IT-Academy

Метрики расстояний и алгоритм KNN

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ

Вопросы занятия

1. Что такое близость объектов и в каких задачах это применяется

2. Идея и особенности алгоритма KNN;

3. Пример решения задачи классификации KNN: практика;

4. Пример решения задачи регрессии через KNN: практика.

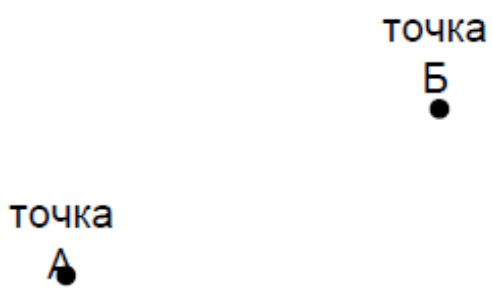
В конце занятия научимся:

- будете знать как выбирать метрики близости;
- познакомитесь с алгоритмом KNN;
- потренируемся на различных метриках;
- реализуете в коде задачу классификации и регрессии с помощью алгоритма KNN.

1. МЕТРИКИ РАССТОЯНИЙ

Евклидово расстояние

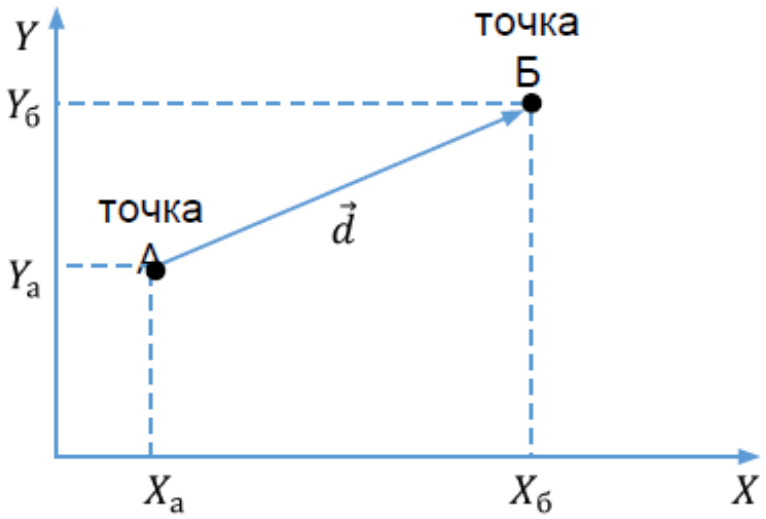
ТОЧКИ НА ПЛОСКОСТИ



1. МЕТРИКИ РАССТОЯНИЙ

Евклидово расстояние

ТОЧКИ НА ПЛОСКОСТИ

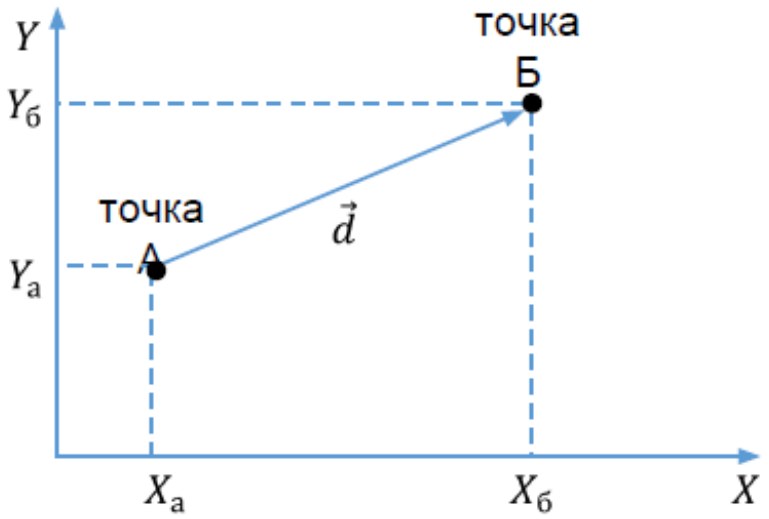


$$d = \sqrt{(X_б - X_а)^2 + (Y_б - Y_а)^2}$$

1. МЕТРИКИ РАССТОЯНИЙ

Евклидово расстояние

ТОЧКИ НА ПЛОСКОСТИ



$$d = \sqrt{(X_б - X_а)^2 + (Y_б - Y_а)^2}$$

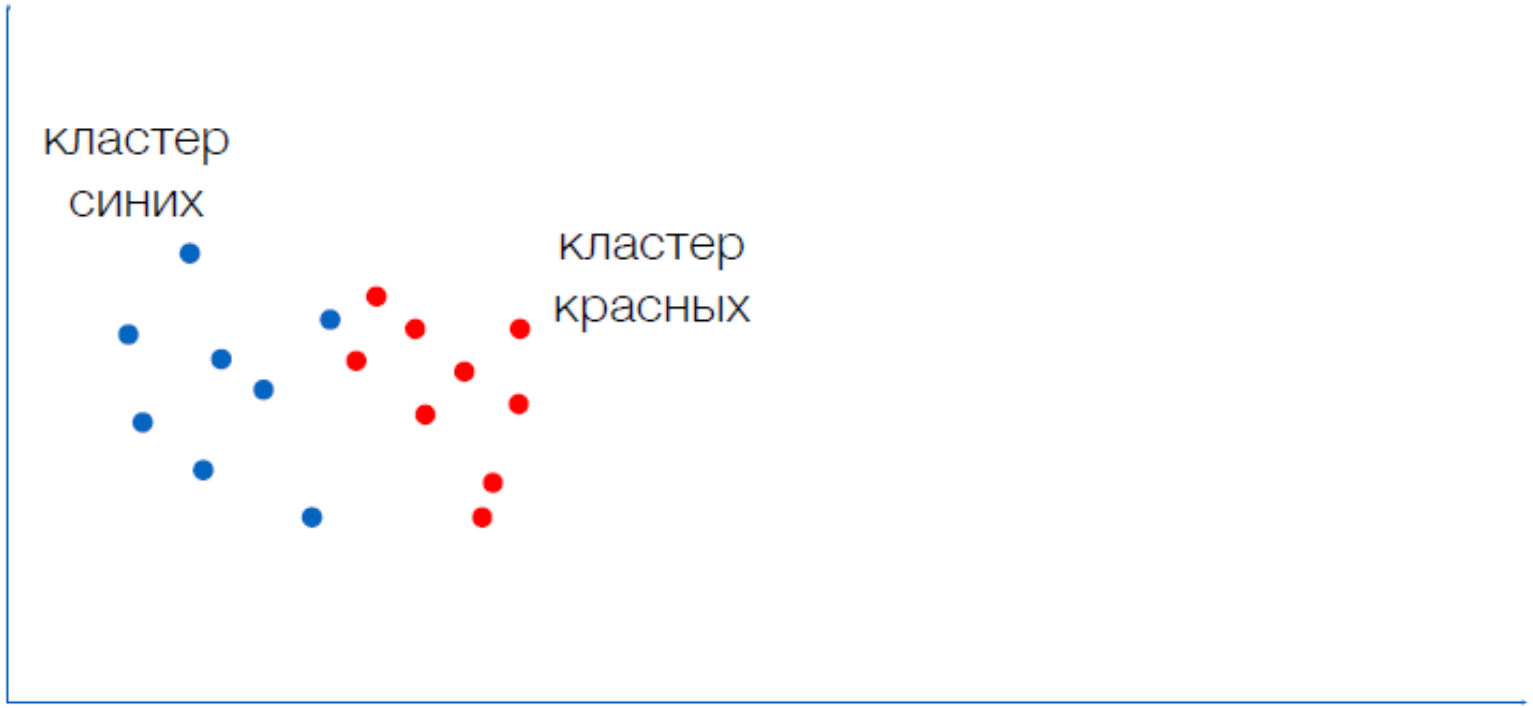
$$d = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

K NEAREST NEIGHBOR

К БЛИЖАЙШИХ СОСЕДЕЙ

IT-Academy

ИДЕЯ АЛГОРИТМА



К БЛИЖАЙШИХ СОСЕДЕЙ

IT-Academy

ИДЕЯ АЛГОРИТМА

к какому кластеру отнести зеленую точку?

IT-Academy

ИДЕЯ АЛГОРИТМА

Берем K ближайших соседей к зеленой точке. Берем класс, наиболее часто встречающийся среди соседей.

Варианты:

- Берем ближайшую точку (k = 1) – группа синих
- Учитываем несколько соседей (k = 4) – группа красных
- Учитываем вес, обратно пропорциональный расстоянию до точки

К БЛИЖАЙШИХ СОСЕДЕЙ

IT-Academy

ПРЕИМУЩЕСТВА И НЕДОСТАТКИ

+ Простая реализация и интерпретация

+ Применим ко многим задачам классификации и регрессии

- Число соседей нужно задавать заранее, что иногда определяет результат

- Плохо работает при сильно пересекающихся данных

IT-Academy

ПРАКТИКА

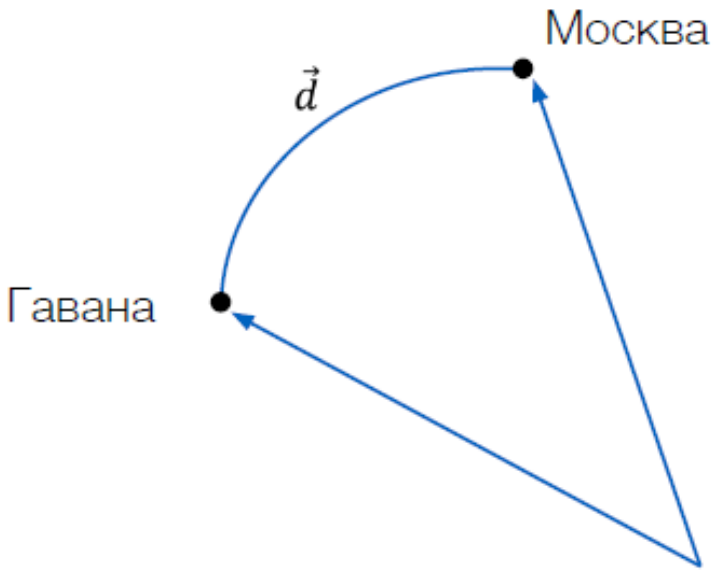
knn.ipynb

1. МЕТРИКИ РАССТОЯНИЙ

Полярные координаты

IT-Academy

УЧЕТ КРИВИЗНЫ ПОВЕРХНОСТИ



d – длина дуги в полярных координатах

1. МЕТРИКИ РАССТОЯНИЙ

Манхэттенское расстояние

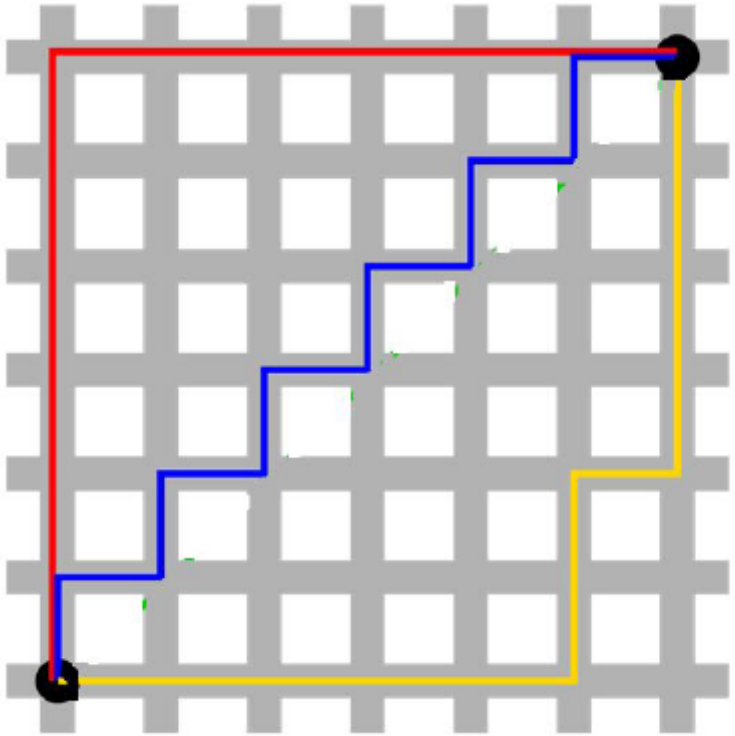
Улицы Манхэттена перпендикулярны друг другу

IT-Academy

1. МЕТРИКИ РАССТОЯНИЙ
Манхэттенское расстояние

IT-Academy

ДЛИНЫ ВСЕХ ПУТЕЙ РАВНЫ



Расстояние городских кварталов

$$d = \sum_{i=1}^{n} |X_i - Y_i|$$

1. МЕТРИКИ РАССТОЯНИЙ

Манхэттенское расстояние

IT-Academy

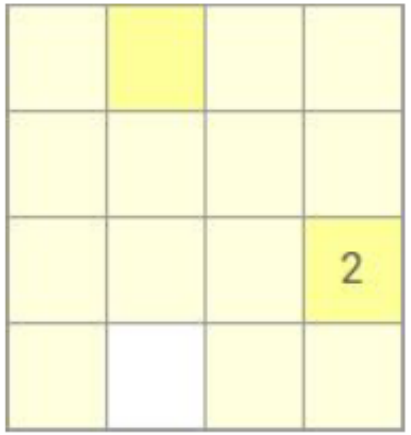
ДЛЯ ПОИСКА ОПТИМАЛЬНОГО РЕШЕНИЯ

Сумма манхэттенских расстояний между костяшками и позициями, в которых они находятся в решённой головоломке «Пятнашки», используется в качестве эвристической функции для поиска оптимального решения

1. МЕТРИКИ РАССТОЯНИЙ

Манхэттенское расстояние

ДЛЯ ПОИСКА ОПТИМАЛЬНОГО РЕШЕНИЯ



В примере манхэттенское расстояние равно 4

IT-Academy

ПРАКТИКА

15.ipynb

МЕТРИКИ БЛИЗОСТИ ОБЪЕКТОВ

IT-Academy

СРАВНЕНИЕ ТЕКСТОВ

МЕТРИКИ БЛИЗОСТИ ОБЪЕКТОВ

IT-Academy

СТАРТОВЫЙ ЛИСТ

```
1   Шехавцова Анна  Ж   1998    РГАУ-МСХА
2   Гречихина Наталья   Ж   1994    МГУ
3   Козлова Алена   Ж   1994    МГУ
4   Груздева Алина  Ж   1998    РГУНГ
5   Кущенко Анна    Ж   1997    МГУ
6   Чистякова Анастасия Ж   1998    РГАУ-МСХА
```

МЕТРИКИ БЛИЗОСТИ ОБЪЕКТОВ

IT-Academy

РАСПОЗНАВАНИЕ РЕЧИ

```
# результат расшифровки речи диктора

speech_recognition = [
    'кучменко она',
    'кущенко оксана',
    'груздь алина',
    'рычихина наталья',
    'шиховцева на',
    'чистова анастасия'
]
```

МЕТРИКИ БЛИЗОСТИ ОБЪЕКТОВ

РАССТОЯНИЕ ХЭММИНГА

В телекоме - для отслеживания ошибок

В биоинформатике - для оценки стабильности цепи

https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.spatial.distance.hamming.html

IT-Academy

РАССТОЯНИЕ ЛЕВЕНШТЕЙНА

Минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую.

M	**M**	**M**	**R**	**I**	**M**	**R**	**R**
C	**O**	**N**	**N**		**E**	**C**	**T**
C	O	N	**E**	**H**	E	**A**	**D**

D — удалить,

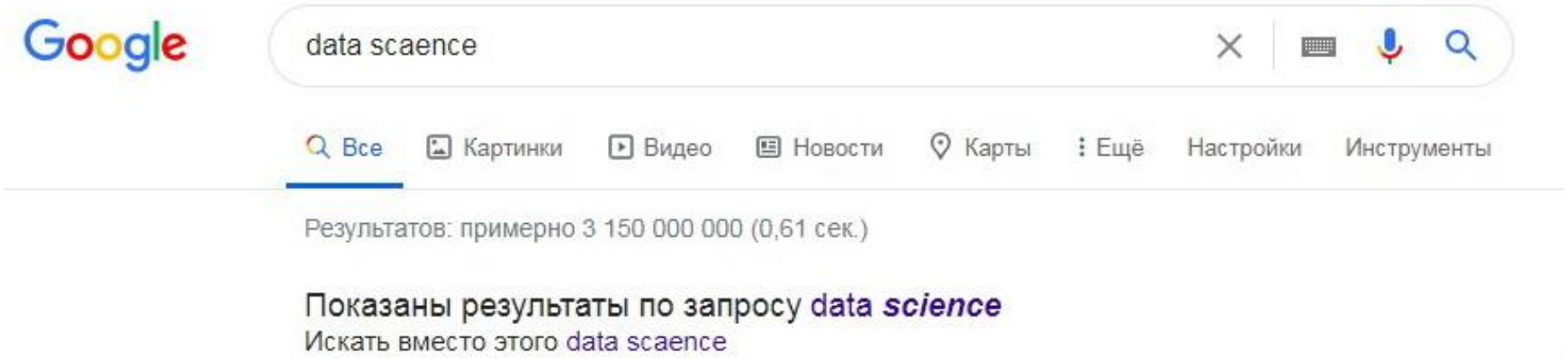
I — вставить,

R — заменить,

M — совпадение

МЕТРИКИ БЛИЗОСТИ ОБЪЕКТОВ

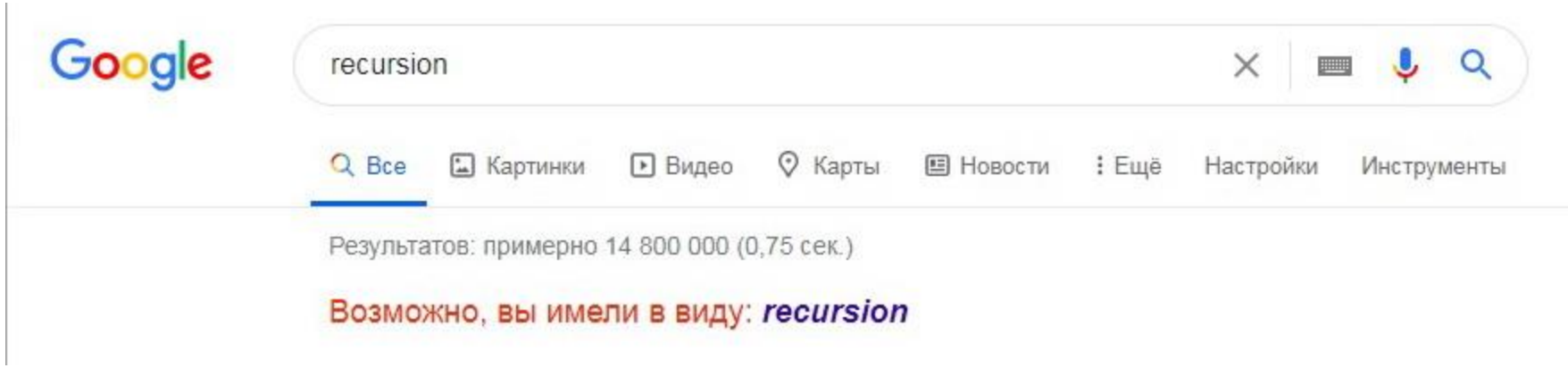
РАССТОЯНИЕ ЛЕВЕНШТЕЙНА



МЕТРИКИ БЛИЗОСТИ ОБЪЕКТОВ

РАССТОЯНИЕ ДАМЕРАУ-ЛЕВЕНШТЕЙНА

То же самое, но с добавлением операции транспозиции (перестановки символов)



Юмор Google

IT-Academy

ПРАКТИКА

Levenshtein distance.ipynb

СХОЖЕСТЬ ПОЛЬЗОВАТЕЛЕЙ

IT-Academy

КОЭФФИЦИЕНТ ЖАККАРА

$$K = \frac{n(A \cap B)}{n(A \cup B)}$$

Отношение количества элементов, общих для множеств A и B, к общему количеству элементов в этих множествах

СХОЖЕСТЬ ПОЛЬЗОВАТЕЛЕЙ

IT-Academy

КОЭФФИЦИЕНТ ЖАККАРА

Удобно использовать в рекомендательных системах

Товары

Признак	Телефон 1 vs 2
Память	совпадает
Экран	разный
Процессор	совпадает

Предпочтения

Фильм	Пользователь 1	Пользователь 2
Гадкий Я	✶ ✶ ✶ ✶	✶
Мумия	✶ ✶	✶ ✶ ✶
Пираты	✶ ✶ ✶ ✶ ✶	✶ ✶ ✶ ✶ ✶

IT-Academy

ПРАКТИКА

Jaccard.ipynb

IT-Academy

ПРАКТИКА

ПРОСМОТР КОДА

KNN REGRESSION.IPYNB

IT-Academy

ПРАКТИКА

KNN NBA.IPYNB

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

IT-Academy

1. Метрики расстояний и близости объектов в применении к различным задачам.
2. Рассмотрели идею алгоритма KNN.
3. Реализовали на практике алгоритм KNN в задачах классификации и регрессии.

IT-Academy

ДОМАШНЕЕ ЗАДАНИЕ

дописать

KNN DIGITS.IPYNB

IT-Academy

Метрики расстояний и алгоритм KNN

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ