

# Машинное обучение

Лекция 2  
Введение в машинное обучение

Андрей Нарцев  
[andrei.nartsev@gmail.com](mailto:andrei.nartsev@gmail.com)  
[anartsev@hse.ru](mailto:anartsev@hse.ru)

НИУ ВШЭ, 2024

# Организационное

Информация о курсе:

- Канал с объявлениями: <https://t.me/+TgIJAAxY7cNhNGYy>
- Материалы на github: <https://github.com/andrewnarts/hse-ml/tree/main/math-faculty-intro-ml/2024>



# Структура курса

## I модуль:

- Метод k ближайших соседей (введем основные понятия)
- Линейная регрессия (поговорим про градиентные методы оптимизации) – **ДЗ 1**
- Линейная классификация (в том числе, логистическая регрессия) – **ДЗ 2**
- Валидация моделей и оценка качества

## КОНТРОЛЬНАЯ РАБОТА

# Структура курса

## II модуль:

- Решающие деревья – **ДЗ 3**
- Композиции моделей (в том числе, градиентный бустинг) – **ДЗ 4**
- Ранжирование (небольшое введение)
- Кластеризация – **ДЗ 5** (бонус)
- Отбор признаков и снижение размерности

## ЭКЗАМЕН

# План лекции

## Введение в машинное обучение:

- Основные понятия (remind)
- Виды задач
- Типы признаков
- Отложенная выборка и переобучение (into)

## Метод k ближайших соседей:

- Гипотеза компактности
- Метрики

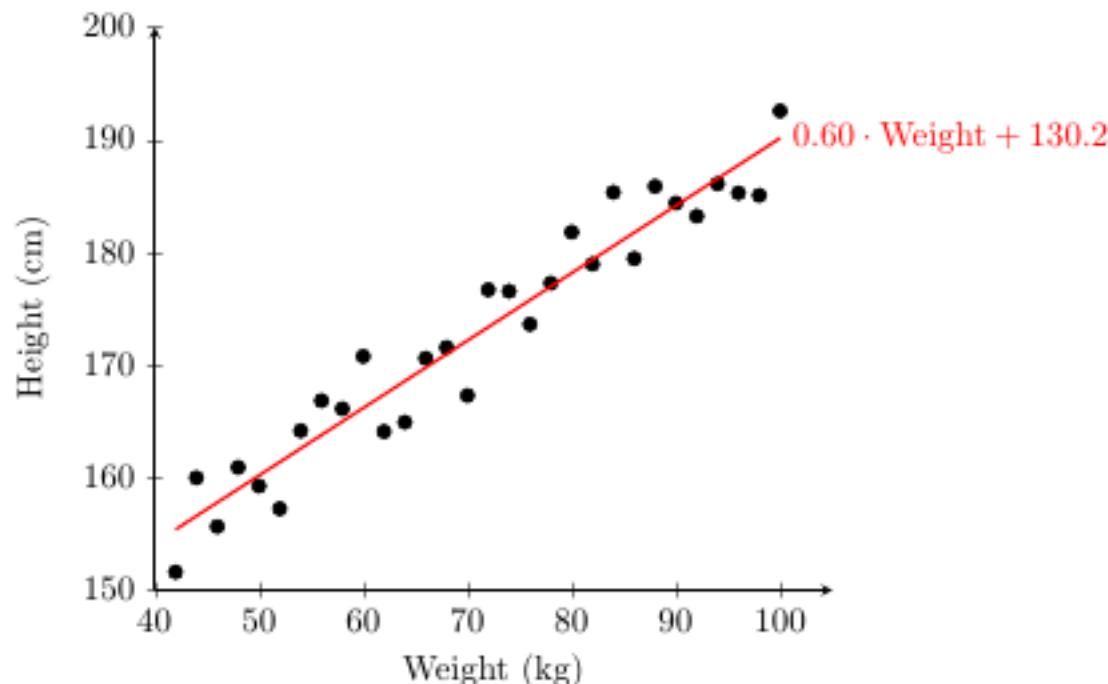
# Напоминание

- $\mathbb{X}$  — пространство объектов,  $\mathbb{Y}$  — пространство ответов
- $x = (x_1, \dots, x_d)$  — признаковое описание
- $X = (x_i, y_i)_{i=1}^\ell$  — обучающая выборка
- $a(x)$  — алгоритм, модель
- $Q(a, X)$  — функционал ошибки алгоритма  $a$  на выборке  $X$
- Обучение:  $a(x) = \arg \min_{a \in \mathcal{A}} Q(a, X)$

# Типы ответов

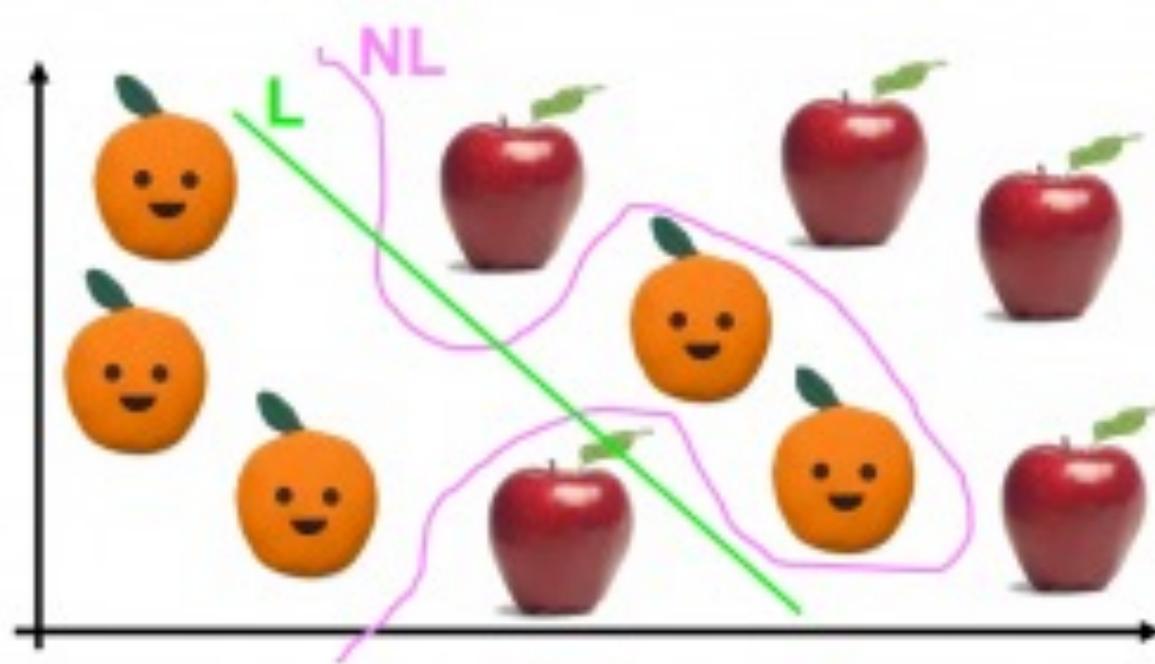
# Регрессия

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



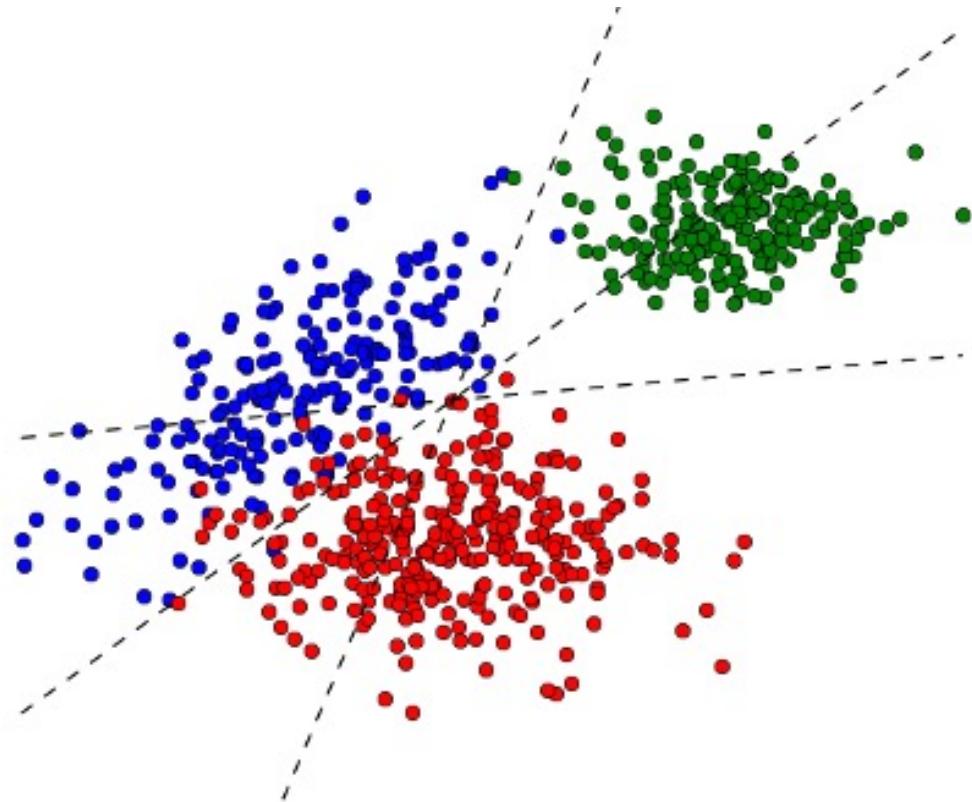
# Классификация

- 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  $K$  нулей и единиц
- $i$ -й элемент ответа — принадлежит ли объект  $i$ -му классу
- Какие темы присутствуют в статье?
- (математика, биология, экономика)

# Ранжирование

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

# Ранжирование

Яндекс

картинки с котиками — 5 млн ответов



Найти

Поиск

[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.

Картинки

Видео

[Уморные котики \(57 фото\) » Бяки.нет | Картинки](#)

[byaki.net > Картинки > 14026-umornye-kotiki-57...](#) ▾

Бяки нет! . NET. Уморные котики (57 фото). 223. Коментариев:9Автор:4ertonok

Просмотров:161 395 Картинки28-10-2008, 00:03.

Карты

Маркет

Ещё

[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.

# Кластеризация

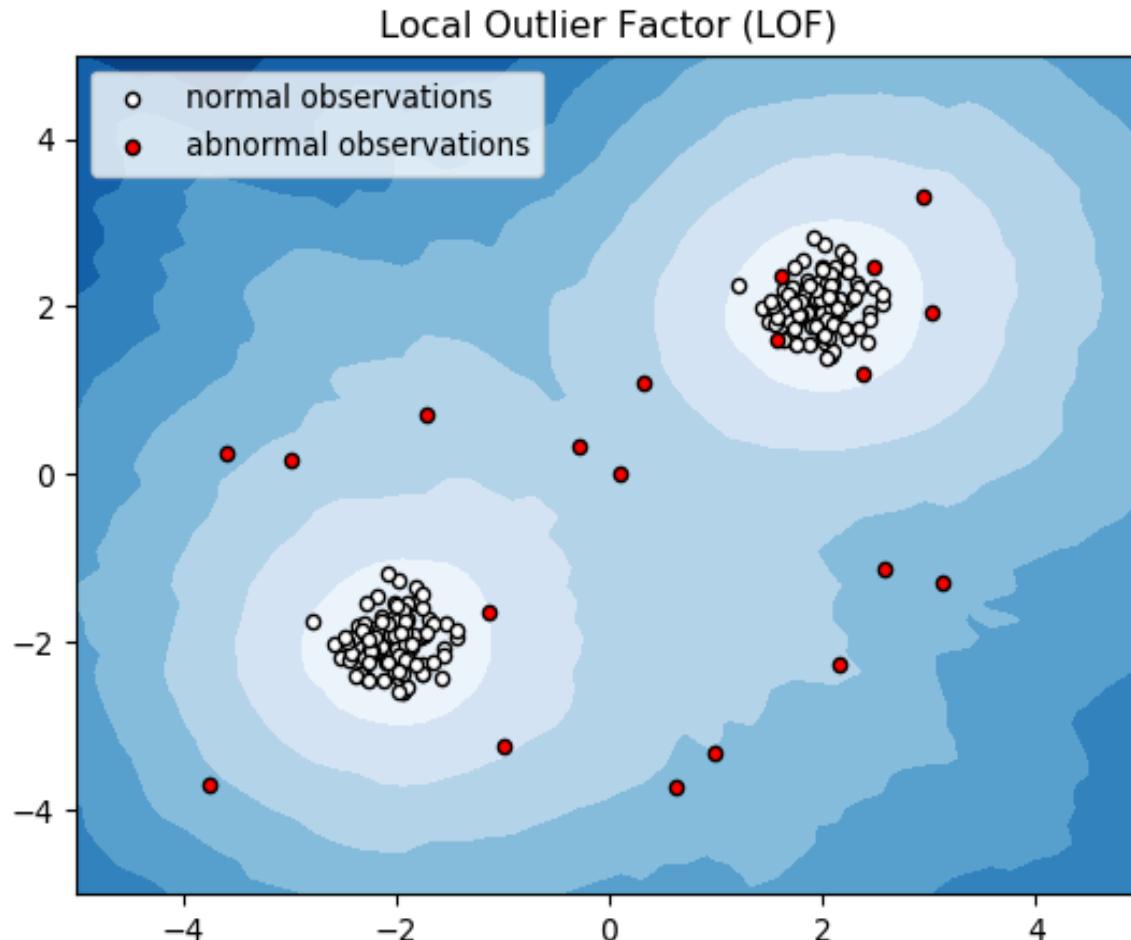
- $\mathbb{Y}$  — отсутствует
- Нужно найти группы похожих объектов
- Сколько таких групп?
- Как измерить качество?
- Пример: сегментация пользователей мобильного оператора

# ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация** – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности** – задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.
- **Оценивание плотности** – задача приближения распределения объектов.

# ПРИМЕР ОЦЕНИВАНИЯ ПЛОТНОСТИ

- Поиск аномалий с помощью оценивания плотностей



# ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация** – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности** – задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.
- **Оценивание плотности** – задача приближения распределения объектов.
- **Визуализация** – задача изображения многомерных объектов в 2х или 3хмерном пространстве с сохранением зависимостей между ними.

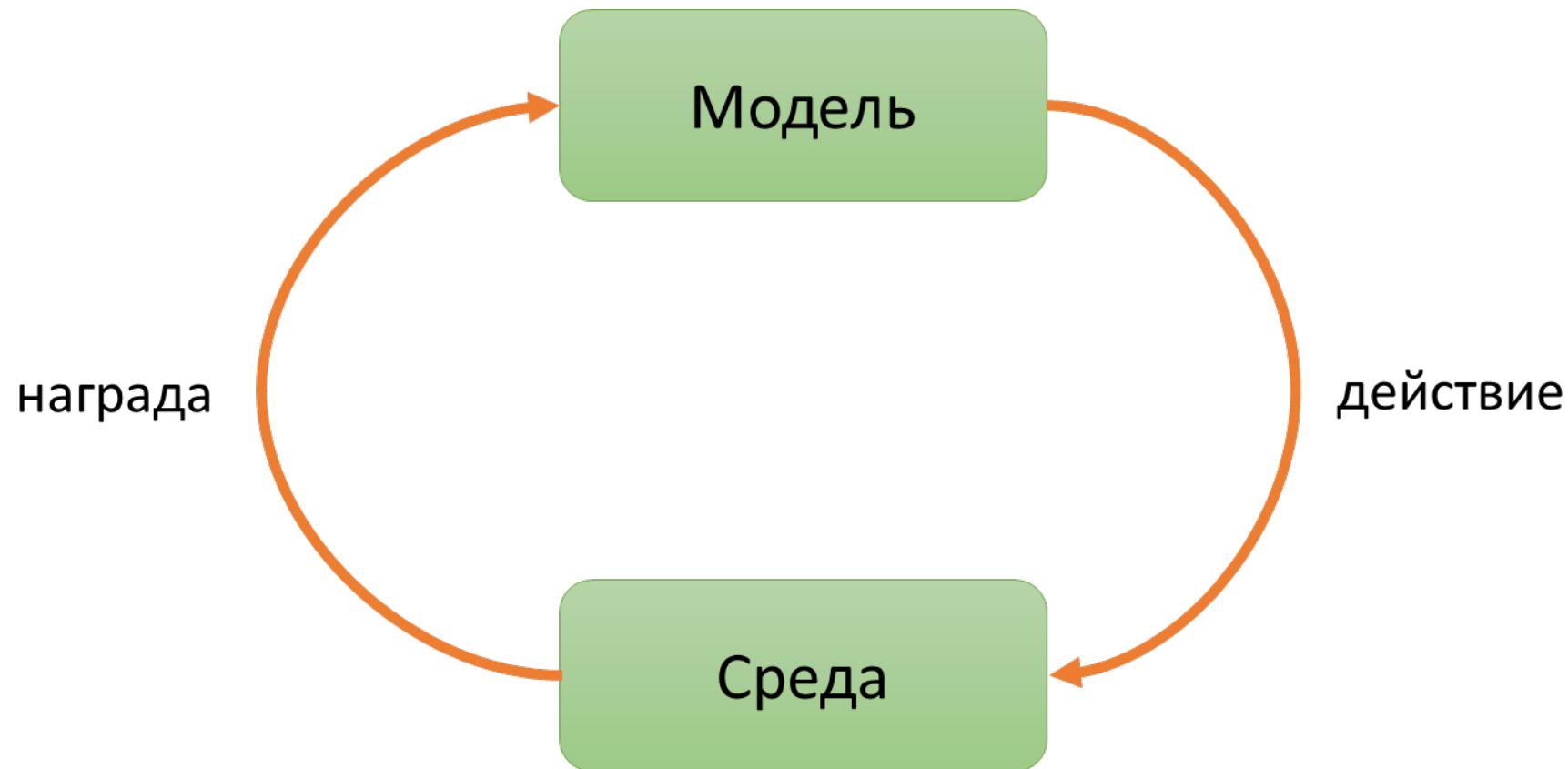
# ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.

# ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.
- Если нам неизвестны значения целевой переменной или целевая переменная вообще отсутствует, то есть алгоритм обучается только по признакам объектов, то это **обучение без учителя**. Примерами обучения с учителем являются кластеризация, понижение размерности и др.

# Обучение с подкреплением



# Типы признаков

# Типы признаков

- $D_j$  — 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

# Типы признаков

- Бинарные
- Числовые
- Категориальные и порядковые
- Есть и более сложные: тексты, изображения, звук и т.д.

# ОТЛОЖЕННАЯ ВЫБОРКА И ПЕРЕОБУЧЕНИЕ

# ОЦЕНКА ПРЕДСКАЗАТЕЛЬНОЙ СПОСОБНОСТИ АЛГОРИТМА

- Пусть мы решаем задачу *предсказания стоимости дома по его признакам.*



- В обучающей выборке 1000 домов.
- Мы обучаем алгоритм по имеющимся 1000 домам. *На каких объектах будем проверять качество алгоритма?*

# ОЦЕНКА ПРЕДСКАЗАТЕЛЬНОЙ СПОСОБНОСТИ АЛГОРИТМА

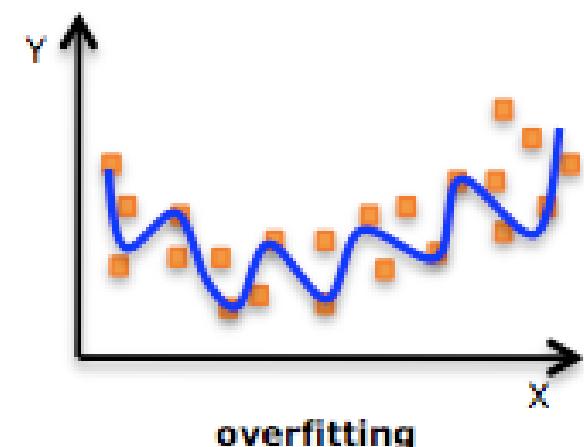
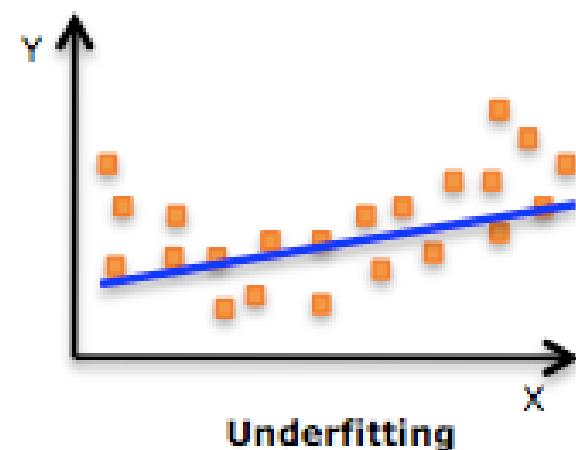
- Перед началом обучения отложим часть обучающих объектов и не будем использовать их для построения модели (отложенная выборка).



# ОТЛОЖЕННАЯ ВЫБОРКА

- Перед началом обучения отложим часть обучающих объектов и не будем использовать их для построения модели (отложенная выборка).
- Тогда можно измерить качество построенной модели на отложенной выборке и оценить ее предсказательную силу.

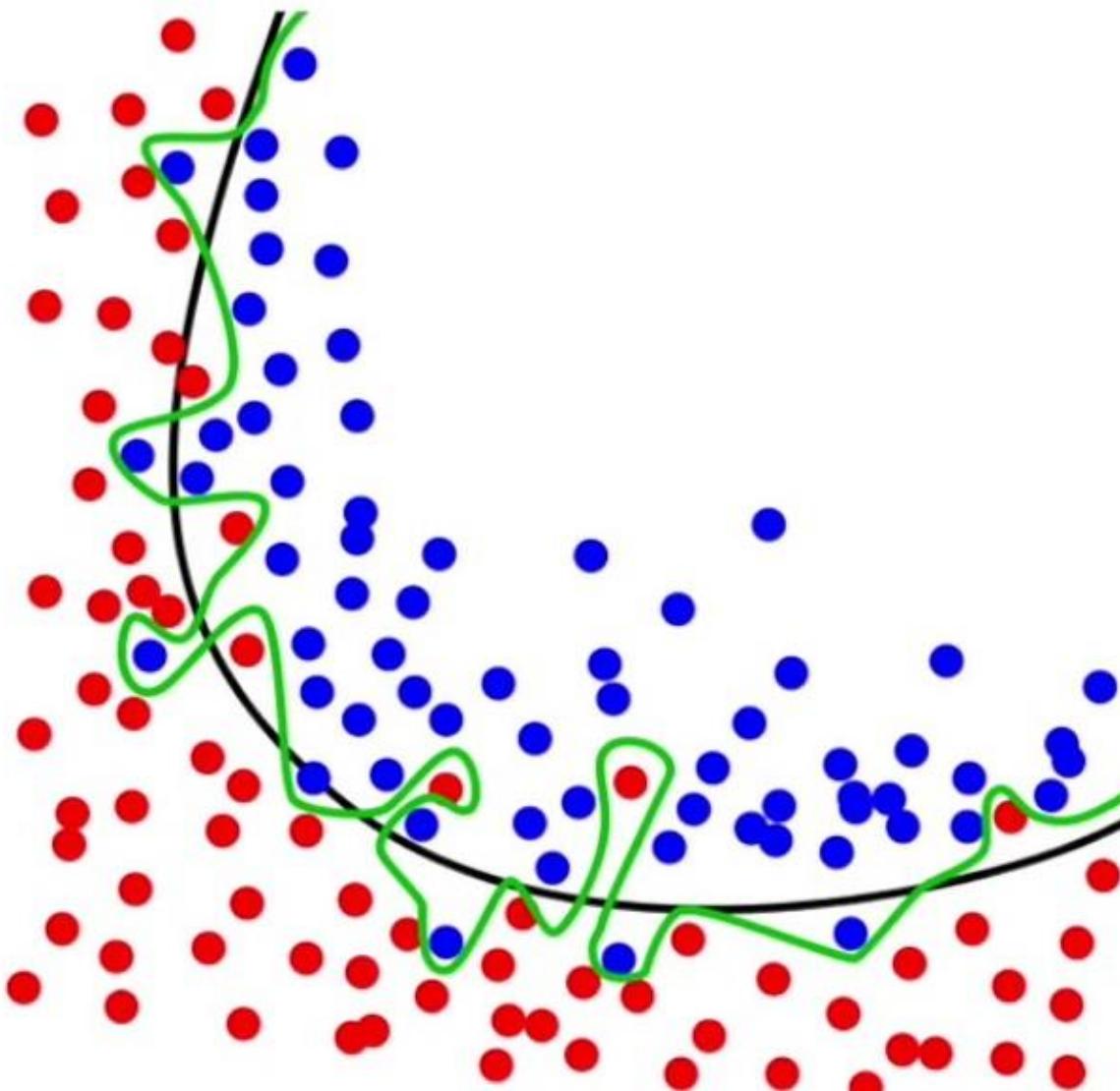
# ПЕРЕОБУЧЕНИЕ И НЕДООБУЧЕНИЕ



# ИЗ-ЗА ЧЕГО ВОЗНИКАЕТ ПЕРЕОБУЧЕНИЕ

- Избыточная сложность модели (большое количество весов). В этом случае лишние степени свободы в модели “тратятся” на чрезмерно точную подгонку под обучающую выборку.
- Переобучение есть всегда, когда есть оптимизация параметров по конечной (заведомо неполной) выборке.

# ПРИМЕР ПЕРЕОБУЧЕНИЯ В ЗАДАЧЕ КЛАССИФИКАЦИИ



Гипотеза компактности и knn

# Как отличить ель от сосны?



# Как отличить ель от сосны?



# Как отличить ель от сосны?



Ель:

- Ветки смотрят вверх
- Ствол не видно
- Густые иголки
- Цвет ближе к зелёному



Сосна:

- Ветки параллельны земле
- Ствол видно
- Иголки более редкие
- Цвет ближе к жёлтому

# Как отличить ель от сосны?



Ветки вверх  
Ствол не видно  
Густые иголки  
Цвет ближе к синему

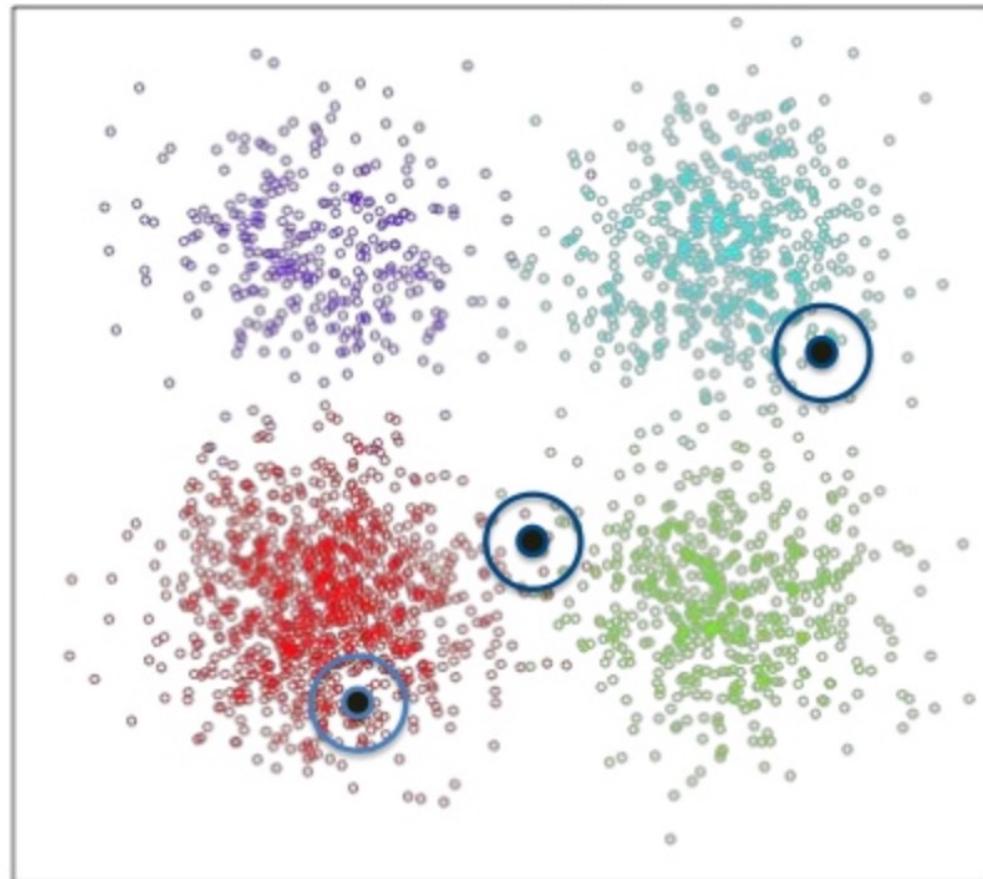


Скорее всего ель

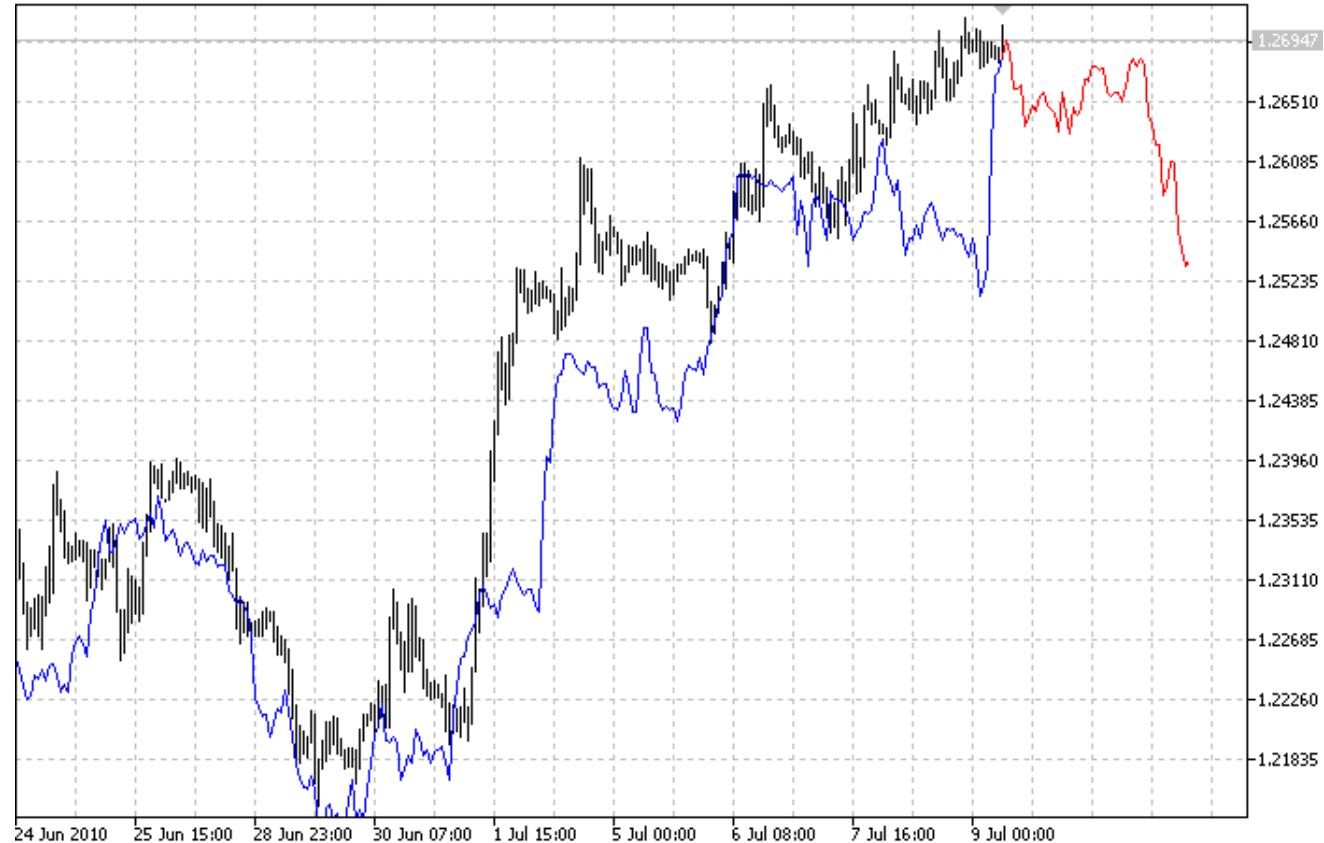
# Что такое обучение?

- Запоминаем примеры (объекты и ответы)
- Когда приходит новый объект, сравниваем с запомненными примерами
- Выдаём ответ от наиболее похожего примера

# Гипотеза компактности



# Гипотеза компактности



# Гипотеза компактности

Если два объекта похожи друг на друга, то ответы на них  
тоже похожи

# kNN: обучение

- Дано: обучающая выборка  $X = (x_i, y_i)_{i=1}^\ell$
- Задача классификация (ответы из множества  $\mathbb{Y} = \{1, \dots, K\}$ )
- Обучение модели:
  - Запоминаем обучающую выборку  $X$

# kNN: применение

Дано: новый объект  $x$

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:  
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем  $k$  ближайших объектов:  $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k [y_{(i)} = y]$$

# kNN: применение

Дано: новый объект  $x$

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:  
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем  $k$  ближайших объектов:  $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k [y_{(i)} = y]$$

# kNN: применение

Дано: новый объект  $x$

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:  
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем  $k$  ближайших объектов:  $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k [y_{(i)} = y]$$

# kNN: применение

Дано: новый объект  $x$

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:  
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем  $k$  ближайших объектов:  $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k [y_{(i)} = y]$$

# kNN: применение

