Курсы по машинному обучению

Тема 7. Ошибки классификации и метрические алгоритмы

Метрические алгоритмы

Что делать если не можем создать матрицу признаков?

Задано:

- Графы
- Фотографии лиц
- Подписи
- Временные ряды
- Структуры белков

Метрические алгоритмы

Что делать если не можем создать матрицу признаков?

Задано:

- Графы
- Фотографии лиц
- Подписи
- Временные ряды
- Структуры белков

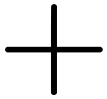
Умеем сравнивать объекты между собой

Метрические алгоритмы

Что делать если не можем создать матрицу признаков?

Задано:

- Графы
- Фотографии лиц
- Подписи
- Временные ряды
- Структуры белков



Умеем сравнивать объекты между собой

Введем метрику сходства объектов

Неотрицательная функция называется метрикой, если:

1. $d(x, y) = 0 \Leftrightarrow x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Метрические алгоритмы

Неотрицательная функция называется метрикой, если:

1. $d(x, y) = 0 \Leftrightarrow x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Примеры:

$d(x, y) = |x - y|$ Числа

$d(x, y) = \sqrt{\sum(x_i - y_i)^2}$ Вектора

$d(f, g) = sup|f(x) - g(x)|$ Функции

Гипотеза компактности

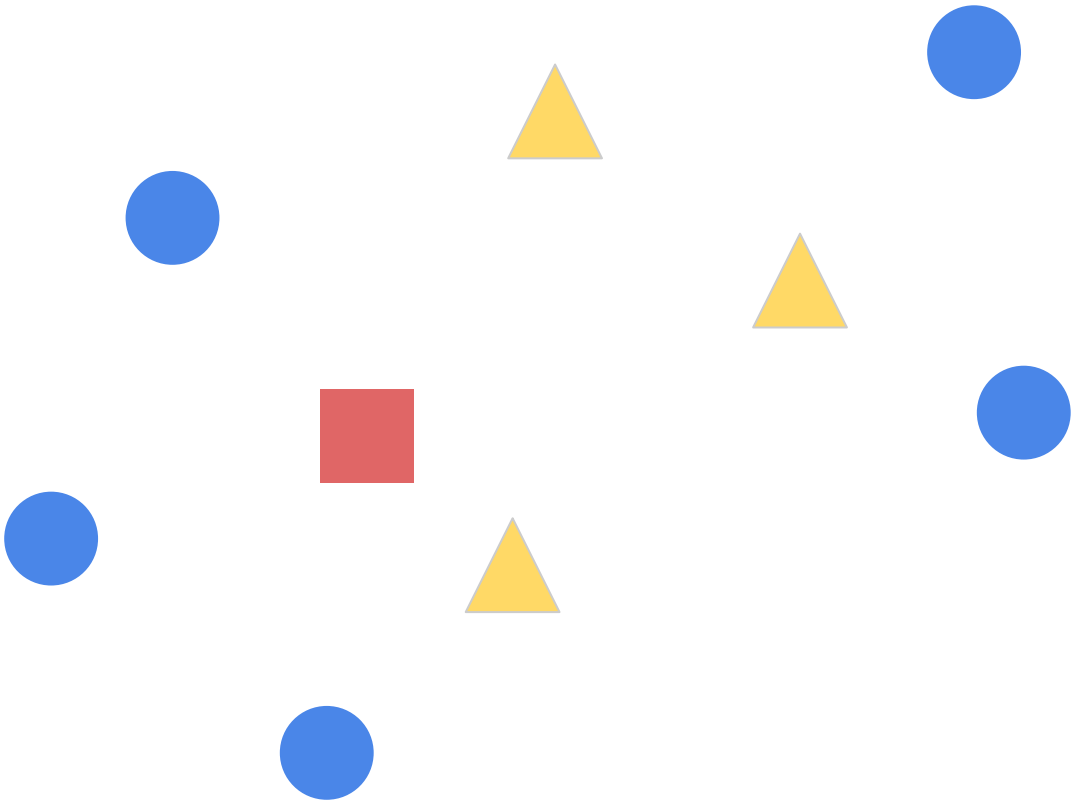
Гипотеза компактности — в задачах классификации предположение о том, что схожие объекты гораздо чаще лежат в одном классе, чем в разных

Другими словами, что классы образуют компактно локализованные подмножества в пространстве объектов.

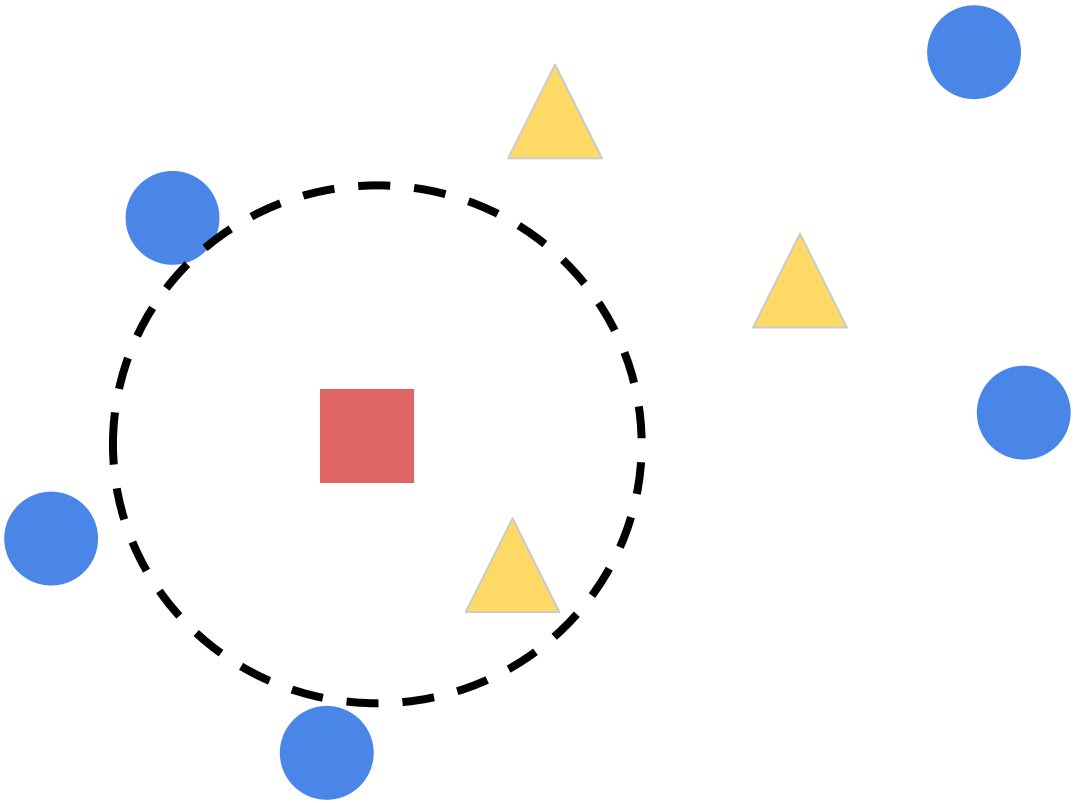
Это также означает, что граница между классами имеет достаточно простую форму.

* Компактные множества тут не при чем

Метод ближайшего соседа



Метод ближайшего соседа



Определяем самый ближний элемент и класс нашего объекта будет совпадать с ближайшим

Метод ближайшего соседа

mail.ru group

Плюсы

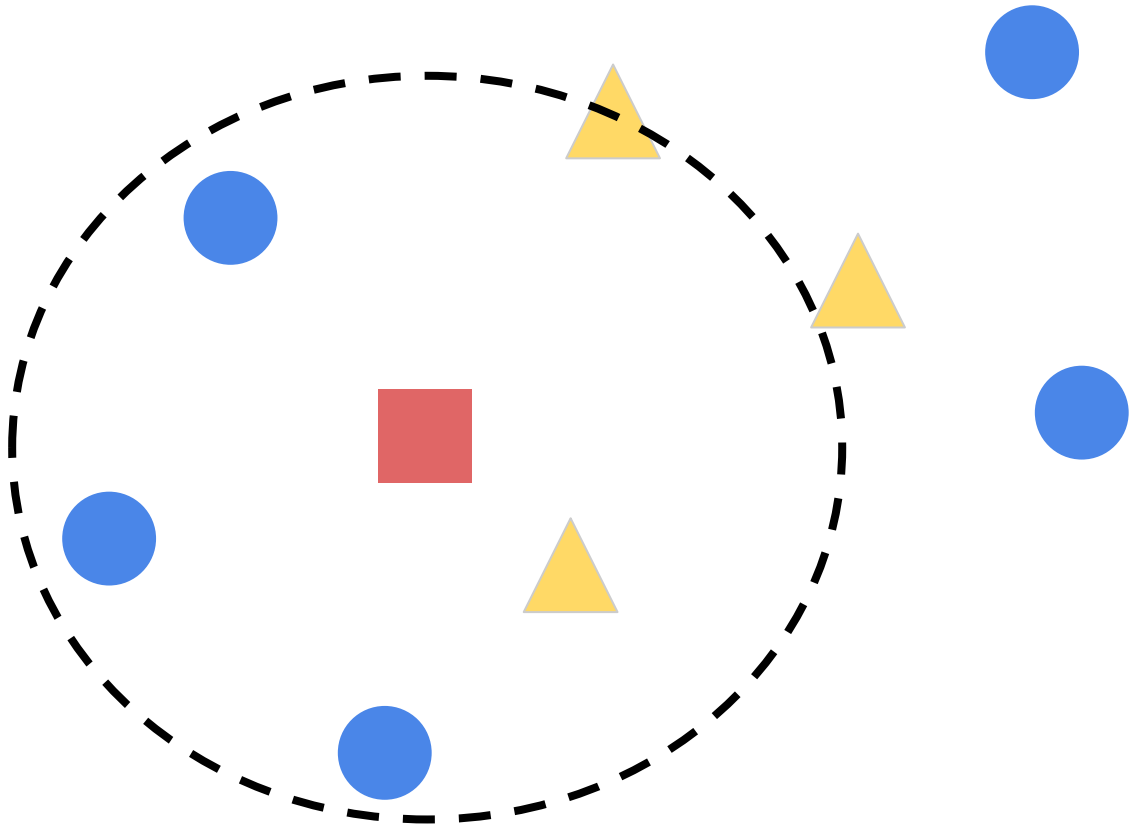
- Простая реализация
- Интерпретируемость

Минусы

- Неустойчивость к выбросам
- Мало гиперпараметров
- Полная зависимость от метрики
- Низкое качество

Метод k ближайших соседей

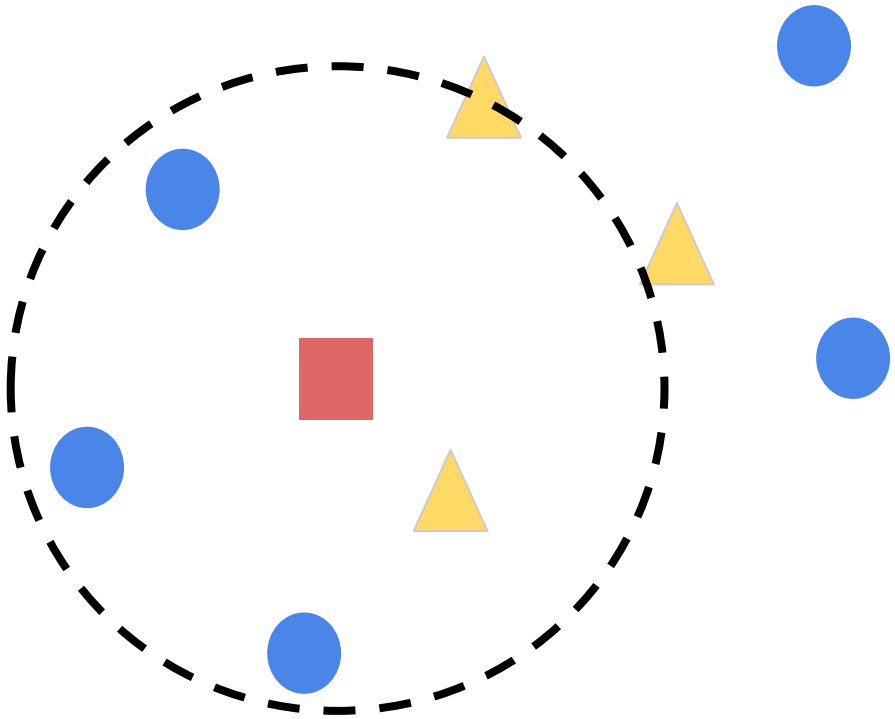
$$arg\max_{y \in Y} \sum_{i=1}^{k} \left[y_u^{(i)} = y \right]$$



Определяем k самых ближайших элементов и класс нашего объекта будет выбран с помощью голосования, аналогично голосованию в случайном лесу

Метод k ближайших соседей

- Выбор класса производится голосованием
- k=1 -> метод ближайшего соседа
- k=L -> метод константа (для всех объектов будет выбран преобладающий класс)
- Подбор параметра по критерию скользящего контроля с исключением объектов по одному (leave-one-out)
- Также можно сравнивать средние расстояния до объектов другого класса



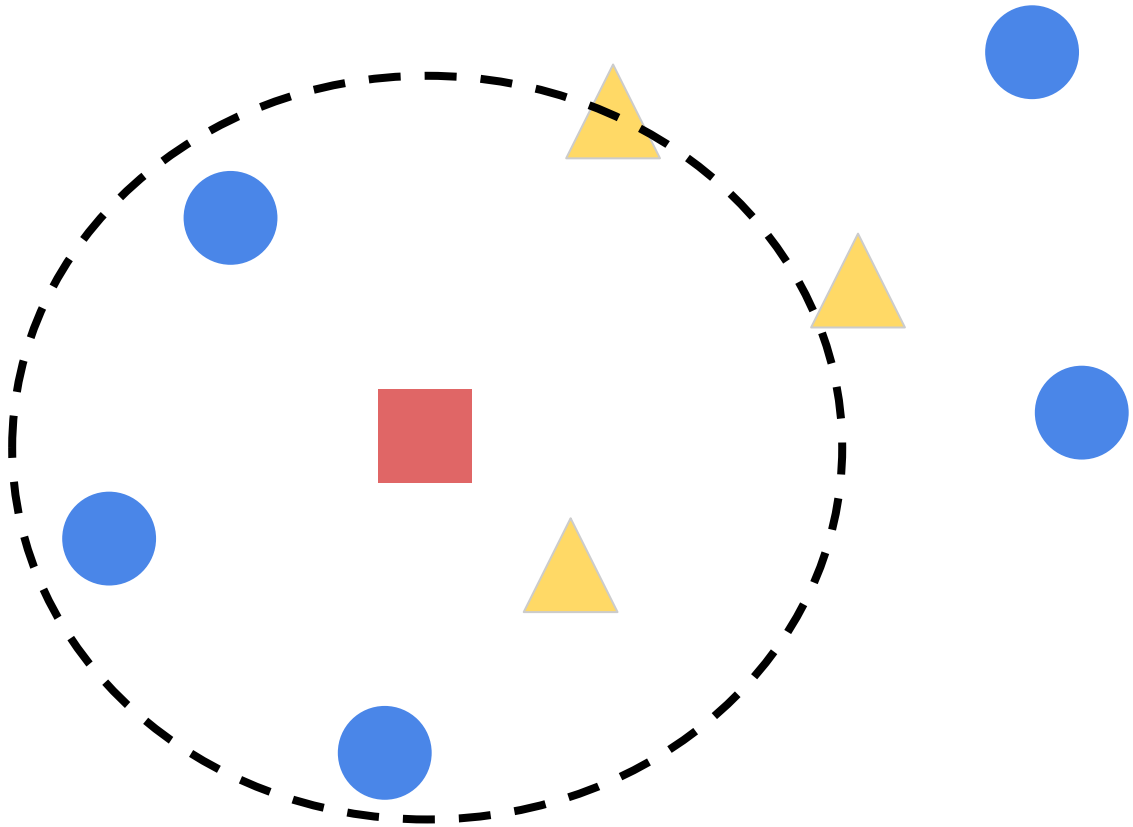
Метод k ближайших соседей

Что делать если голосование не выявило лидирующий класс?

- Для двух классов брать нечетные k
- Использовать веса
- Выбирать случайно

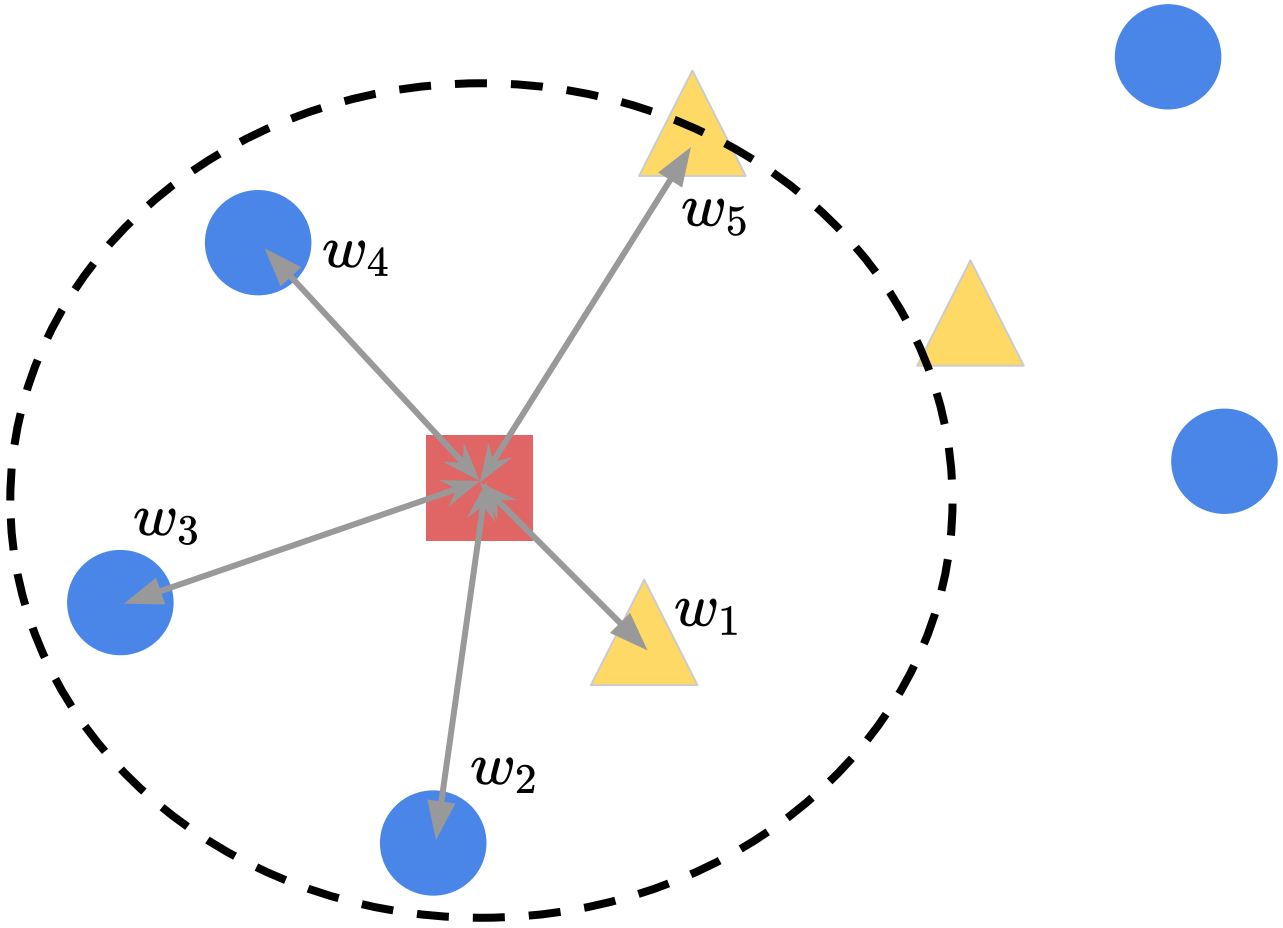
Метод k ближайших соседей

$$arg\max_{y \in Y} \sum_{i=1}^{k} \left[y_u^{(i)} = y \right] w_i$$



Определяем k самых ближайших элементов и класс нашего объекта будет выбран с помощью голосования, но с учетом веса каждого элемента

Метод парзеновского окна



Метод парзеновского окна

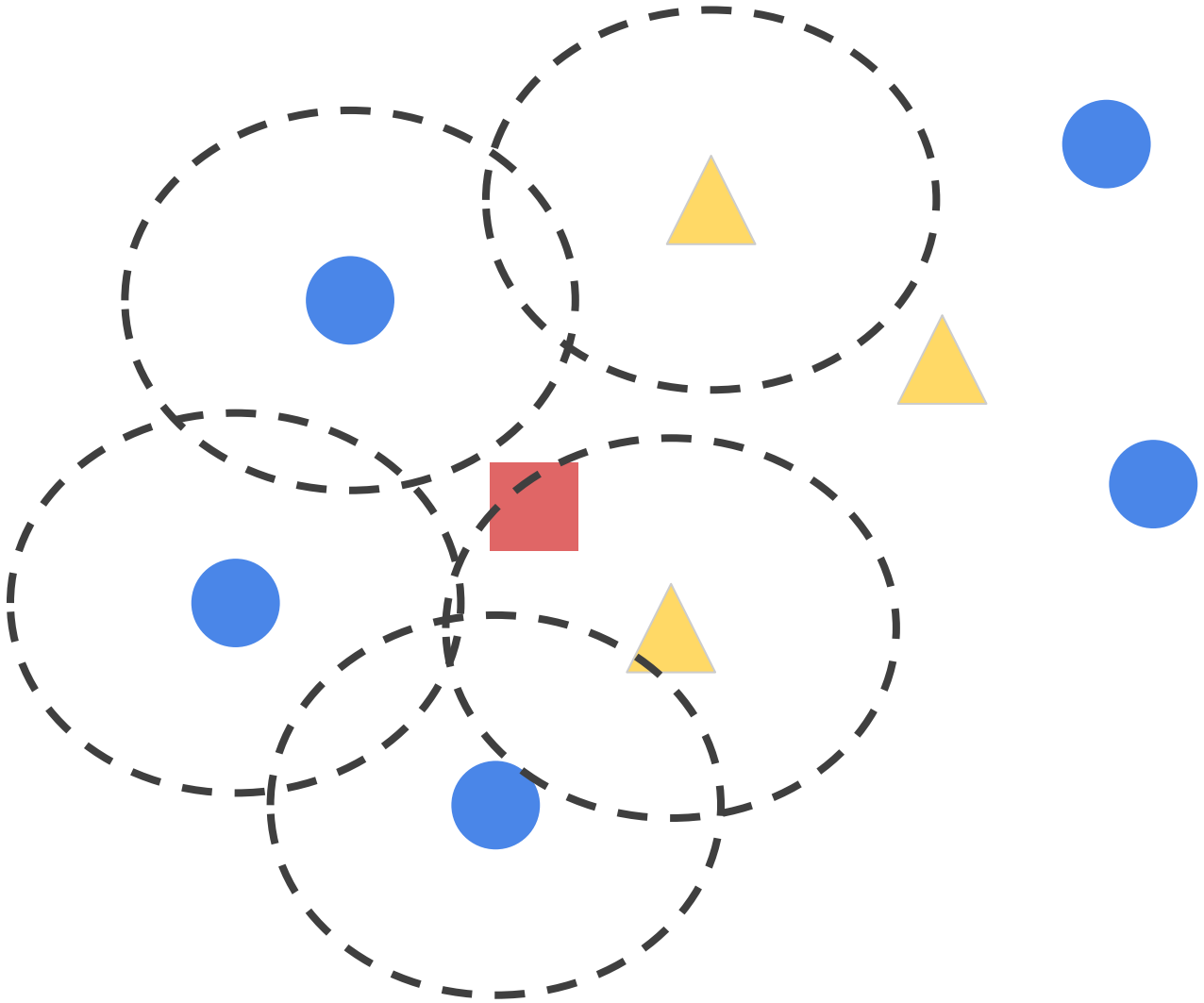
- $\rho(u, x_u^{(i)})$ – функция расстояния до соседа
- Функция ядра $K(z)$ – не возрастает на $[0, \infty)$
- Постоянное ширина окна

$$a(u; X^l, h, K) = arg \max_{y \in Y} \sum_{i=1}^{l} \left[y_u^{(i)} = y \right] K\left(\frac{\rho(u, x_u^{(i)})}{h} \right)$$

- Переменная ширина окна (неравномерное распределение)

$$a(u; X^l, k, K) = arg \max_{y \in Y} \sum_{i=1}^{k} \left[y_u^{(i)} = y \right] K\left(\frac{\rho(u, x_u^{(i)})}{\rho(u, x_u^{(k+1)})} \right)$$

Метод потенциальных функций



Проблемы метрических алгоритмов

- Выбор числа соседей k
- Отсев шума
- Большие выборки
- При погрешностях (в данных или метрике) снижается точность по границам классов

- Поиск ближайшего соседа = сравнение с каждым элементом выборки
- Проблема выбора метрики
- Проклятье размерности
- Малое число параметров

Проблемы метрических алгоритмов

- Отступ объекта $x_i \epsilon X^i$ относительно алгоритма классификации, имеющего вид $a(u) = arg \max_{y \in Y} \Gamma_y(u)$, называется величина

$$M(x_i) = \Gamma_{y_i}(x_i) - \max_{y \in Y \setminus y_i} \Gamma_y(x_i)$$

Эталонные	M >> 0	По ним можно классифицировать
Неинформативные	M > 0	Не несут важной информации
Пограничные	M ~ 0	Чувствительны к изменениям
Ошибочные	M < 0	Ошибка в метрике
Шумовые (выбросы)	M << 0	Ошибка в данных

Проклятие размерности

- В n-мерном шаре весь объем сосредоточен на сфере
- Пример. Рассмотрим сферу в 20-мерном пространстве. Формула для объема шара в 20-мерном пространстве:

$$V = \frac{\pi^{10}\pi}{10!}R^{20}$$

Найдем отношение объема шара радиуса 1 и радиуса 0.9:

$$\frac{V_{0,9}}{V_1} = \frac{0,9^{20}}{1} = 0.12$$

- 88% точек лежит на сфере, а значит почти до всех ближайших соседей расстояние одинаковое

Проклятие размерности

- Также данные крайне разреженные
- Пример 1. Рассмотрим единичный интервал [0; 1]. 100 равномерно разбросанных точек будет достаточно, чтобы покрыть этот интервал с частотой не менее 0.01
- Пример 2. Теперь рассмотрим 10-мерный куб. Для 10^{20} достижения той же степени покрытия потребуется уже точек. То есть, по сравнению с одномерным пространством, требуется в 10^{18} раз больше точек.

Проклятие размерности

Решение - снижение размерности.
Например, PCA - метод главных компонент