

Основы машинного обучения

Лекция 2

Метод k ближайших соседей

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2023

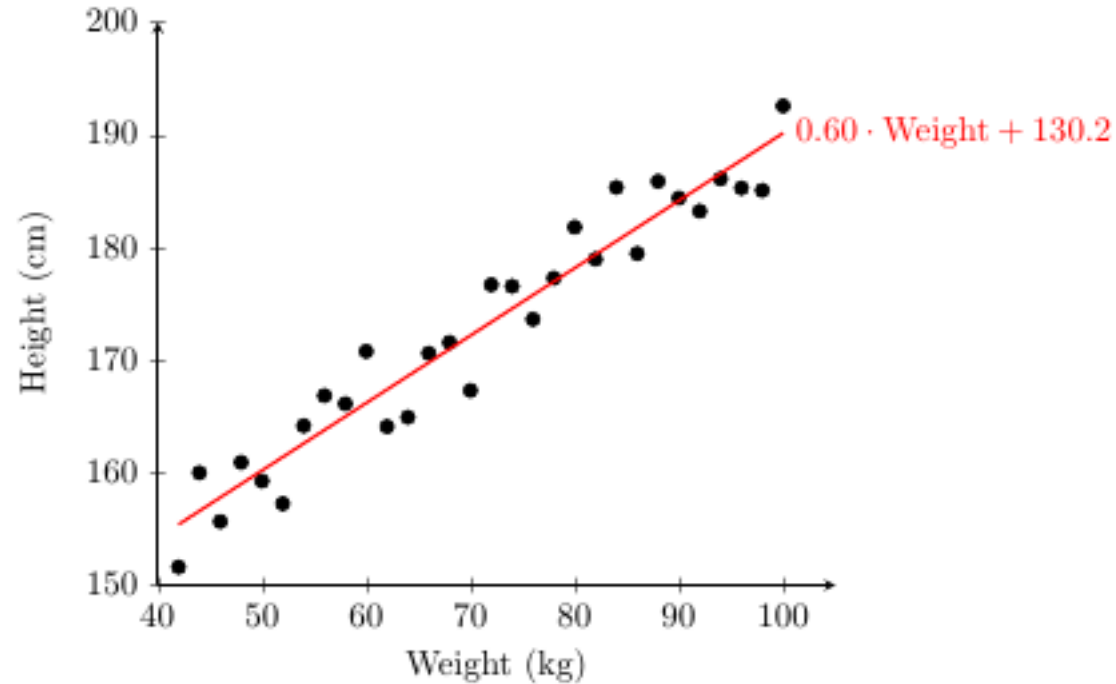
Напоминание

- \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

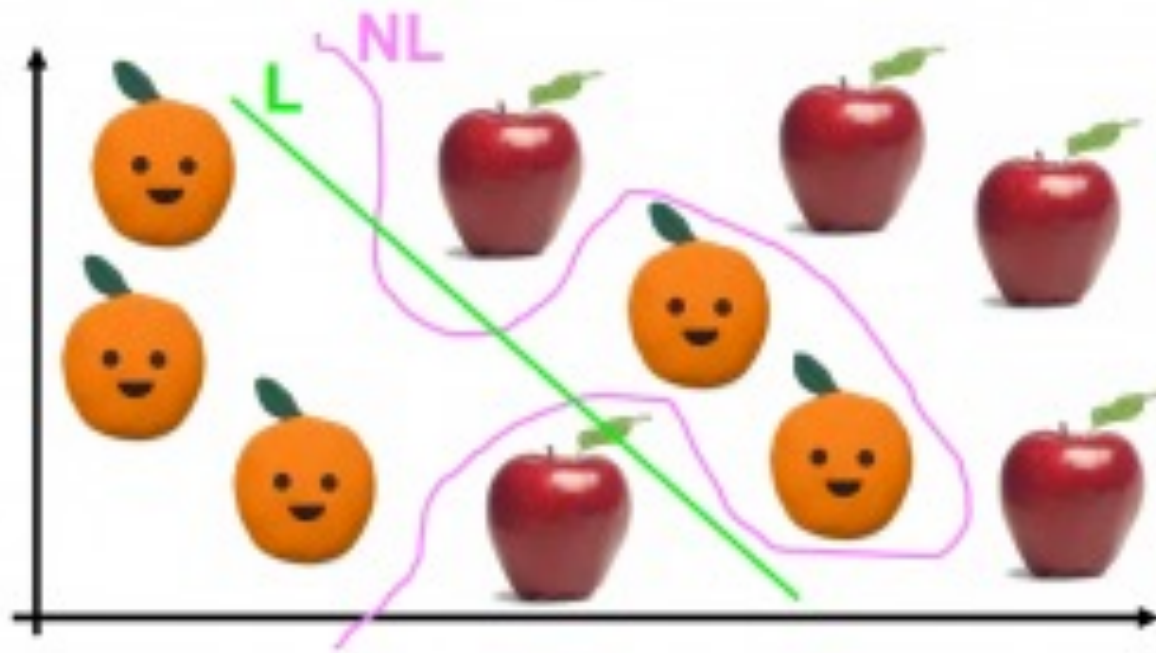
Регрессия

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



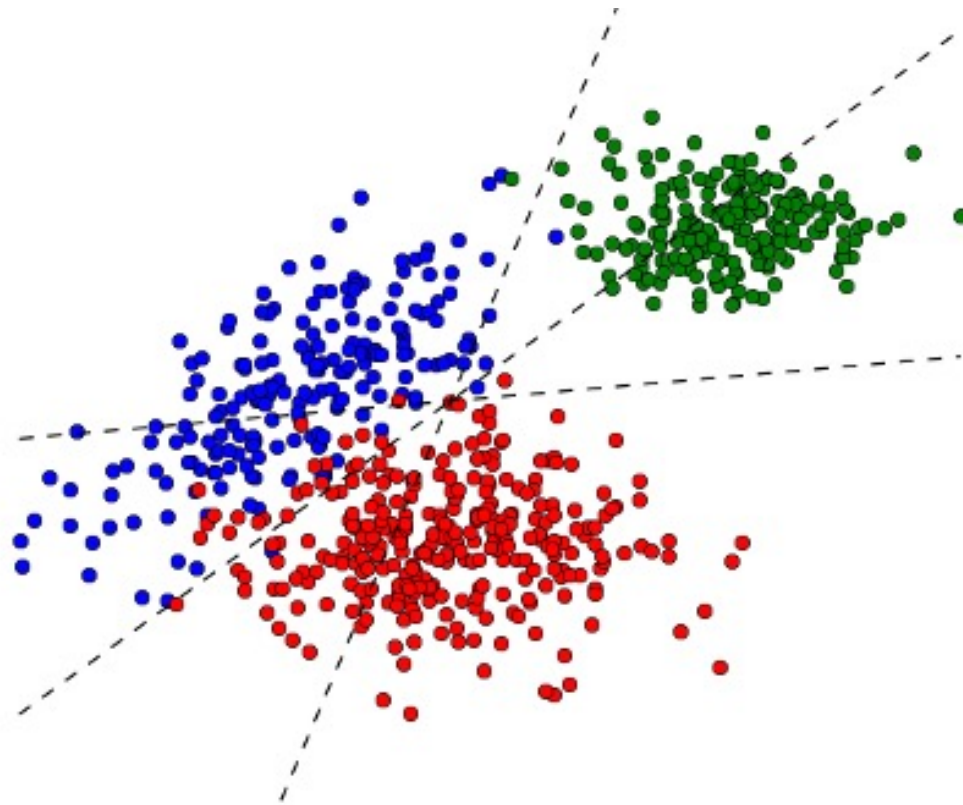
Классификация

- 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

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. Комментарии:9Автор:4ertonok
Просмотров:161 395 **Картинки**28-10-2008, 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 Котят | картинки на стол Котят

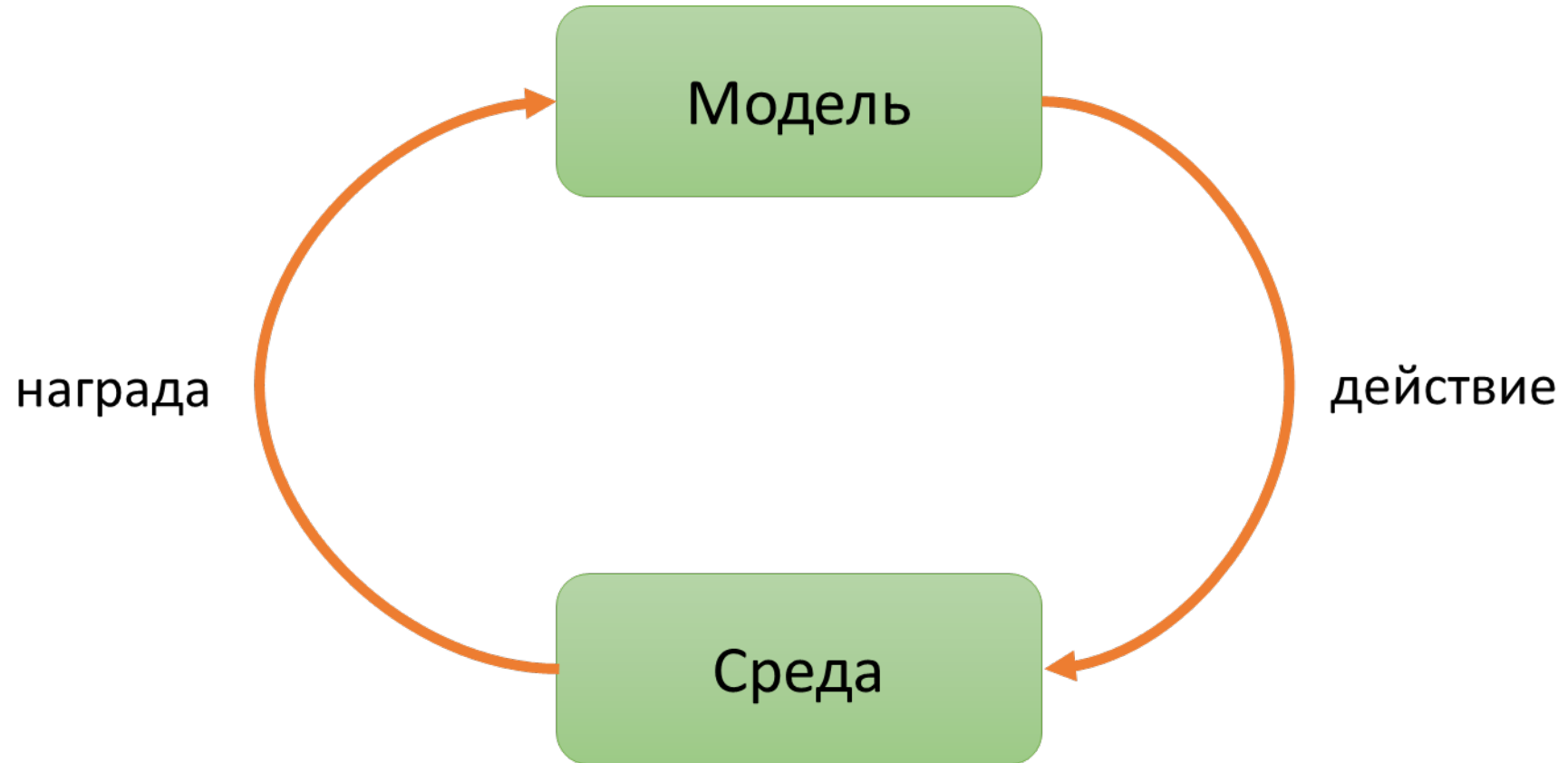
[7fon.ru](#) > Чёрные обои и **картинки** > Обои котят ▼

Картинки Котят с 1 по 15. **Обои** для рабочего стола Котят. ... Скачать **Картинки** Котят на рабочий стол бесплатно.

Кластеризация

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

Обучение с подкреплением



Типы задач

- Регрессия
- Классификация
- Кластеризация
- Много других: ранжирование, поиск аномалий и т.д.

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

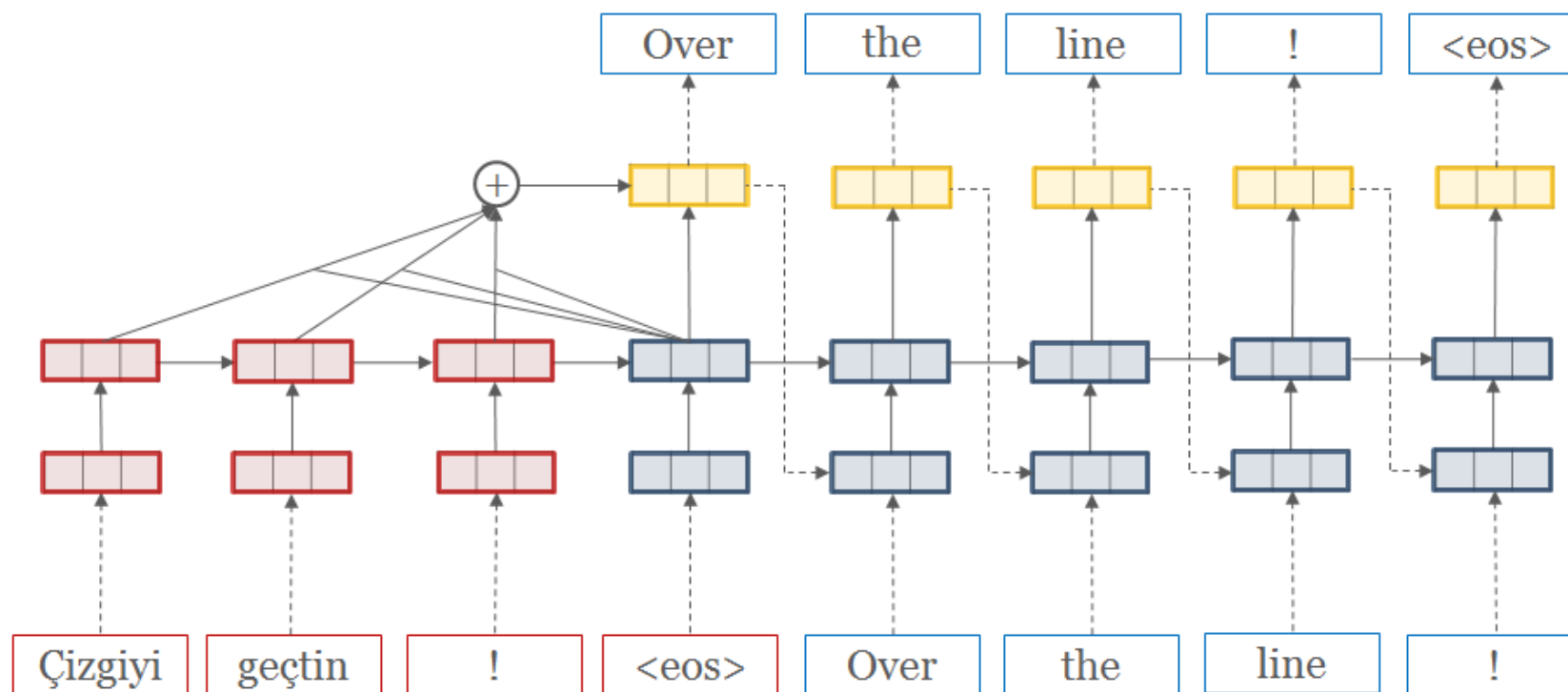
- D_j — упорядоченное множество
- Военское звание
- Роль в фильме (первого плана, второго плана, массовка)
- Тип населенного пункта

Типы признаков

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

Зачем это нужно?

Машинный перевод

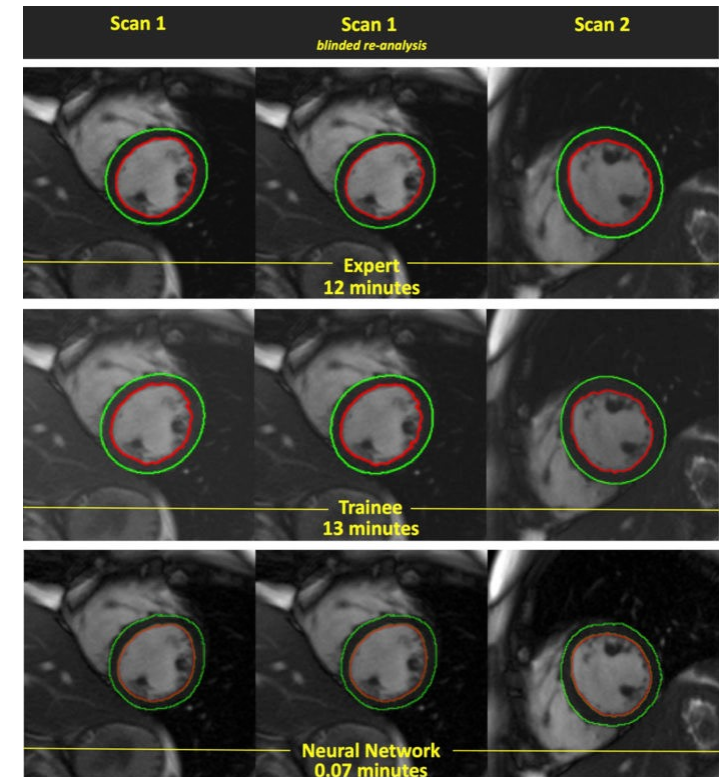


Генерация текста

- GPT-3 от OpenAI
- <https://arxiv.org/abs/2005.14165>
- <https://talktotransformer.com>

Биоинформатика и медицина

- Поиск связей между ДНК и заболеваниями (23andme и другие)
- Таргетные лекарства
- Анализ медицинских снимков



Сельское хозяйство

- Робототехника
- Мониторинг посевов и почвы
- Прогнозирование болезней и урожайности



Рекомендательные системы

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

Frequently Bought Together



Price For All Three: **\$86.01**

[Add all three to Cart](#) [Add all three to Wish List](#)

[Show availability and shipping details](#)

- ☒ **This item:** Machine Learning for Hackers by Drew Conway Paperback **\$33.87**
- ☒ Machine Learning in Action by Peter Harrington Paperback **\$25.75**
- ☒ Programming Collective Intelligence: Building Smart Web 2.0 Applications by Toby Segaran Paperback **\$26.39**

Customers Who Bought This Item Also Bought

Page 1 of 17

Programming Collective Intelligence: Building ... Toby Segaran ★★★★☆ (84) Paperback \$26.39	Machine Learning in Action Peter Harrington ★★★★☆ (10) Paperback \$25.75	Mining the Social Web: Analyzing Data from ... Matthew A. Russell ★★★★☆ (19) Paperback \$26.36	Data Analysis with Open Source Tools Philipp K. Janert ★★★★☆ (29) Paperback \$24.05	R Cookbook (O'Reilly Cookbooks) Paul Teetor ★★★★☆ (18) Paperback \$32.43	The Art of R Programming: A Tour of Statistical ... Norman Matloff ★★★★☆ (29) Paperback \$25.06

Are any of these items inappropriate for this page? [Let us know](#)

Зачем это нужно?

- Это круто
 - Сложные задачи
 - Движение к искусственному интеллекту (что бы это ни значило)
- Это полезно
 - Извлечение прибыли из данных
 - Data-driven companies

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

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

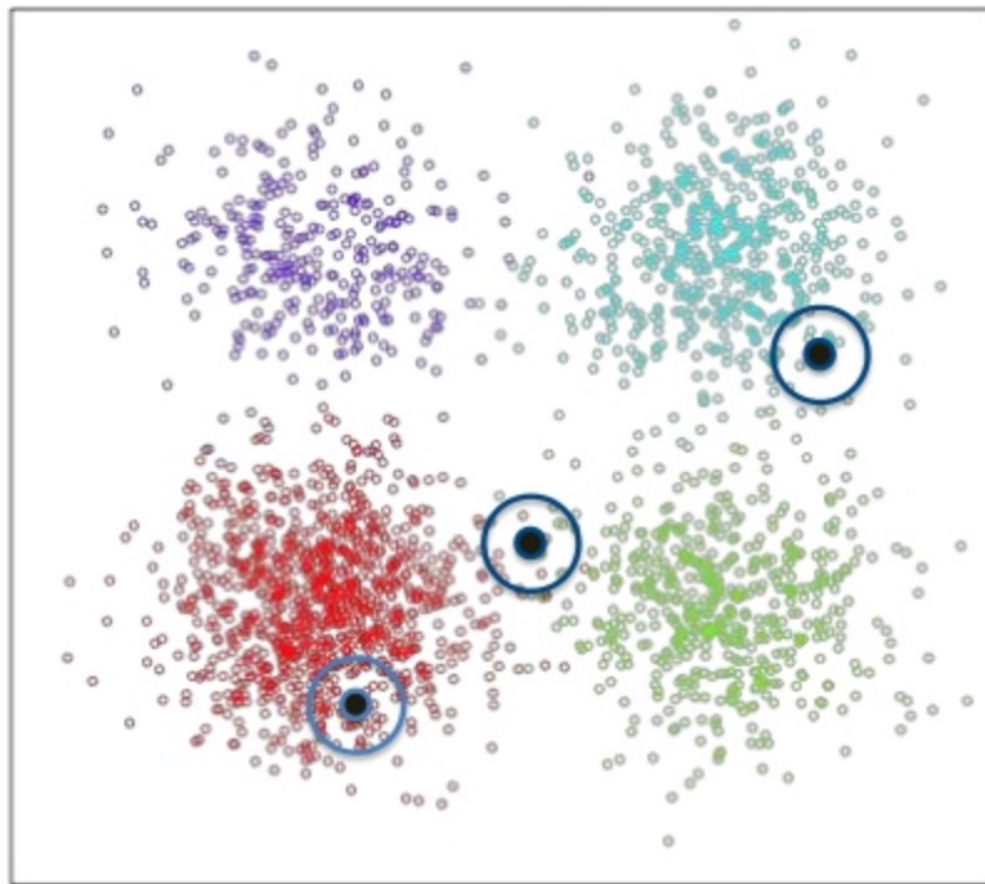


Скорее всего ель

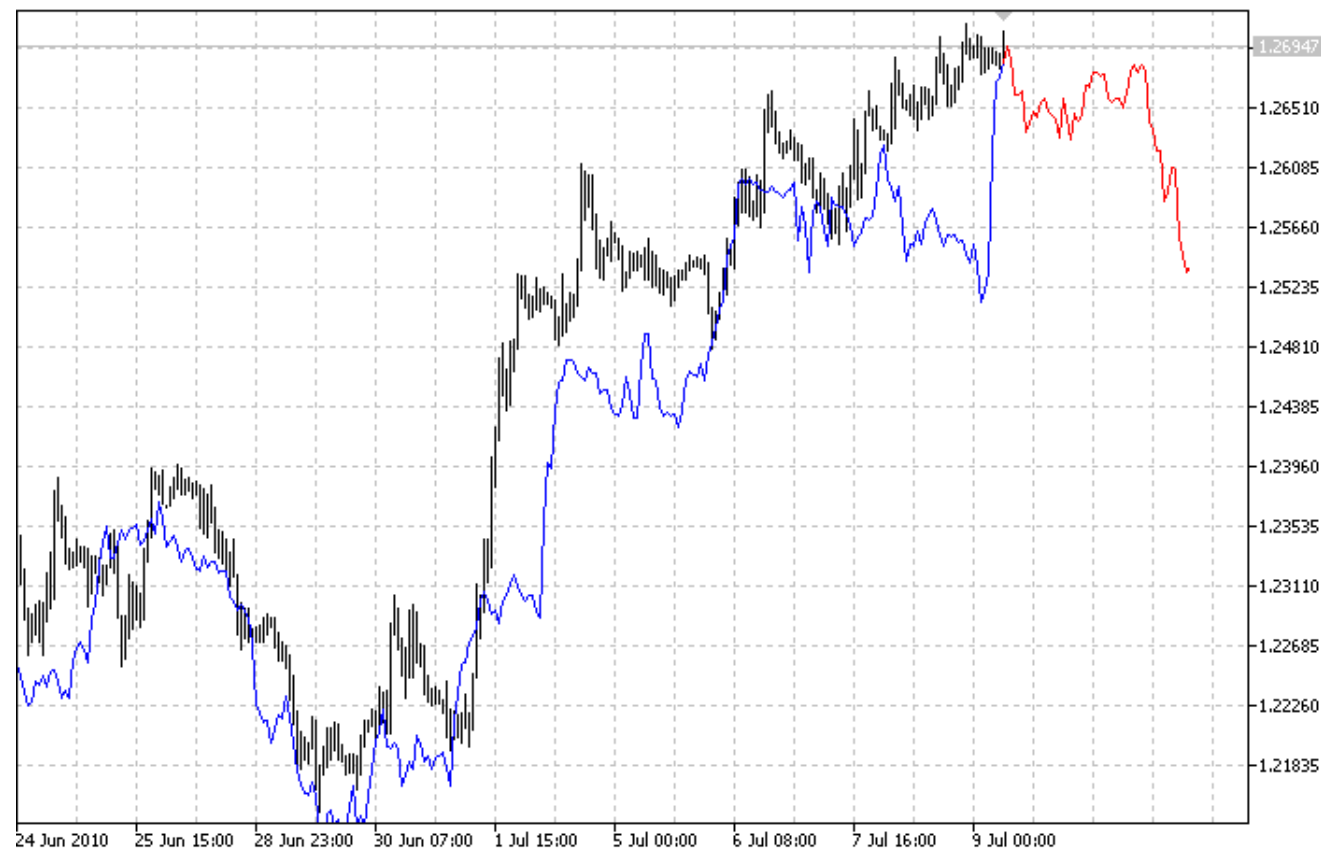
Что такое обучение?

- 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:

- Сертируем 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

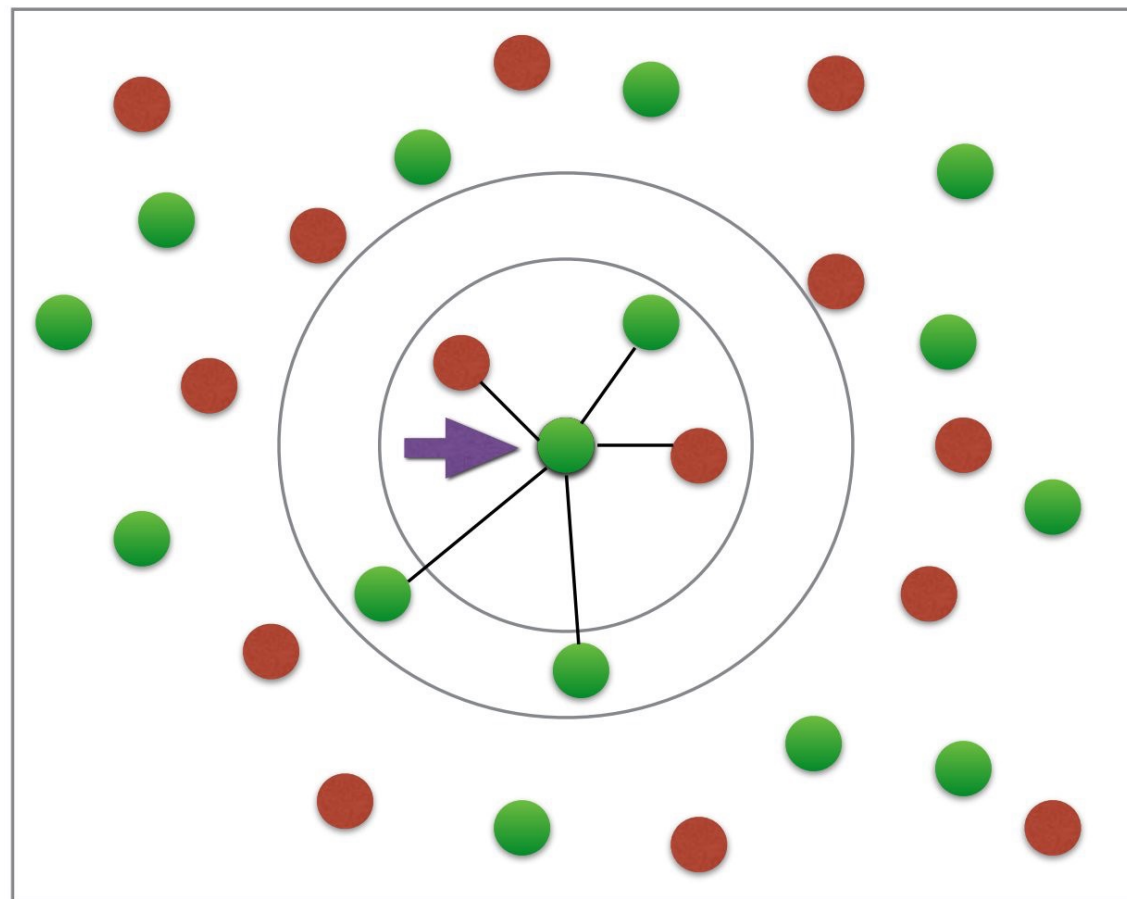
Дано: новый объект x

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



Сравнение объектов и метрики

Числовые данные

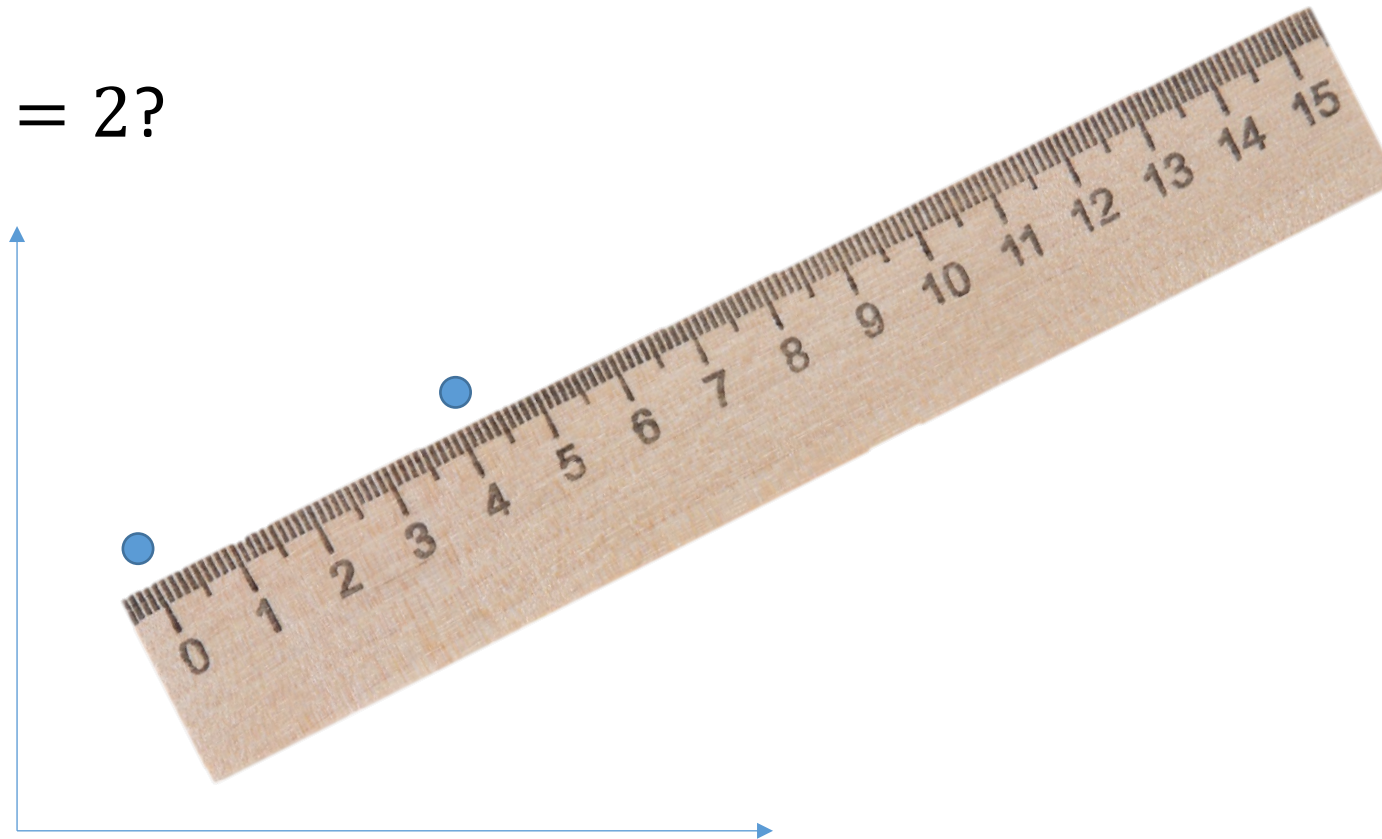
Сколько раз в день вызывает такси	Средние расходы на такси в день	Как часто вызывал комфорт	Возраст	Согласился повысить категорию?
2	400	0.3	29	да
0.3	80	0	28	нет
...

Числовые данные

- Каждый объект описывается набором из d чисел — **вектором**
- Если x — вектор, то x_i — его i -я координата
- Если x_i — вектор, то x_{ij} — его j -я координата

Числовые данные

- Каждый объект описывается набором из d чисел — **вектором**
- Что, если $d = 2$?



Метрика

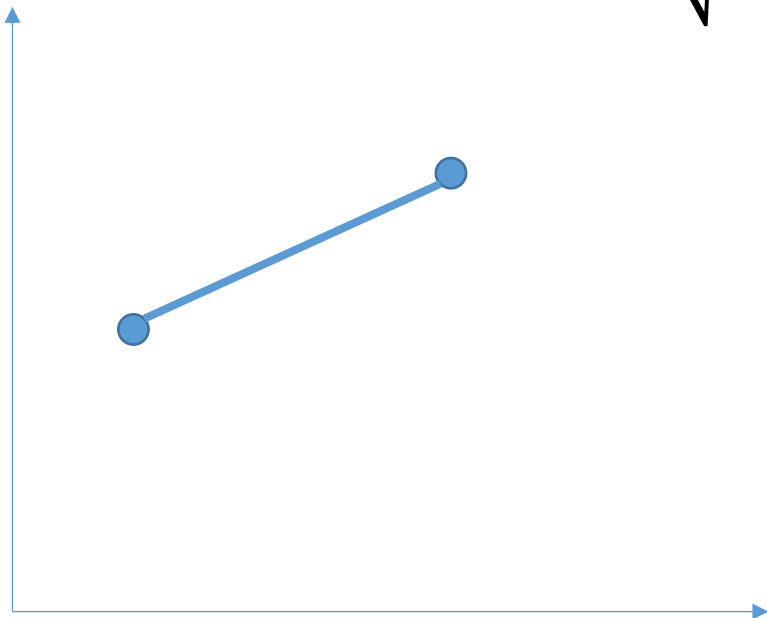
Метрика — обобщение расстояния на многомерные пространства

Метрика — это функция ρ с двумя аргументами, удовлетворяющая трём требованиям:

- $\rho(x, z) = 0$ тогда и только тогда, когда $x = z$
- $\rho(x, z) = \rho(z, x)$
- $\rho(x, z) \leq \rho(x, v) + \rho(v, z)$ — неравенство треугольника

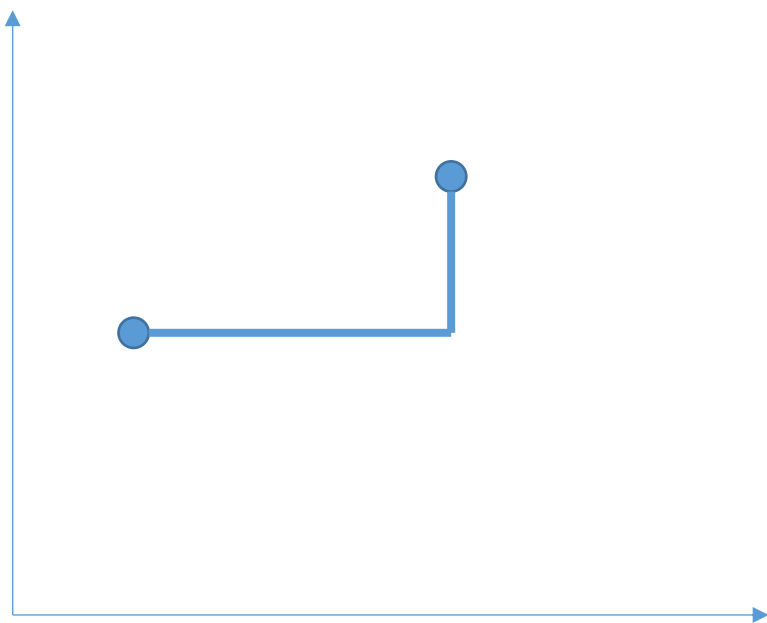
Евклидова метрика

$$\rho(x, z) = \sqrt{\sum_{j=1}^d (x_j - z_j)^2}$$

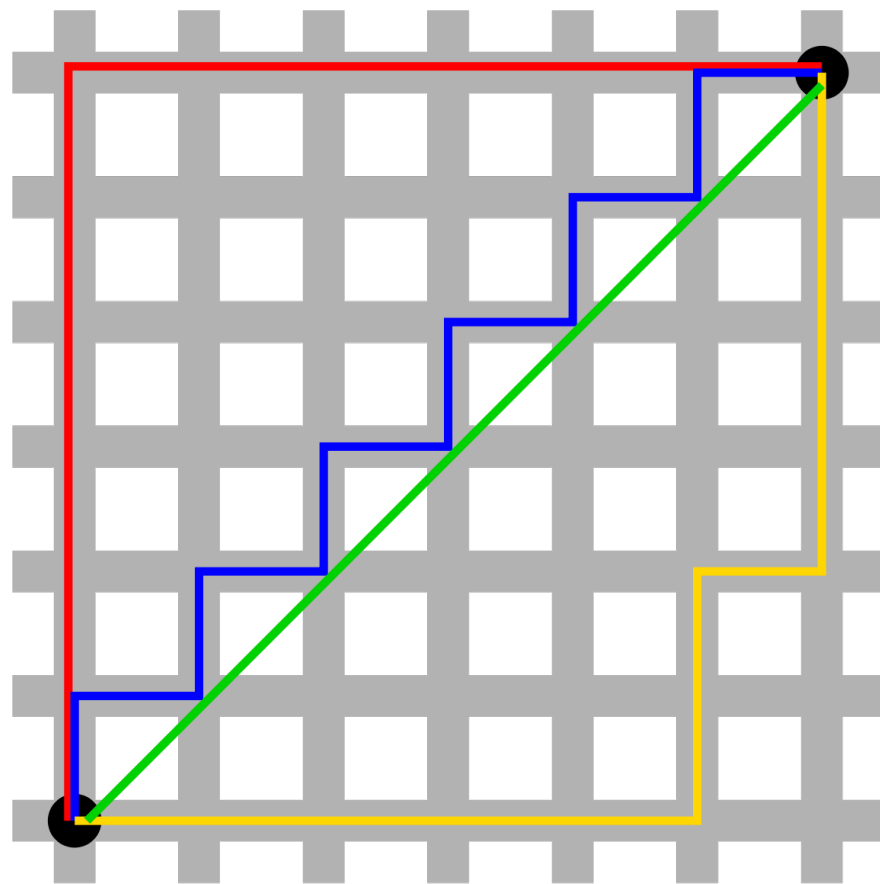


Манхэттенская метрика

$$\rho(x, z) = \sum_{j=1}^d |x_j - z_j|$$



Сравнение



Обобщение

$$\rho(x, z) = \sqrt[p]{\sum_{j=1}^d |x_j - z_j|^p}$$

- Метрика Минковского
- Можно подбирать p под конкретную задачу