Метод k ближайших соседей

Дмитрий Корнев
tswr@yandex-team.ru

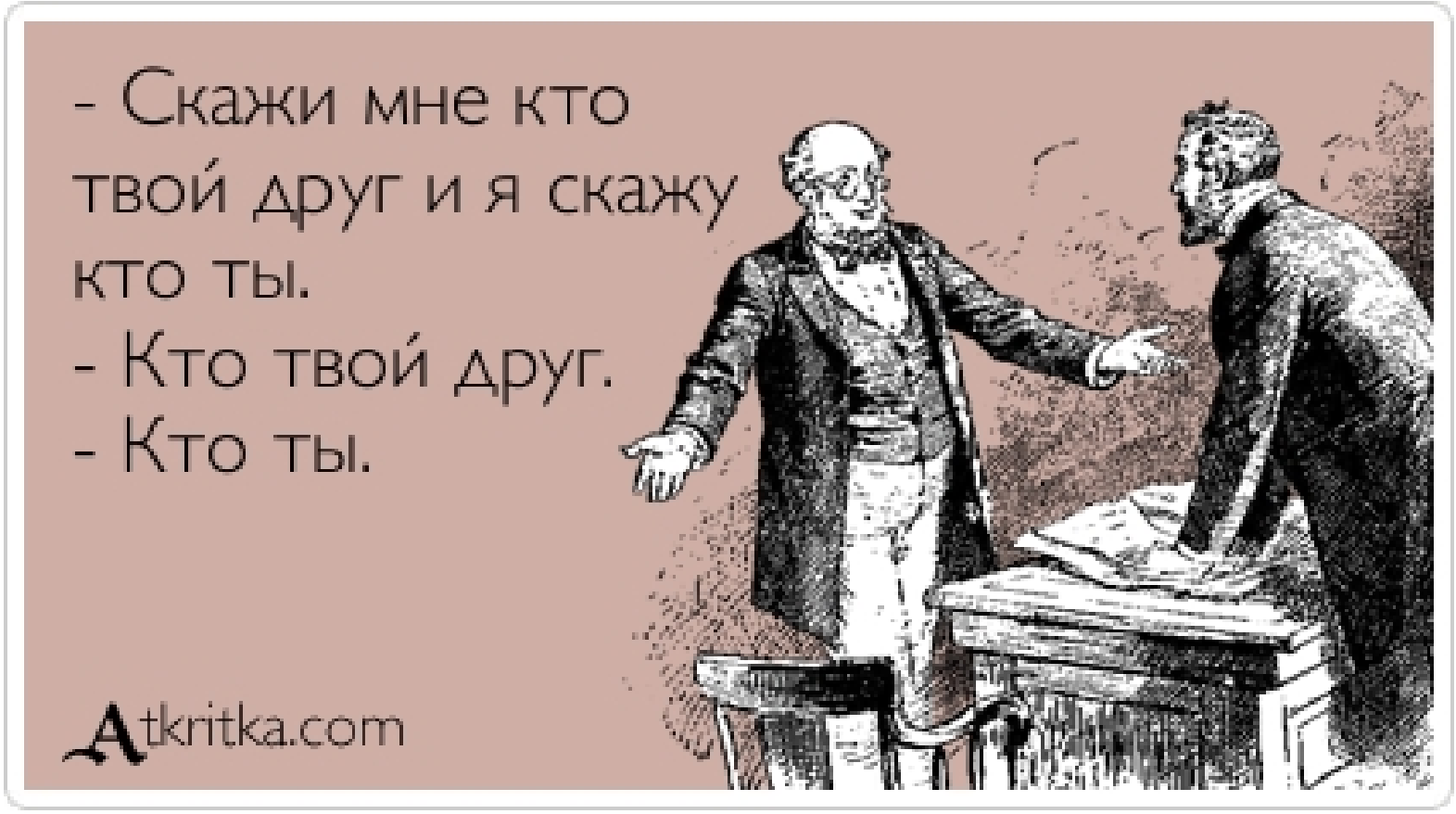
Школа анализа данных

Весна 2020

Метод k ближайших соседей

- Теория
- Практика
  - Методы поиска ближайших соседей

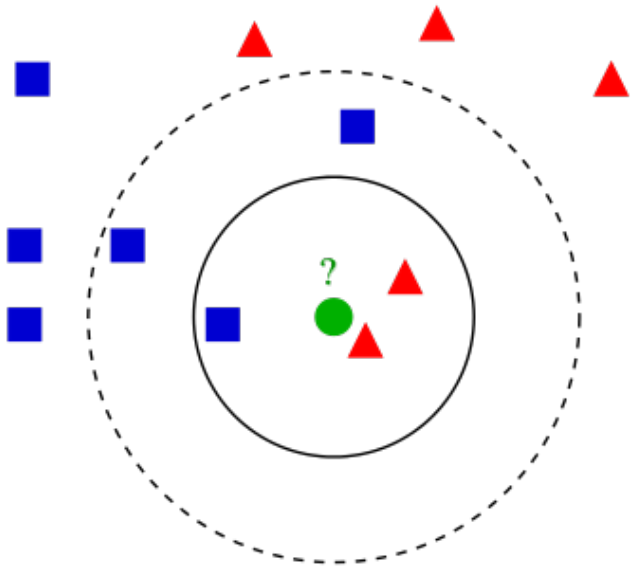
«Скажи мне, кто твой друг, и я скажу тебе, кто ты»
Еврипид



Метод k ближайших соседей — классификация

- Голосование среди k соседей

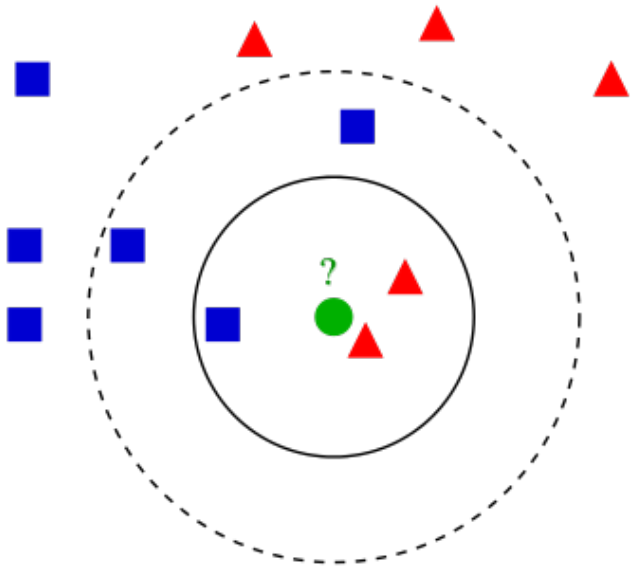
$$a(x) = \arg\max_{y \in Y} \sum_{i \in \text{nearest}_k(x)} [y_i = y]$$



Метод k ближайших соседей — классификация

- Голосование среди k соседей

$$a(x) = \arg\max_{y \in Y} \sum_{i \in \text{nearest}_k(x)} [y_i = y]$$



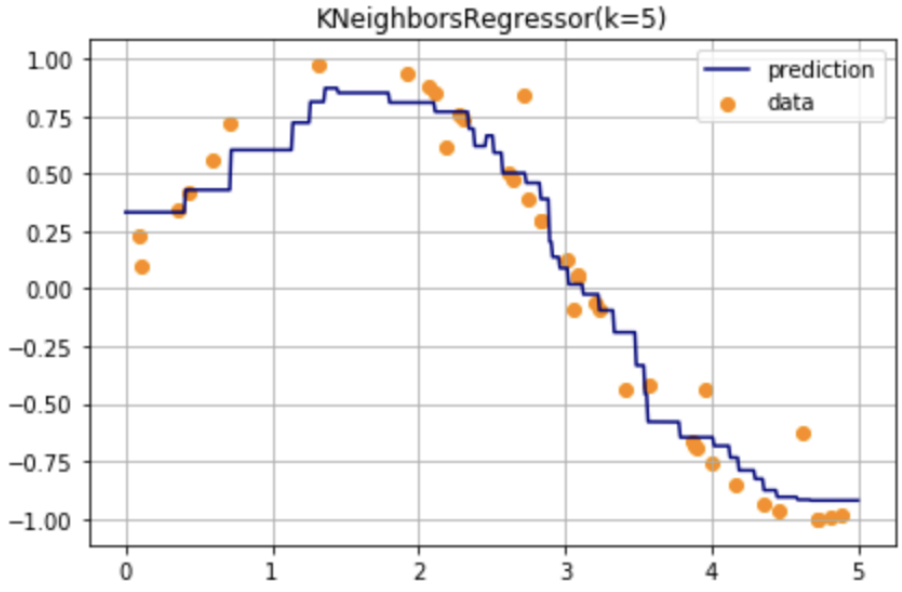
- Метрический алгоритм классификации

$$a(x) = \arg\max_{y \in Y} \sum_{i=1}^{l} [y_i = y]\, w_i(x)$$

Метод k ближайших соседей — регрессия

- Голосование среди k соседей

$$a(x) = \frac{1}{k} \sum_{i \in \text{nearest}_k(x)} y_i$$



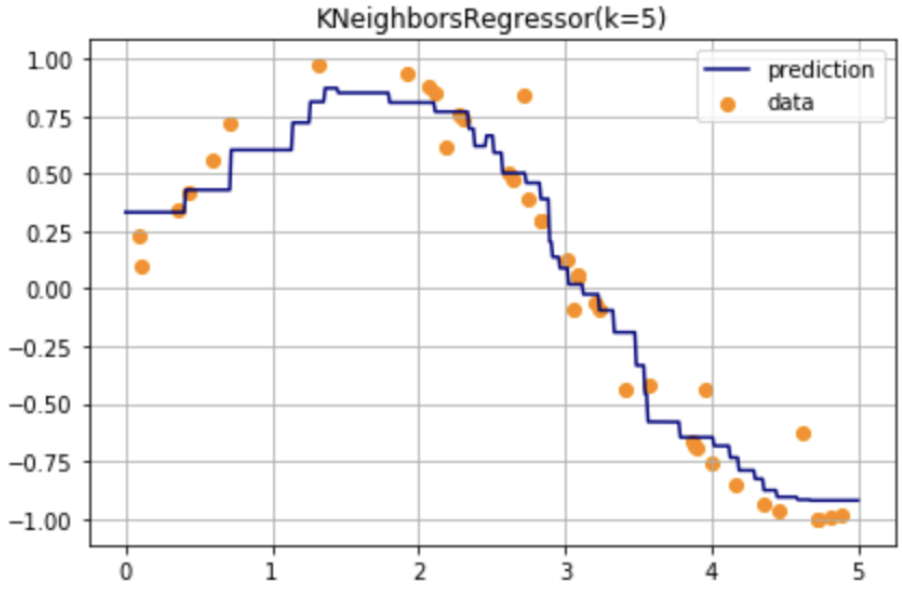
Метод k ближайших соседей — регрессия

- Голосование среди k соседей

$$a(x) = \frac{1}{k} \sum_{i \in \text{nearest}_k(x)} y_i$$

- Формула Надарая-Ватсона

$$a(x) = \frac{\sum_{i=1}^{l} w_i(x)\, y_i}{\sum_{i=1}^{l} w_i(x)}$$



Метрические алгоритмы

- Метод k ближних соседей

$$w_i(x) = [i \leqslant k]$$

Метрические алгоритмы

- Метод k ближних соседей

$$w_i(x) = [i \leqslant k]$$

- Метрические алгоритмы

$$w_i(x) = f(\rho(x_i, x))$$

Метрические алгоритмы

- Метод k ближних соседей

$$w_i(x) = [i \leqslant k]$$

- Метрические алгоритмы

$$w_i(x) = f(\rho(x_i, x))$$

$$w_i(x) = K\left(\frac{\rho(x_i, x)}{h}\right)$$

Метрические алгоритмы

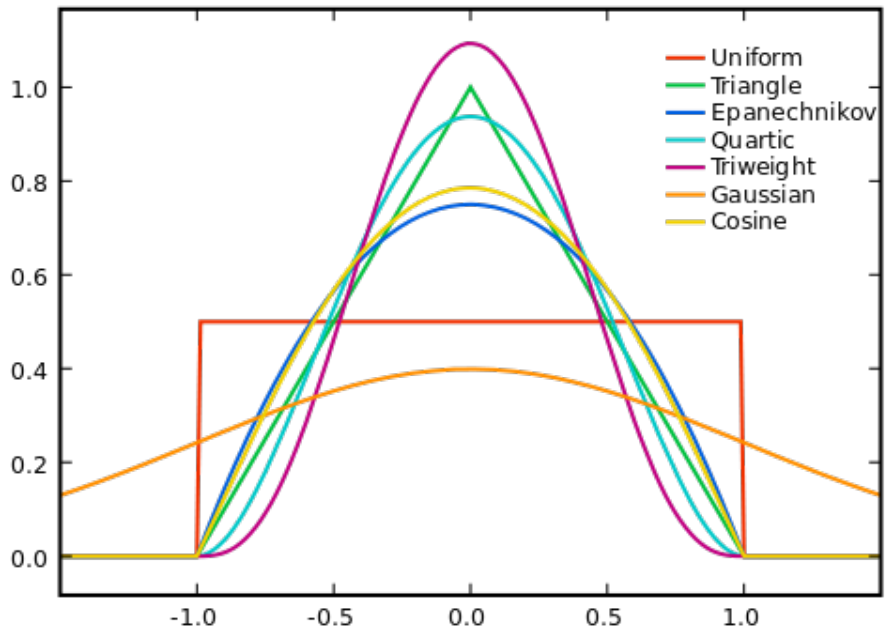
- Метод k ближних соседей

$$w_i(x) = [i \leqslant k]$$

- Метрические алгоритмы

$$w_i(x) = f(\rho(x_i, x))$$

$$w_i(x) = K\left(\frac{\rho(x_i, x)}{h}\right)$$



Особенности использования k-NN на практике

- Определить функцию близости — сложная задача, важны преобразования признаков (масштабирование)

Особенности использования k-NN на практике

- Определить функцию близости — сложная задача, важны преобразования признаков (масштабирование)
- Алгоритм работает лучше на признаках одной природы

Особенности использования k-NN на практике

- Определить функцию близости — сложная задача, важны преобразования признаков (масштабирование)
- Алгоритм работает лучше на признаках одной природы
- Часто k-NN используют как предварительный фильтр для поиска подходящих кандидатов

Особенности использования k-NN на практике

- Определить функцию близости — сложная задача, важны преобразования признаков (масштабирование)
- Алгоритм работает лучше на признаках одной природы
- Часто k-NN используют как предварительный фильтр для поиска подходящих кандидатов
- Примеры

Метод k ближайших соседей

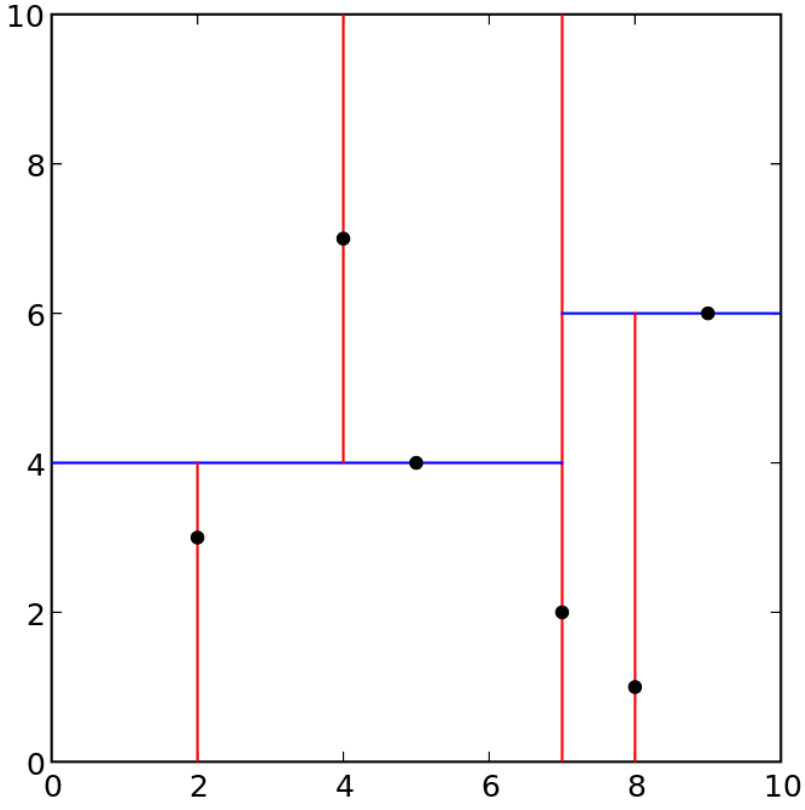
Дана обучающая выборка

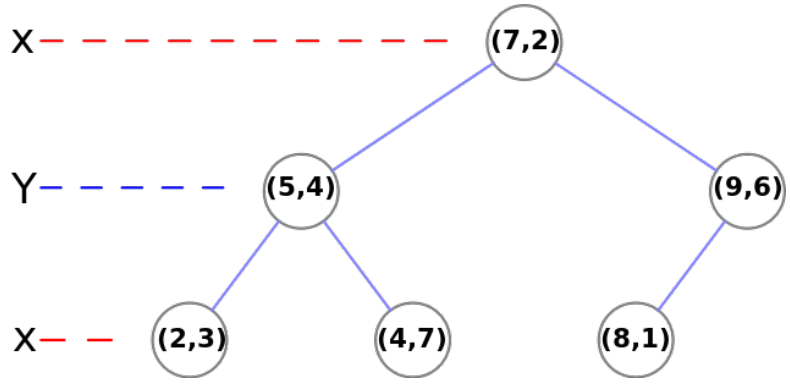
$$X^l = (x_i, y_i)_{i=1}^l, \qquad x_i \in \mathrm{R}^n.$$

Какова алгоритмическая сложность fit/predict?

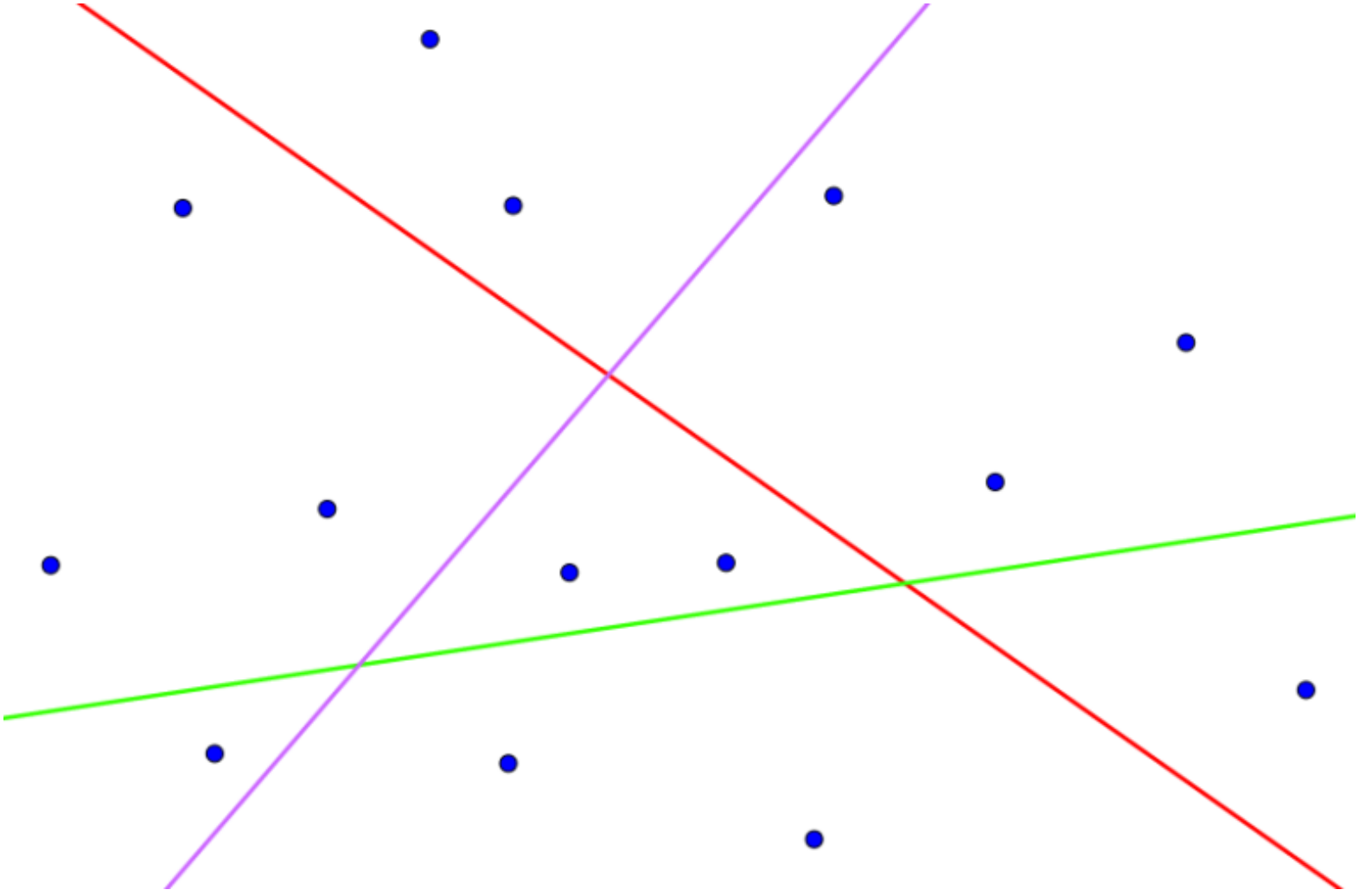
K-d деревья

$$X = ((2,3),(5,4),(9,6),(4,7),(8,1),(7,2))$$

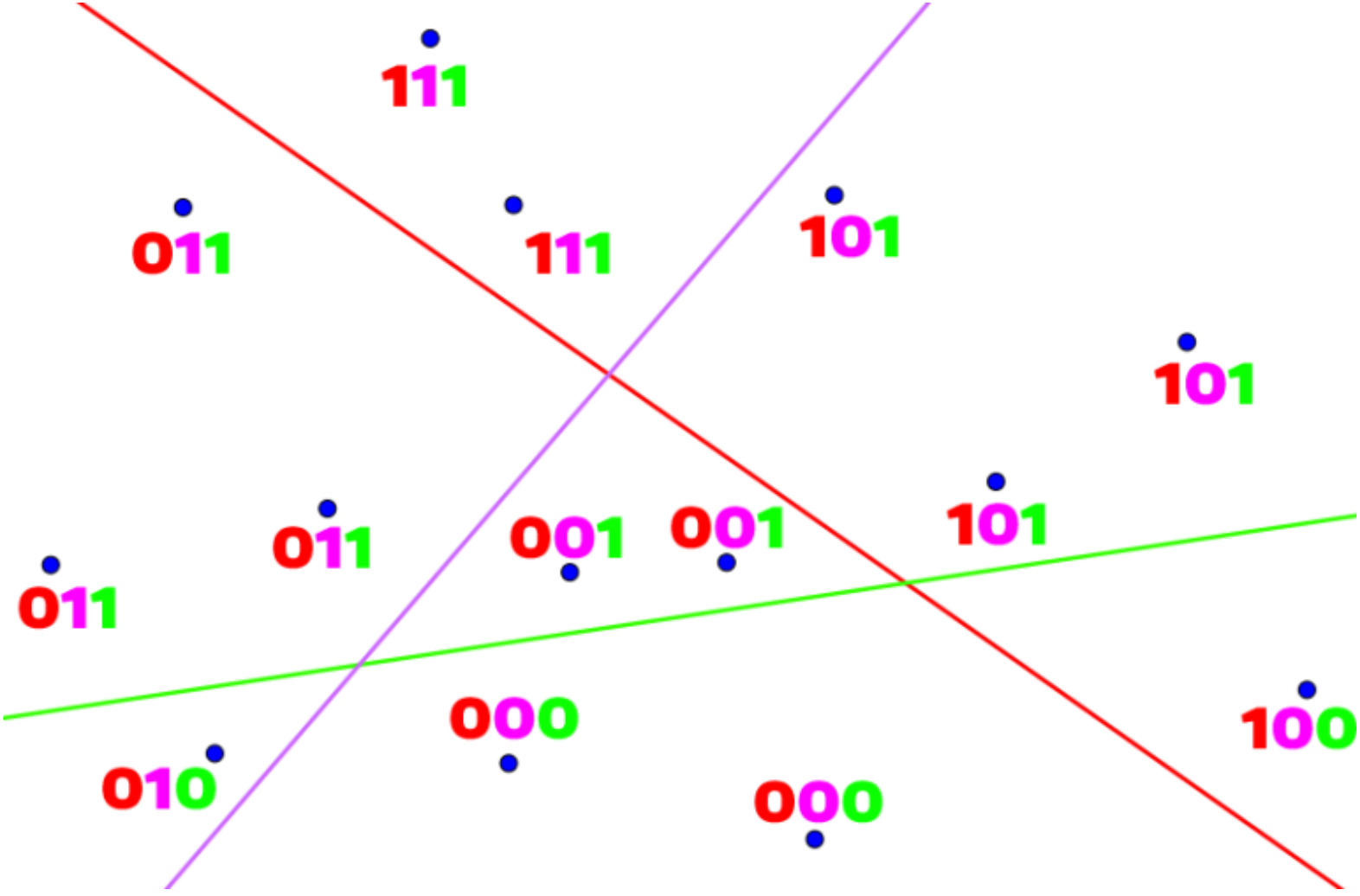




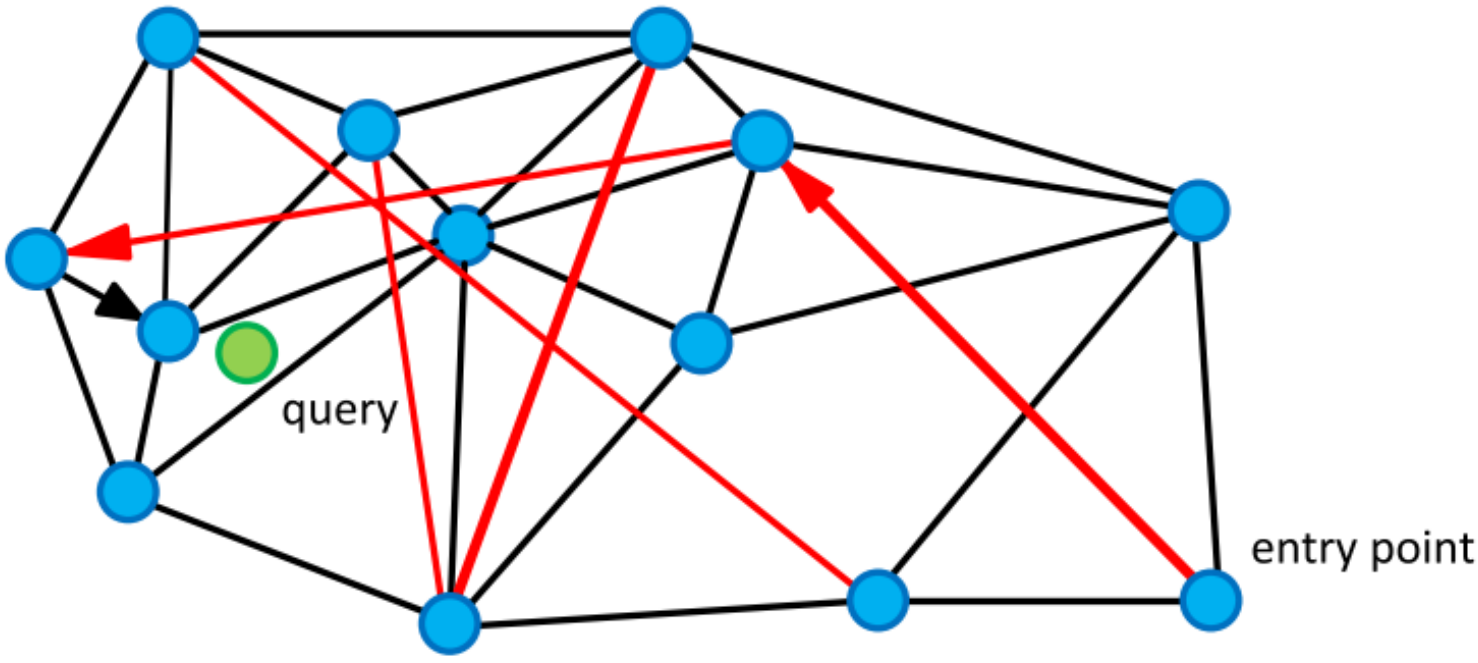
Locality-Sensitive Hashing



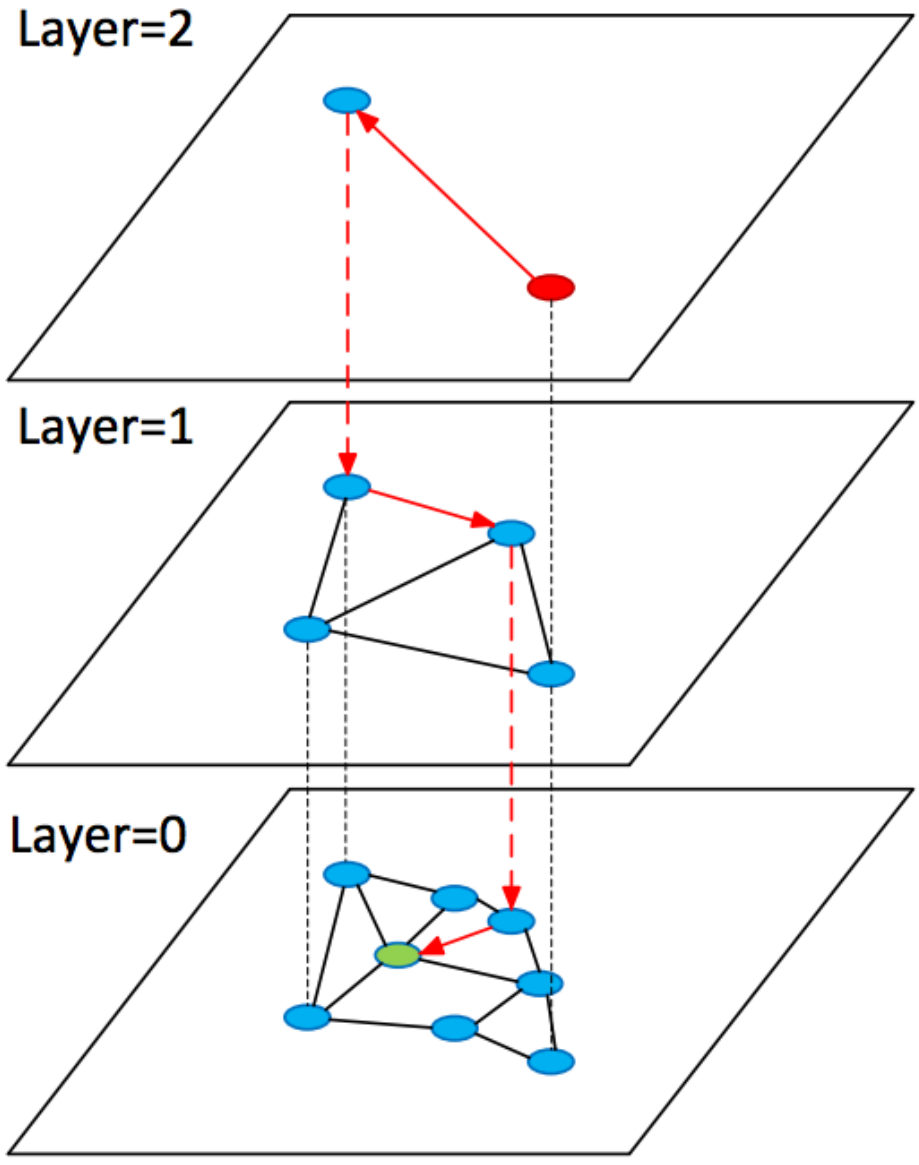
Locality-Sensitive Hashing



Navigable Small World (NSW)



Hierarchical Navigable Small World (HNSW)



Faiss

- Кластеризация
- Инвертированный индекс
- Product quantizers

Реализации

- K-d tree, ball tree — sklearn
- LSH — annoy, datasketch, NearPy
- HSNW — nmslib
- Clustering based — faiss