Метрические методы классификации (k Nearest Neighbors)

Автор: Альмухаметова Гузель

Группа:152

Аксиомы тождества, симметрии и неравенства треугольника

$d(x, y) = 0$ *тогда и только тогда, когда* $x = y$;

$d(x, y) = d(y, x)$ *для всех* $x, y \in Y$;

$d(x, y) \leqslant d(x, z) + d(z, y)$ *для всех* $x, y, z \in Y$.

Примеры функций расстояния

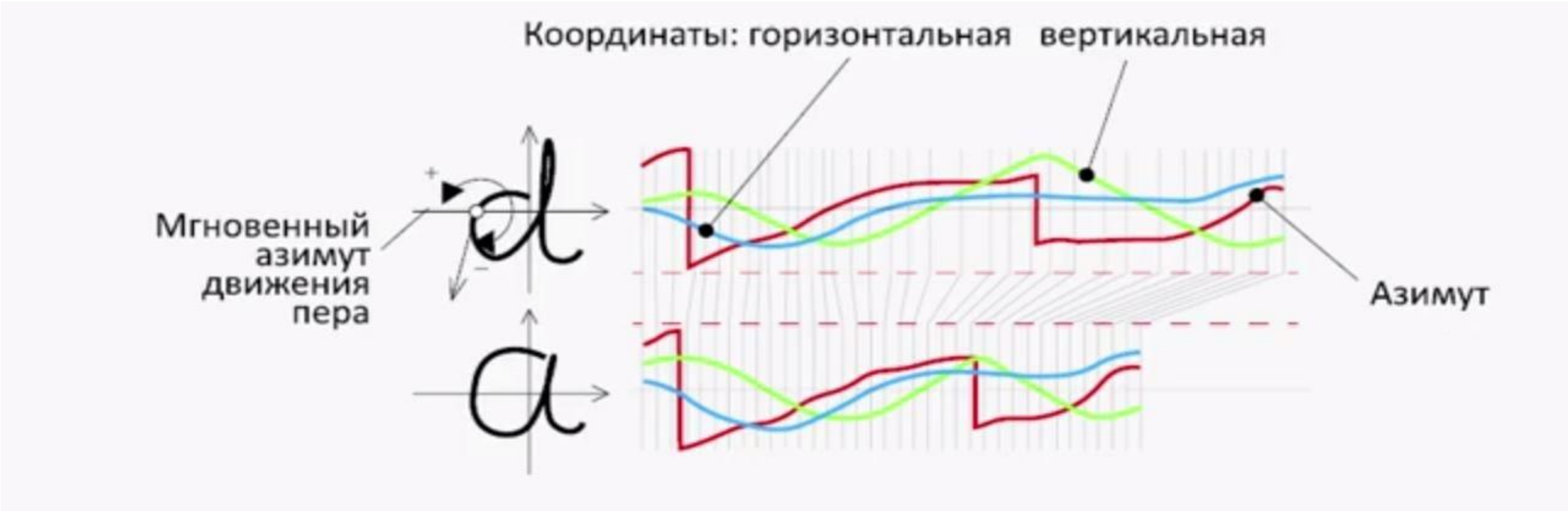
- Метрика Минковского $\rho_p(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{1/p}$ для $p \geqslant 1$
- Евклидова метрика (p = 2)

Distance Metric Learning

- Манхэттенское расстояние (p = 1) $\rho_0(x,y) = \sum_{i=1}^{d} [x_i \neq y_i].$
- «Считающее» расстояние (p = 0)

- Редакторское расстояние Левенштейна

```
CTGGGCTAAAAGGTCCCTTAGCC..TTTAGAAAAA.GGGCCATTAGGAAATTGC
CTGGGACTAAA....CCTTAGCCTATTTACAAAAATGGGCCATTAGG...TTGC
```



Временной ряд — собранный в разные моменты времени статистический материал о значении каких-либо параметров (в простейшем случае одного) исследуемого процесса.

- Косинусная мера

Пусть заданы векторы x и y. Известно, что их скалярное произведение и косинус угла θ между ними связаны следующим соотношением:

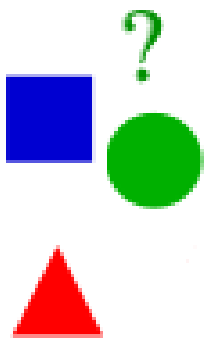
$$\langle x, y \rangle = \|x\| \|y\| \cos(\theta).$$

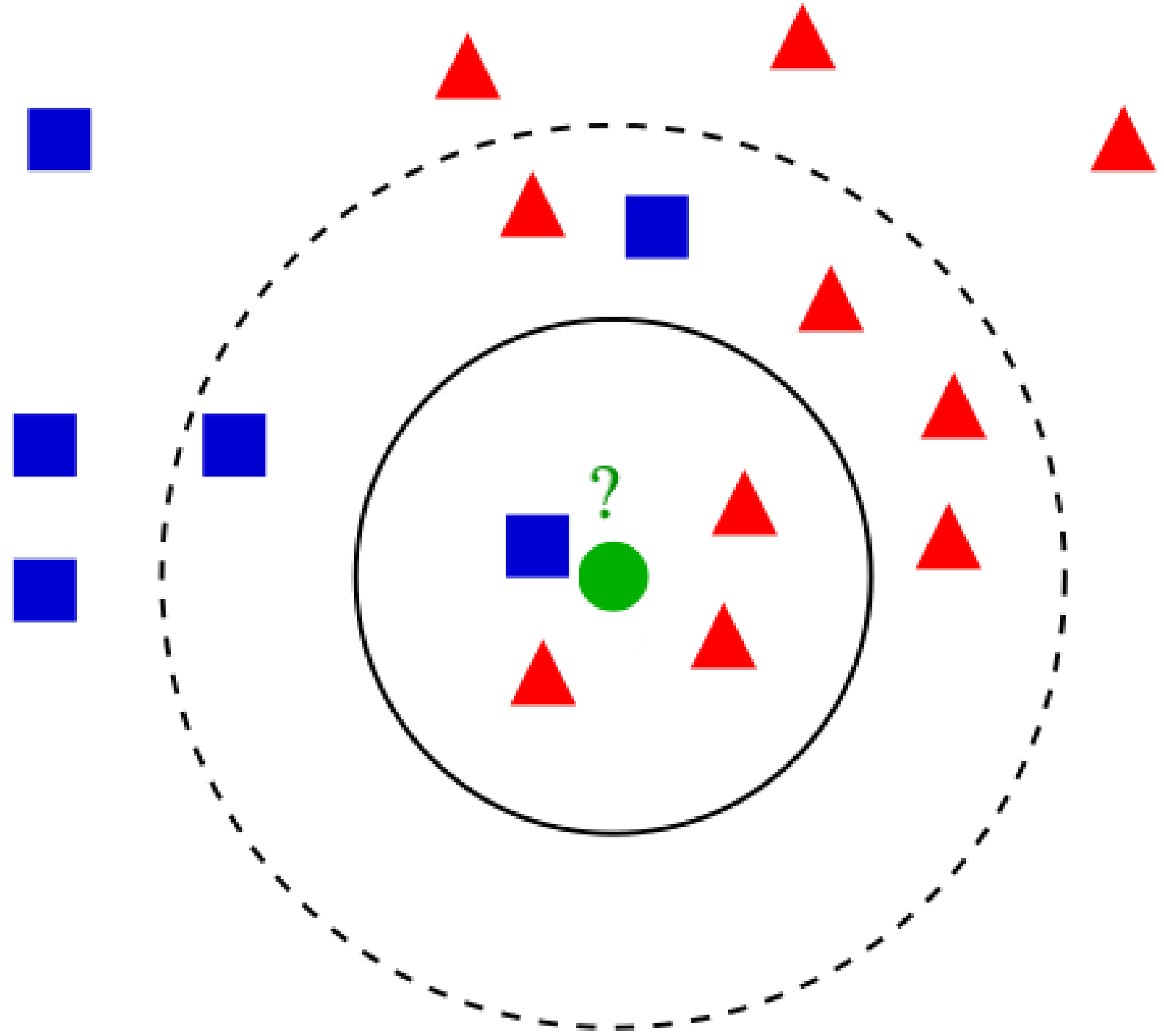
Соответственно, косинусное расстояние определяется как

$$\rho_{\cos}(x, y) = \arccos\left(\frac{\langle x, y \rangle}{\|x\| \|y\|}\right) = \arccos\left(\frac{\sum_{i=1}^{d} x_i y_i}{\left(\sum_{i=1}^{d} x_i^2\right)^{1/2} \left(\sum_{i=1}^{d} y_i^2\right)^{1/2}}\right).$$

Путин провел встречу в Башкирии

Президент созвал конференцию в Уфе



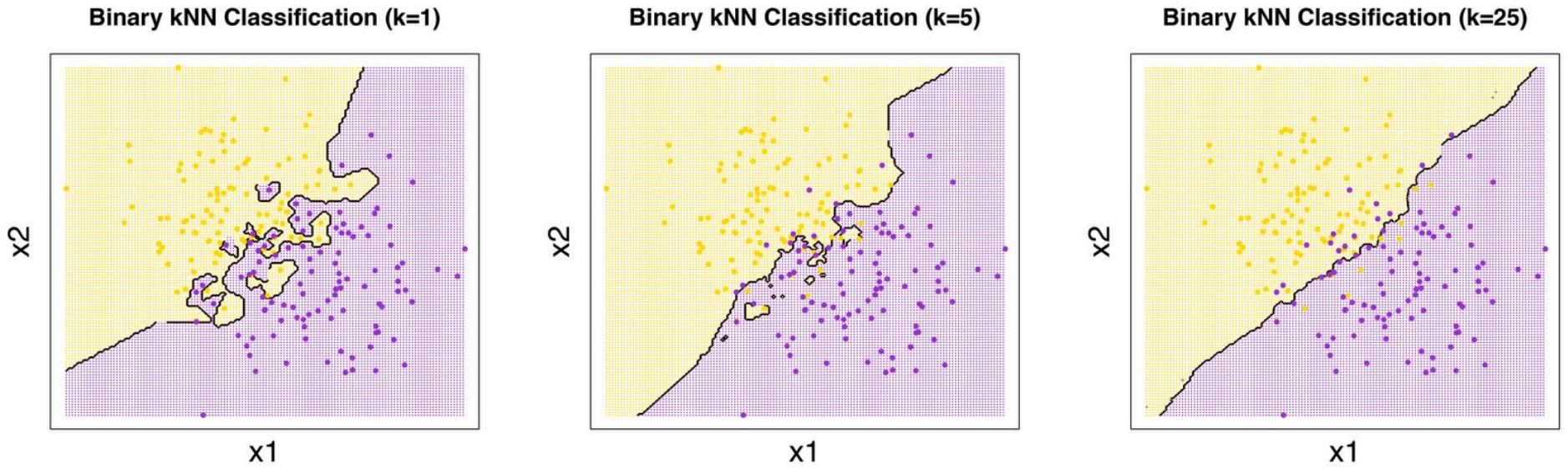


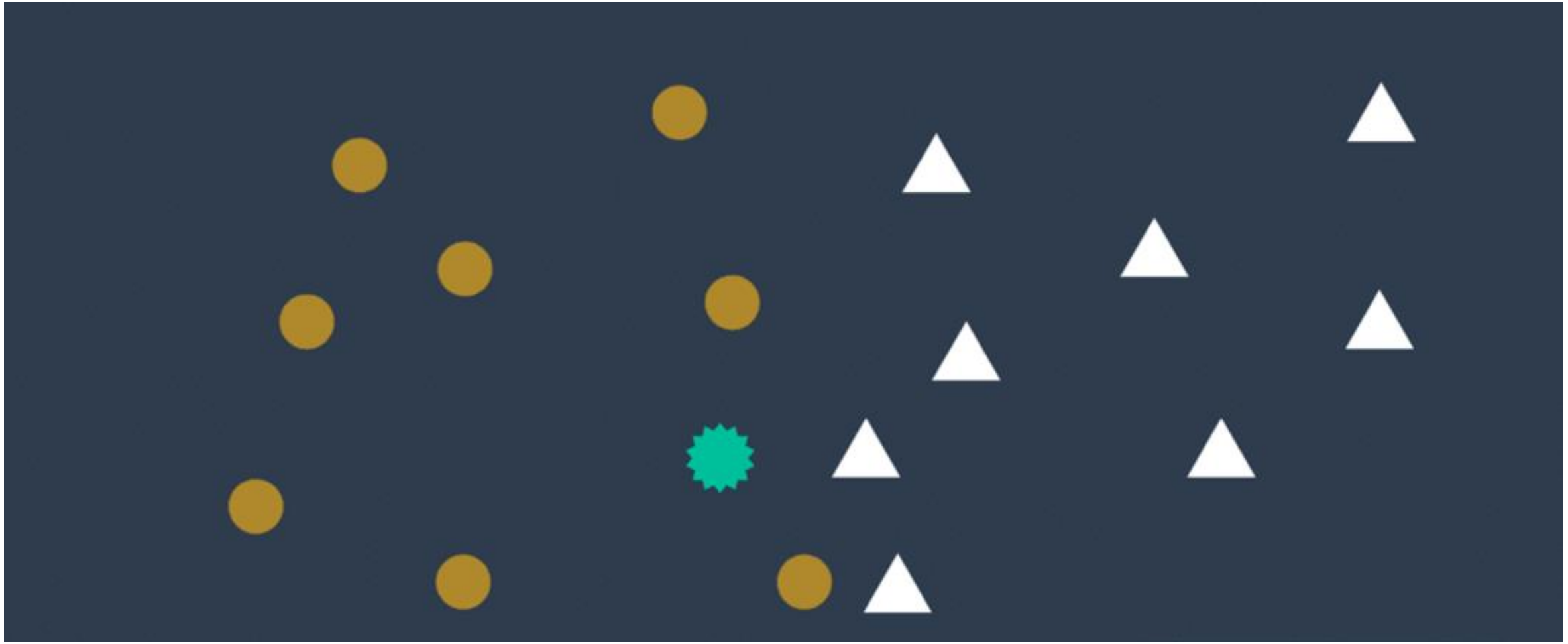
Идеи

Для произвольного $x \in X$ отранжируем объекты $x_1, \ldots, x_\ell$:

$$\rho(x, x^{(1)}) \leqslant \rho(x, x^{(2)}) \leqslant \cdots \leqslant \rho(x, x^{(\ell)}),$$

$w(i, x)$ — вес, оценка сходства объекта x с его i-м соседом, неотрицательная, не возрастающая по i.





Можно попробовать изменить параметр k – количество рассматриваемых соседей или усреднить значение по нескольким различным k.

$w(i,x) = K\left(\frac{\rho(x,x^{(i)})}{h}\right)$, где h — ширина окна,
$K(r)$ — ядро, не возрастает и положительно на $[0,1]$.

Метод парзеновского окна *фиксированной ширины:*

$$a(x; X^\ell, h, K) = \arg\max_{y \in Y} \sum_{i=1}^{\ell} [y_i = y]\, K\left(\frac{\rho(x, x_i)}{h}\right)$$

Метод парзеновского окна *переменной ширины:*

$$a(x; X^\ell, k, K) = \arg\max_{y \in Y} \sum_{i=1}^{\ell} [y_i = y]\, K\left(\frac{\rho(x, x_i)}{\rho(x, x^{(k+1)})}\right)$$

Проклятие размерности

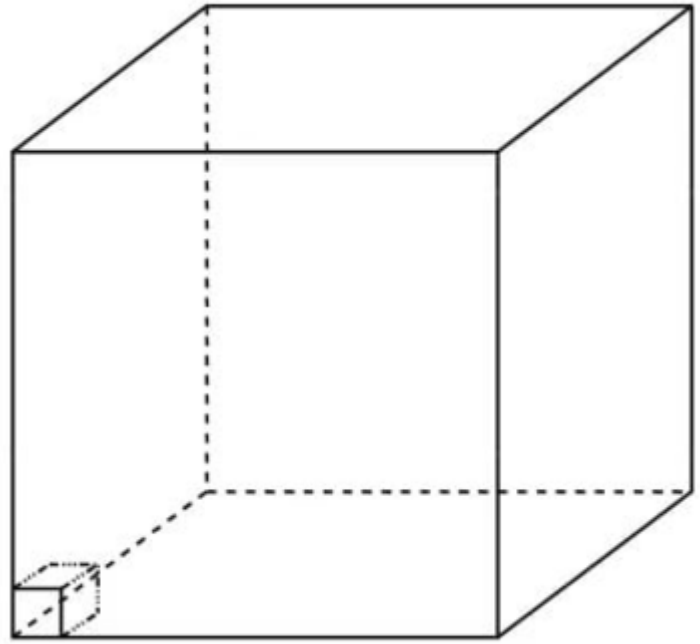
Вероятность = 0,95

$$\min\left\{\delta \mid \sum_{k=5}^{5000}\binom{5000}{k}\delta^k(1-\delta)^{5000-k} \geqslant 0.95\right\}.$$

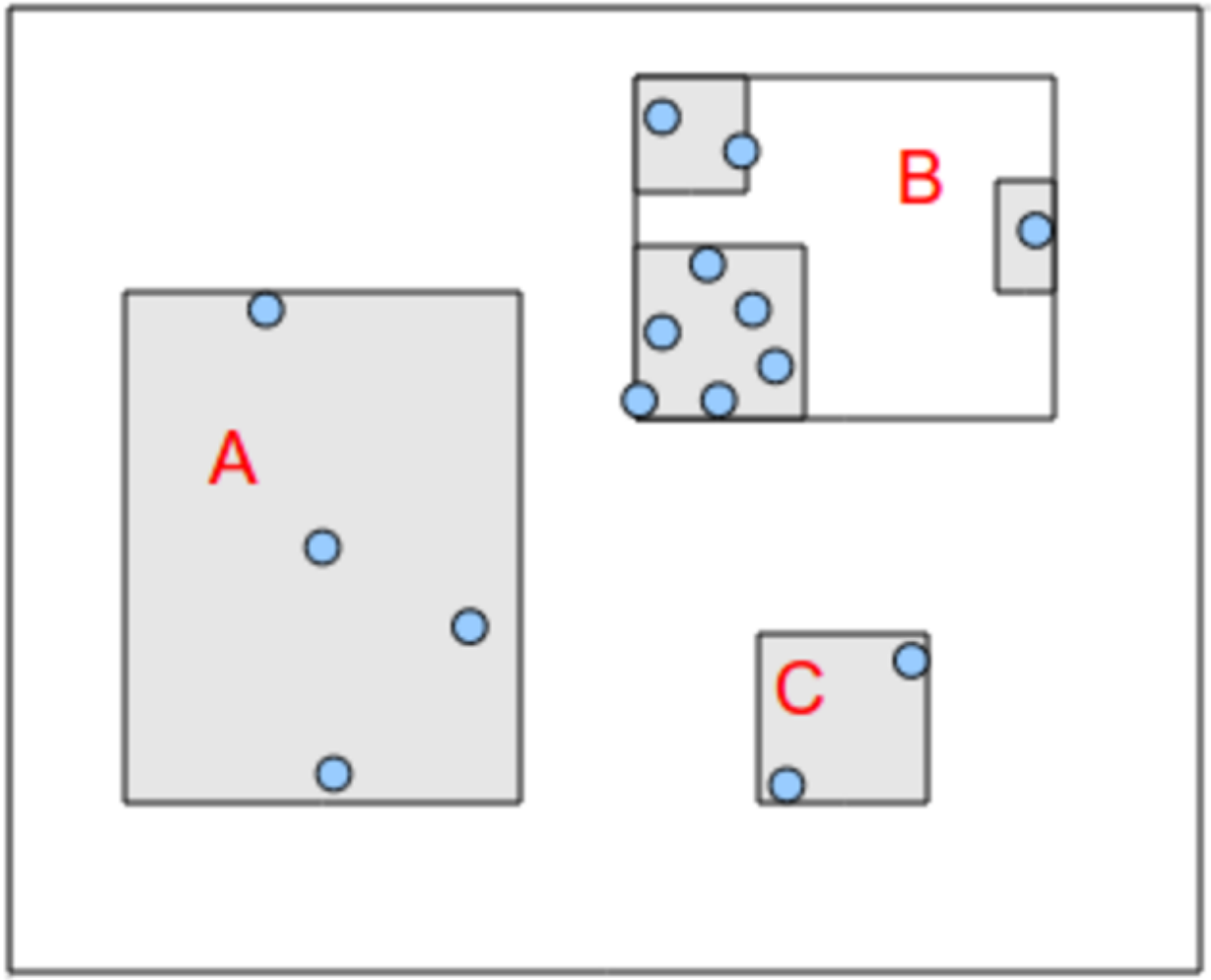
Минимальное значение δ, удовлетворяющее этому уравнению, приблизительно равно приблизительно 0.0018

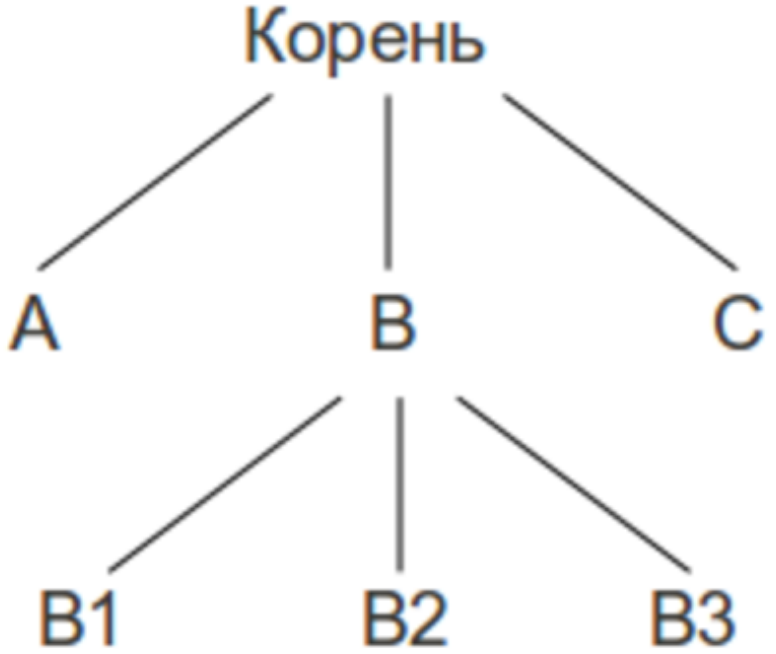
Итого, для того, чтобы найти пять соседей объекта u, нужно по каждой координате отступить на 0.0018^(1/d).

Уже при d = 10 получаем, что нужно отступить на 0.53, при d = 100 — на 0.94



KD деревья





Для классификации каждого из объектов тестовой выборки необходимо последовательно выполнить следующие операции:

- Вычислить расстояние до каждого из объектов обучающей выборки
- Отобрать k объектов обучающей выборки, расстояние до которых минимально
- Классифицировать объект

Плюсы и минусы метода ближайших соседей

Плюсы:

- Простая реализация;
- Неплохо изучен теоретически;
- Как правило, метод хорош для решения задач классификации
- Можно адаптировать под нужную задачу выбором метрики
- Неплохая интерпретация, можно объяснить, почему тестовый пример был классифицирован именно так.

Плюсы и минусы метода ближайших соседей

Минусы:

- Если в наборе данных много признаков, то трудно подобрать подходящие веса и определить, какие признаки не важны для решения
- Зависимость от выбранной метрики расстояния между примерами.
- Нет теоретических оснований выбора определенного числа соседей — только перебор
- В случае малого числа соседей метод чувствителен к выбросам, то есть склонен переобучаться

Используемая литература

- *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer, 2014. — 739 p.
- *Bishop C. M.* Pattern Recognition and Machine Learning. — Springer, 2006. — 738 p.
- Weber, R., Schek, H. J., Blott, S. (1998). A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. // Proceedings of the 24th VLDB Conference, New York C, 194–205.
- https://www.coursera.org/learn/ml-clustering-and-retrieval/lecture/S0gfp/kd-tree-representation
- https://habrahabr.ru/company/ods/blog/322534/
- http://www.machinelearning.ru/wiki/images/9/9a/Sem1_knn.pdf
- https://habrahabr.ru/post/312882/
- http://www.machinelearning.ru/wiki/index.php?title=Метод_ближайших_соседей
- http://www.cs.cornell.edu/~kilian/papers/NIPS2005_0265.pdf