МАШИННОЕ ОБУЧЕНИЕ
метрические методы классификации, регрессии и кластеризации

Воронцов Константин Вячеславович
ФУПМ МФТИ • ВМК МГУ • Яндекс • FORECSYS

8 июля 2016
Сочи, Сириус • Проектная смена • 1–24 июля 2016

Содержание

Восстановление зависимости по эмпирическим данным

Задача восстановления зависимости $y = f(x)$
по точкам *обучающей выборки* (x_i, y_i), $i = 1, \ldots, \ell$.

Дано: векторы $x_i = (x_i^1, \ldots, x_i^n)$ — объекты обучающей выборки,
$y_i = f(x_i)$ — правильные ответы, $i = 1, \ldots, \ell$:

$$\begin{pmatrix} x_1^1 & \ldots & x_1^n \\ \ldots & \ldots & \ldots \\ x_\ell^1 & \ldots & x_\ell^n \end{pmatrix} \xrightarrow{f} \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}$$

Найти: функцию $a(x)$, способную давать правильные ответы
на *тестовых объектах* $\tilde{x}_i = (\tilde{x}_i^1, \ldots, \tilde{x}_i^n)$, $i = 1, \ldots, q$:

$$\begin{pmatrix} \tilde{x}_1^1 & \ldots & \tilde{x}_1^n \\ \ldots & \ldots & \ldots \\ \tilde{x}_q^1 & \ldots & \tilde{x}_q^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \ldots \\ a(\tilde{x}_q) \end{pmatrix}$$

Гипотезы компактности и непрерывности

Гипотеза непрерывности (для регрессии):
близким объектам соответствуют близкие ответы.

Гипотеза компактности (для классификации):
близкие объекты, как правило, лежат в одном классе.

Формализация понятия «близости»:
задана функция расстояния $\rho(x, x_i)$.

Пример. Евклидово расстояние и его обобщение:

$$\rho(x, x_i) = \left(\sum_{j=1}^{n} \left|x^j - x_i^j\right|^2\right)^{1/2} \qquad \rho(x, x_i) = \left(\sum_{j=1}^{n} w_j \left|x^j - x_i^j\right|^p\right)^{1/p}$$

$x = (x^1, \ldots, x^n)$ — вектор признаков объекта x,
$x_i = (x_i^1, \ldots, x_i^n)$ — вектор признаков объекта x_i.

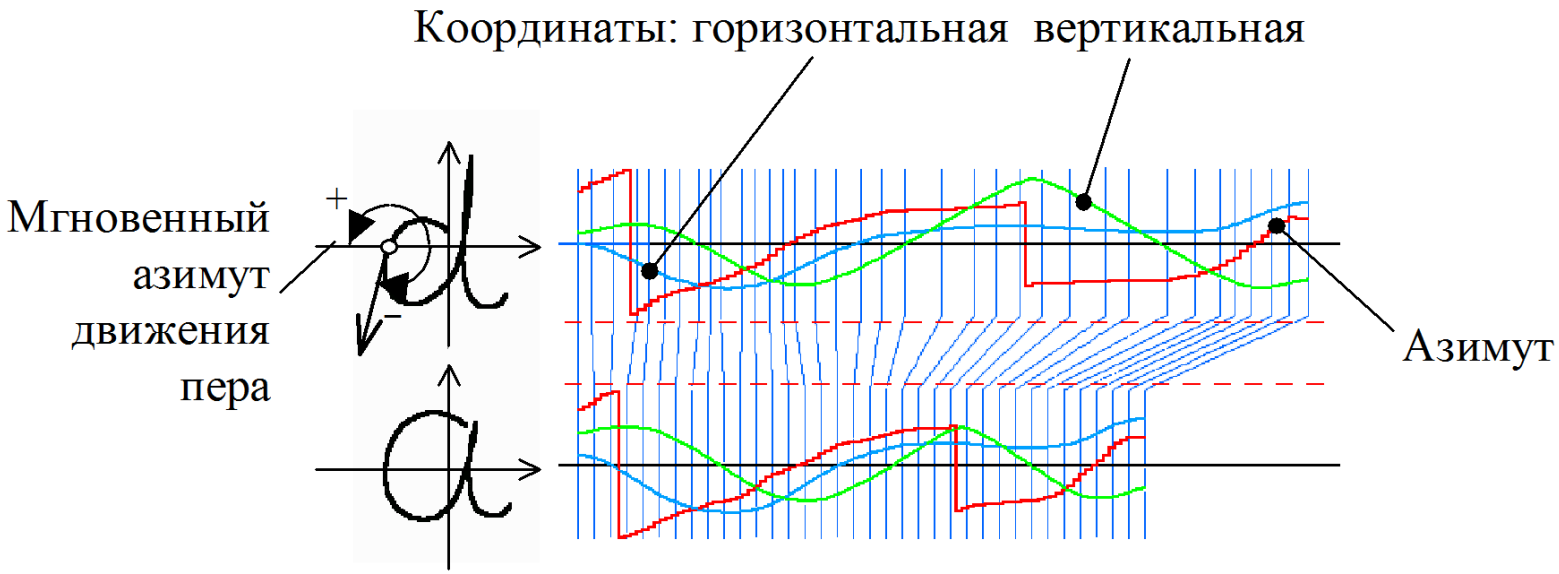
Беспризнаковые способы вычисления расстояний

Расстояния на основе выравниваний:

- **между текстами** (редакторское расстояние Левенштейна):

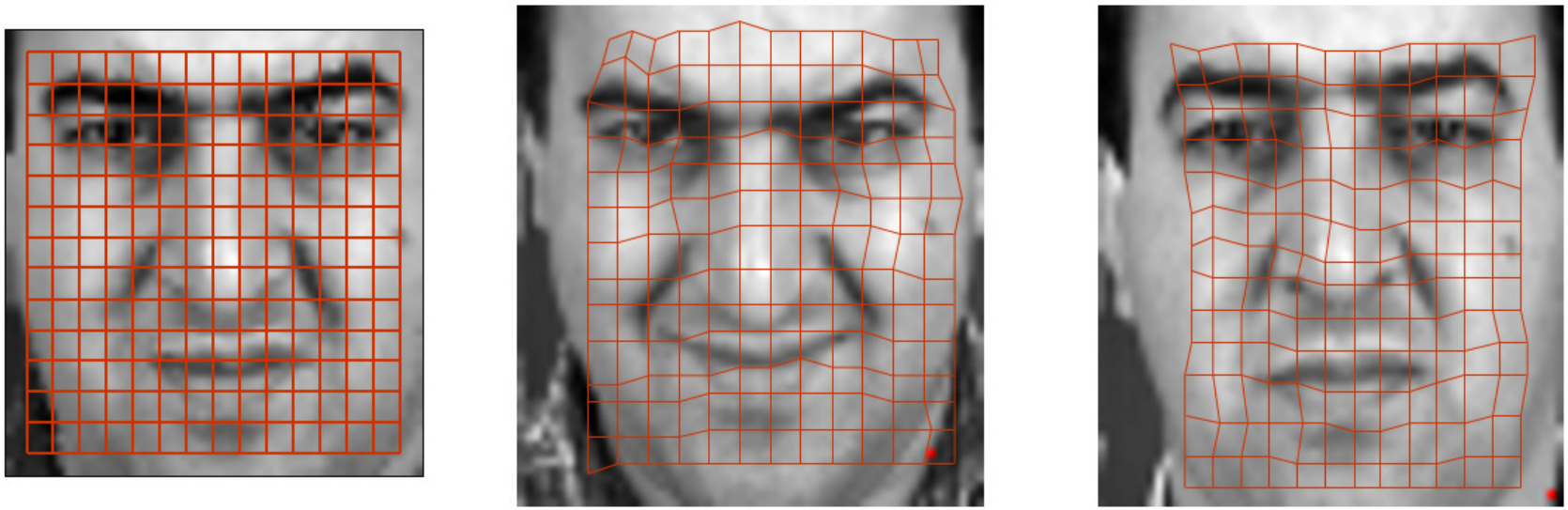
```
CTGGGCTAAAAGGTCCCTTAGCC--TTTAGAAAAA-GGGCCATTAGGAAATTGCA
CTGGGACTAAA----CCTTAGCCTATTTACAAAAATGGGCCATTAGG---TTGCA
```

- **между сигналами** (энергия сжатий и растяжений):



Беспризнаковые способы вычисления расстояний

Расстояния **между изображениями** на основе выравнивания:



Оценивается энергия растяжения прямоугольной сетки

Обобщённый метрический классификатор

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

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

$x^{(i)}$ — i-й сосед объекта x среди $x_1, \ldots, x_\ell$;
$y^{(i)}$ — ответ на i-м соседе объекта x.

Метрический алгоритм классификации:

$$a(x; X^\ell) = \arg\max_{y \in Y} \underbrace{\sum_{i=1}^{\ell} \left[y^{(i)} = y\right] w(i, x)}_{\Gamma_y(x)},$$

$w(i, x)$ — вес, оценка сходства объекта x с его i-м соседом,
неотрицательная, не возрастающая по i.
$\Gamma_y(x)$ — *оценка близости* объекта x к классу y.

Метод k ближайших соседей (k nearest neighbors, kNN)

$w(i,x) = [i \leqslant 1]$ — метод ближайшего соседа
$w(i,x) = [i \leqslant k]$ — метод k ближайших соседей

Преимущества:

- простота реализации (lazy learning);
- параметр k можно оптимизировать по критерию скользящего контроля (leave-one-out):

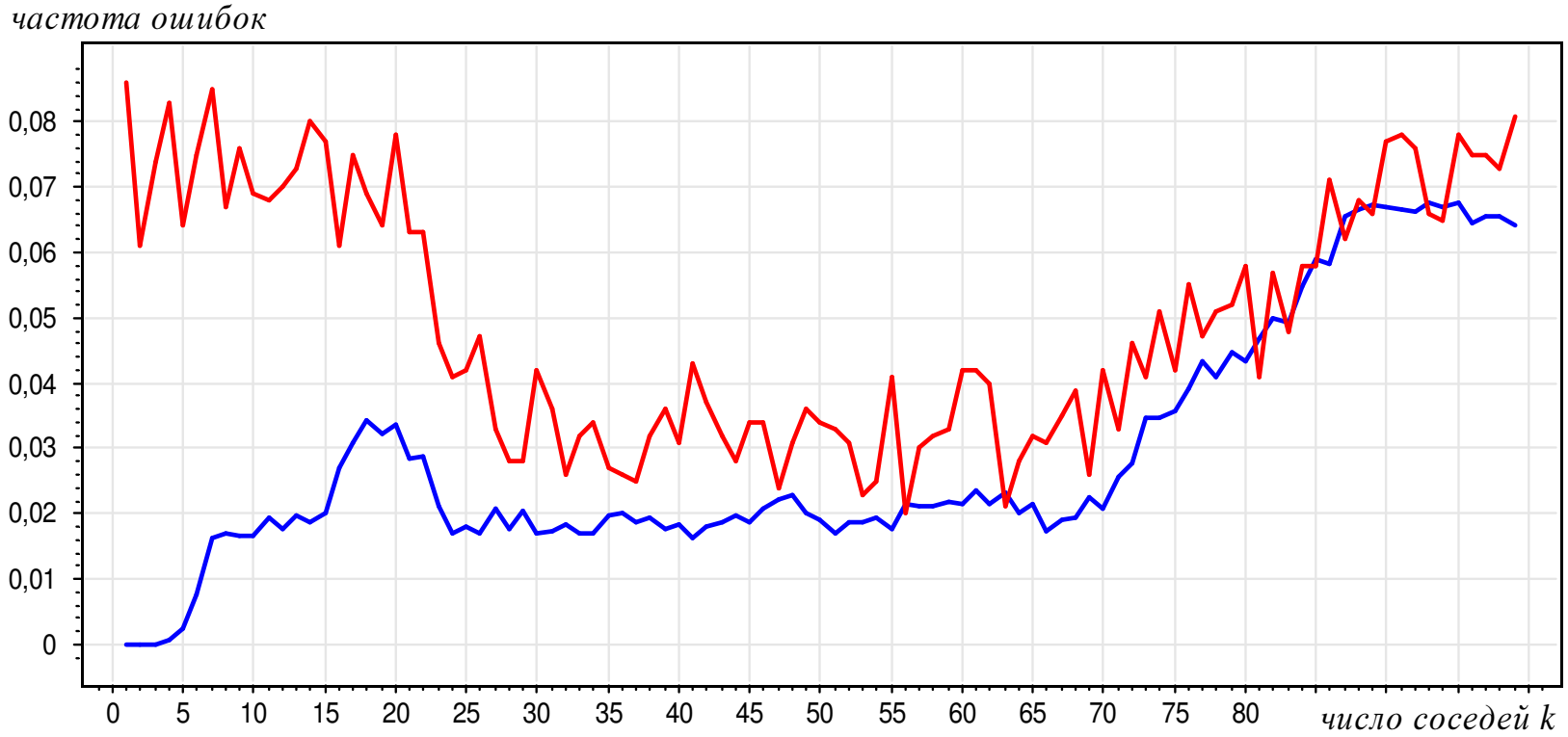
$$\mathrm{LOO}(k, X^\ell) = \sum_{i=1}^{\ell} \Big[a\big(x_i; X^\ell \setminus \{x_i\}, k\big) \neq y_i \Big] \to \min_k .$$

Недостатки:

- неоднозначность классификации при $\Gamma_y(x) = \Gamma_s(x), \quad y \neq s.$
- не учитываются значения расстояний

Пример зависимости LOO от числа соседей

Пример. Задача Iris, усреднение по 50 случайным разбиениям



— несмещённое число ошибок LOO

— смещённое число ошибок, когда объект учитывается как сосед самого себя

Метод окна Парзена

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

Метод парзеновского окна *заданной ширины* h:

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

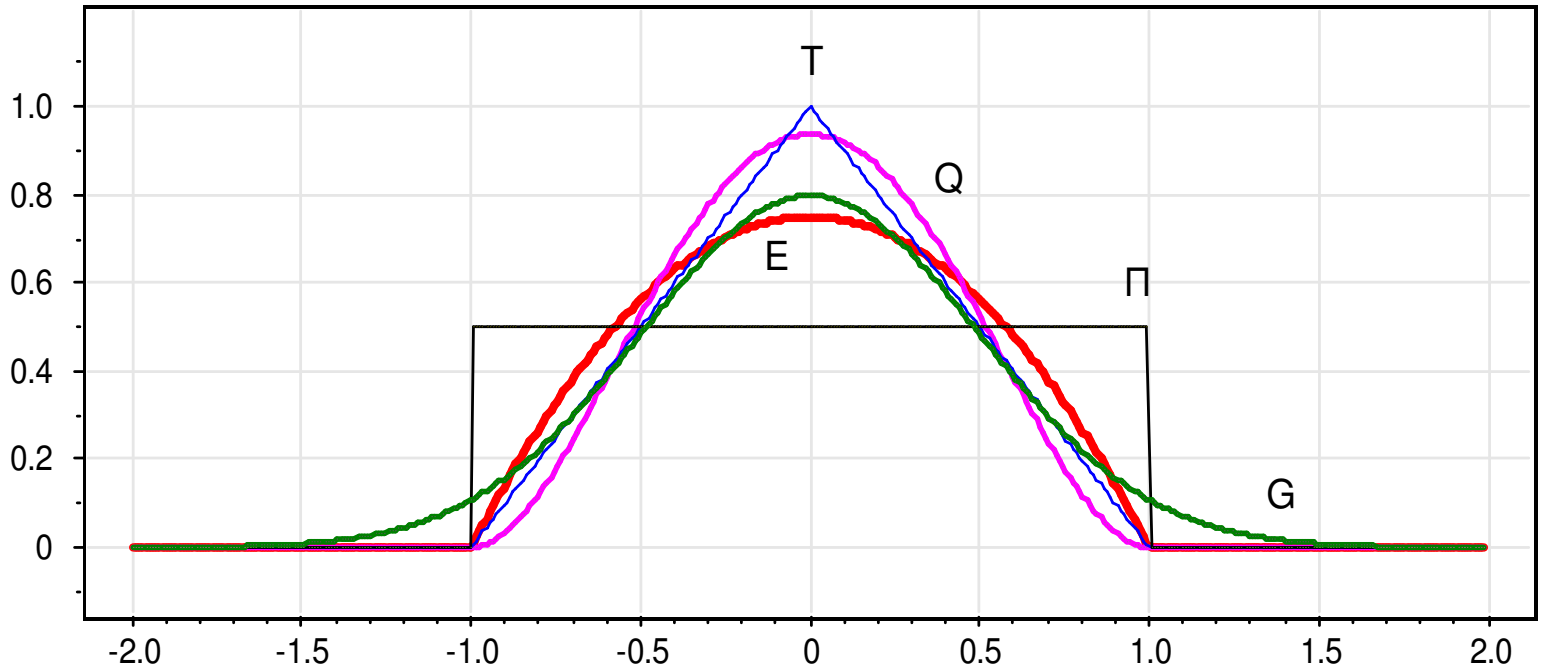
Метод парзеновского окна *переменной ширины* $h = \rho(x, x^{(k+1)})$:

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

Оптимизация параметров — по критерию LOO:

- выбор ширины окна h или числа соседей k
- выбор ядра K слабо влияет на качество классификации

Часто используемые ядра $K(r)$



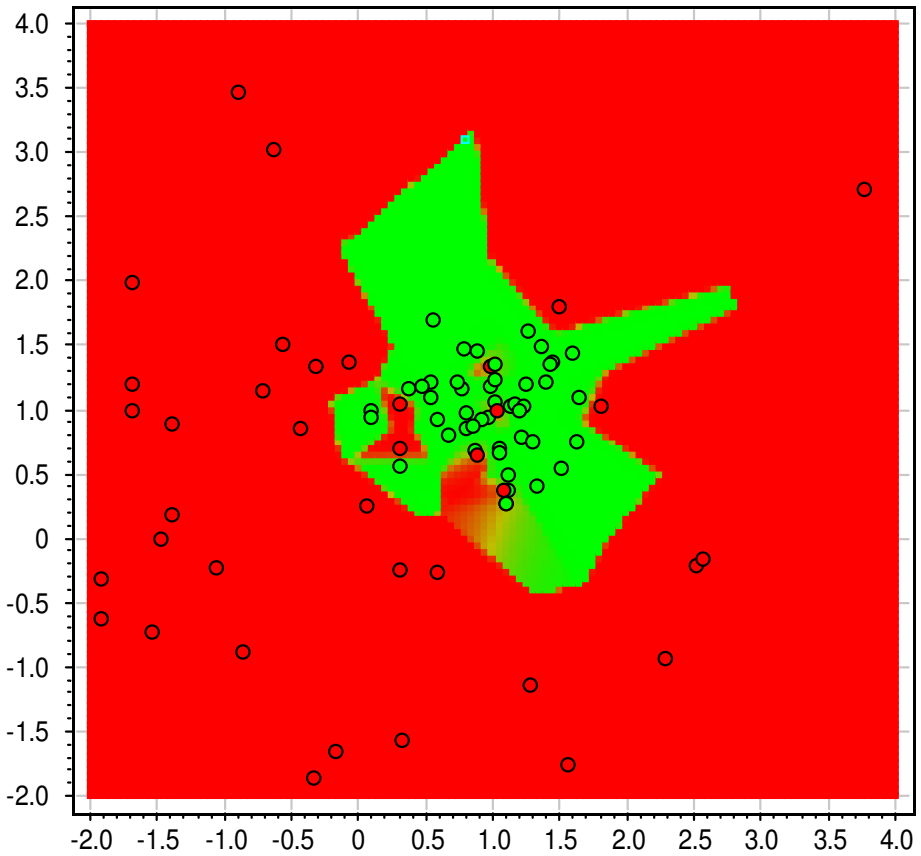
$\Pi(r) = \big[|r| \leqslant 1\big]$ — прямоугольное
$T(r) = \big(1 - |r|\big)\big[|r| \leqslant 1\big]$ — треугольное
$E(r) = (1 - r^2)\big[|r| \leqslant 1\big]$ — квадратичное (Епанечникова)
$Q(r) = (1 - r^2)^2\big[|r| \leqslant 1\big]$ — квартическое
$G(r) = \exp\big(-2r^2\big)$ — гауссовское

Парзеновское окно ширины h

Пример: двумерная выборка, два класса $Y = \{-1, +1\}$.

$$a(x) = \arg\max_{y \in Y} \Gamma_y(x) = \text{sign}\big(\underbrace{\Gamma_{+1}(x) - \Gamma_{-1}(x)}\big)$$

$h = 0.05$

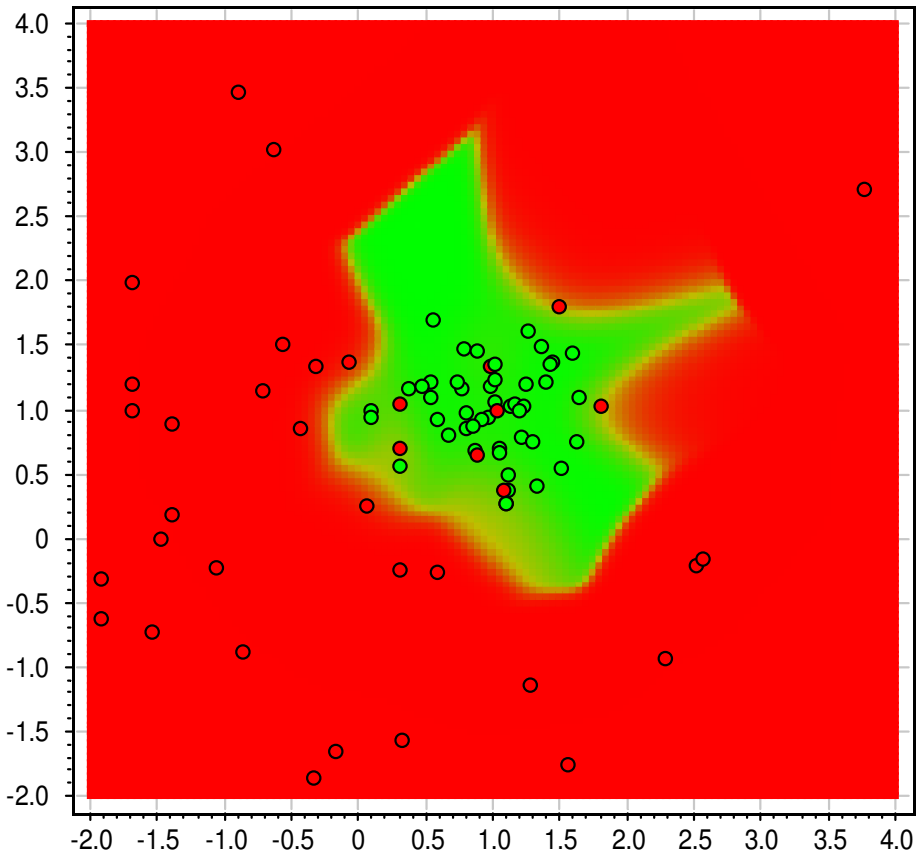


Парзеновское окно ширины h

Пример: двумерная выборка, два класса $Y = \{-1, +1\}$.

$$a(x) = \arg\max_{y \in Y} \Gamma_y(x) = \text{sign}\big(\underbrace{\Gamma_{+1}(x) - \Gamma_{-1}(x)}\big)$$

$h = 0.2$

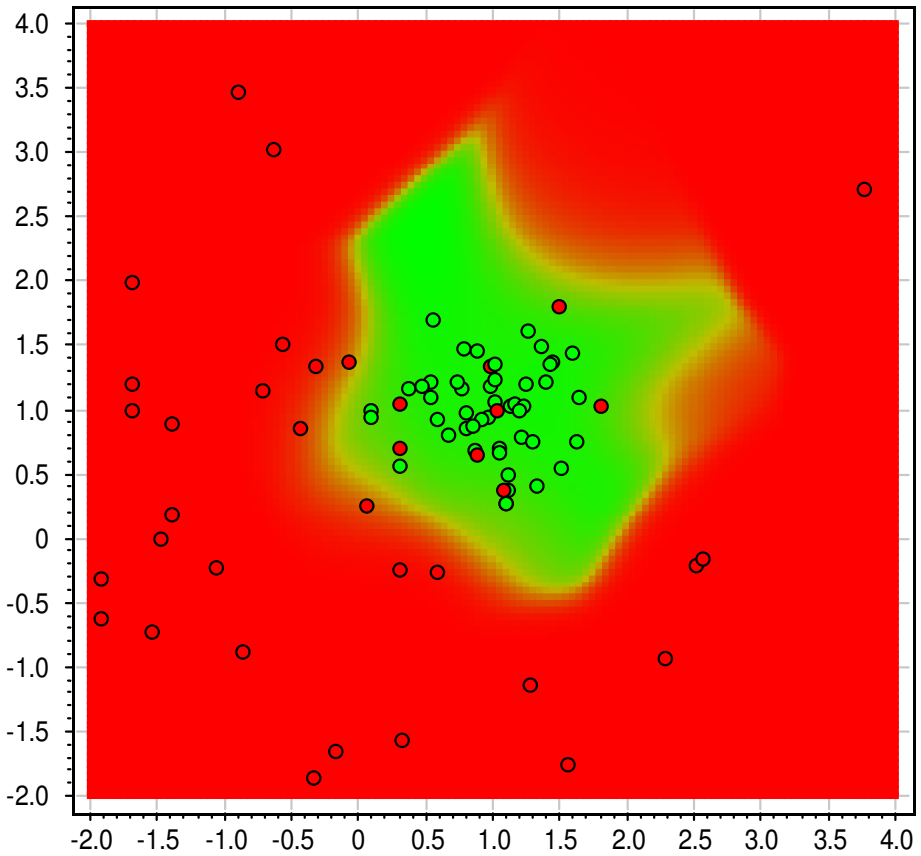


Парзеновское окно ширины h

Пример: двумерная выборка, два класса $Y = \{-1, +1\}$.

$$a(x) = \arg\max_{y \in Y} \Gamma_y(x) = \text{sign}\big(\underbrace{\Gamma_{+1}(x) - \Gamma_{-1}(x)}\big)$$

$h = 0.3$

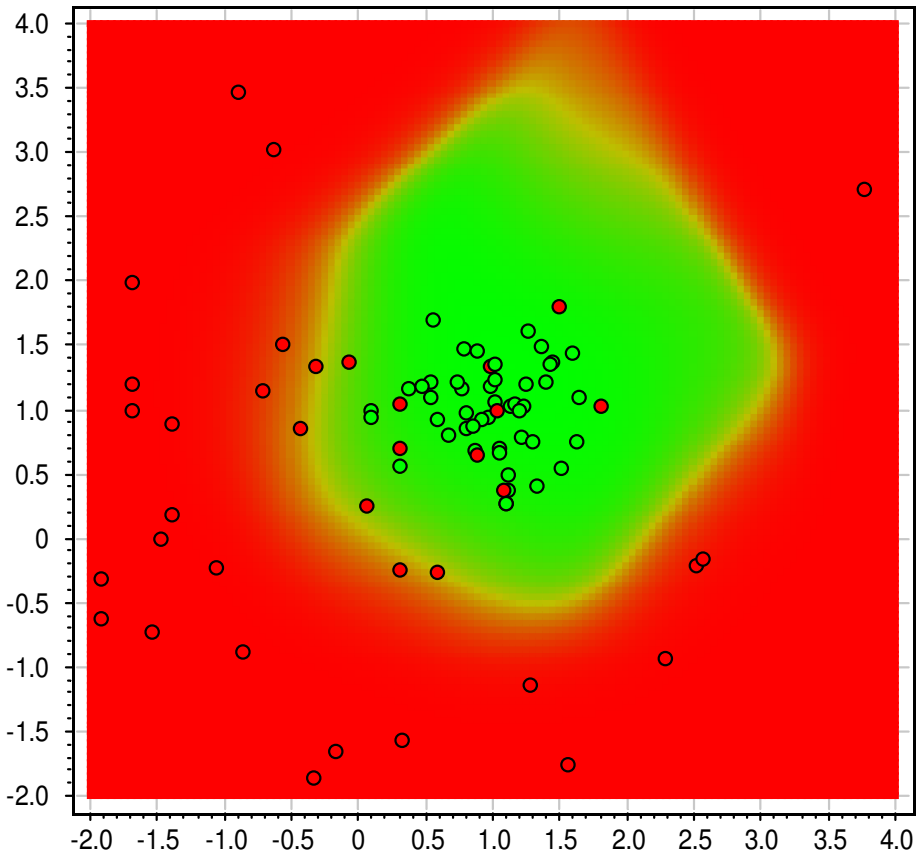


Парзеновское окно ширины h

Пример: двумерная выборка, два класса $Y = \{-1, +1\}$.

$$a(x) = \arg\max_{y \in Y} \Gamma_y(x) = \text{sign}\big(\underbrace{\Gamma_{+1}(x) - \Gamma_{-1}(x)}\big)$$

$h = 0.5$

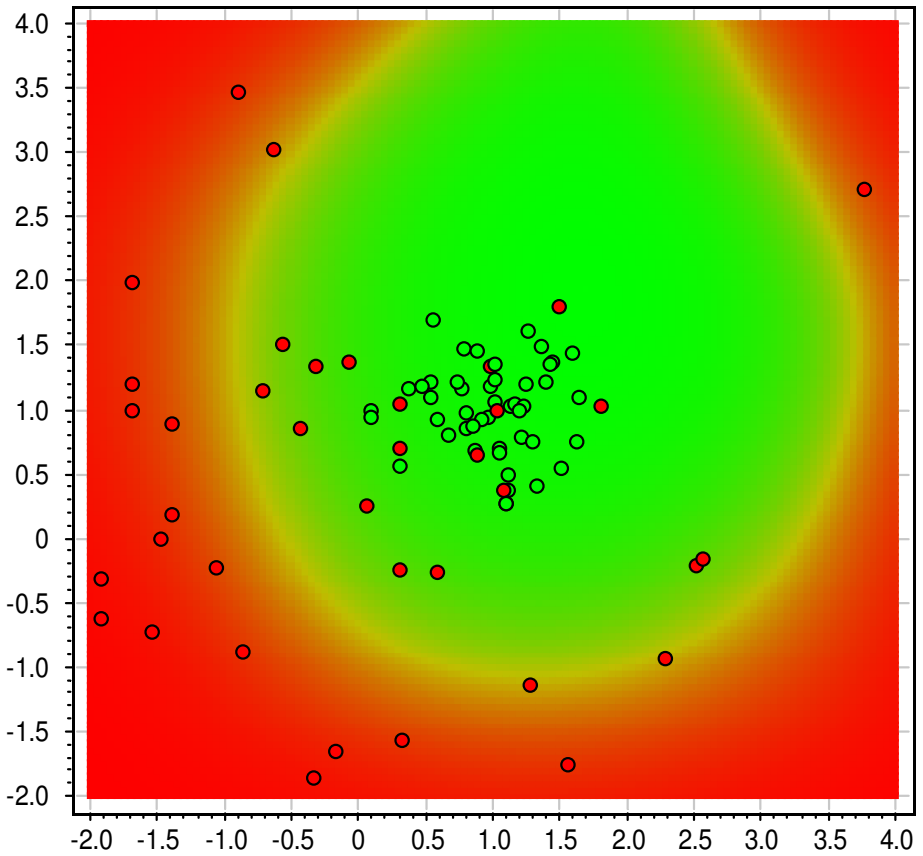


Парзеновское окно ширины h

Пример: двумерная выборка, два класса $Y = \{-1, +1\}$.

$$a(x) = \arg\max_{y \in Y} \Gamma_y(x) = \text{sign}\big(\underbrace{\Gamma_{+1}(x) - \Gamma_{-1}(x)}\big)$$

$h = 1.0$

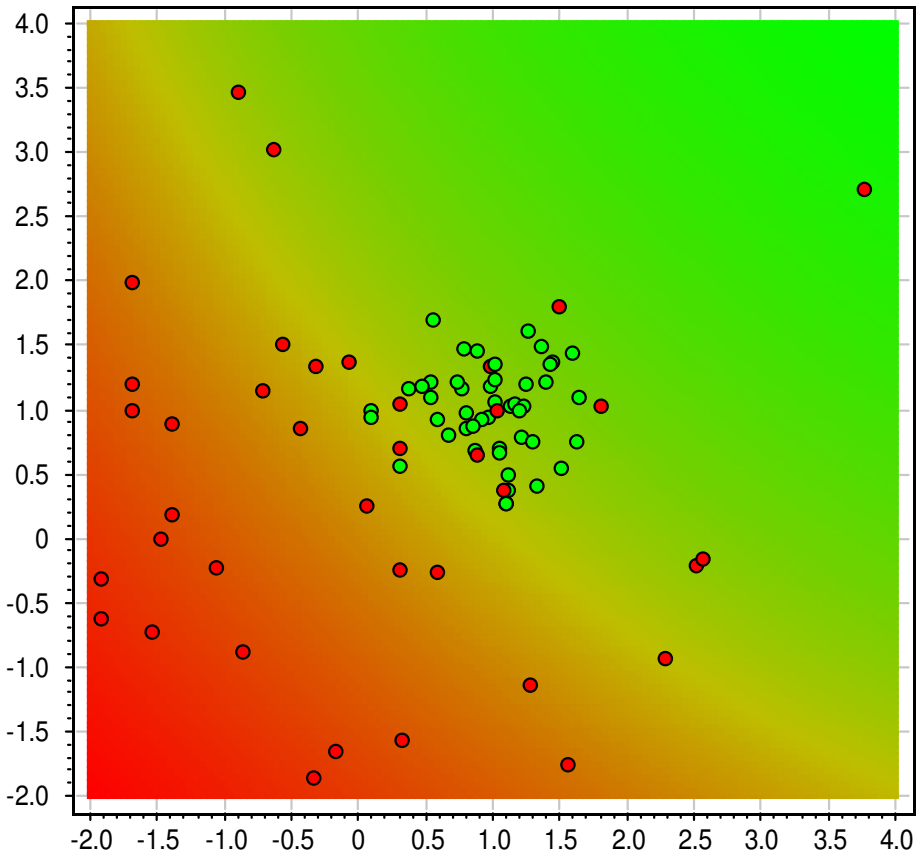


Парзеновское окно ширины h

Пример: двумерная выборка, два класса $Y = \{-1, +1\}$.

$$a(x) = \arg\max_{y \in Y} \Gamma_y(x) = \text{sign}\big(\underbrace{\Gamma_{+1}(x) - \Gamma_{-1}(x)}\big)$$

$h = 5.0$



Метод потенциальных функций

$$w(i,x) = \gamma^{(i)} K\left(\frac{\rho(x,x^{(i)})}{h^{(i)}}\right)$$

Более простая запись (без ранжирования объектов):

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

где γ_i — веса объектов, $\gamma_i \geqslant 0$, $h_i > 0$.

Физическая аналогия:

γ_i — величина «заряда» в точке x_i;

h_i — «радиус действия» потенциала с центром в точке x_i;

y_i — знак «заряда» (в случае двух классов $Y = \{-1, +1\}$);

в электростатике $K(r) = \frac{1}{r}$ или $\frac{1}{r+a}$,

для задач классификации нет таких ограничений на K.

Метод потенциальных функций = линейный классификатор

Два класса: $Y = \{-1, +1\}$.

$$a(x; X^\ell) = \arg\max_{y \in Y} \Gamma_y(x) = \operatorname{sign}\big(\Gamma_{+1}(x) - \Gamma_{-1}(x)\big) =$$

$$= \operatorname{sign} \sum_{i=1}^{\ell} \gamma_i y_i \, K\left(\frac{\rho(x, x_i)}{h_i}\right).$$

Сравним с линейной моделью классификации:

$$a(x) = \operatorname{sign} \sum_{j=1}^{n} \gamma_j f_j(x).$$

- функции $f_j(x) = y_j K\big(\frac{1}{h_j}\rho(x, x_j)\big)$ — признаки объекта x
- γ_j — веса линейного классификатора
- $n = \ell$ — число признаков равно числу объектов обучения

Задача регрессии и метод наименьших квадратов

- $X = \mathbb{R}^n$ — объекты; $Y = \mathbb{R}$ — ответы;
  $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка;
  $y_i = f(x_i)$, $f : X \to Y$ — неизвестная зависимость;

- $a(x) = g(x, \alpha)$ — параметрическая модель зависимости,
  $\alpha \in \mathbb{R}^p$ — вектор параметров модели.

- Метод наименьших квадратов (МНК):

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} w_i \big(g(x_i, \alpha) - y_i\big)^2 \to \min_\alpha,$$

  где w_i — вес, степень важности i-го объекта.

- **Недостаток:**
  надо иметь хорошую параметрическую модель $g(x, \alpha)$

Непараметрическая регрессия. Формула Надарая–Ватсона

Приближение константой $g(x,\alpha)=\alpha$ в окрестности $x \in X$:

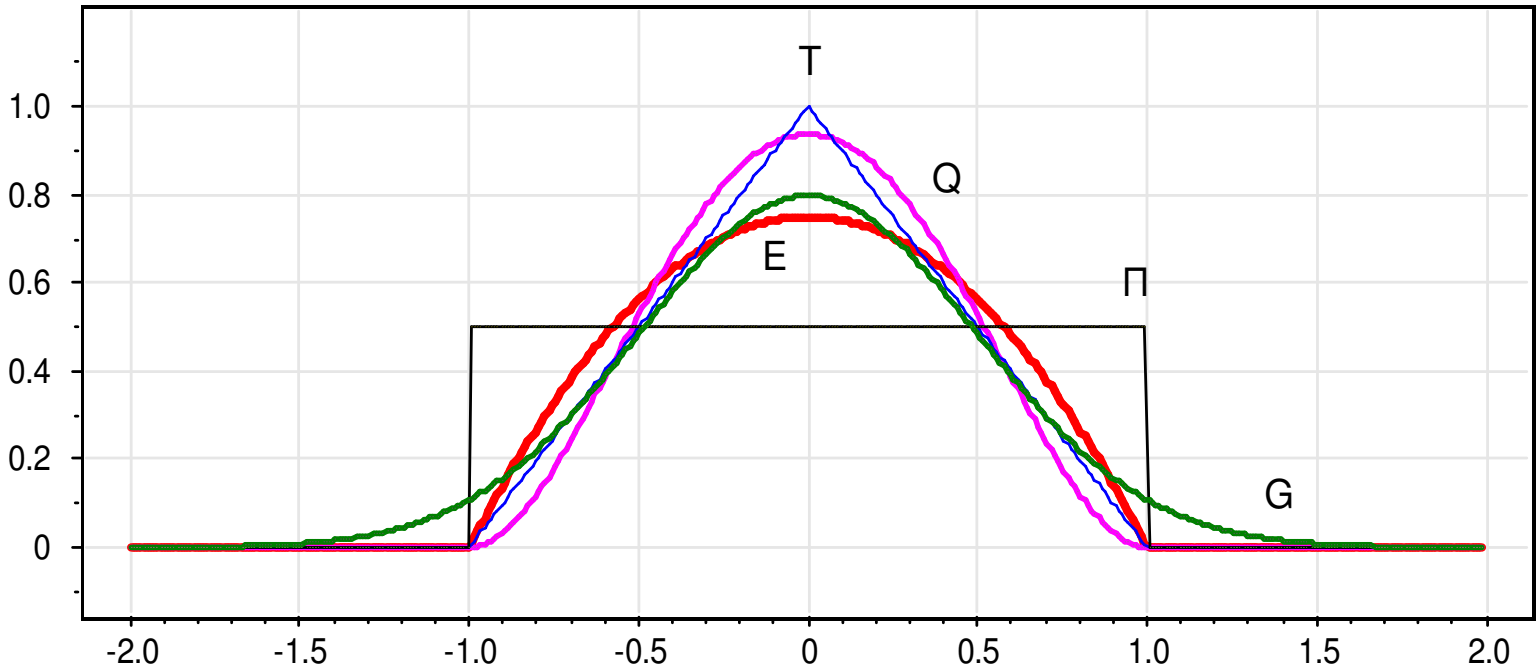
$$Q(\alpha; X^\ell) = \sum_{i=1}^{\ell} w_i(x)\big(\alpha - y_i\big)^2 \to \min_{\alpha\in\mathbb{R}};$$

где $w_i(x) = K\left(\frac{\rho(x,x_i)}{h}\right)$ — веса объектов x_i относительно x;
$K(r)$ — *ядро*, невозрастающее, ограниченное, гладкое;
h — ширина окна сглаживания.

Формула *ядерного сглаживания* Надарая–Ватсона:

$$a_h(x; X^\ell) = \frac{\sum\limits_{i=1}^{\ell} y_i w_i(x)}{\sum\limits_{i=1}^{\ell} w_i(x)} = \frac{\sum\limits_{i=1}^{\ell} y_i K\left(\frac{\rho(x,x_i)}{h}\right)}{\sum\limits_{i=1}^{\ell} K\left(\frac{\rho(x,x_i)}{h}\right)}.$$

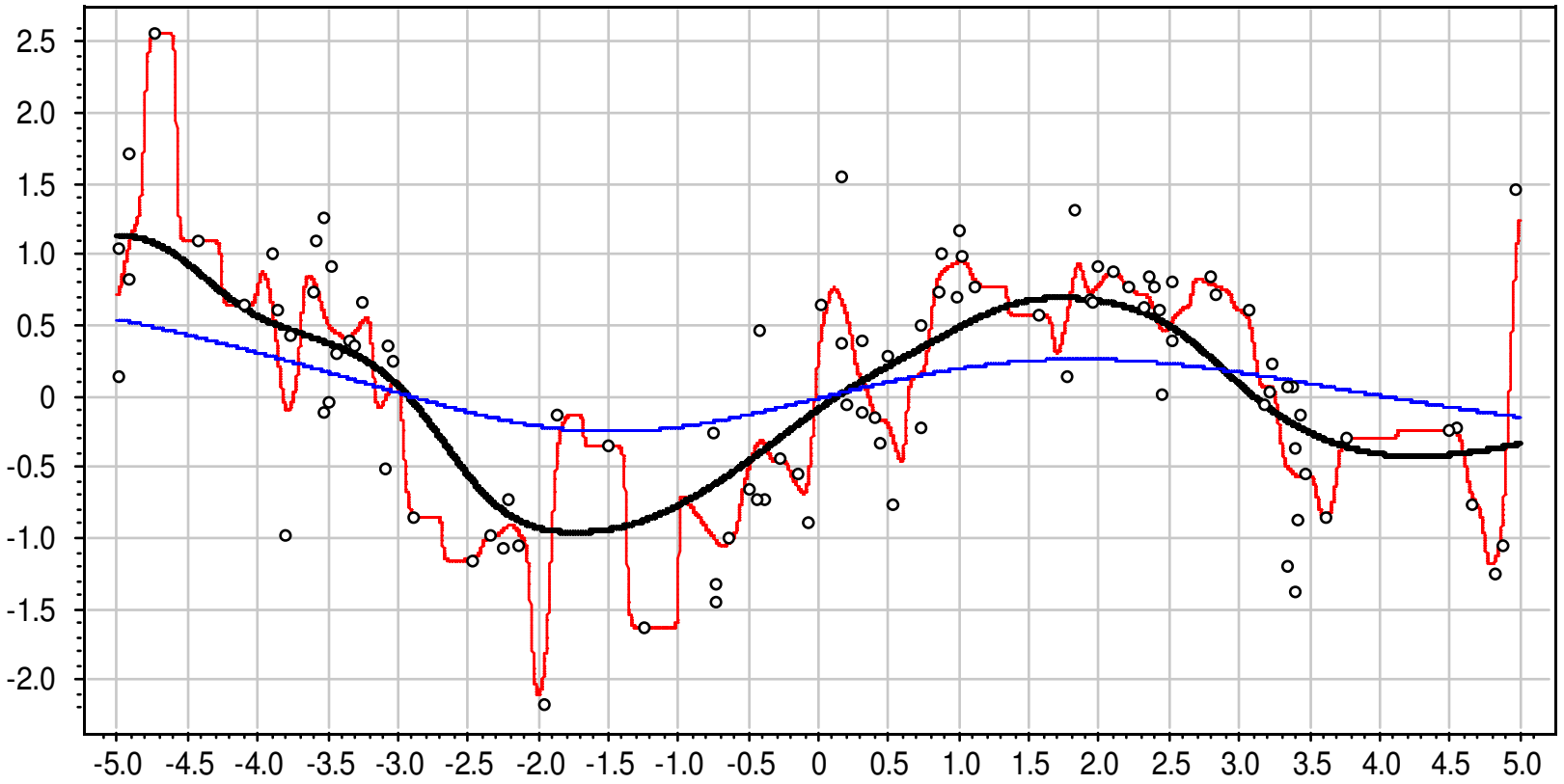
Часто используемые ядра $K(r)$ — те же, что в методе Парзена



$\Pi(r) = \big[|r| \leqslant 1\big]$ — прямоугольное
$T(r) = \big(1 - |r|\big)\big[|r| \leqslant 1\big]$ — треугольное
$E(r) = (1 - r^2)\big[|r| \leqslant 1\big]$ — квадратичное (Епанечникова)
$Q(r) = (1 - r^2)^2\big[|r| \leqslant 1\big]$ — квартическое
$G(r) = \exp\big(-2r^2\big)$ — гауссовское

Выбор ядра K и ширины окна h

$h \in \{0.1, 1.0, 3.0\}$, гауссовское ядро $K(r) = \exp\left(-2r^2\right)$

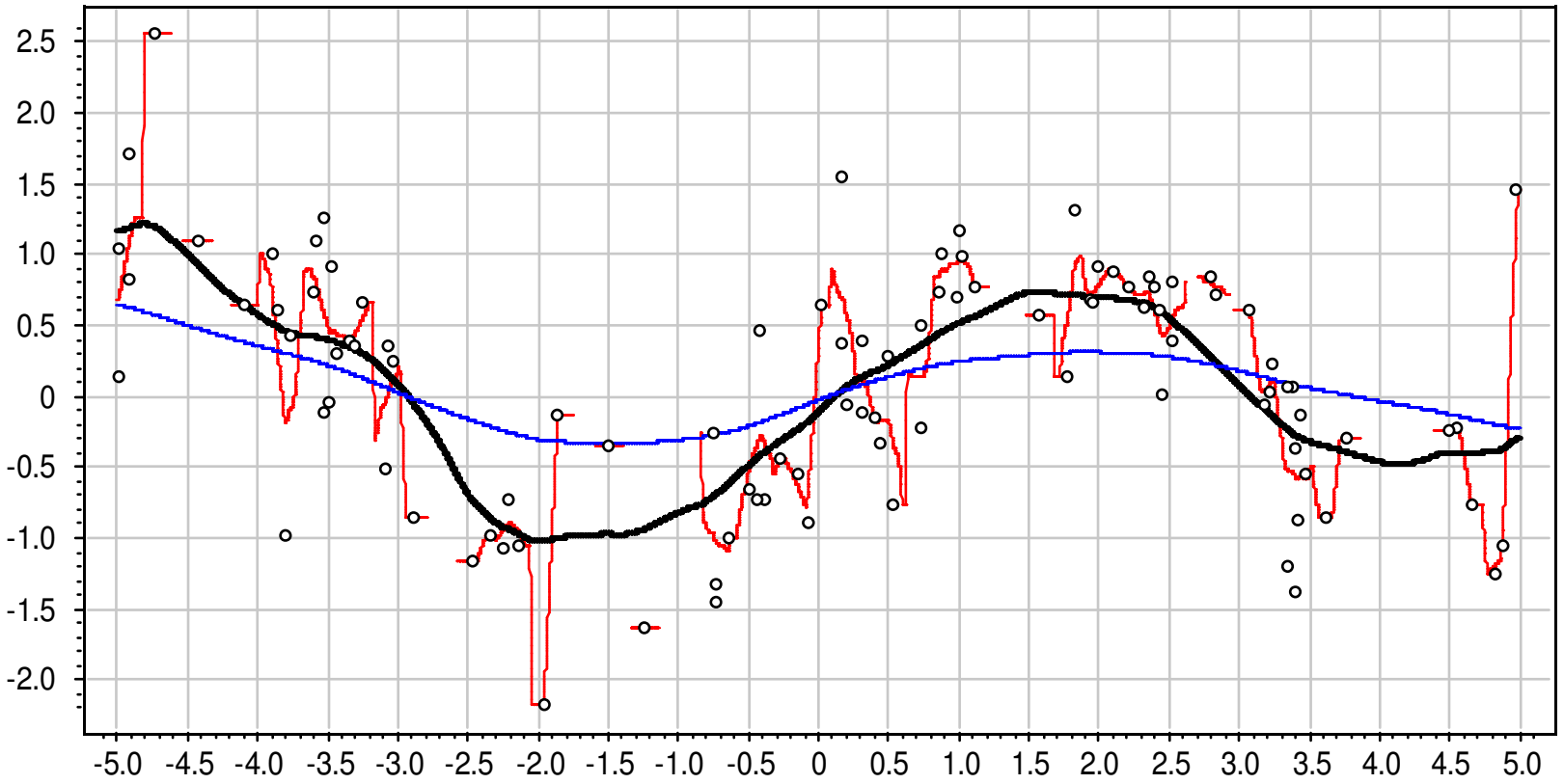


Гауссовское ядро $\Rightarrow$ гладкая аппроксимация

Ширина окна существенно влияет на точность аппроксимации

Выбор ядра K и ширины окна h

$h \in \{0.1, 1.0, 3.0\}$, треугольное ядро $K(r) = \big(1 - |r|\big)\big[|r| \leqslant 1\big]$

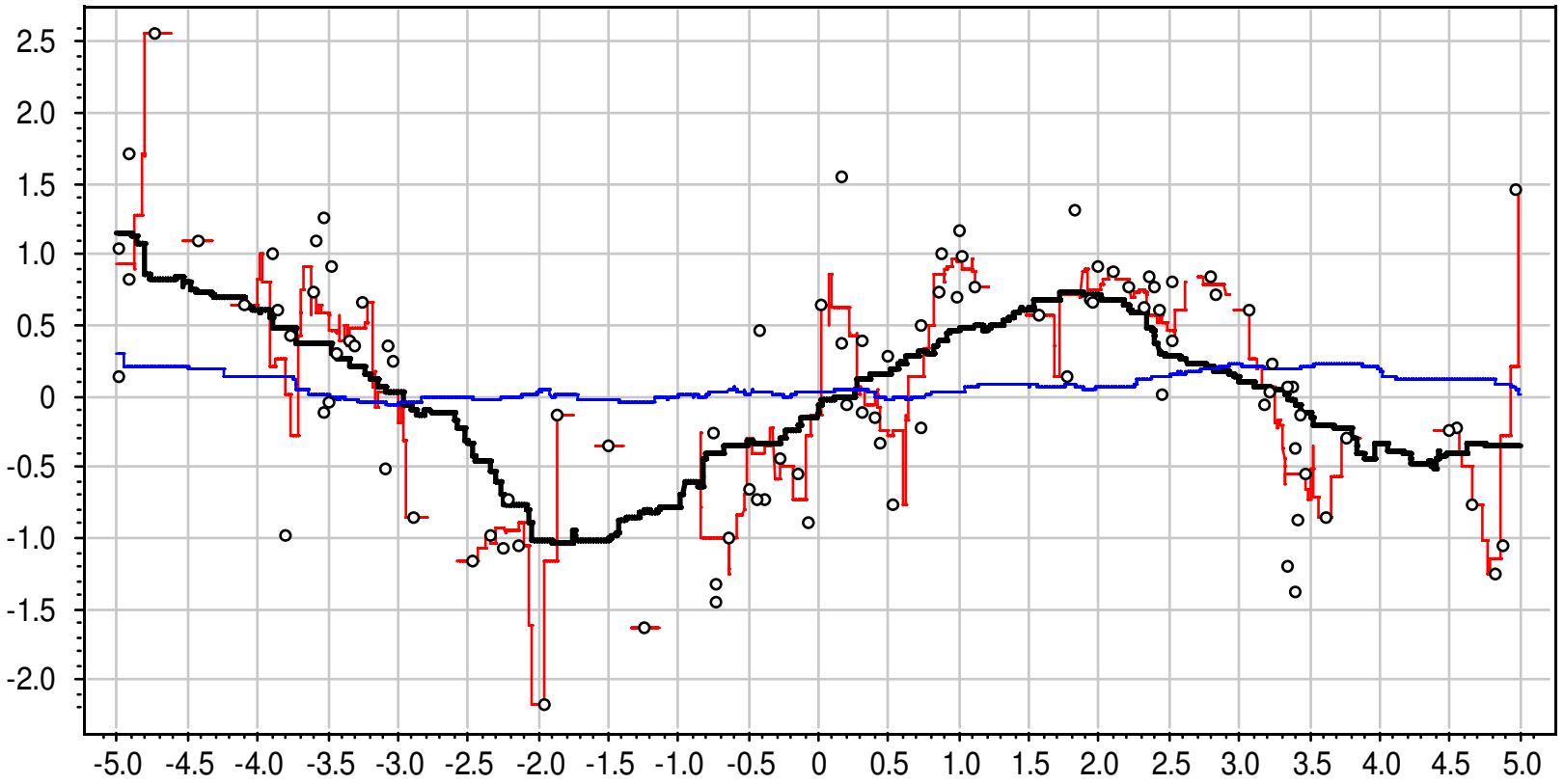


Треугольное ядро $\Rightarrow$ кусочно-линейная аппроксимация

Аппроксимация не определена, если в окне нет точек выборки

Выбор ядра K и ширины окна h

$h \in \{0.1, 1.0, 3.0\}$, прямоугольное ядро $K(r) = [|r| \leqslant 1]$



Прямоугольное ядро $\Rightarrow$ кусочно-постоянная аппроксимация

Выбор ядра слабо влияет на точность аппроксимации

Выбор ядра K и ширины окна h

- Ядро $K(r)$
  — существенно влияет на гладкость функции $a_h(x)$,
  — слабо влияет на качество аппроксимации.
- Ширина окна h
  — существенно влияет на качество аппроксимации.
- Переменная ширина окна $h(x)$ по k ближайшим соседям:

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

  где $h(x) = \rho(x, x^{(k+1)})$, $x^{(k)}$ — k-й сосед объекта x.
- Оптимизация h (или k) по скользящему контролю:

$$\text{LOO}(h, X^\ell) = \sum_{i=1}^{\ell}\Big(a_h\big(x_i; X^\ell \setminus \{x_i\}\big) - y_i\Big)^2 \to \min_h.$$

Постановка задачи кластеризации

Дано:

$X^\ell = \{x_1, \ldots, x_\ell\}$ — обучающая выборка;

$\rho(x_i, x_s)$ — функция расстояния между объектами.

Найти:

$y_i \in Y$ — метки кластеров объектов:

— каждый кластер состоит из близких объектов;

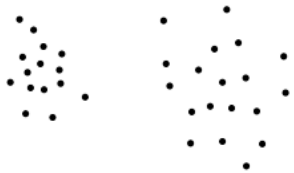
— объекты разных кластеров существенно различны.

Кластеризация — это *обучение без учителя*.

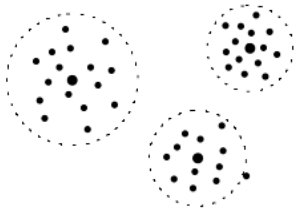
Решение задачи кластеризации принципиально неоднозначно:

- различные критерии качества кластеризации
- различные эвристические методы кластеризации
- различные варианты функции расстоянии ρ

Типы кластерных структур



внутрикластерные расстояния, как правило, меньше межкластерных



кластеры с центром

кластеры могут соединяться перемычками

кластеры могут накладываться на разреженный фон из редко расположенных объектов

Типы кластерных структур

ленточные кластеры

перекрывающиеся кластеры

кластеры могут образовываться не по сходству, а по иным типам регулярностей

кластеры могут вообще отсутствовать

Метод k-средних (k-means)

Объекты x_i задаются векторами признаков $\left(x_i^1, \ldots, x_i^n\right)$.

Вход: X^ℓ — обучающая выборка, параметр k;

Выход: центры кластеров μ_y, $y \in Y$ и кластеризация $y_i \in Y$;

1 начальное приближение центров μ_y, $y \in Y$;

2 **повторять**

3 отнести каждый x_i к ближайшему центру:

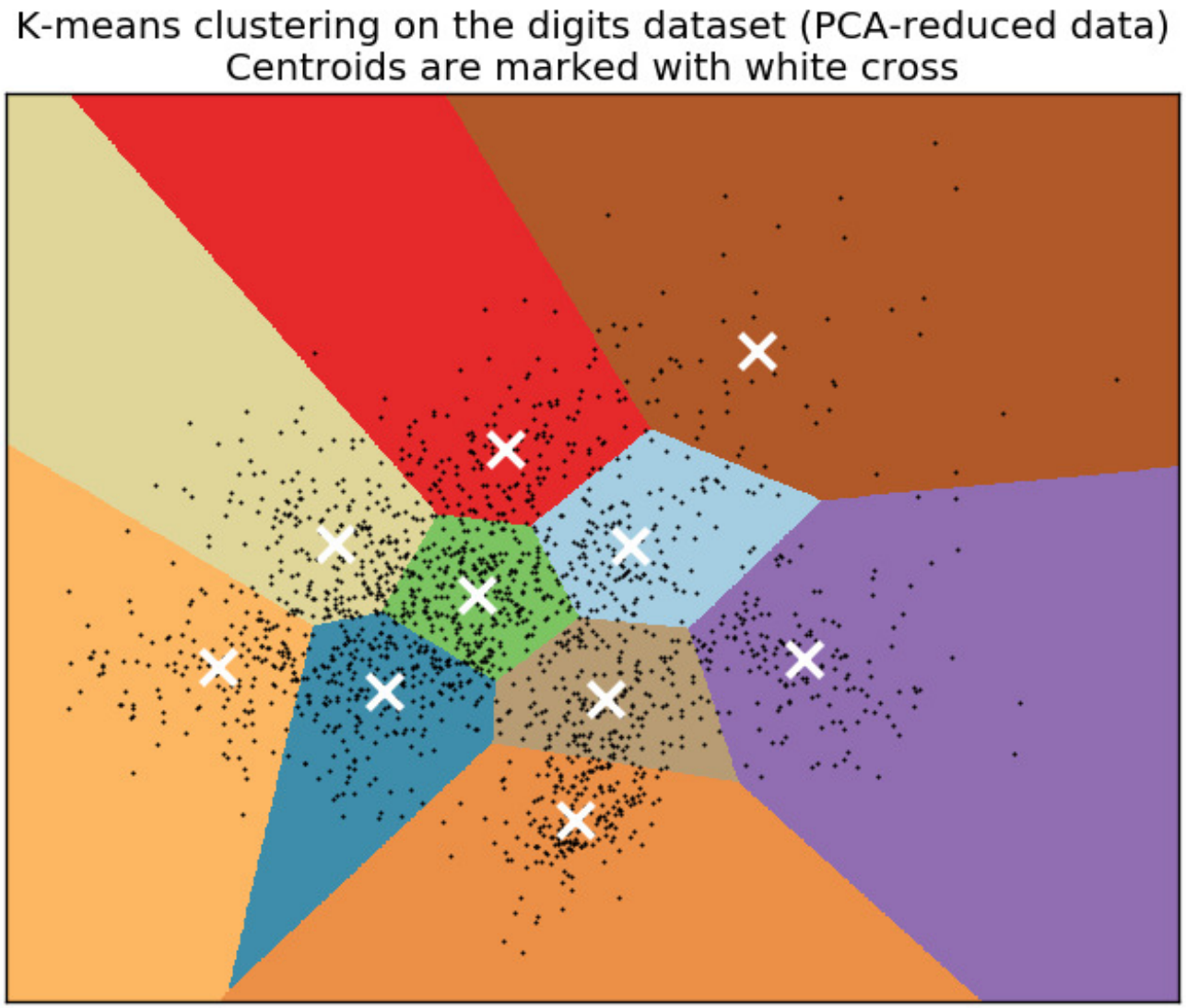
$$y_i := \arg\min_{y \in Y} \rho(x_i, \mu_y), \quad i = 1, \ldots, \ell;$$

4 вычислить новые положения центров:

$$\mu_{yj} := \frac{\sum_{i=1}^{\ell} [y_i = y]\, x_i^j}{\sum_{i=1}^{\ell} [y_i = y]}, \quad y \in Y, \quad j = 1, \ldots, n;$$

5 **пока** y_i не перестанут изменяться;

Пример. Результат кластеризации методом k-средних



Агломеративная иерархическая кластеризация

Алгоритм Ланса-Уильямса [1967] основан на оценивании расстояний $R(U, V)$ между парами кластеров U, V.

сначала все кластеры одноэлементные: $C_1 := \{\{x_1\}, \ldots, \{x_\ell\}\}$;
расстояния между ними: $R(\{x_i\}, \{x_s\}) := \rho(x_i, x_s)$;
для всех $t = 2, \ldots, \ell$ (t — номер итерации):
    найти в C_{t-1} два ближайших кластера:
    $(U, V) := \arg\min_{U \neq V} R(U, V)$;
    $R_t := R(U, V)$;
    слить их в один кластер $W := U \cup V$;
    $C_t := C_{t-1} \cup \{W\} \setminus \{U, V\}$;
    для всех $S \in C_t$
        вычислить $R(W, S)$ по формуле Ланса-Уильямса;

Формула Ланса-Уильямса

Как определить расстояние $R(W, S)$
между кластерами $W = U \cup V$ и S,
зная расстояния $R(U, S)$, $R(V, S)$, $R(U, V)$?

Формула, обобщающая большинство разумных способов определить это расстояние [Ланс, Уильямс, 1967]:

$$\begin{aligned} R(U \cup V, S) = {} & \alpha_U \cdot R(U, S) + \\ & + \alpha_V \cdot R(V, S) + \\ & + \beta \cdot R(U, V) + \\ & + \gamma \cdot \big|R(U, S) - R(V, S)\big|, \end{aligned}$$

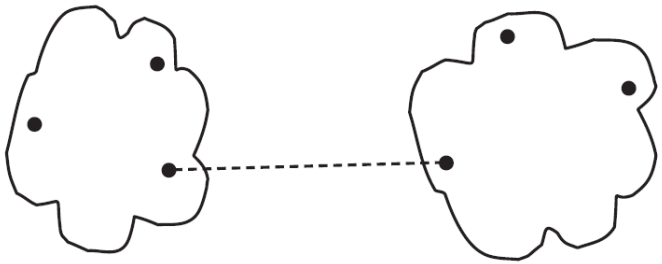
где α_U, α_V, β, γ — числовые параметры.

Частные случаи формулы Ланса-Уильямса

1. Расстояние ближнего соседа:

$$R^{\text{б}}(W,S) = \min_{w\in W, s\in S} \rho(w,s);$$

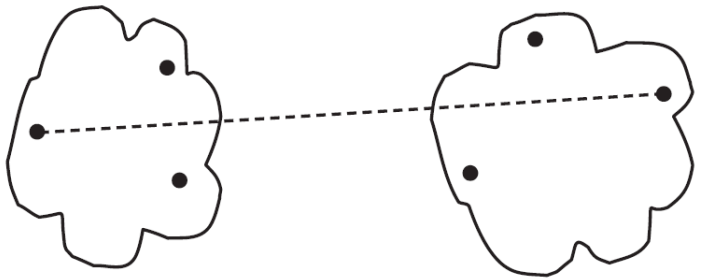
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = -\tfrac{1}{2}.$$



2. Расстояние дальнего соседа:

$$R^{\text{д}}(W,S) = \max_{w\in W, s\in S} \rho(w,s);$$

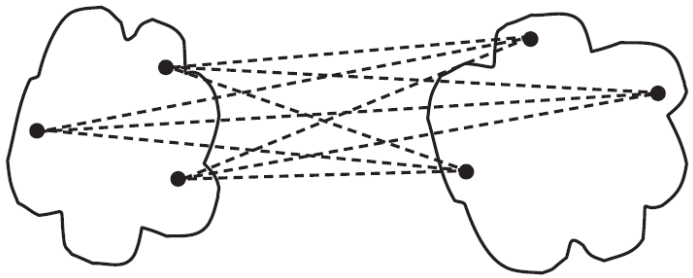
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = \tfrac{1}{2}.$$



3. Групповое среднее расстояние:

$$R^{\text{г}}(W,S) = \frac{1}{|W||S|} \sum_{w\in W} \sum_{s\in S} \rho(w,s);$$

$$\alpha_U = \frac{|U|}{|W|},\ \ \alpha_V = \frac{|V|}{|W|},\ \ \beta = \gamma = 0.$$



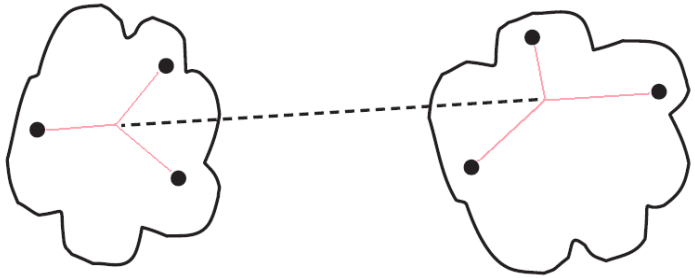
Частные случаи формулы Ланса-Уильямса

4. Расстояние между центрами:

$$R^{ц}(W,S) = \rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|U|}{|W|}, \quad \alpha_V = \frac{|V|}{|W|},$$

$$\beta = -\alpha_U\alpha_V, \quad \gamma = 0.$$



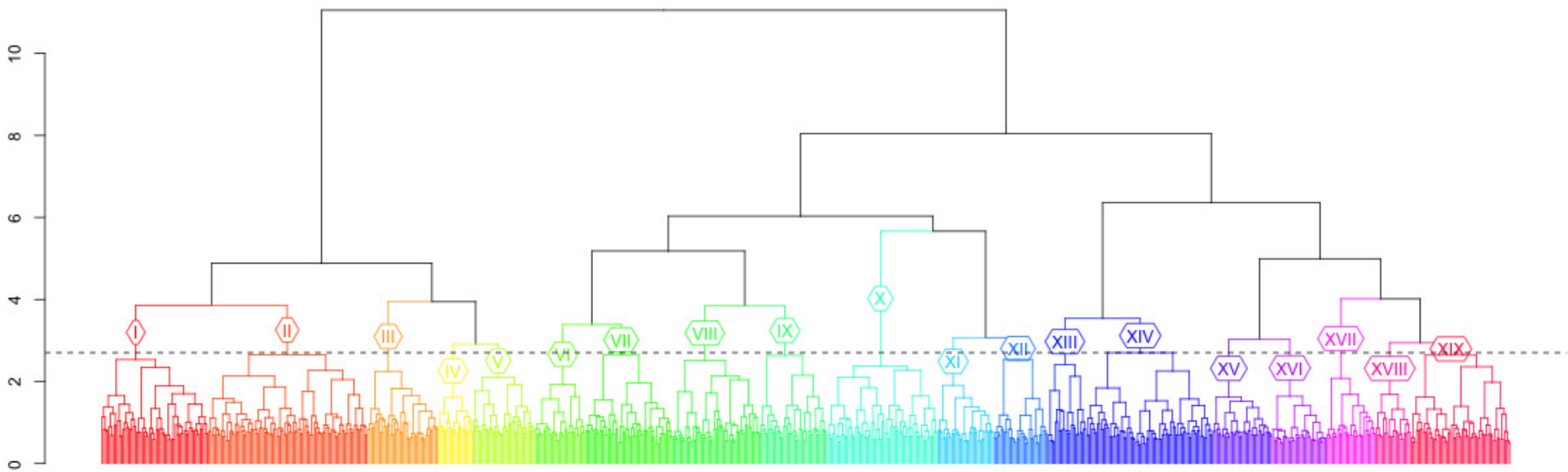
5. Расстояние Уорда — рекомендуется для использования

$$R^{у}(W,S) = \frac{|S||W|}{|S|+|W|}\,\rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

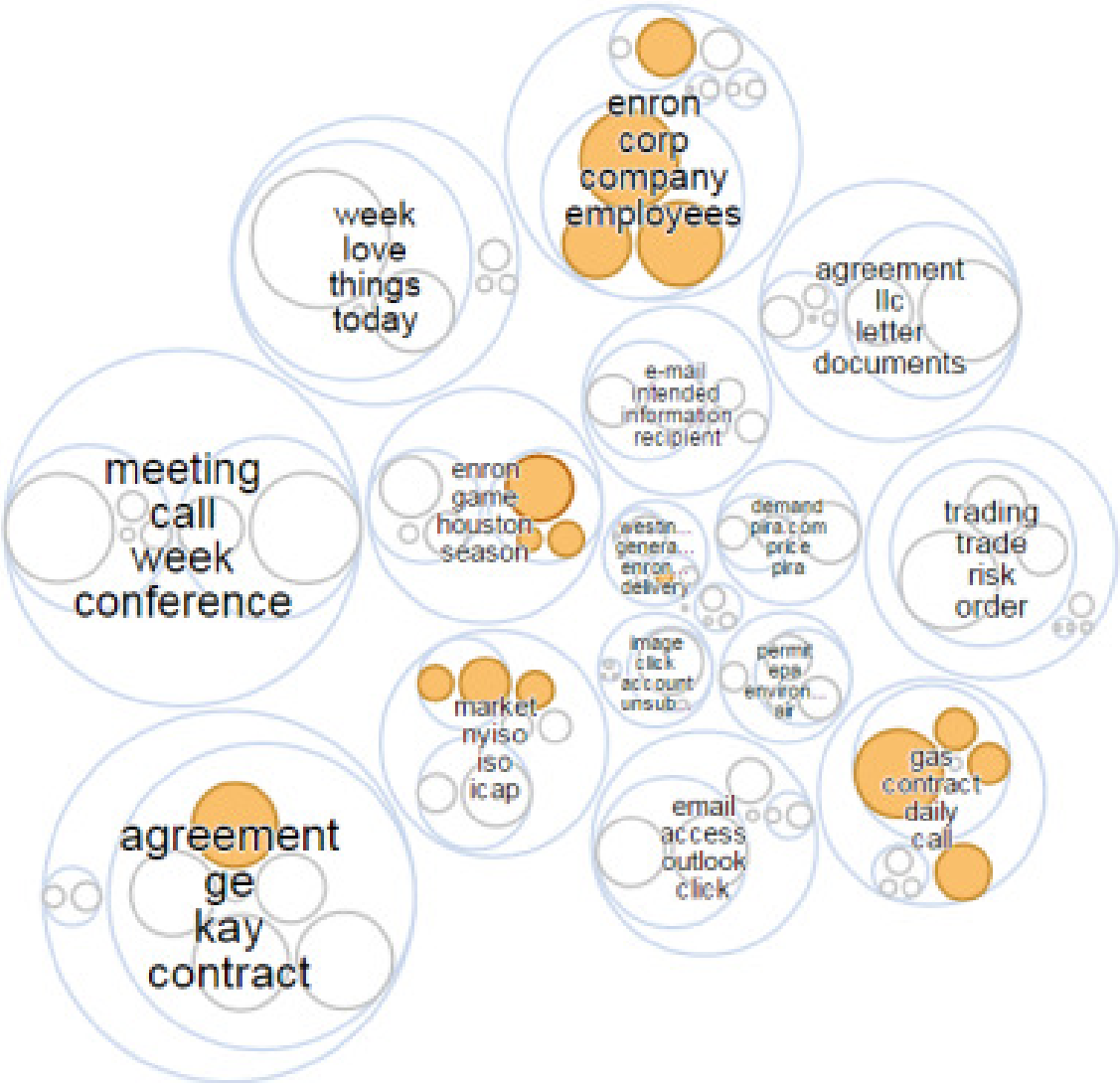
$$\alpha_U = \frac{|S|+|U|}{|S|+|W|}, \quad \alpha_V = \frac{|S|+|V|}{|S|+|W|}, \quad \beta = \frac{-|S|}{|S|+|W|}, \quad \gamma = 0.$$

Дендрограмма — визуализация иерархической кластеризации

- По вертикальной оси откладываются расстояния R_t
- Уровень отсечения определяет число кластеров
- Дендрограмма не имеет самопересечений и группирует объекты в кластеры вдоль горизонтальной оси



Ещё один способ визуализации иерархической кластеризации



Резюме

- Функции расстояния полезны для решения задач классификации, регрессии, кластеризации.
- Ширину окна h оптимизируют по критерию LOO:
  — при $h \to 0$ модель усложняется, начинает настраиваться на шум и переобучается;
  — при $h \to \infty$ модель упрощается, начинает стремиться к константе и недообучается;
  — «золотую середину» находят перебором h по сетке.
- Кластеризация неоднозначна, результат может зависеть от метрики, от критерия, от метода
- Кластеризация используется для выделения структурных особенностей выборки объектов

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov

Если что-то было не понятно,
не стесняйтесь подходить и спрашивать :)