kNN и приближённая задача поиска ближайшего соседа

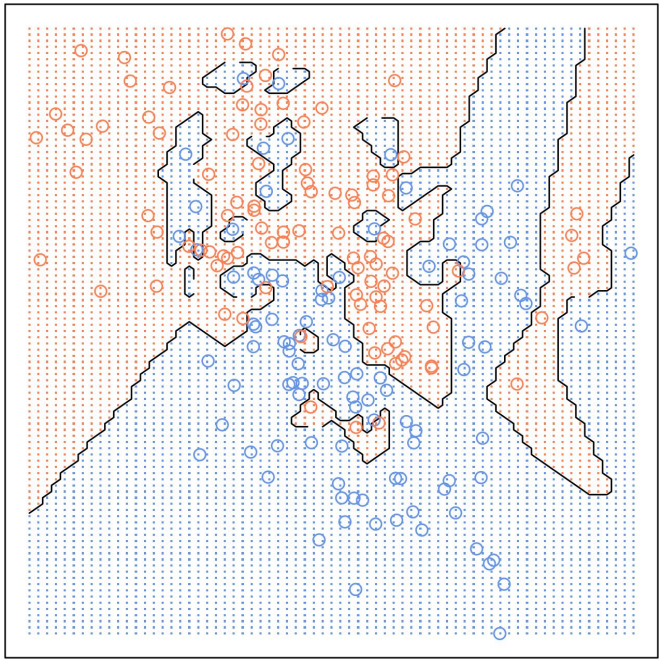
Гусева Полина, Булатова Екатерина, БПМИ181

Метод k ближайших соседей

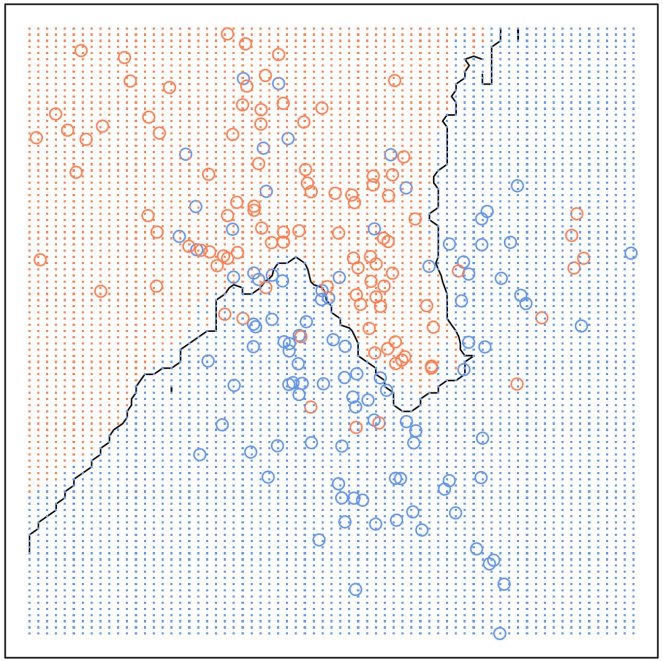
- Объекты представляются в виде многомерных векторов
- Опирается на “гипотезу компактности”
- Нормализация значений
- Метод используется для решения задач классификации и регрессии:
  - Классификация - наиболее распространённый среди соседей
  - Регрессия - среднее из соседей
- Взвешенный метод:
  - Простейшее решение: $$Q_j = \sum_{i=1}^{n} \frac{1}{d(x, a_i)^2}$$

Метод k ближайших соседей

1-nearest neighbours



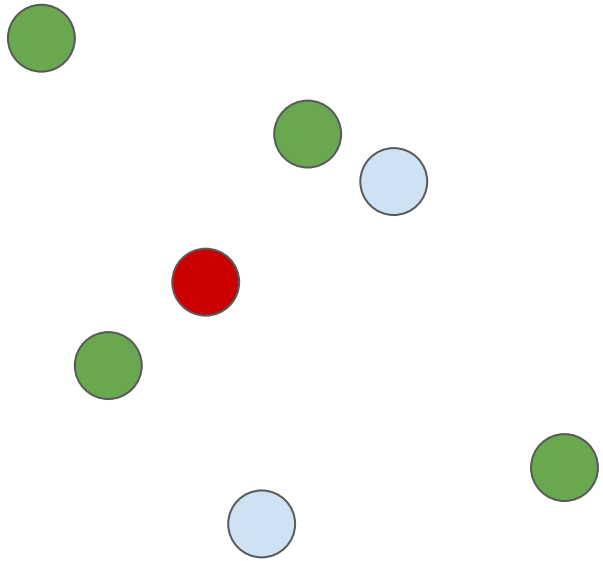
20-nearest neighbours



$$k=[\sqrt{N}]$$

Приближенный метод поиска k ближайших соседей

- “Проклятие размерности”
- k-Approximate NN
- Его точность



k = 4
из найденных точек
действительно
ближайшие соседи r = 2
точность = r / k = 0.5

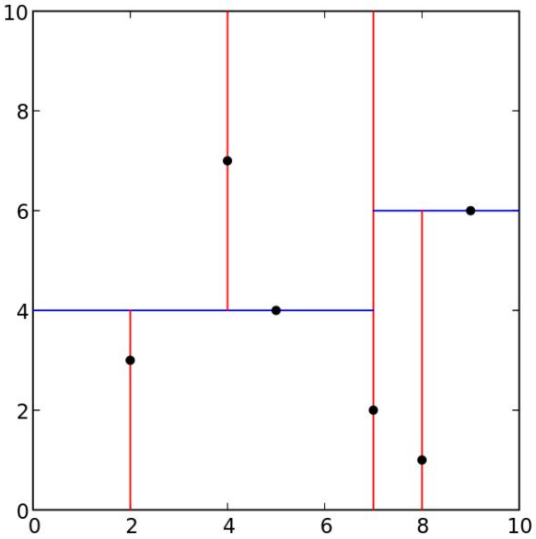
Методы решения задачи k-NNS

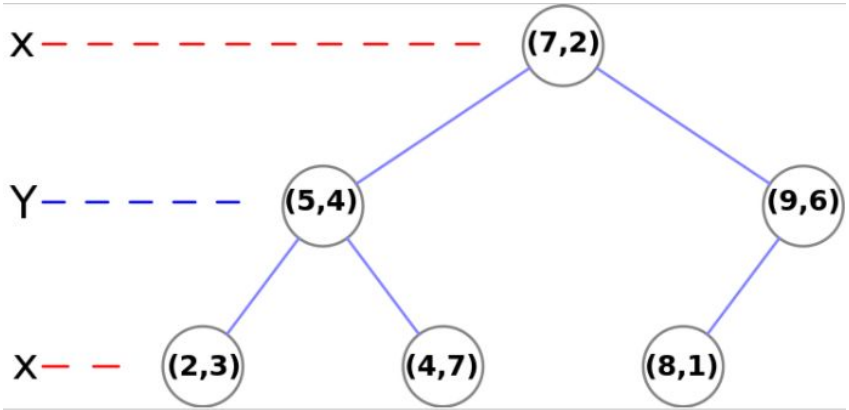
Линейный поиск.

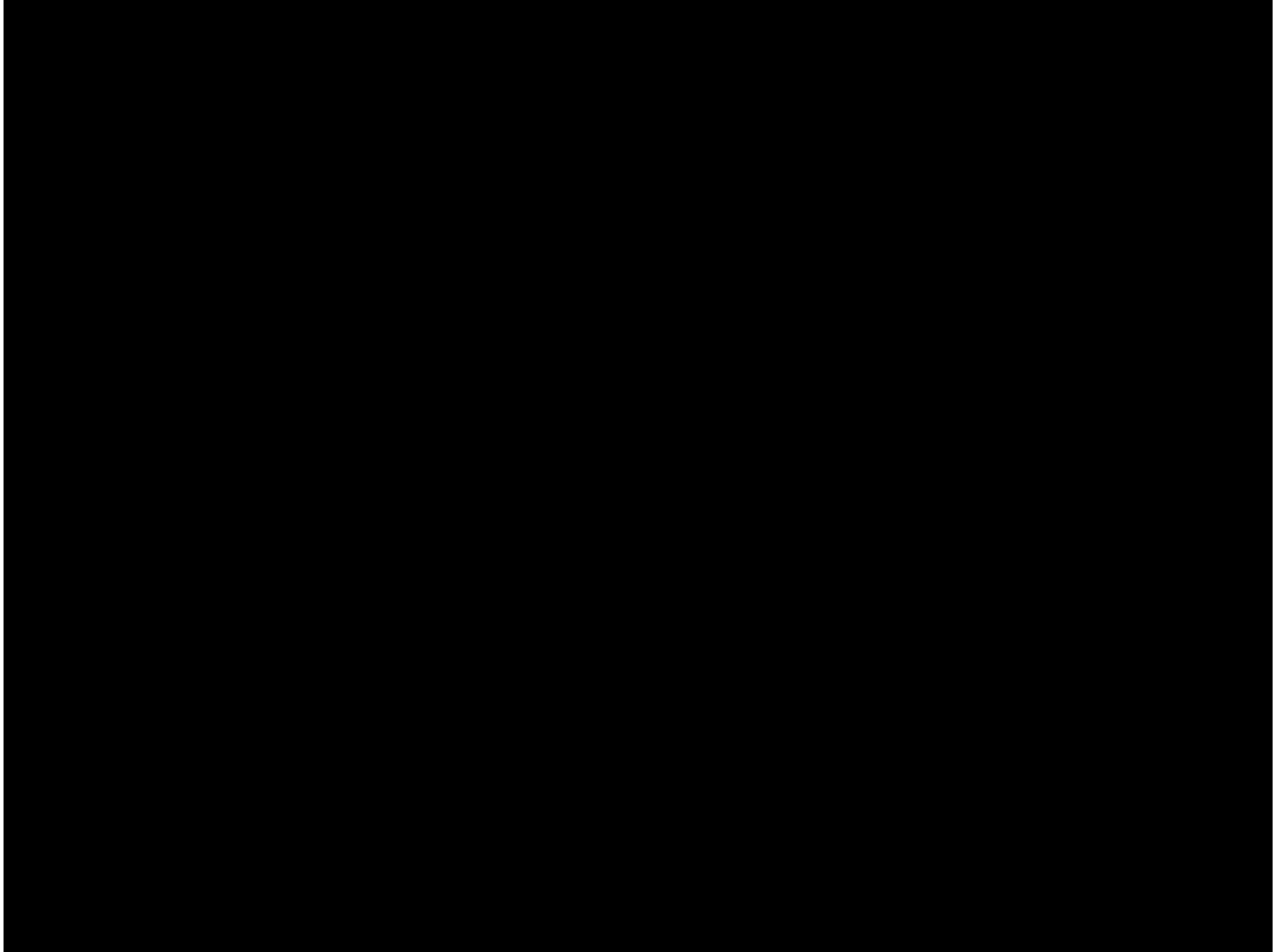
Дерево поиска k-d.

- k-d деревья $O(n \log n)$
- Применение к k-NNs
- Асимптотика

$$O(h) + O(h \cdot log(h))$$



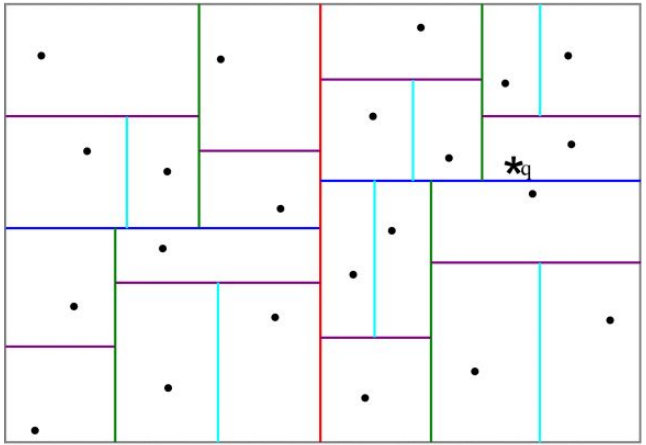




Методы решения задачи k-NNS

k-d дерево поиска. Алгоритм:

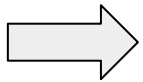
- Начинаем в корне;
- Доходим до нужного листа, его добавляем в приоритетную очередь;
- Двигаемся вверх, каждую вершину:
  - кладем в нужное место в очереди;
  - если нужно, идем в другого ребенка.



Методы решения задачи k-NNS

Инвертированные индексы.

1. Это - кот
2. Это - то, что это есть
3. Что есть кот?



это : {1, 2}

кот : {1, 3}

есть, что : {2, 3}

то : {2}

обрабатываем “есть кот”:

{1, 3} & {2, 3} = {3}

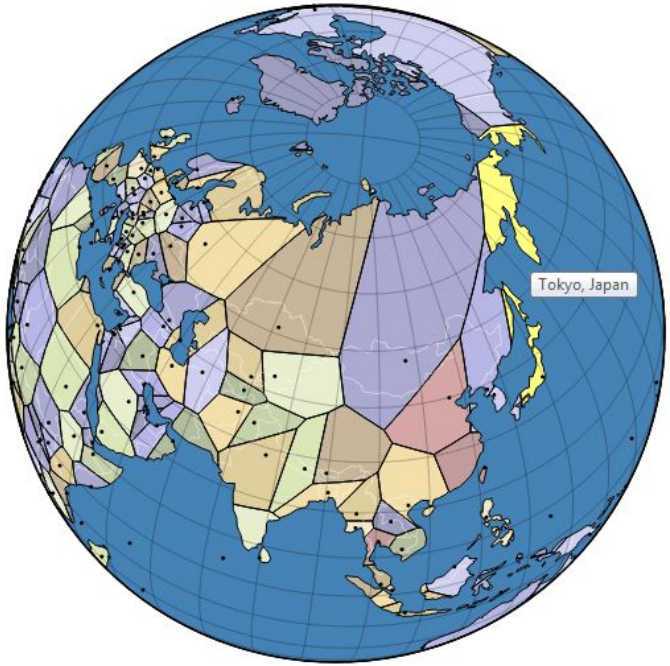
{1, 3} | {2, 3} = {1, 2, 3}

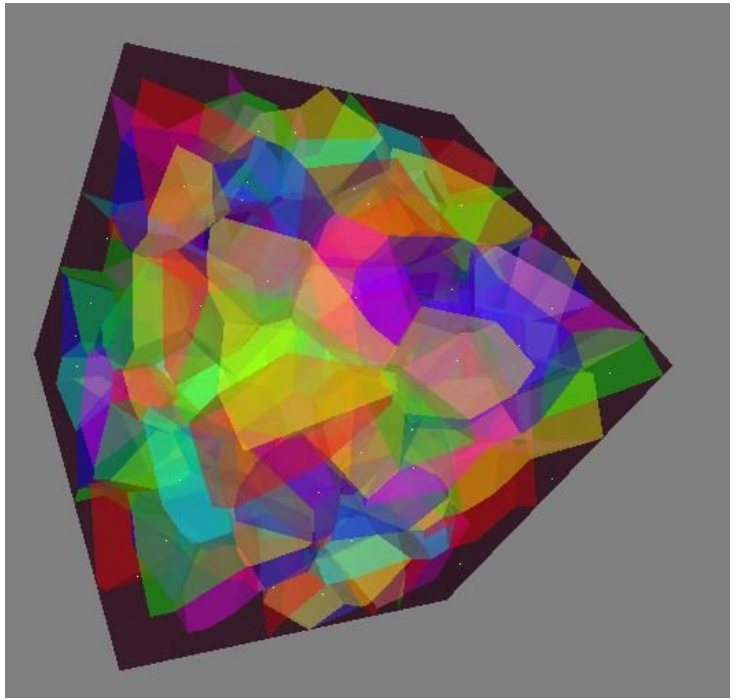
Сложность: O($\sqrt{n}$)
(средняя длина списка)

Методы решения задачи k-NNS

Диаграмма Вороного.

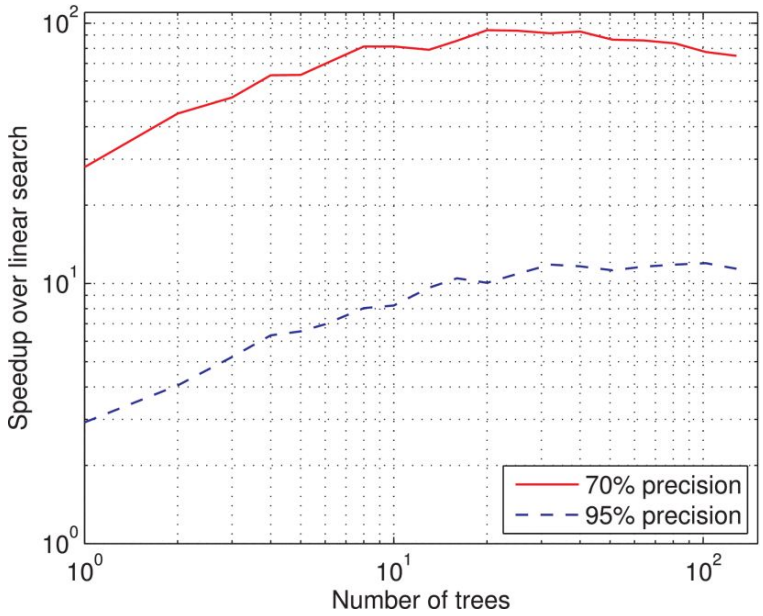
Разбиение плоскости. $O(n \log n)$

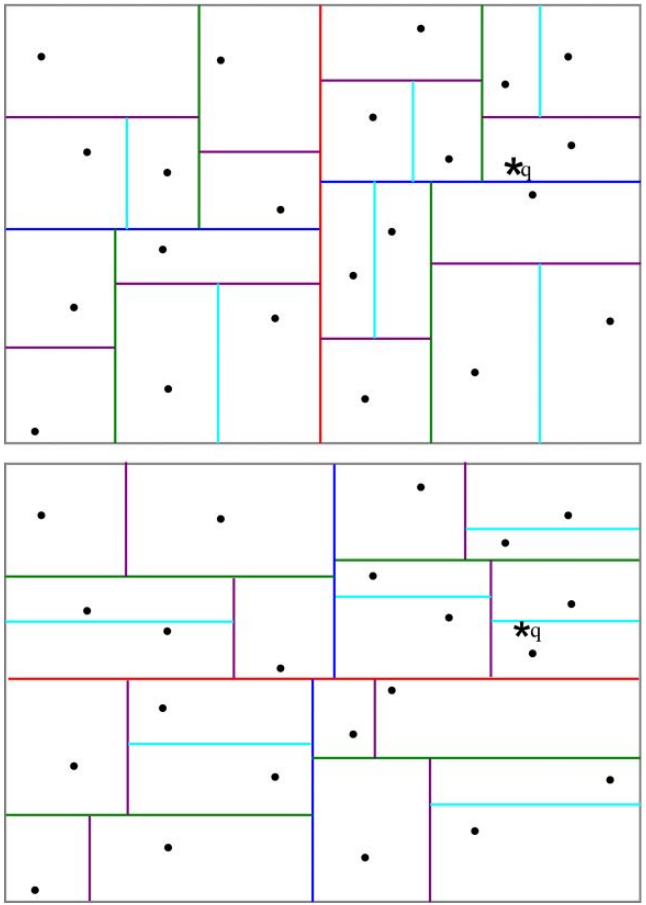




Методы решения задачи k-ANNS

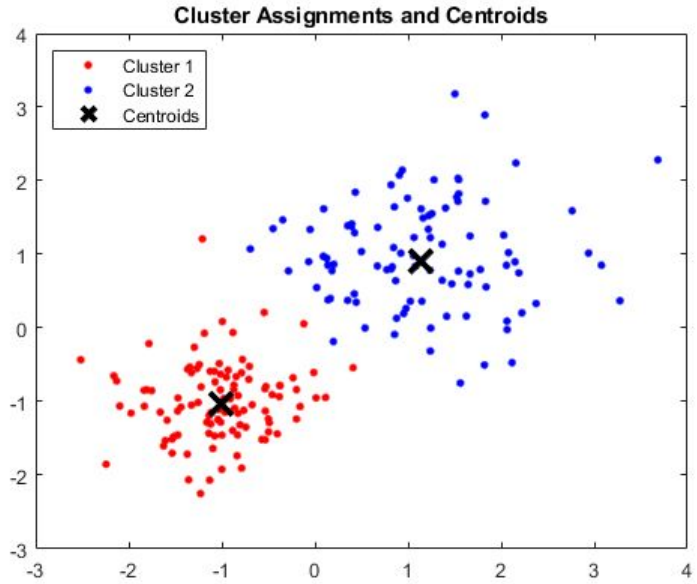
Случайные k-d деревья поиска.



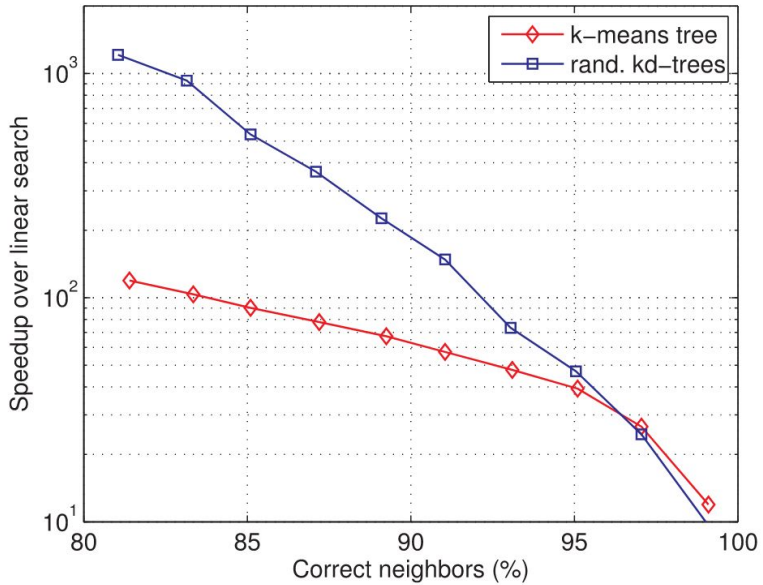


Методы решения задачи k-ANNS

Дерево k-средних с приоритетами.



Разбиение на кластеры методом k-средних



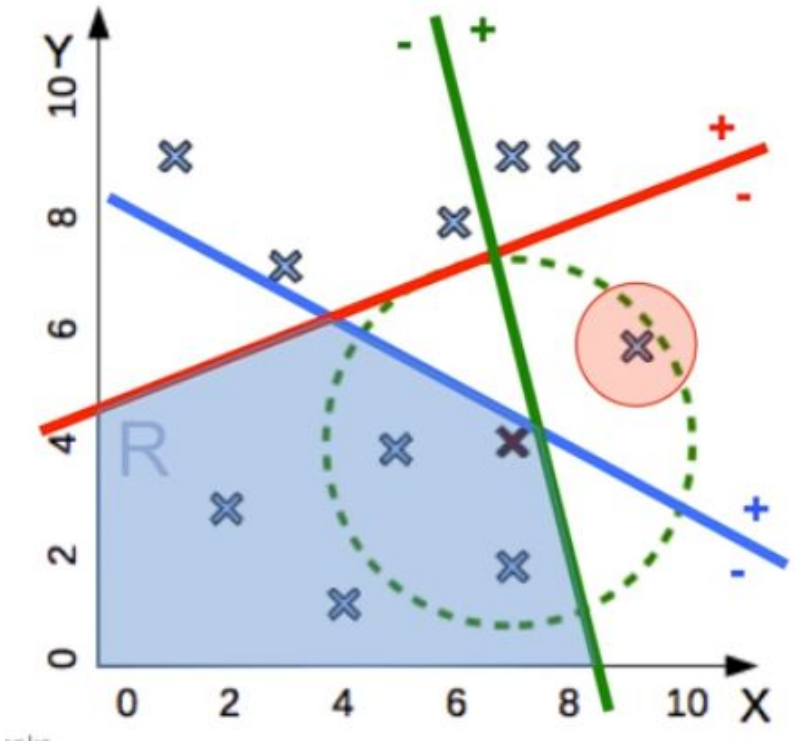
Методы решения задачи k-ANNS

Локально-чувствительное хеширование.

Случайное разбиение пространства в 2^H областей.

Довольно низкая точность.

Сложность поиска: $O\left(Hd + \frac{dn}{2^H}\right)$



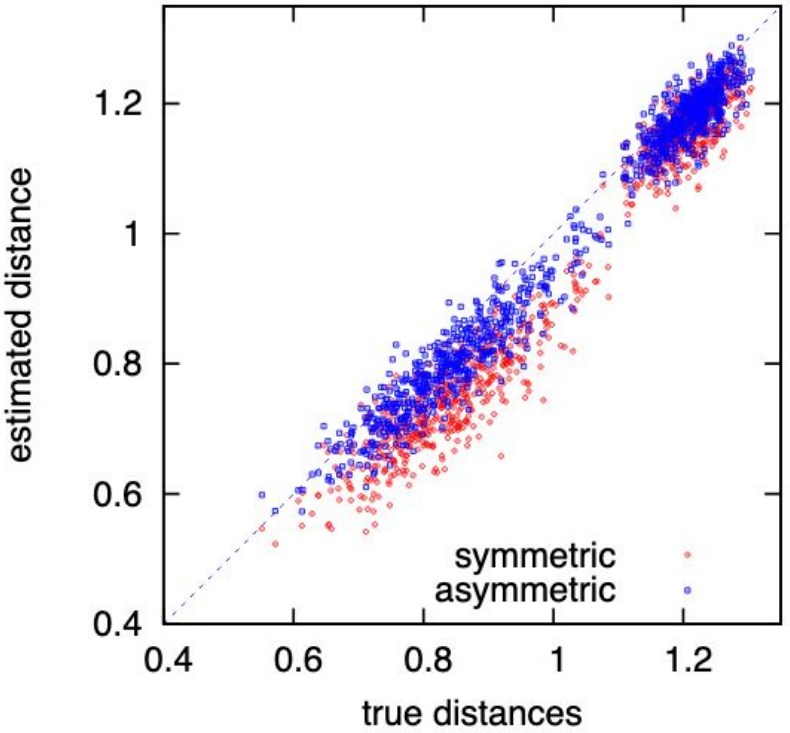
Методы решения задачи k-ANNS

Разложение пространства в прямую сумму.

Упрощающая функция Q отображает точки в конечное множество точек S.

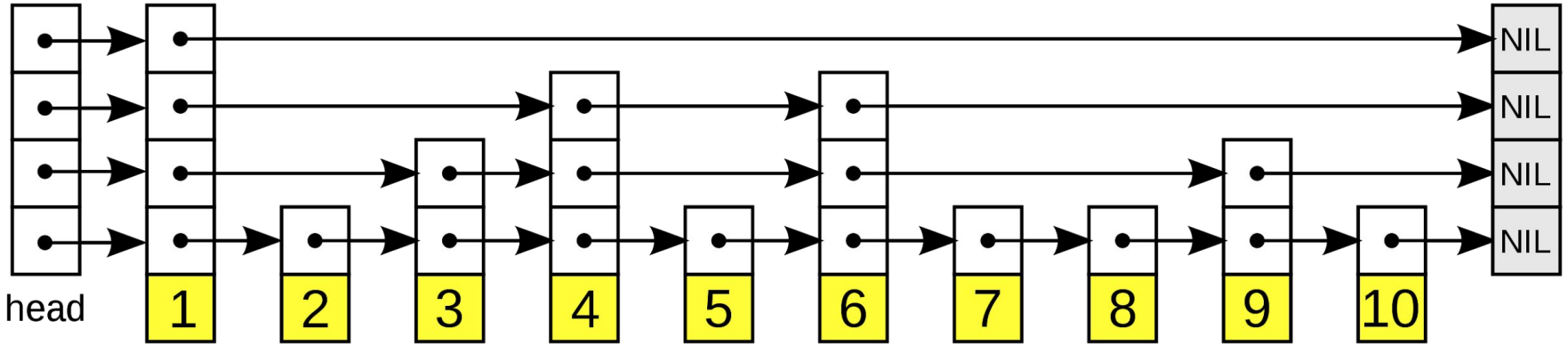
Условия оптимальности для Q:

1) каждая точка переводится в ближайшую к ней точку из S
2) каждая точка из S равна матожиданию ее прообразов



Методы решения задачи k-ANNS

Список с пропусками.



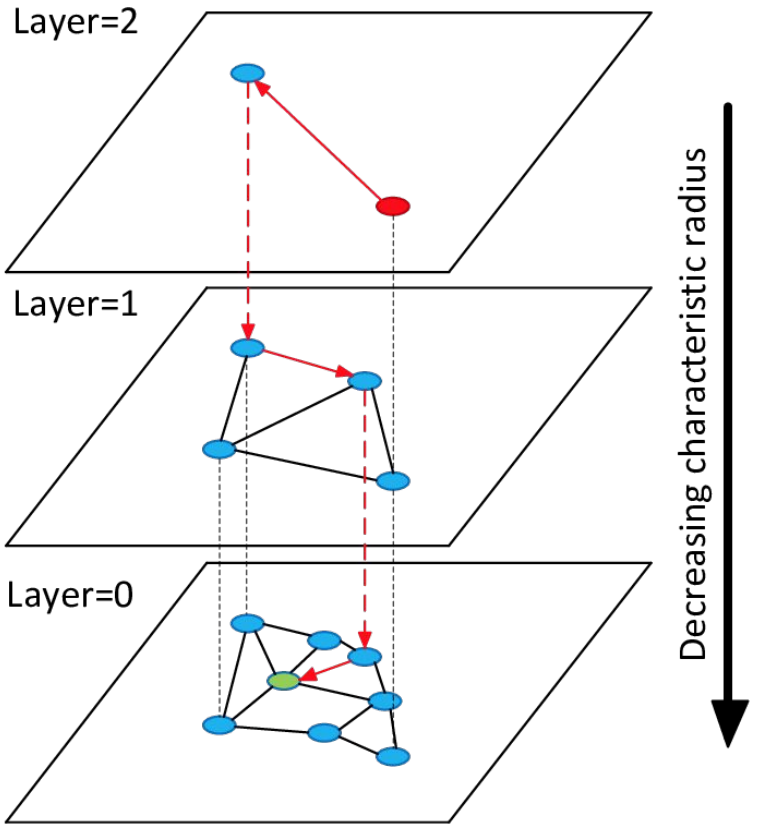
Построение за $O(n \log n)$

Поиск за $O(\log n)$

Методы решения задачи k-ANNS

Иерархический граф типа “мир тесен”.

Обобщение списка с пропусками.



Итог

- K-NNS
  - Линейный поиск
  - Инвертированные индексы
  - Диаграмма Вороного
- K-ANNS
  - Лес случайных k-d деревьев
  - Деревья поиска k-средних с приоритетами
  - Локально-чувствительное хеширование
  - Разложение в прямое произведение
  - Иерархические графы типа “мир тесен”

Ссылки

- Marius Muja. Scalable Nearest Neighbor Algorithmsfor High Dimensional Data
- Kevin Zakka. A Complete Guide to K-Nearest-Neighbors with Applications in Python and R
- University of Colorado. K-D Trees and KNN Searches
- Диаграммы Вороного для аэропортов и столиц
- Exploring data visualisation. Voronoi Diagrams
- Victor Lavrenko. Locality sensitive hashing
- Jia Pan, Dinesh Manocha. Fast GPU-based Locality Sensitive Hashing for K-NearestNeighbor Computation
- Hervé Jégou, Matthijs Douze, Cordelia Schmid. Product Quantization for Nearest Neighbor Search
- Yu. A. Malkov, D. A. Yashunin. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs