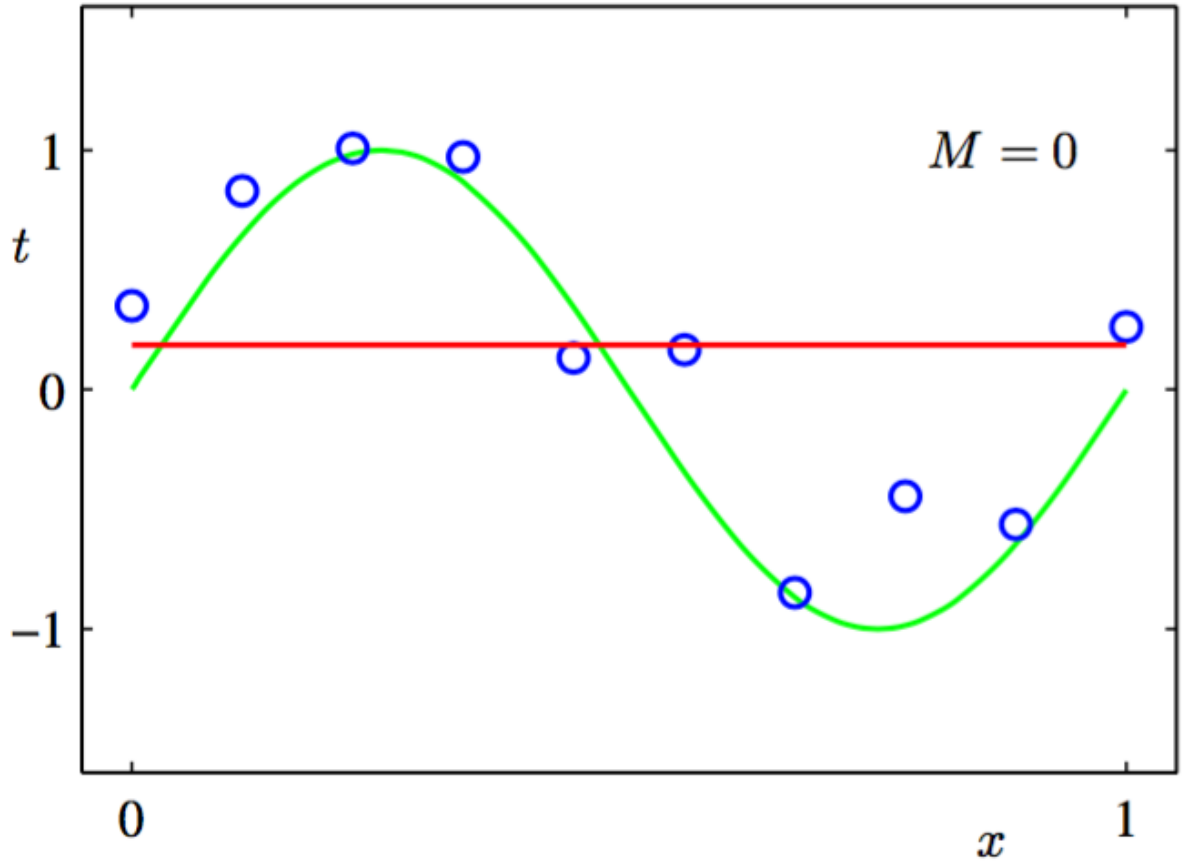
Обобщающая способность

- Зеленый — истинная зависимость
- Красный — прогноз алгоритма
- Синий — выборка
- Линейный алгоритм



Обобщающая способность

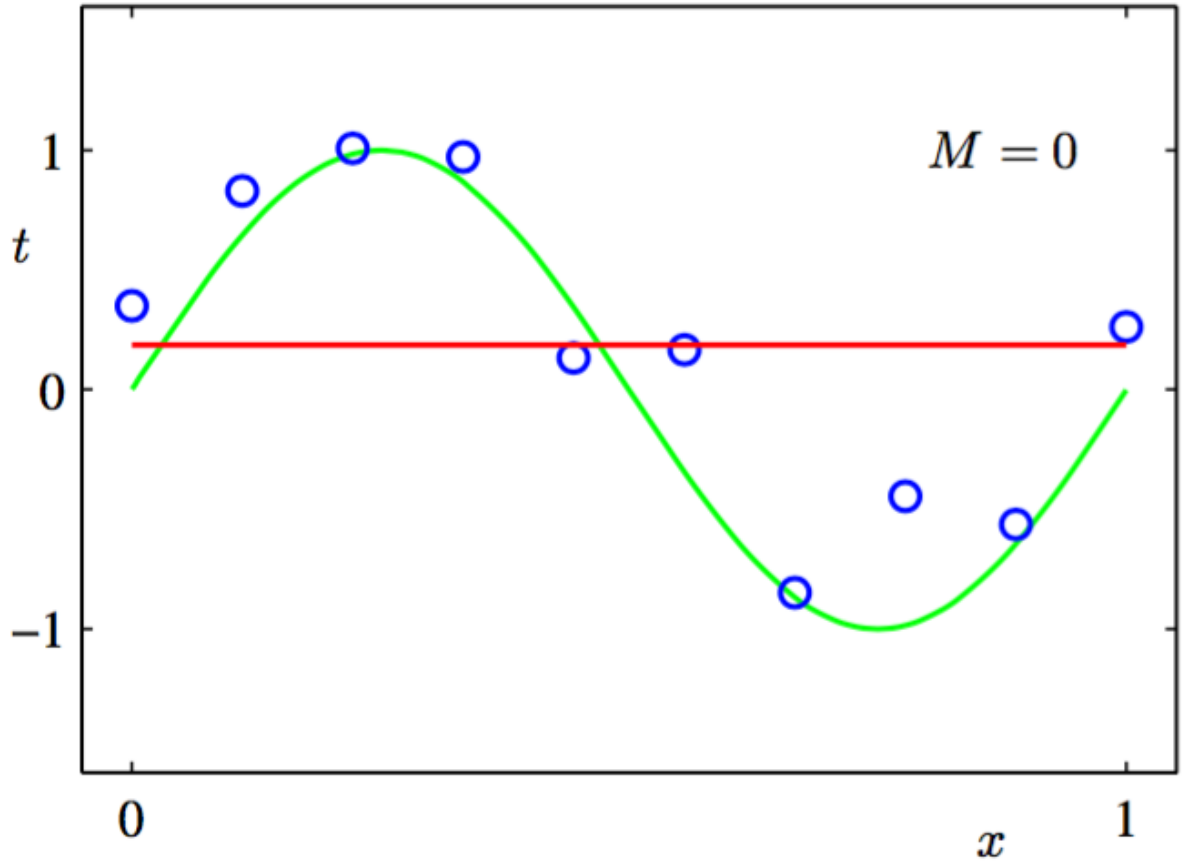
- Без признаков
- Константный алгоритм



Обобщающая способность

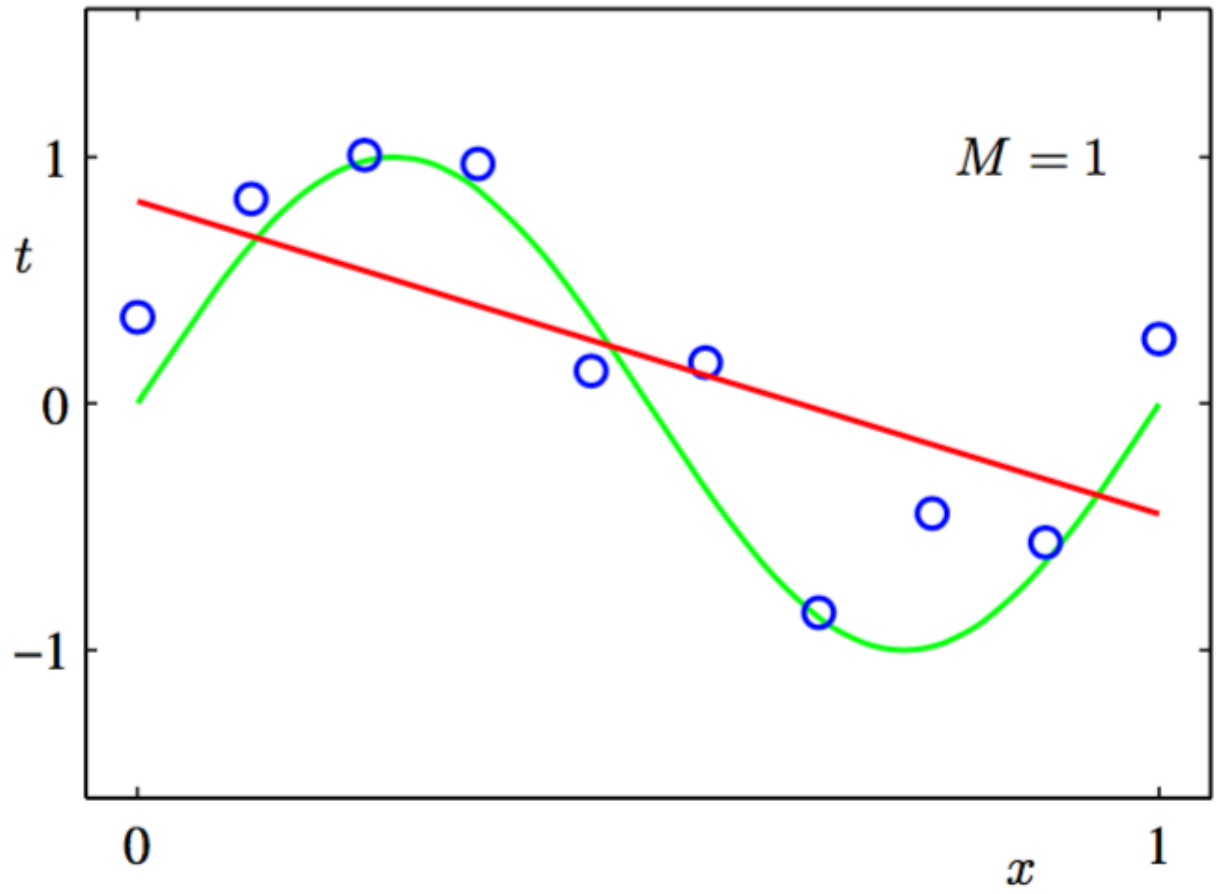
- Без признаков
- Константный алгоритм

Недообучение



Обобщающая способность

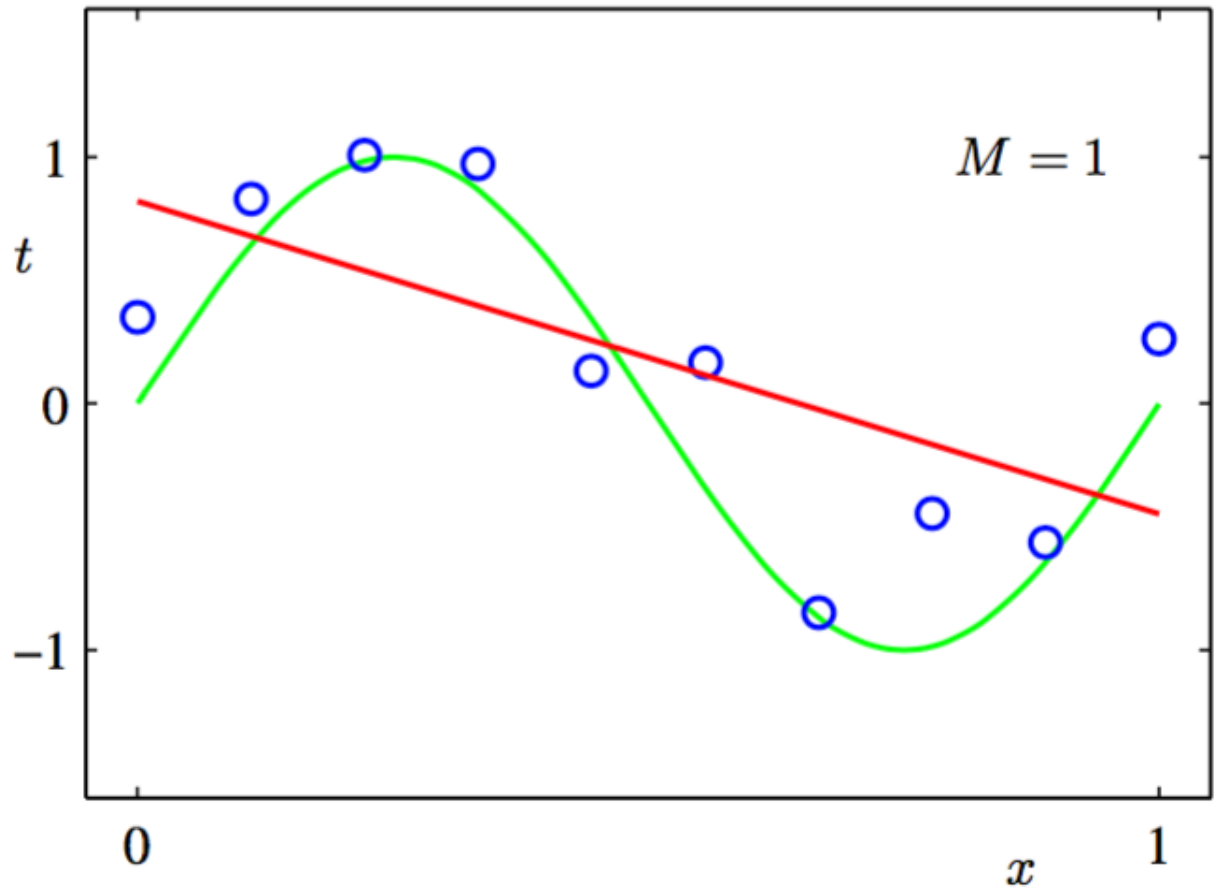
- 1 признак
- x



Обобщающая способность

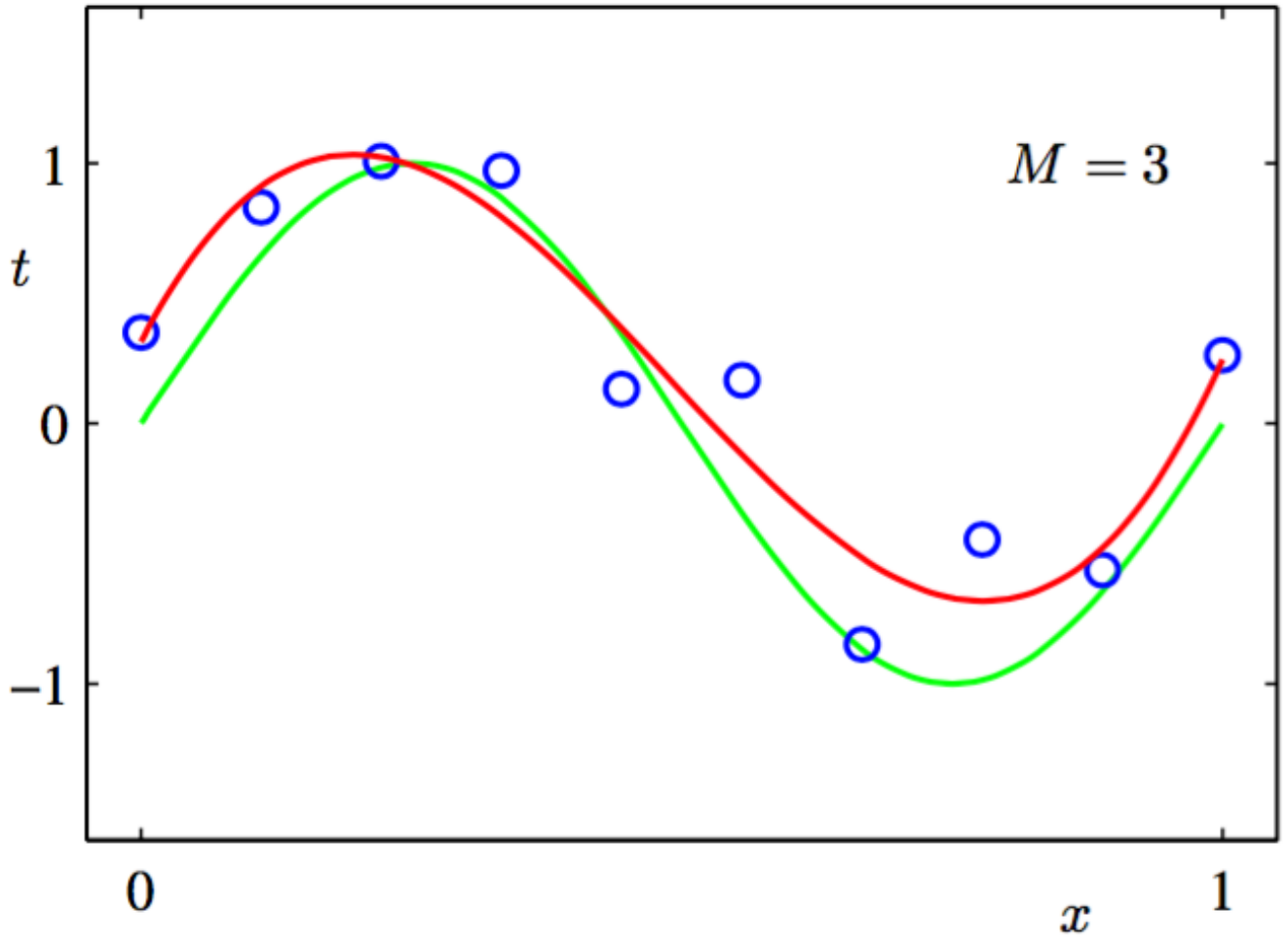
- 1 признак
- x

Недообучение



Обобщающая способность

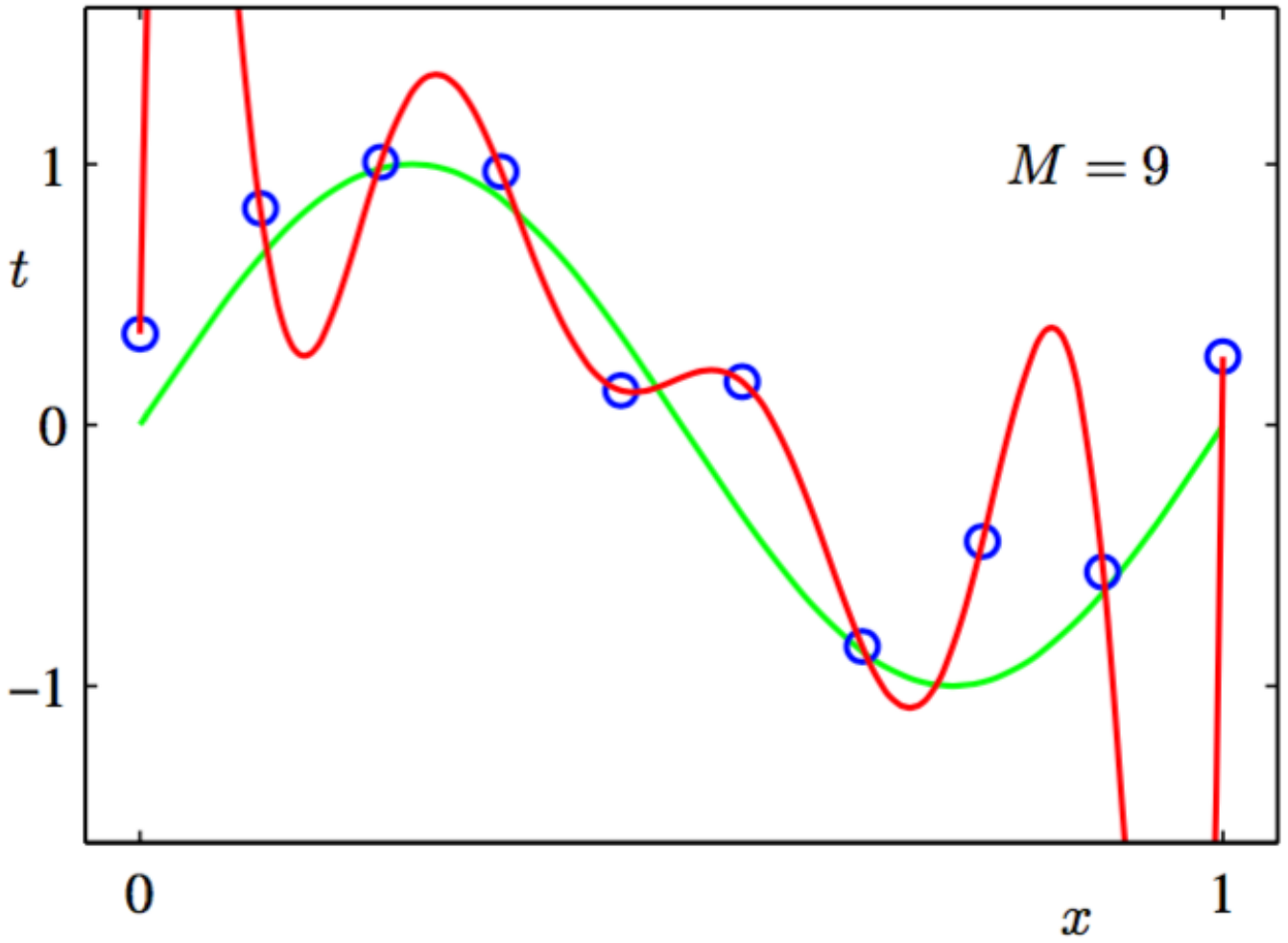
- 3 признака
- x, x^2, x^3



Обобщающая способность

- 9 признаков
- $x, x^2, x^3, x^4, \ldots, x^9$

Переобучение (overfitting)



Обобщающая способность

- Недообучение — плохое качество на обучении и на новых данных
- Переобучение — хорошее качество на обучении, плохое на новых данных

- Переобучение — алгоритм запоминает ответы, а не находит закономерности

Как выявить переобучение?

- Хороший алгоритм — хорошее качество на обучении
- Переобученный алгоритм — хорошее качество на обучении
- По обучающей выборке очень сложно выявить переобучение

Как выявить переобучение?

- Отложенная выборка — данные, на которых не обучались
- Кросс-валидация
- Меры сложности модели

Задачи анализа данных

Медицинская диагностика

- Объект — пациент в определенный момент времени
- Ответ — диагноз
- Классификация с пересекающимися классами

Медицинская диагностика — признаки

- Бинарные: пол, головная боль, слабость, и т.д.
- Порядковые: тяжесть состояния, желтушность, и т.д.
- Вещественные: возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т.д.

Медицинская диагностика — особенности

- Много пропусков в данных (missing data)
- Недостаточный объем данных
- Алгоритм должен быть интерпретируемым
- Нужна оценка вероятности для каждого заболевания

Кредитный скоринг

- Объект — заявка на выдачу кредита банком
- Ответ — вернет ли клиент кредит
- Бинарная классификация

Кредитный скоринг — признаки

- Бинарные: пол, наличие телефона, и т.д.
- Категориальные: место жительства, профессия, семейный статус, работодатель, и т.д.
- Порядковые: образование, должность, и т.д.
- Вещественные: возраст, зарплата, стаж работы, доход семьи, сумма кредита, и т.д.

Кредитный скоринг — особенности

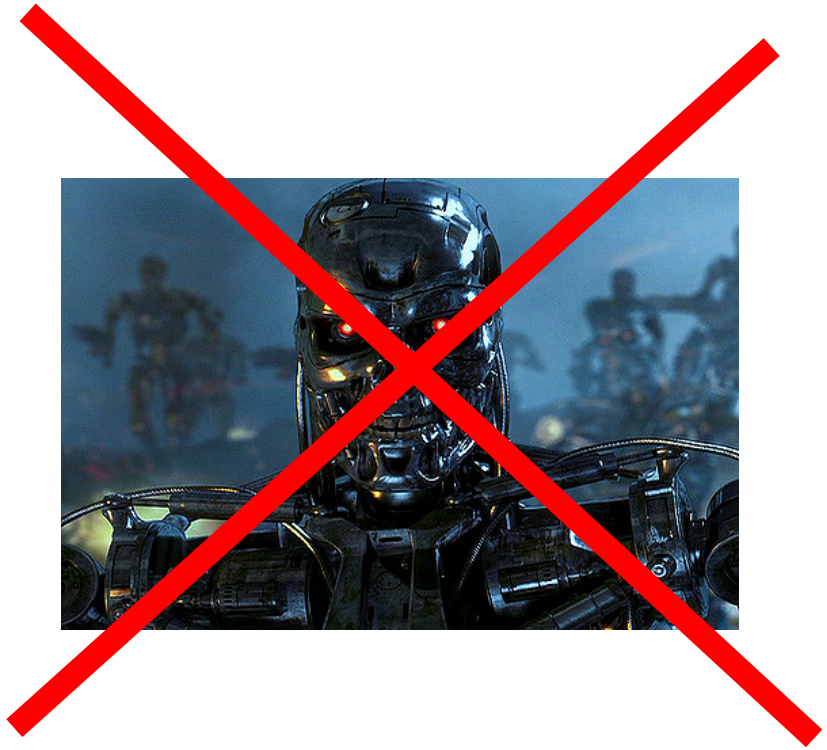
- Нужно оценивать вероятность дефолта

Откуда взять данные?

- **Социология**: лайки, комменты, посты в соцсетях и блогах
- **Международные отношения**: новости из разных источников, высказывания официальных лиц, курсы валют и акций
- **Экономика**: трансакции физических и юридических лиц, макроэкономические показатели, отчетность компаний
- **Лингвистика**: литературные произведения, общение в Интернете
- **Логистика**: данные по грузоперевозкам, загруженности складов, розничным продажам
- **Маркетинг**: поведение пользователей на сайте, онлайн-покупки и поисковые запросы, электронная переписка
- **Юриспруденция**: протоколы судов, данные о правоохранительной деятельности

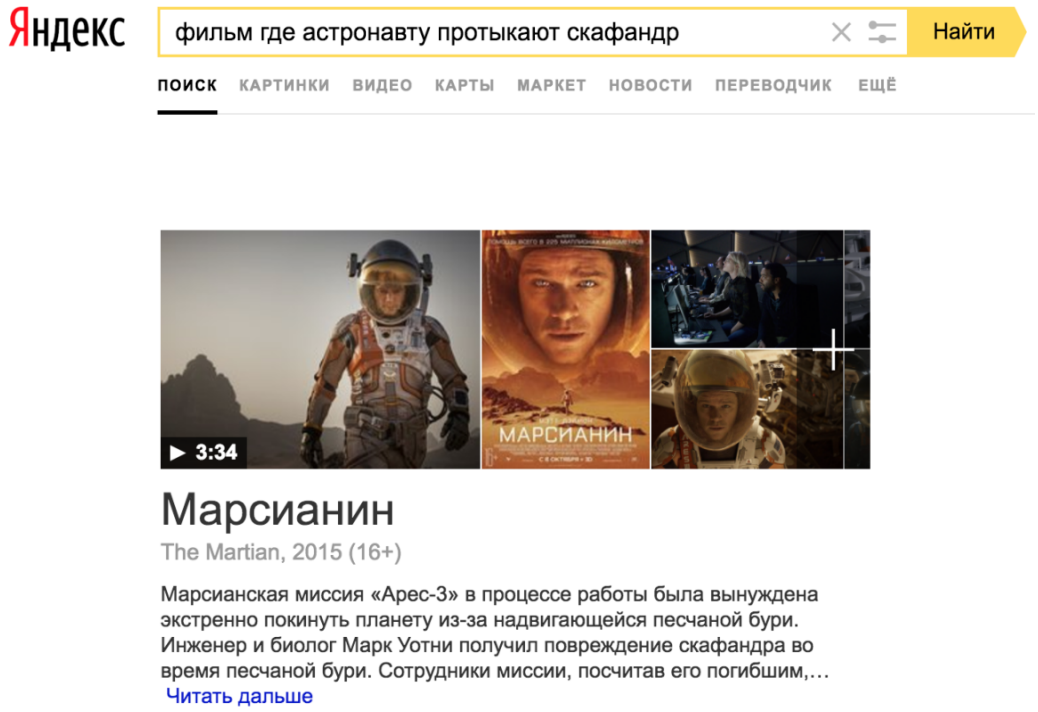
Зачем это нужно?

Искусственный интеллект



Сильный ИИ

через 20-100 лет



Специализированный ИИ

уже сейчас

Как можно заниматься анализом данных?

- Data scientist
  - Работа с данными
  - Знание инструментов и методов
  - Опыт решения задач
- Менеджер
  - Понимание, как работает машинное обучение
  - Понимание узких мест, оценивание сроков
- Заказчик
  - Метрики качества
  - Требования к данным
  - Ограничения современных подходов