• Введение в машинное обучение •

Метрические методы машинного обучения

Воронцов Константин Вячеславович
k.v.vorontsov@phystech.edu
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Введение в машинное обучение (курс лекций, К.В.Воронцов)»

МФТИ.ФПМИ.ИС.ИАД • 9 апреля 2026

Основные научные школы машинного обучения

1. *символизм* – поиск логических закономерностей
   - Decision Tree, Rule Induction
2. *коннекционизм* – обучаемые нейронные сети
   - BackPropagation, Deep Belief Nets, Deep Learning
     CNN, ResNet, LSTM, GRU, Attention, Transformer
3. *эволюционизм* – саморазвитие сложных моделей
   - Genetic Algorithms, Genetic Programming, Symbolic Regression
4. *байесионизм* и вероятностно-статистические методы
   - MLE, EM, GLM, LR, OBC, Naive Bayes, QD, LDF
     Bayesian Networks, Bayesian Learning, Graphical Models
5. *аналогизм* – «близким объектам близкие ответы»
   - kNN, RBF, SVM, KDE, Kernel Smoothing

⊕ *композиционизм* – кооперация моделей
   - Weighted Voting, Boosting, Bagging, Stacking,
     Random Forest, Яндекс.CatBoost

Педро Домингос. Верховный алгоритм. 2016. 336 с.

Содержание

Гипотезы непрерывности и компактности

Задачи классификации и регрессии:

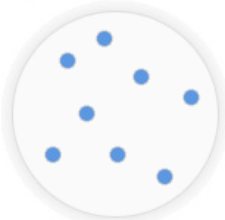
X — объекты, Y — ответы;

$X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка;

Гипотеза непрерывности (для регрессии):

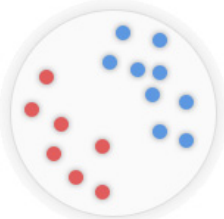
близким объектам соответствуют близкие ответы.

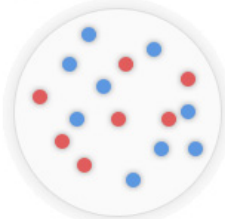
выполнена: 

не выполнена: 

Гипотеза «компактности» (для классификации):

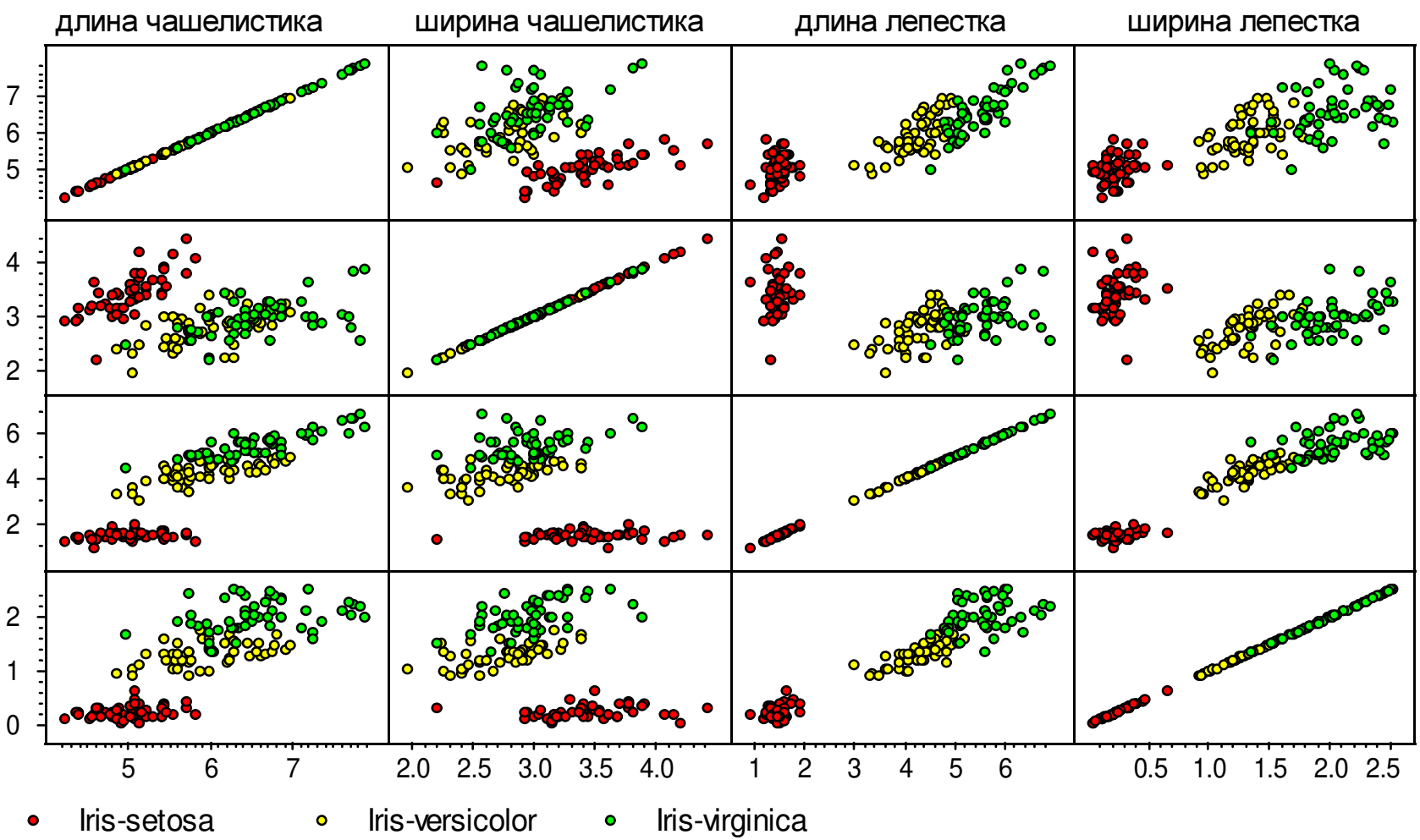
близкие объекты, как правило, лежат в одном классе.

выполнена: 

не выполнена: 

Пример: задача классификации цветков ириса [Фишер, 1936]

Классы — компактные сгустки точек (3 класса по 50 объектов)



Формализация понятия «расстояние» (distance)

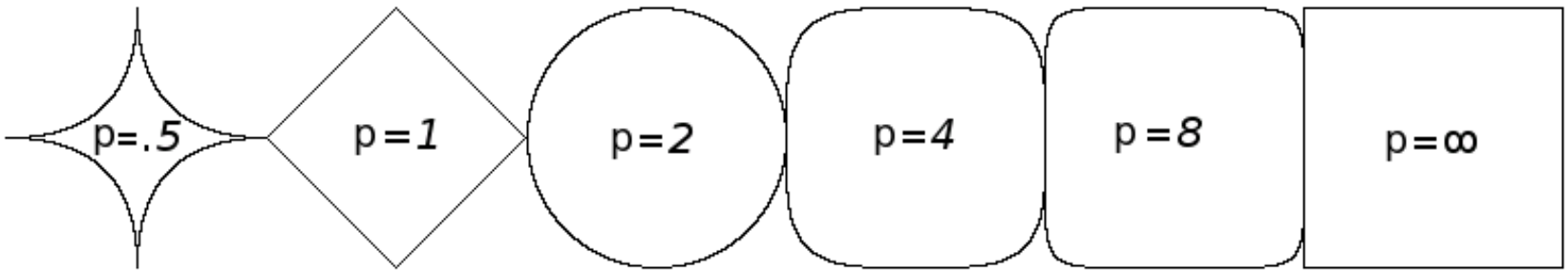
Евклидова метрика и *обобщённая метрика Минковского*:

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

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

w_j — веса признаков (возможно, обучаемые) играют две роли:
— нормировка, приведение к общему масштабу
— подавление неинформативных (мешающих) признаков

Линии уровня (*эквидистантные поверхности*) при различных p:

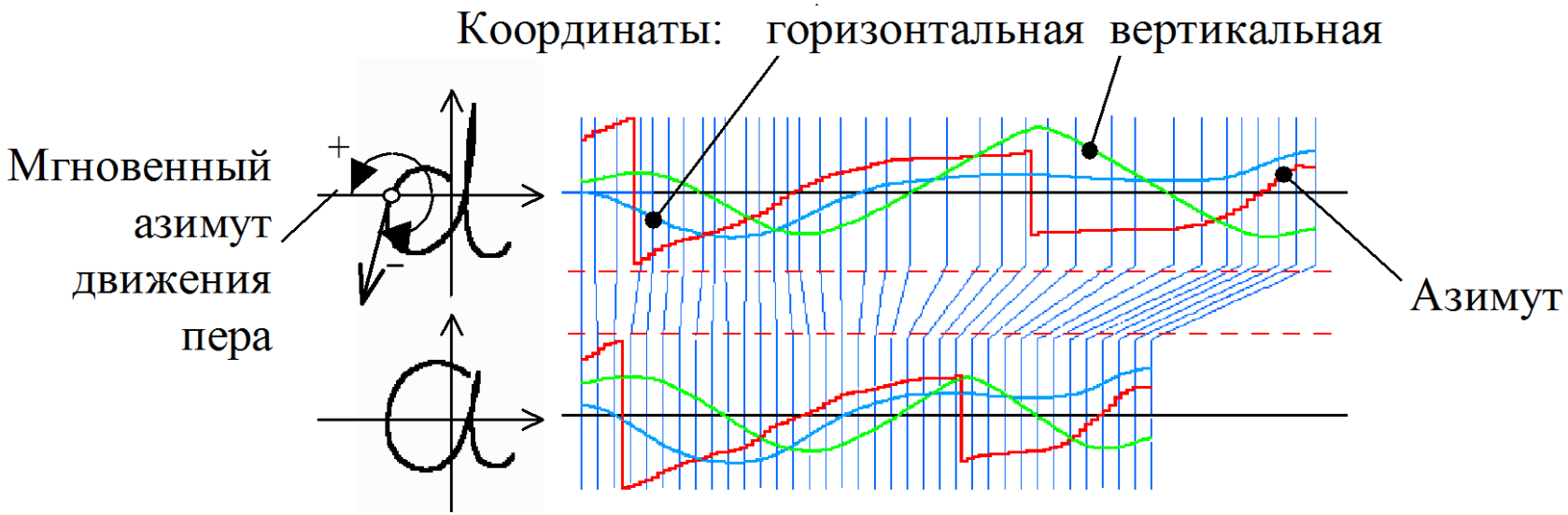


Расстояния между строками / сигналами

Для строк — редакторское расстояние Левенштейна:

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

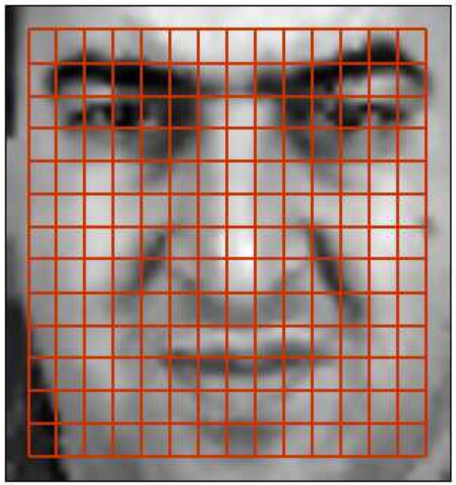
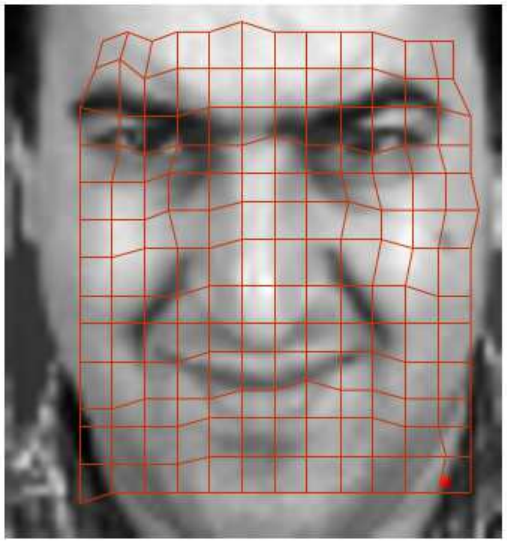
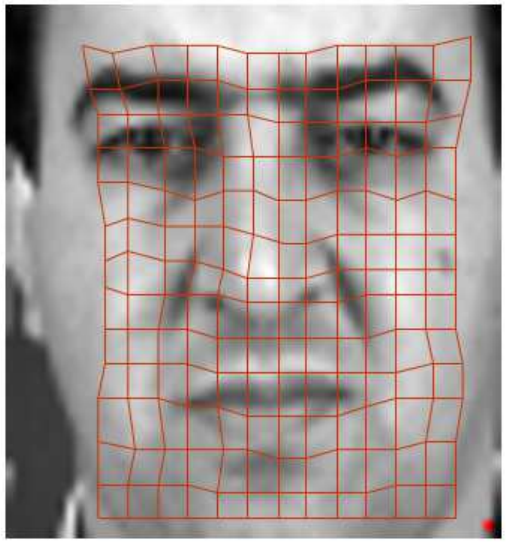
Для сигналов — энергия сжатий и растяжений:



В.В.Моттль, О.С.Середин, В.В.Сулимова. Потенциальные функции для беспризнакового восстановления зависимостей на множествах сигналов и символьных последовательностей. Искусственный интеллект. 2004.

Расстояния между изображениями

Расстояние **между изображениями** на основе выравнивания:

Оценивается энергия растяжения прямоугольной сетки

V.Mottl, A.Kopylov, A.Kostin, A.Yermakov, J.Kittler. Elastic transformation of the image pixel grid for similarity based face identification. ICPR-2002

Задача обучения метрического классификатора

Дано: $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка, $|Y| < \infty$

Модель: *метрический классификатор*, относящий объект x к тому классу, которому принадлежат его ближайшие соседи:

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

$S(x, x_i)$ — *функция близости* (similarity) пары объектов x и x_i

$\Gamma_y(x)$ — *оценка близости* объекта x к объектам класса y

Найти: параметры функции близости $S(x, x_i)$

Критерий: минимум эмпирического риска

$r(x, x_i)$ — ранг объекта $x_i = x^{(r)}$ в ранжированной выборке X^ℓ:
$\rho(x, x^{(1)}) \leqslant \rho(x, x^{(2)}) \leqslant \cdots \leqslant \rho(x, x^{(\ell)})$

Варианты параметризации функции близости

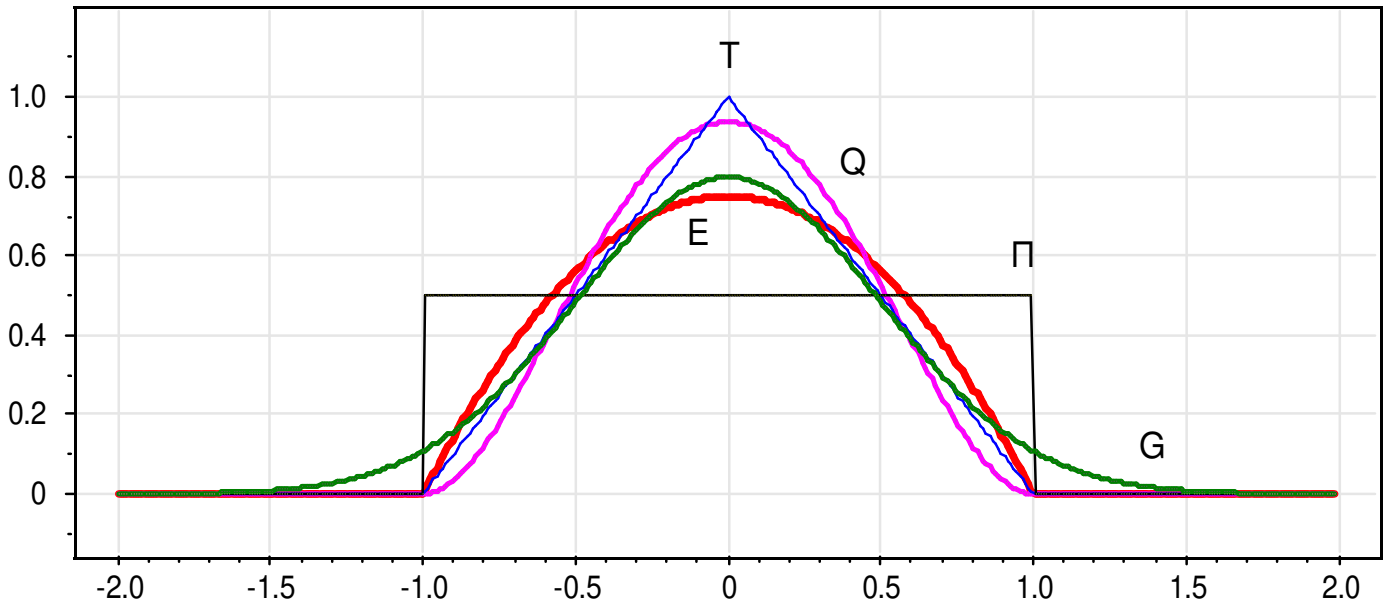
$S(x, x_i) = [r(x, x_i) = 1]$ — метод *ближайшего соседа* (1NN)
$S(x, x_i) = [r(x, x_i) \leqslant k]$ — метод k *ближайших соседей* (kNN)
$S(x, x_i) = [r(x, x_i) \leqslant k] w_i$ — метод *взвешенных kNN*

$S(x, x_i) = K\left(\frac{1}{h}\rho(x, x_i)\right)$ — метод *окна Парзена* (Parzen Window),
$K(r)$ — *ядро* (kernel), не возрастает и положительно на $[0, 1]$
h — фиксированная *ширина окна* (bandwidth)

$S(x, x_i) = K\left(\frac{1}{h(x)}\rho(x, x_i)\right)$ — метод окна Парзена с переменной шириной окна по k соседям $h(x) = \rho(x, x^{(k+1)})$

$S(x, x_i) = \alpha_i K\left(\frac{1}{h_i}\rho(x, x_i)\right)$ — метод *потенциальных функций*,
каждому x_i соответствуют два обучаемых параметра:
α_i — вес или «сила» потенциала с центром в точке x_i
h_i — «радиус действия» потенциала с центром в точке x_i

Часто используемые ядра $K(r)$



$$\left.\begin{aligned}
&\Pi(r) = \big[|r| \leqslant 1\big] \text{ — прямоугольное}\\
&T(r) = \big(1-|r|\big)\big[|r| \leqslant 1\big] \text{ — треугольное}\\
&E(r) = \big(1-r^2\big)\big[|r| \leqslant 1\big] \text{ — квадратичное}\\
&Q(r) = \big(1-r^2\big)^2\big[|r| \leqslant 1\big] \text{ — квартическое}
\end{aligned}\right\} \text{ — финитные ядра}$$

$G(r) = \exp\big(-2r^2\big)$ — гауссовское — нефинитное ядро

Научная школа М. А. Айзермана

- *Гипотеза компактности*: схожие объекты, как правило, находятся в одном классе
- Идея *метода потенциальных функций* заимствуется из физики (электростатики)
- Линейная модель классификации с ℓ признаками $f_j(x) = K\left(\frac{1}{h_j}\rho(x, x_j)\right)$:

$$a(x; X^\ell) = \arg\max_{y \in Y} \sum_{j=1}^{\ell} [y_j = y]\, \alpha_j f_j(x),$$

где α_j, h_j — обучаемые параметры модели, $j = 1, \ldots, \ell$

Марк Аронович Айзерман (1913–1992)

Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Теоретические основы метода потенциальных функций в задаче об обучении автоматов разделению входных ситуаций на классы. 1964.
Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Метод потенциальных функций в теории обучения машин. 1970.
Аркадьев А. Г., Браверман Э. М. Обучение машин распознаванию образов. 1964.

Влияние ширины окна h на вид разделяющей поверхности

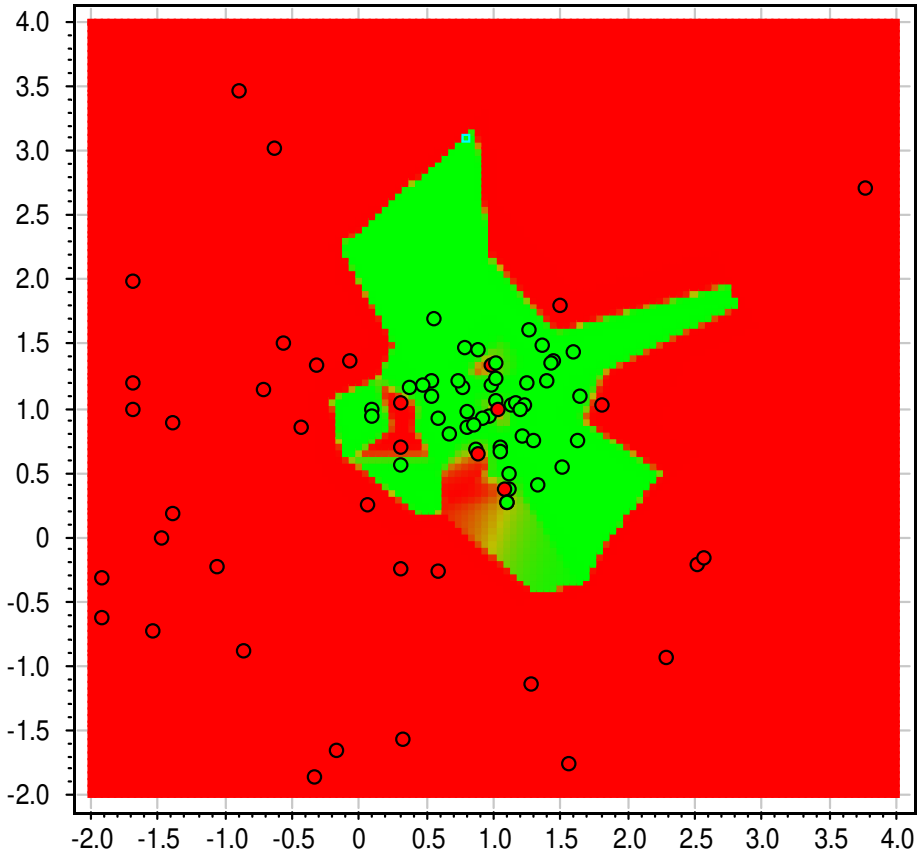
Пример: $x_i \in \mathbb{R}^2$, $y_i \in \{-1, +1\}$, гауссовское $K(r) = \exp(-2r^2)$

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

ширина окна: $h = 0.05$

чем меньше ширина окна h,
тем сложнее форма границы,
возможно переобучение

чем больше ширина окна h,
тем проще форма границы,
возможно недообучение



Влияние ширины окна h на вид разделяющей поверхности

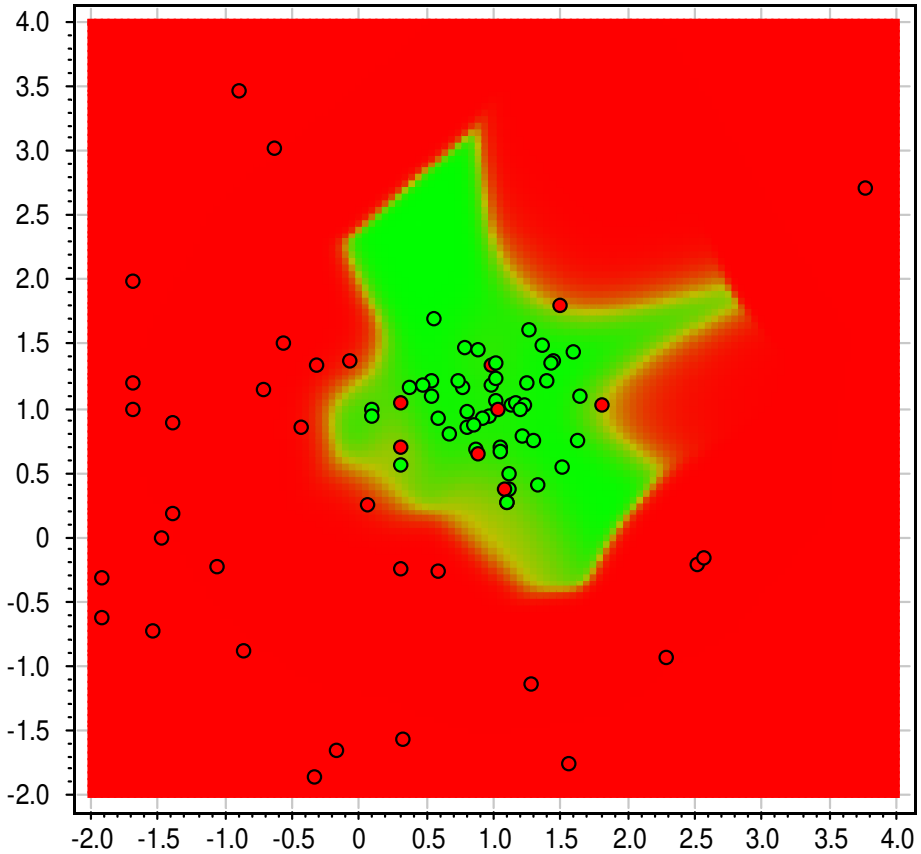
Пример: $x_i \in \mathbb{R}^2$, $y_i \in \{-1, +1\}$, гауссовское $K(r) = \exp(-2r^2)$

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

ширина окна: $h = 0.2$

чем меньше ширина окна h,
тем сложнее форма границы,
возможно переобучение

чем больше ширина окна h,
тем проще форма границы,
возможно недообучение



Влияние ширины окна h на вид разделяющей поверхности

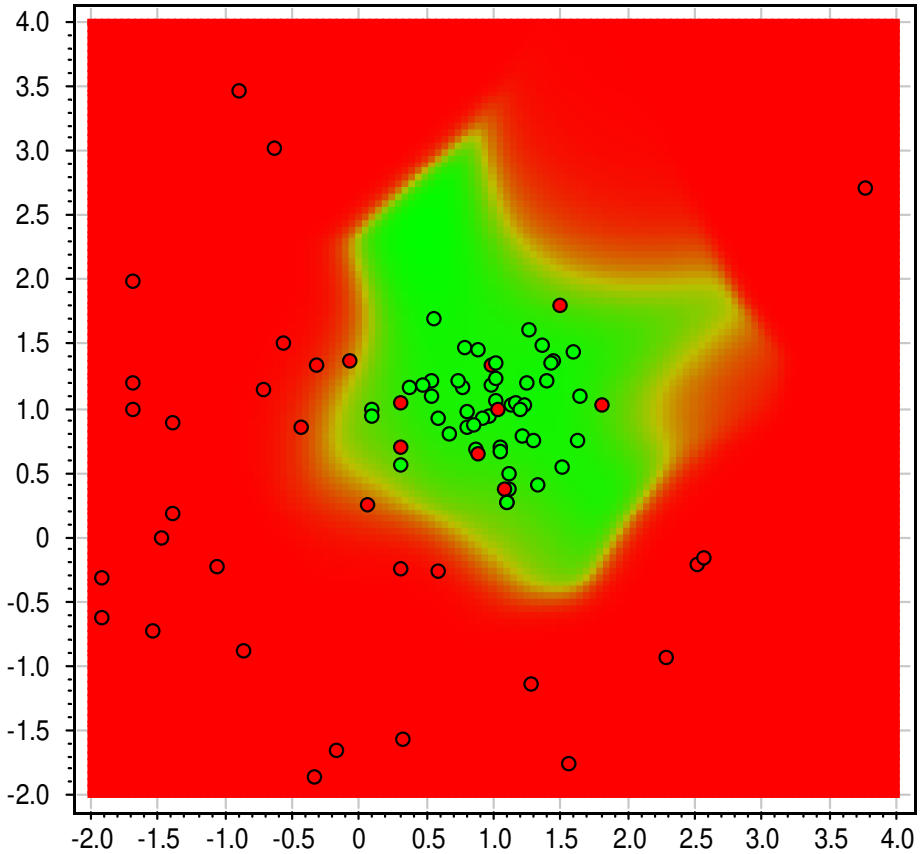
Пример: $x_i \in \mathbb{R}^2$, $y_i \in \{-1, +1\}$, гауссовское $K(r) = \exp(-2r^2)$

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

ширина окна: $h = 0.3$

чем меньше ширина окна h,
тем сложнее форма границы,
возможно переобучение

чем больше ширина окна h,
тем проще форма границы,
возможно недообучение



Влияние ширины окна h на вид разделяющей поверхности

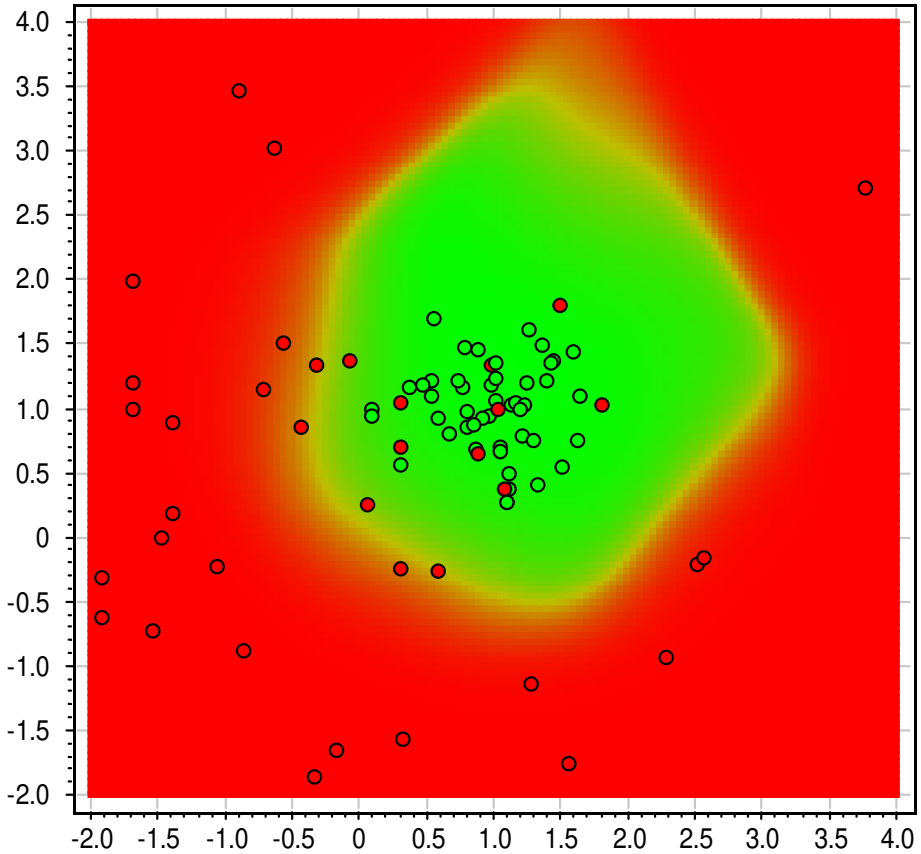
Пример: $x_i \in \mathbb{R}^2$, $y_i \in \{-1, +1\}$, гауссовское $K(r) = \exp(-2r^2)$

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

ширина окна: $h = 0.5$

чем меньше ширина окна h,
тем сложнее форма границы,
возможно переобучение

чем больше ширина окна h,
тем проще форма границы,
возможно недообучение



Влияние ширины окна h на вид разделяющей поверхности

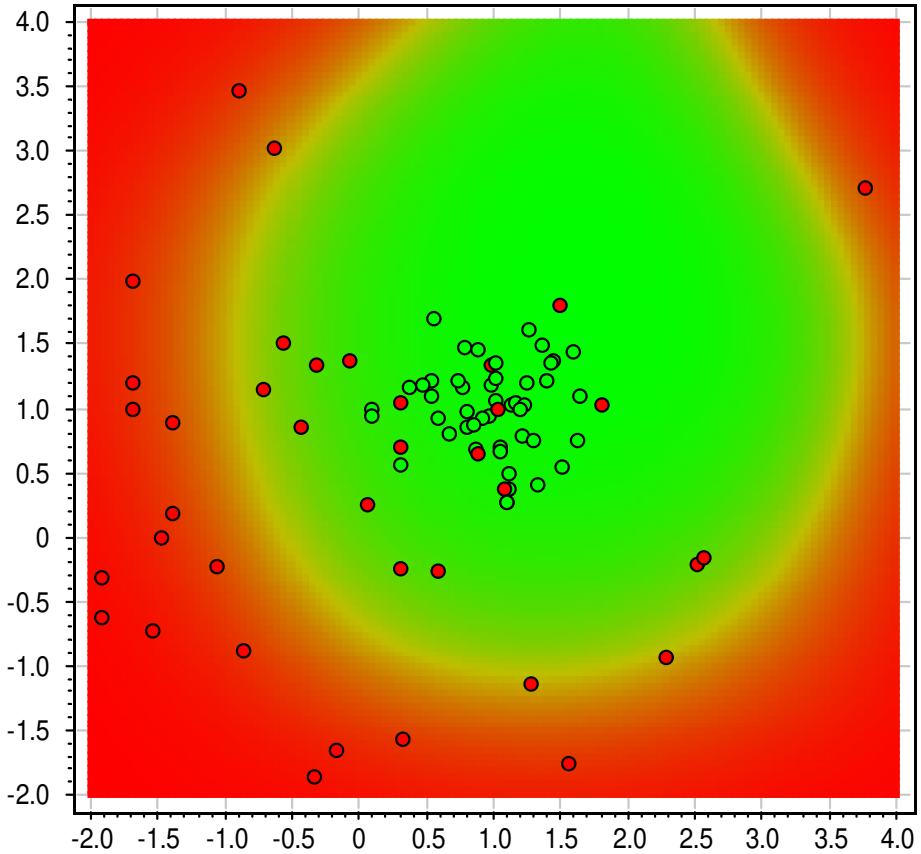
Пример: $x_i \in \mathbb{R}^2$, $y_i \in \{-1, +1\}$, гауссовское $K(r) = \exp(-2r^2)$

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

ширина окна: $h = 1.0$

чем меньше ширина окна h,
тем сложнее форма границы,
возможно переобучение

чем больше ширина окна h,
тем проще форма границы,
возможно недообучение



Влияние ширины окна h на вид разделяющей поверхности

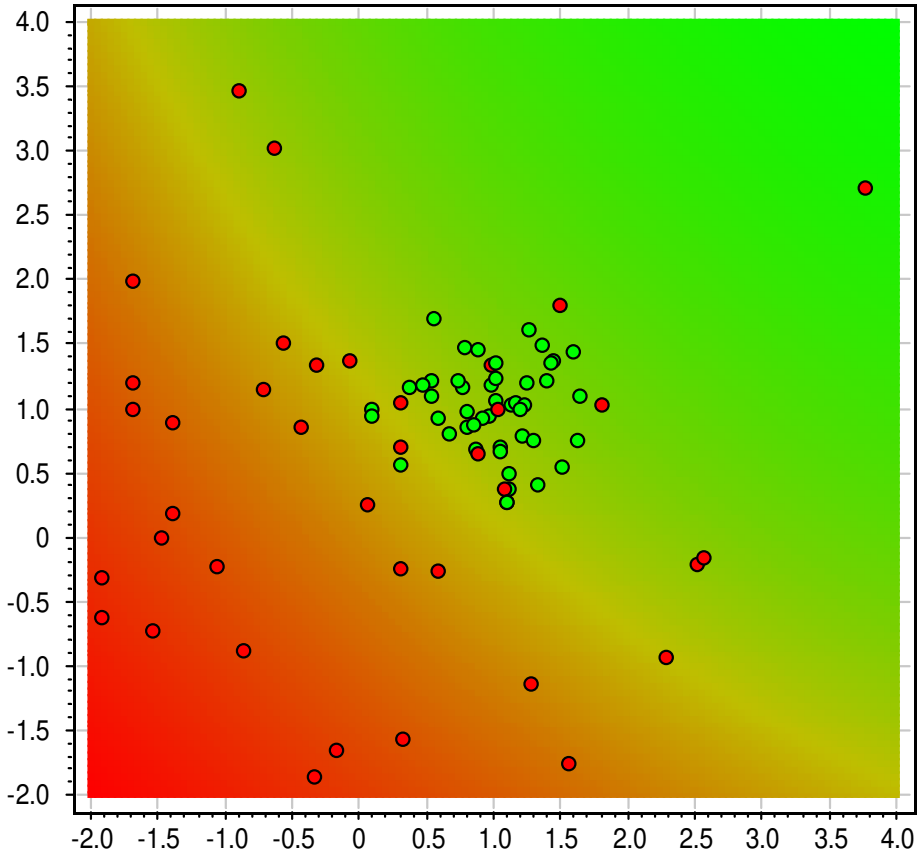
Пример: $x_i \in \mathbb{R}^2$, $y_i \in \{-1, +1\}$, гауссовское $K(r) = \exp(-2r^2)$

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

ширина окна: $h = 5.0$

чем меньше ширина окна h,
тем сложнее форма границы,
возможно переобучение

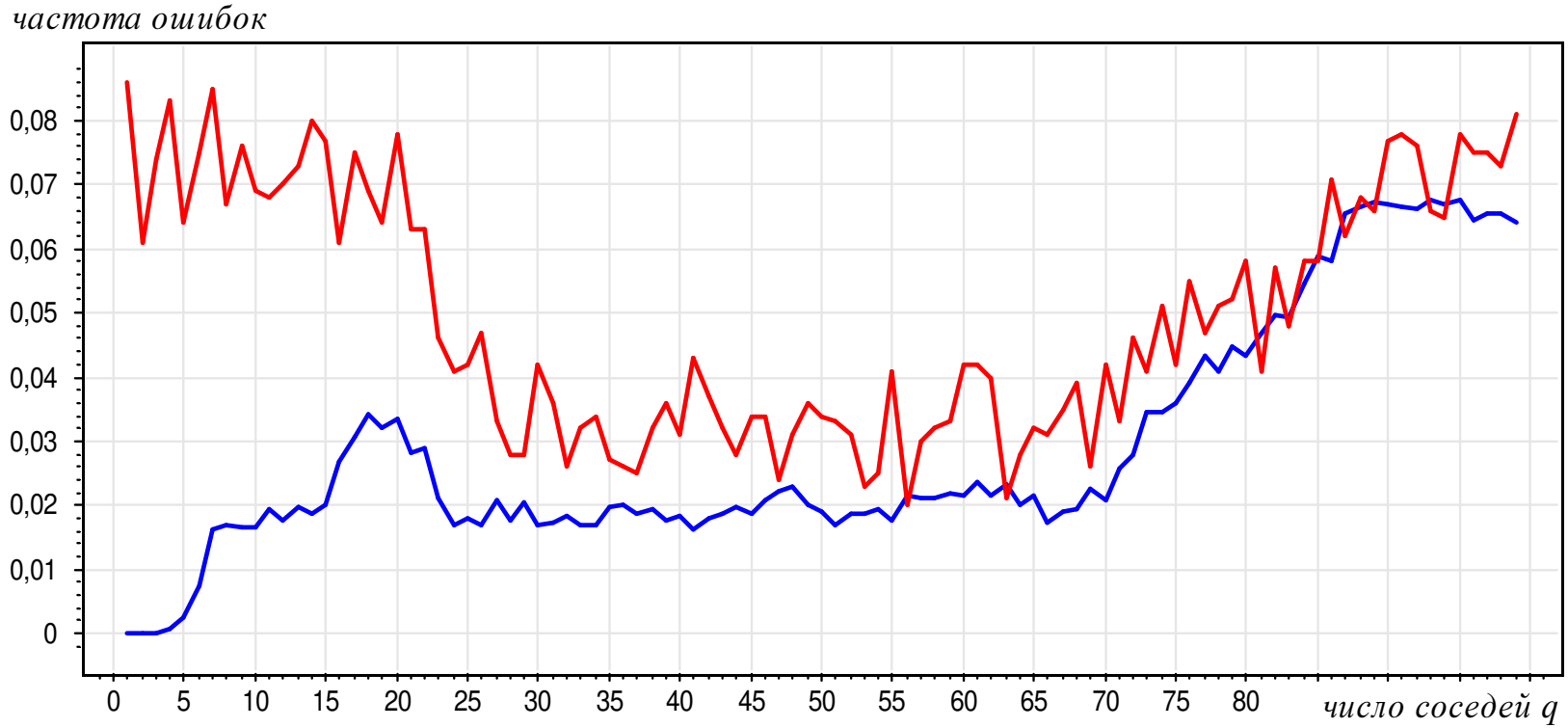
чем больше ширина окна h,
тем проще форма границы,
возможно недообучение



Влияние ширины окна h на качество классификации

Пример.
Задача UCI: Iris.

$$\mathrm{LOO}(h) = \sum_{i=1}^{\ell} \left[a_h\left(x_i; X^\ell \setminus \{x_i\}\right) \neq y_i \right] \to \min_h$$



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

— несмещённое число ошибок LOO

Выбор ядра K и ширины окна h

- Ядро $K(r)$
  — влияет на гладкость разделяющей поверхности
  — почти не влияет на качество классификации
- Ширина окна h
  — существенно влияет на качество классификации
- Переменная ширина окна по k ближайшим соседям:

$$S(x, x_i) = K\left(\frac{\rho(x, x_i)}{h(x)}\right), \qquad h(x) = \rho(x, x^{(k+1)})$$

  где $x^{(k)}$ — k-й сосед объекта x.
- Оптимизация ширины окна (h или k) по leave-one-out:

$$\mathrm{LOO}(h, X^\ell) = \sum_{i=1}^{\ell} \left[a_h\left(x_i; X^\ell \setminus \{x_i\}\right) \neq y_i\right] \to \min_h$$

Задачи регрессии и метод наименьших квадратов

Дано: $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка, $y_i \in \mathbb{R}$

Найти регрессионную модель для аппроксимации $y(x)$:

$$a(x) = f(x, \theta)$$

где θ — вектор параметров, f — фиксированная функция

Критерий — метод наименьших квадратов:

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

где w_i — весовой коэффициент, степень важности объекта x_i

Мотивация перехода к непараметрическим моделям:
когда нет теории, дающей «физичную» модель $f(x, \theta)$

Непараметрическая регрессия, формула Надарая–Ватсона

Приближение константой $f(x,\theta)=\theta$ в окрестности точки x:

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

где $w_i(x)=K\big(\frac{1}{h}\rho(x,x_i)\big)$ — веса объектов x_i в окрестности x;
$K(r)$ — *ядро* (kernel), невозрастающее, ограниченное, гладкое;
h — ширина окна сглаживания (bandwidth).

Формула *ядерного сглаживания* Надарая–Ватсона:

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

Обоснование формулы Надарая–Ватсона (одномерный случай)

Теорема

Пусть выполнены следующие условия:

1) *выборка* $X^\ell = (x_i, y_i)_{i=1}^\ell$ *простая, из распределения* $p(x, y)$;
2) *ядро* $K(r)$ *ограничено:* $\int_0^\infty K(r)\,dr < \infty$, $\lim\limits_{r\to\infty} rK(r) = 0$;
3) *зависимость* $\mathsf{E}(y|x)$ *не имеет вертикальных асимптот:*
$\mathsf{E}(y^2|x) = \int_Y y^2 p(y|x)\,dy < \infty$ *при любом* $x \in X$;
4) *последовательность* h_ℓ *убывает, но не слишком быстро:*
$\lim\limits_{\ell\to\infty} h_\ell = 0$, $\lim\limits_{\ell\to\infty} \ell h_\ell = \infty$.

Тогда имеет место сходимость по вероятности:

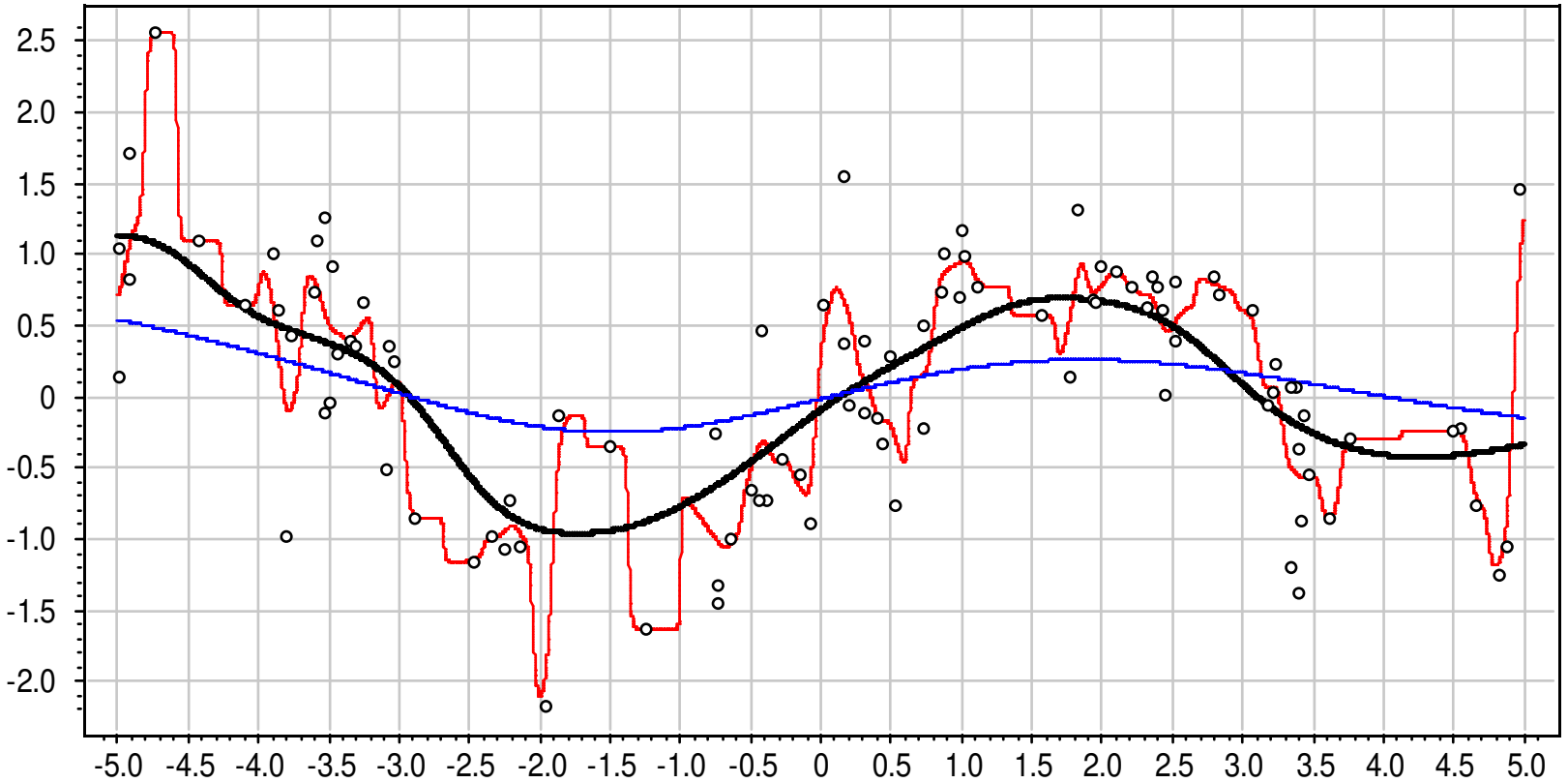
$a_{h_\ell}(x; X^\ell) \xrightarrow{P} \mathsf{E}(y|x)$ *в любой точке* $x \in X$,

в которой $\mathsf{E}(y|x)$, $p(x)$ *и* $\mathsf{D}(y|x)$ *непрерывны и* $p(x) > 0$.

В.Хардле. Прикладная непараметрическая регрессия. 1993.

Влияние ядра K и ширины окна h на вид аппроксимации

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

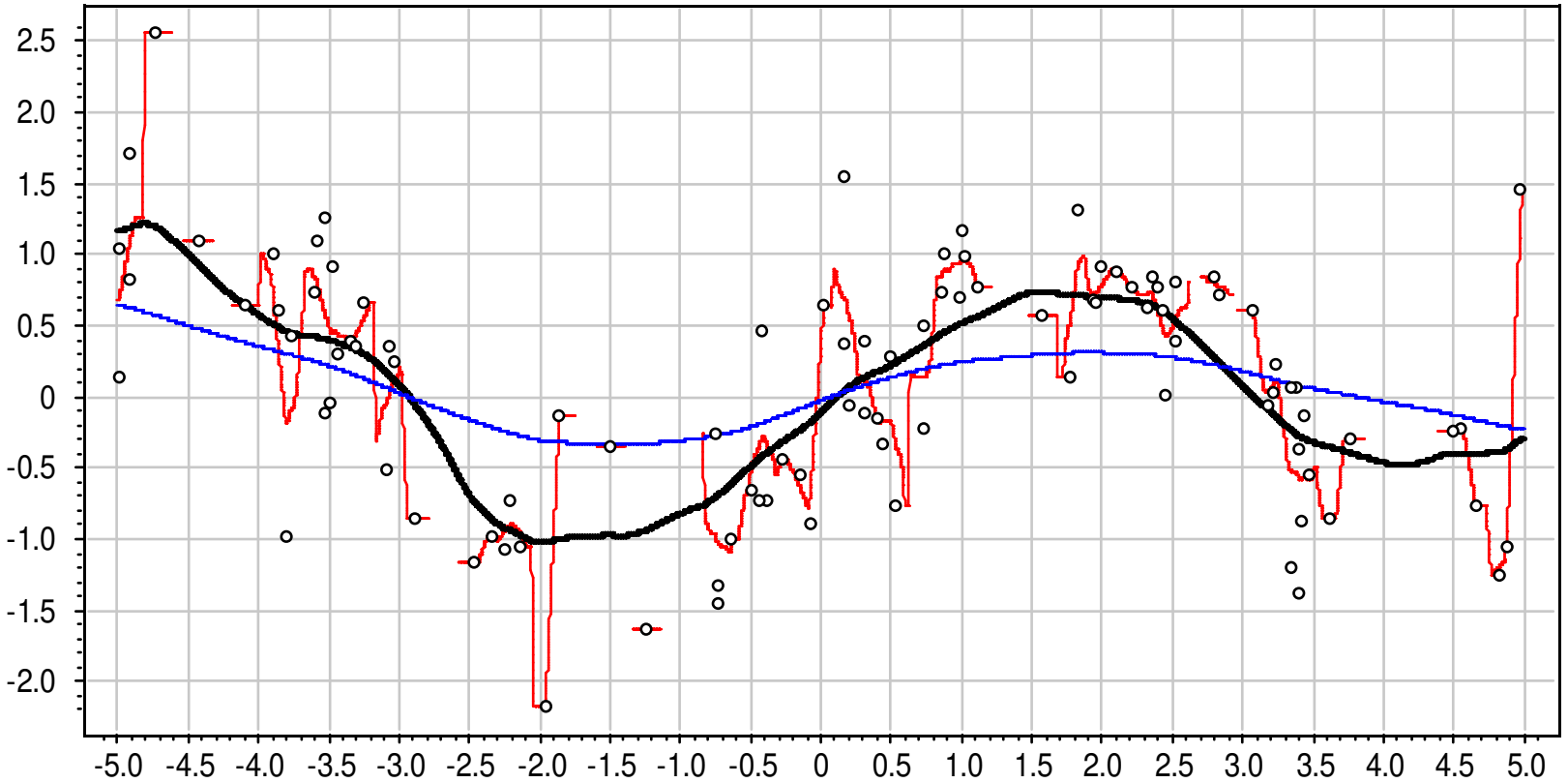


Гауссовское ядро $\Rightarrow$ гладкая аппроксимация

Ширина окна существенно влияет на точность аппроксимации

Влияние ядра K и ширины окна h на вид аппроксимации

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

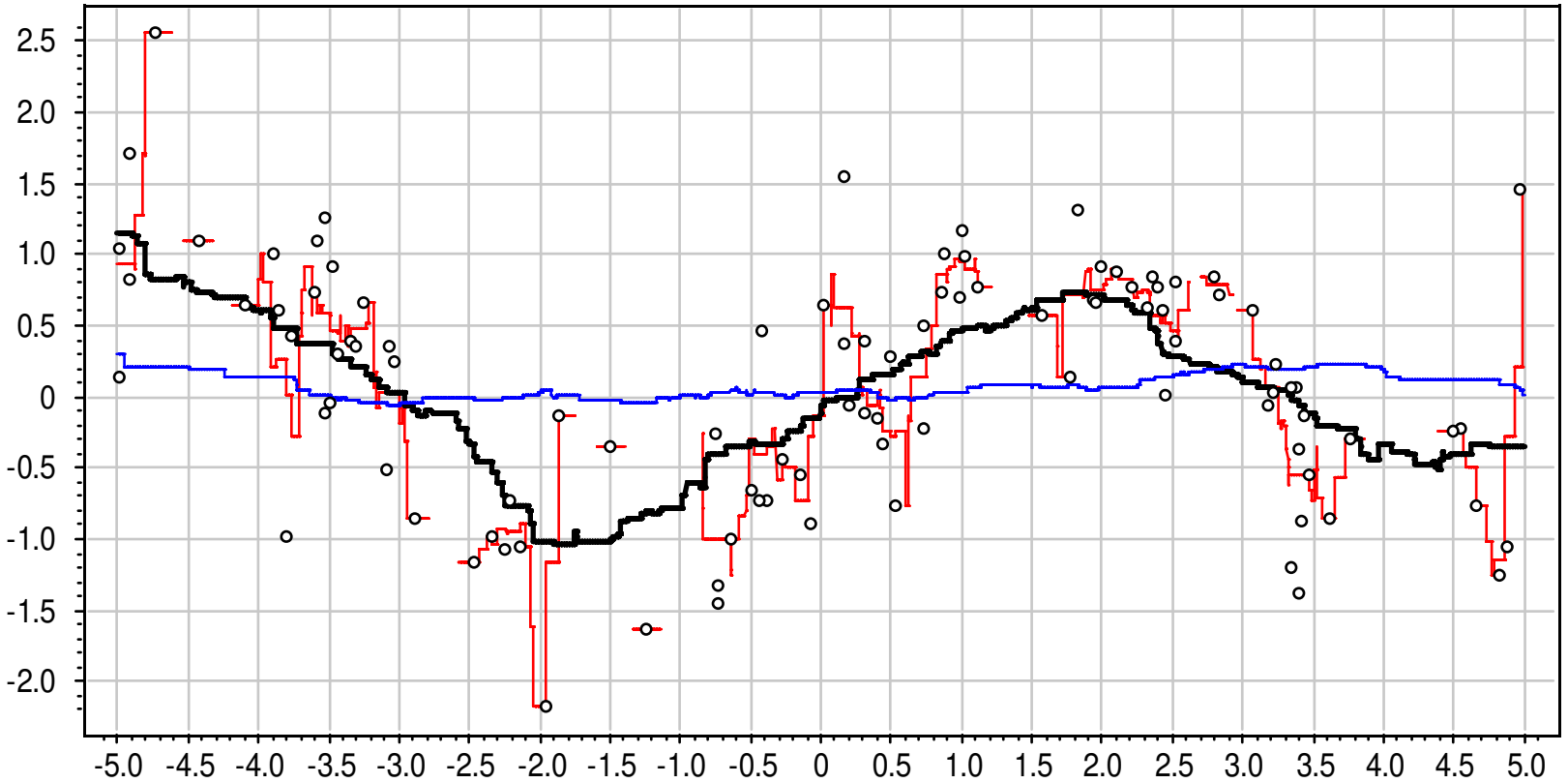


Треугольное ядро $\Rightarrow$ кусочно-линейная аппроксимация

Аппроксимация не определена, если в окне нет точек выборки

Влияние ядра K и ширины окна h на вид аппроксимации

$h \in \{0.1, 1.0, 3.0\}$, 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 $K(r)$
  — влияет на гладкость аппроксимирующей функции $a_h(x)$
  — почти не влияет на качество аппроксимации
- Ширина окна h
  — существенно влияет на качество аппроксимации
- Переменная ширина окна по k ближайшим соседям:

$$w_i(x) = K\left(\frac{\rho(x, x_i)}{h(x)}\right), \qquad h(x) = \rho(x, x^{(k+1)})$$

  где $x^{(k)}$ — k-й сосед объекта x
- Оптимизация ширины окна (h или k) по leave-one-out:

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

Задача отбора эталонов (prototype selection)

Дано: $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка
$a(x; X^\ell)$ — модель, хранящая выборку (lazy learning)

Найти подмножество эталонных объектов $U \subseteq X^\ell$

Критерий — минимизация числа эталонов
при минимальном ухудшении качества модели:

$$Q(a, U) = \sum_{i=1}^{\ell} \mathscr{L}\big(a(x_i; U), y_i\big) + \lambda|U| \to \min_{U \subseteq X^\ell}$$

Цели отбора эталонов в метрических алгоритмах:

- уменьшить объём хранимых данных
- избавиться от объектов-выбросов
- улучшить качество (обобщающую способность) модели

Отбор эталонов в линейных метрических моделях

$f_j(x) = K\left(\frac{1}{h_j}\rho(x, x_j)\right)$, $j = 1, \ldots, \ell$ — признаки объекта x

Линейная модель классификации, $Y = \{-1, +1\}$, обучение с убывающей функцией отступа $L(M)$ и L_1-регуляризацией:

$$a(x, w) = \text{sign} \sum_{j=1}^{\ell} w_j f_j(x) = \text{sign}\langle w, f(x)\rangle;$$

$$\sum_{i=1}^{\ell} L\left(y_i\langle w, f(x_i)\rangle\right) + \tau \sum_{i=1}^{\ell} |w_i| \to \min_w;$$

Линейная модель регрессии, $Y = \mathbb{R}$, обучение методом НК:

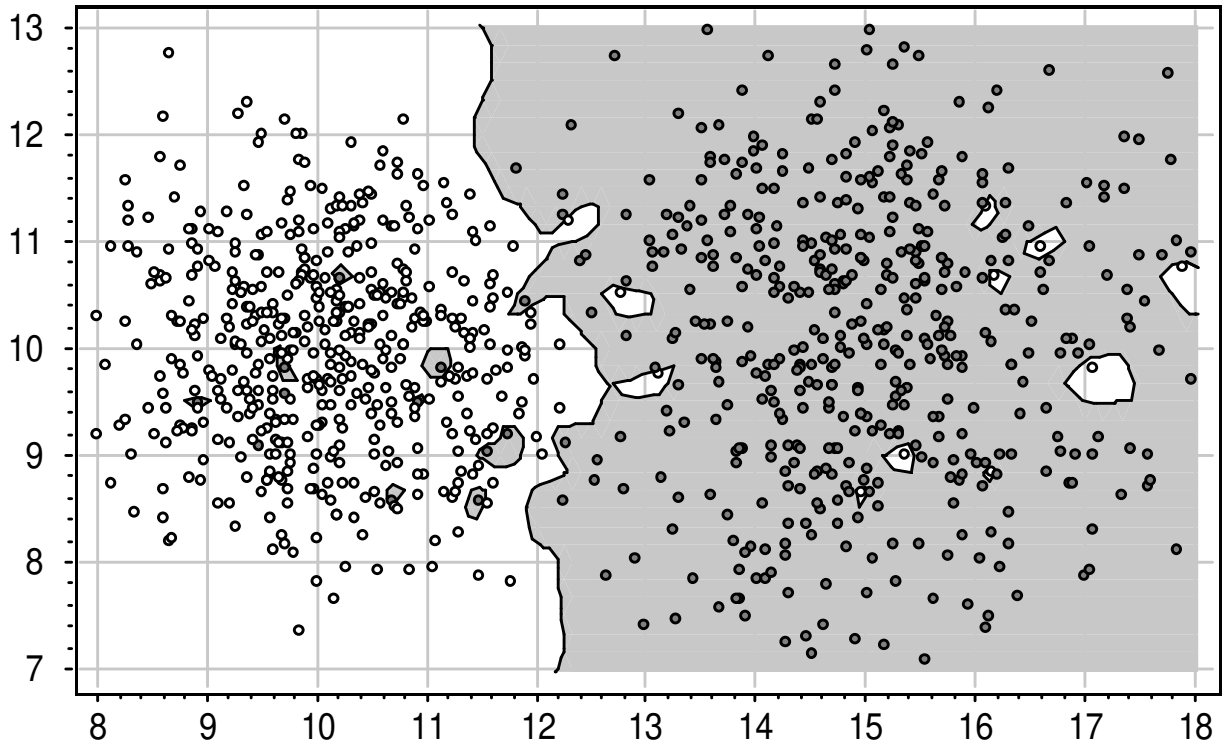
$$a(x, w) = \sum_{j=1}^{\ell} w_j f_j(x) = \langle w, f(x)\rangle$$

$$\sum_{i=1}^{\ell} \left(a(x_i, w) - y_i\right)^2 + \tau \sum_{i=1}^{\ell} |w_i| \to \min_w;$$

Чем больше τ, тем меньше остаётся эталонов

Оптимизация h_j: взять несколько f_j с разными h_j для каждого x_j

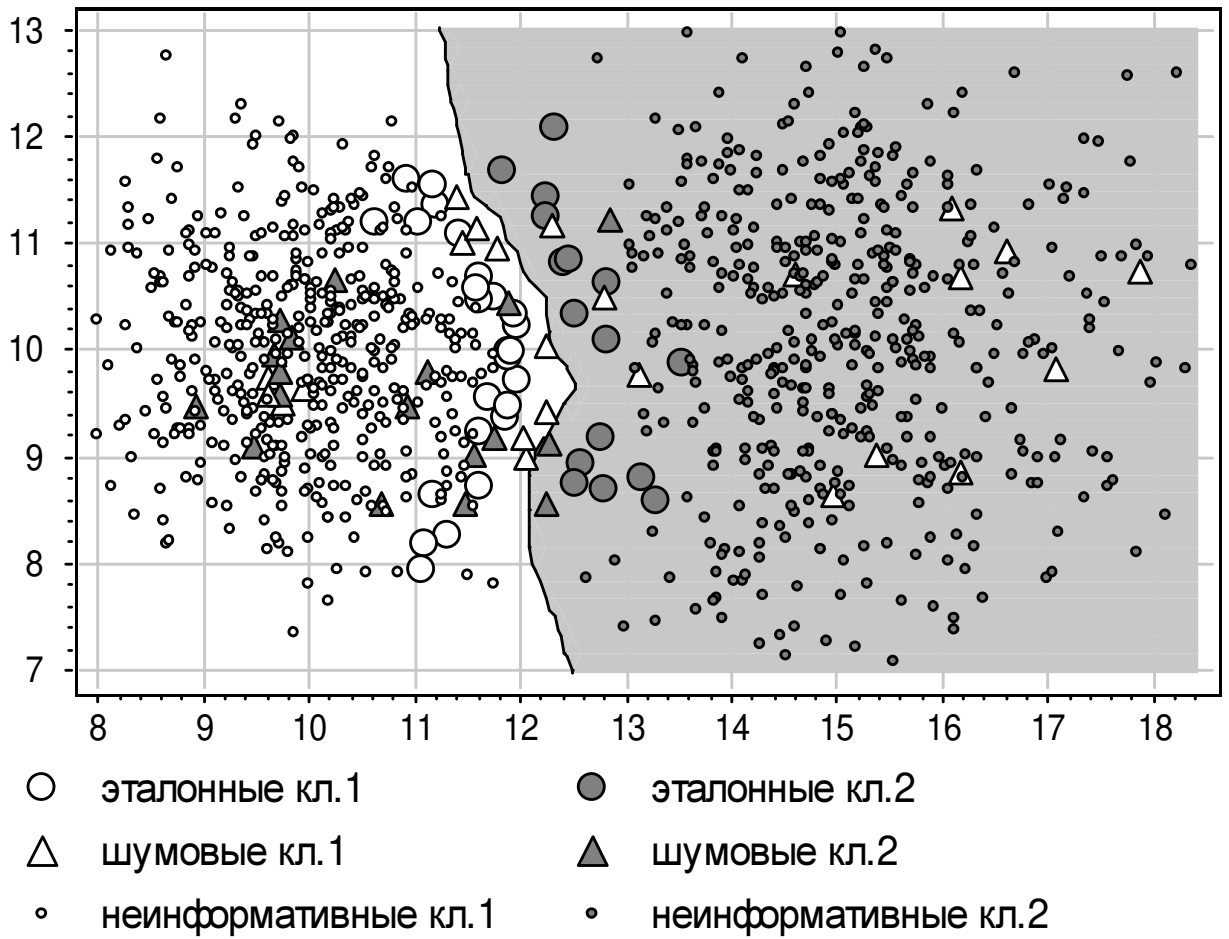
Пример. Отбор эталонов в задаче бинарной классификации



Синтетическая задача классификации:
2 класса по 500 объектов, добавлено 30 шумовых объектов

М.Н.Иванов, К.В.Воронцов. Отбор эталонов, основанный на минимизации функционала полного скользящего контроля. ММРО-2009

Пример. Отбор эталонов в задаче бинарной классификации



М.Н.Иванов, К.В.Воронцов. Отбор эталонов, основанный на минимизации функционала полного скользящего контроля. ММРО-2009

Задача непараметрического восстановления плотности

Задача: по выборке $X^\ell = (x_i)_{i=1}^\ell$ оценить плотность $\hat{p}(x)$, без введения параметрической модели плотности

Дискретный случай: $x_i \in X$, $|X| \ll \ell$. Частотная оценка:

$$\hat{p}(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} [x_i = x]$$

Одномерный непрерывный случай: $x_i \in \mathbb{R}$. По определению плотности, если $P[a, b]$ — вероятностная мера отрезка $[a, b]$:

$$p(x) = \lim_{h \to 0} \frac{1}{2h} P[x - h, x + h]$$

Эмпирическая частотная оценка плотности по окну ширины h (заменяем вероятность долей объектов выборки):

$$\hat{p}_h(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} \frac{1}{2} \left[\frac{|x - x_i|}{h} < 1 \right]$$

Локальная непараметрическая оценка Парзена-Розенблатта

Другое название — Kernel Density Estimate (KDE)

Непрерывная или гладкая оценка плотности

Вводится *ядро* $K(r)$ — чётная неотрицательная функция, невозрастающая при $r > 0$, нормированная $\int K(r)\,dr = 1$

$$\hat{p}_h(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} K\left(\frac{x - x_i}{h}\right)$$

при $K(r) = \frac{1}{2}\left[|r| < 1\right]$ имеем частотную оценку плотности

Многомерная оценка плотности

$\rho(x, x_i)$ — функция расстояния в пространстве объектов X

$$\hat{p}_h(x) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)$$

где $V(h) = \int_X K\left(\frac{\rho(x,x_i)}{h}\right) dx$ — нормировочный множитель

Обоснование оценки Парзена-Розенблатта

Теорема (одномерный случай, $x_i \in \mathbb{R}$)

Пусть *выполнены следующие условия:*

1) X^ℓ — *простая выборка из распределения* $p(x)$;
2) *ядро* $K(r)$ *непрерывно и ограничено:* $\int_X K^2(r)\,dr < \infty$;
3) *последовательность* h_ℓ: $\lim\limits_{\ell\to\infty} h_\ell = 0$ *и* $\lim\limits_{\ell\to\infty} \ell h_\ell = \infty$.

Тогда:

1) $\hat{p}_{h_\ell}(x) \to p(x)$ *при* $\ell \to \infty$ *для почти всех* $x \in X$;
2) *скорость сходимости имеет порядок* $O(\ell^{-2/5})$.

Пример. Сферическое гауссовское ядро в $\mathbb{R}^n$ с евклидовой метрикой образуется произведением одномерных ядер:

$$\hat{p}_h(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}h}\exp\left(-\frac{(f_j(x)-f_j(x_i))^2}{2h^2}\right)$$

Влияние ядра на качество восстановления плотности

Функционал качества восстановления плотности:

$$J(K) = \int_{-\infty}^{+\infty} \mathsf{E}\big(\hat{p}_h(x) - p(x)\big)^2\, dx.$$

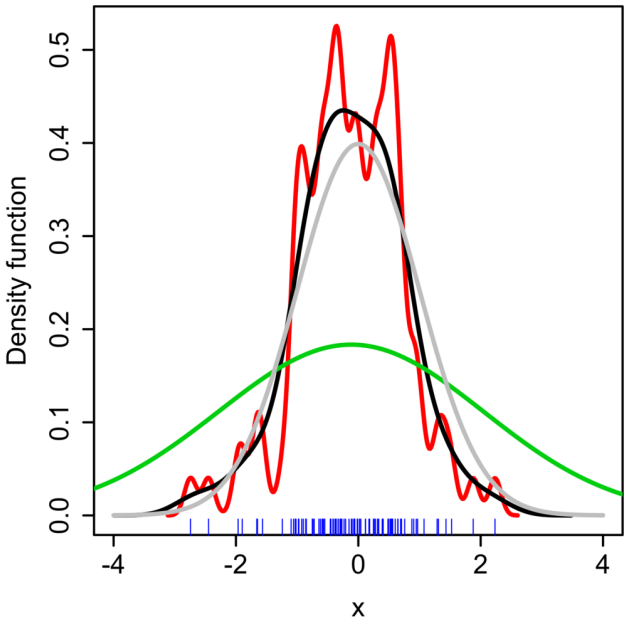
Асимптотические значения отношения $J(K^*)/J(K)$ при $\ell \to \infty$ не зависят от вида распределения $p(x)$.

ядро $K(r)$	степень гладкости	$J(K^*)/J(K)$
Епанечникова $K^*(r)$	$\hat{p}'_h$ разрывна	1.000
Квартическое	$\hat{p}''_h$ разрывна	0.995
Треугольное	$\hat{p}'_h$ разрывна	0.989
Гауссовское	∞ дифференцируема	0.961
Прямоугольное	$\hat{p}_h$ разрывна	0.943

М.Б.Лагутин. Наглядная математическая статистика. 2015

Влияние ширины окна на качество восстановления

Оценка $\hat{p}_h(x)$ при различных значениях ширины окна h:



истинная плотность
(стандартная гауссовская)

$h = 0.05$ — переобучение
$h = 0.337$ — оптимальная
$h = 2.0$ — недообучение

- Качество восстановления плотности существенно зависит от ширины окна h, но слабо зависит от вида ядра K
- При неоднородности локальных сгущений плотности можно задавать $h_k(x) = \rho\left(x, x^{(k+1)}\right)$, где k — число соседей

Выбор ядра K и ширины окна h

- Ядро $K(r)$
  — влияет на гладкость оценки плотности $\hat{p}_h(x)$
  — почти не влияет на качество восстановления плотности
- Ширина окна h
  — существенно влияет на качество восстановления
- Переменная ширина окна по k ближайшим соседям:

$$K\left(\frac{\rho(x,x_i)}{h(x)}\right), \qquad h(x) = \rho(x, x^{(k+1)})$$

  где $x^{(k)}$ — k-й сосед объекта x
- Оптимизация ширины окна (h или k) по критерию максимума правдоподобия в режиме leave-one-out:

$$\mathrm{LOO}(h, X^\ell) = -\sum_{i=1}^{\ell} \ln \hat{p}_h(x_i; X^\ell \setminus x_i) \to \min_h$$

Резюмируем: (непара)метрические методы анализа данных

Восстановление плотности: *оценка Парзена–Розенблатта*

$$\hat{p}_h(x) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{1}{h}\rho(x, x_i)\right)$$

Классификация: *метод парзеновского окна* — байесовский классификатор с парзеновскими плотностями классов

$$a_h(x) = \arg\max_{y \in Y} P(y)\hat{p}(x|y) = \arg\max_{y \in Y} \sum_{i=1}^{\ell} [y_i = y]\, K\left(\frac{1}{h}\rho(x, x_i)\right)$$

Регрессия: *метод ядерного сглаживания* Надарая–Ватсона

$$a_h(x) = \sum_{i=1}^{\ell} y_i K\left(\frac{1}{h}\rho(x, x_i)\right) \Big/ \sum_{i=1}^{\ell} K\left(\frac{1}{h}\rho(x, x_i)\right)$$

Общая проблематика:

- выбор ядра $K(r)$ и ширины окна h, h_i, $h(x)$
- отбор эталонных объектов

Постановка задачи кластеризации (обучение без учителя)

Дано:

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

$\rho\colon X \times X \to [0, \infty)$ — функция расстояния между объектами.

Найти:

Y — множество кластеров,

$a\colon X \to Y$ — алгоритм кластеризации,

такие, что:

— каждый кластер состоит из близких объектов;

— объекты разных кластеров существенно различны.

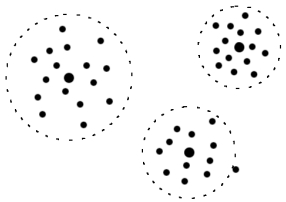
Критерий?

- однозначной постановки задачи кластеризации нет
- есть много эвристических методов, даже не оптимизации
- число кластеров $|Y|$ тоже может быть не известно
- результат кластеризации зависит от метрики ρ

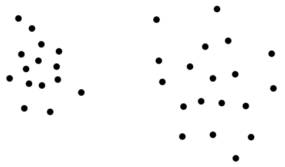
Цели кластеризации

- **Упростить дальнейшую обработку данных**,
  разбить выборку X^ℓ на подвыборки схожих объектов,
  далее работать с ними по принципу «разделяй и властвуй»
- **Сократить объём хранимых данных**,
  оставив по одному эталону от каждого кластера,
  получить максимально представительную подвыборку
- **Выделить нетипичные объекты**,
  которые не подходят ни к одному из кластеров
  (выделение аномалий, одноклассовая классификация)
- **Построить иерархию множества объектов,**
  пример — классификация животных и растений К.Линнея
  (задачи таксономии, иерархической кластеризации)

Типы кластерных структур

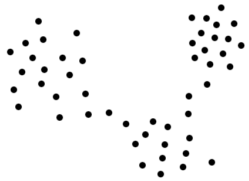


кластеры
с центрами



расстояния внутри
кластеров меньше
межкластерных

ленточные
кластеры



перемычки
между кластерами

разреженный фон
из нетипичных
объектов

перекрывающиеся
кластеры

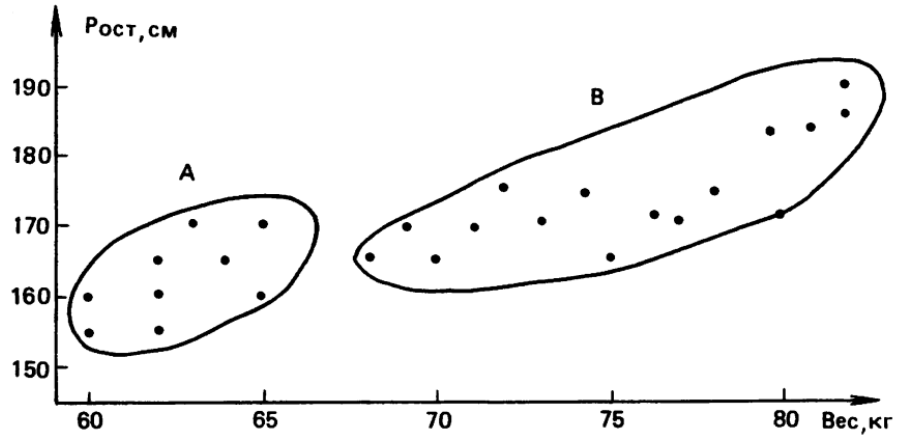
кластеры могут
вообще
отсутствовать

это не кластеры,
на практике такое
не встречается

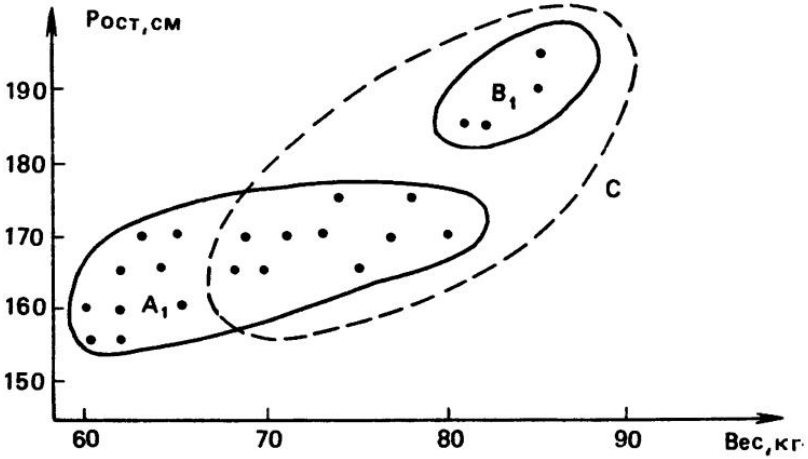
Каждый метод выделяет лишь определённые типы структур

Проблема чувствительности к выбору метрики

Результат зависит от нормировки признаков:



A — студентки,
B — студенты



после перенормировки
(сжали ось «вес» вдвое)

И.Д.Мандель. Кластерный анализ. 1988

Задача кластеризации (clustering)

Дано: $X^\ell = \{x_1, \ldots, x_\ell\}$ — обучающая выборка, $x_i \in \mathbb{R}^n$

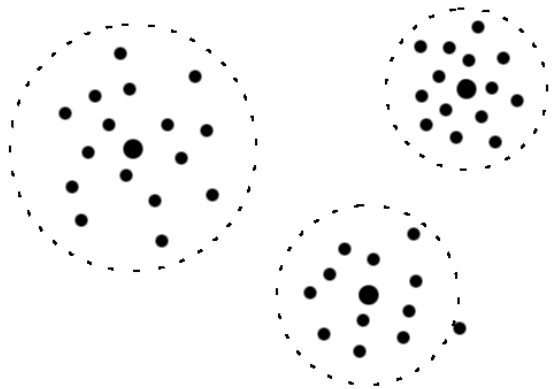
Найти:

— центры кластеров — параметры $\mu_a \in \mathbb{R}^n$, $a = 1, \ldots, K$

— какому кластеру принадлежит каждый объект $a_i \in \{1, \ldots, K\}$

Критерий: минимум суммы внутрикластерных расстояний

$$\sum_{i=1}^{\ell} \rho(x_i, \mu_{a_i})^2 \to \min_{\{a_i\}, \{\mu_a\}}$$



Метрика, как правило, евклидова (но может быть и другая):

$$\rho(x, \mu_a)^2 = \sum_{d=1}^{n} \big(f_d(x) - \mu_{ad}\big)^2$$

Метод K-средних (K-means) для кластеризации

Минимизация суммы квадратов внутрикластерных расстояний:

$$\sum_{i=1}^{\ell} \rho(x_i, \mu_{a_i})^2 \to \min_{\{a_i\}, \{\mu_a\}}, \quad \rho(x_i, \mu_a)^2 = \sum_{j=1}^{n} \big(f_j(x_i) - \mu_{aj}\big)^2$$

Алгоритм Ллойда (сильно упрощённый EM-алгоритм)

вход: X^ℓ, K; **выход:** центры μ_a, $a \in \{1, \ldots, K\}$;

$\mu_a :=$ начальное приближение центра, $a \in \{1, \ldots, K\}$;

повторять

отнести каждый x_i к ближайшему центру:

$a_i := \arg\min\limits_{a \in Y} \rho(x_i, \mu_a)$, $i = 1, \ldots, \ell$;

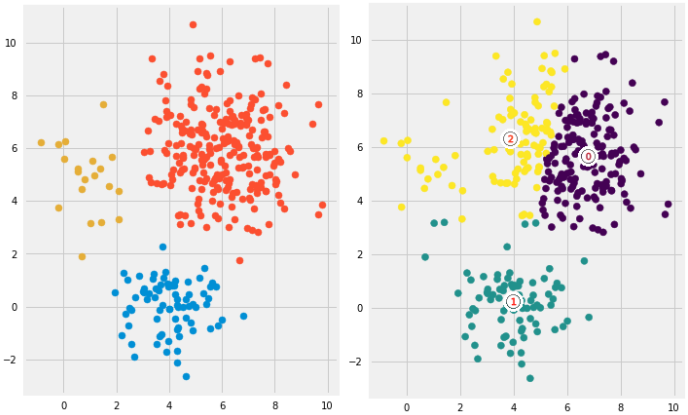
вычислить новые положения центров:

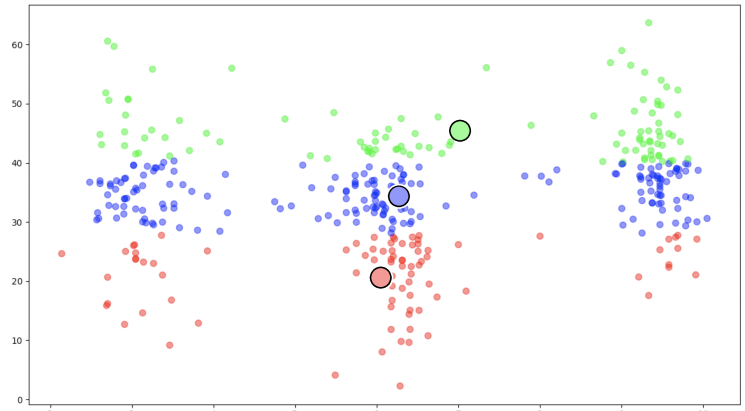
$$\mu_a := \frac{\sum_{i=1}^{\ell} [a_i = a] x_i}{\sum_{i=1}^{\ell} [a_i = a]}, \quad a \in \{1, \ldots, K\};$$

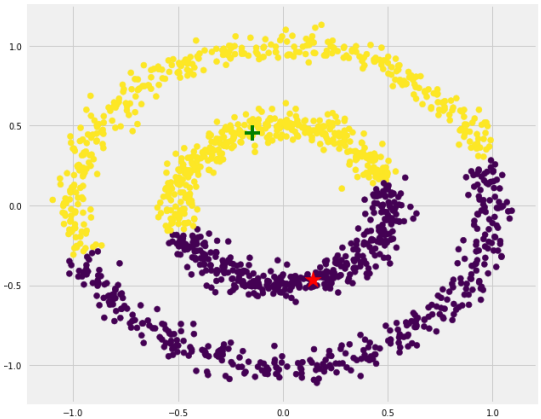
пока a_i не перестанут изменяться;

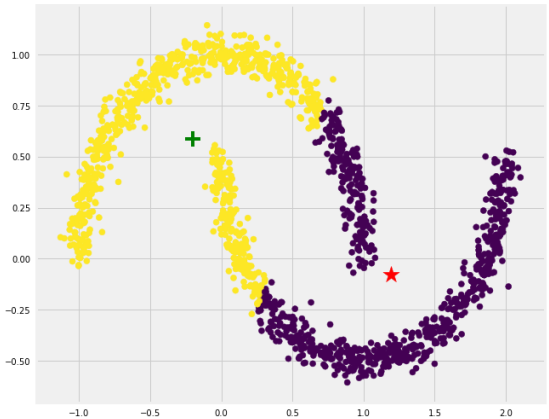
Примеры неудачной кластеризации k-means

Причина — неудачное начальное приближение или форма кластеров, существенно отличная от сферической









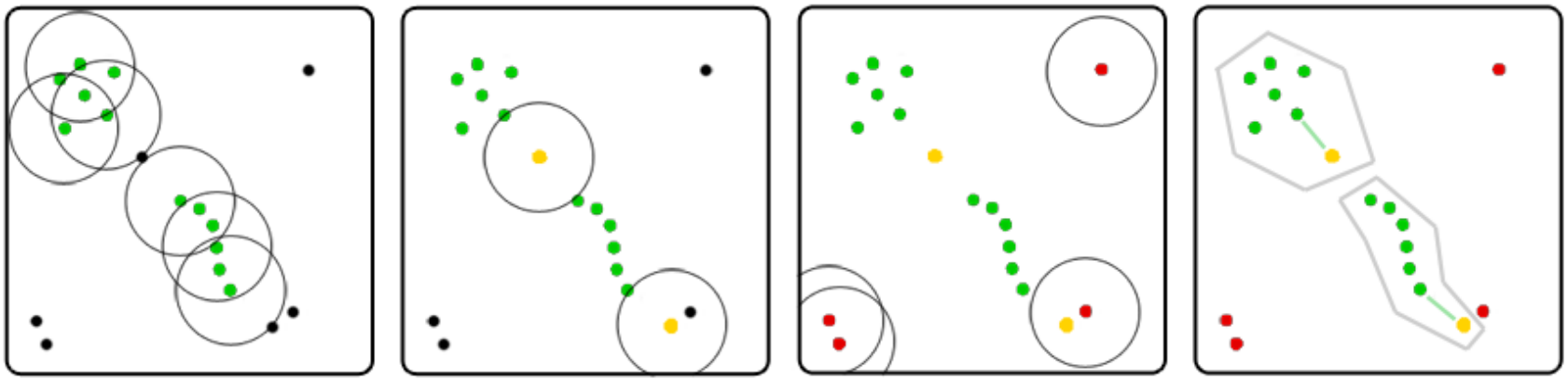
Алгоритм кластеризации DBSCAN

(Density-Based Spatial Clustering of Applications with Noise)

Объект $x \in U$, его ε-окрестность $U_\varepsilon(x) = \{u \in U : \rho(x,u) \leqslant \varepsilon\}$

Каждый объект может быть одного из трёх типов:

- корневой: имеющий плотную окрестность, $|U_\varepsilon(x)| \geqslant m$
- граничный: не корневой, но в окрестности корневого
- шумовой (выброс): не корневой и не граничный



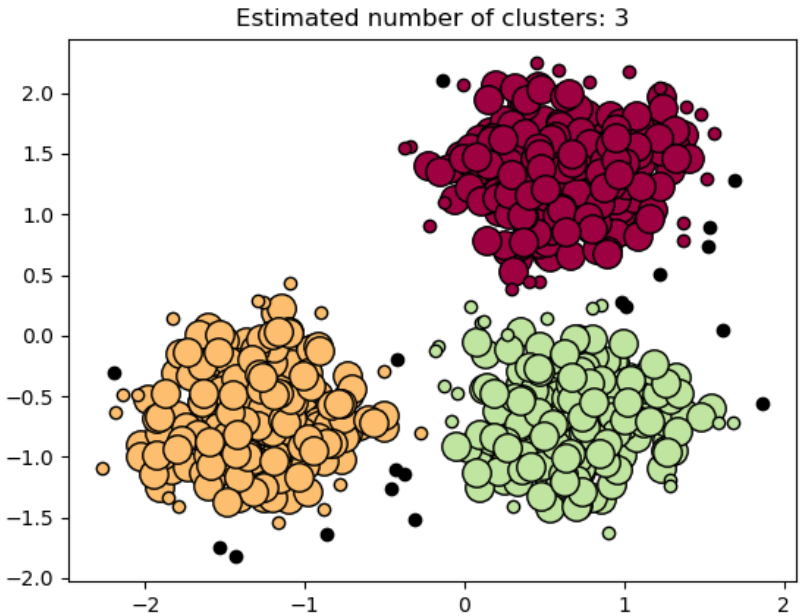
Ester, Kriegel, Sander, Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD-1996.

Алгоритм кластеризации DBSCAN

Вход: выборка $X^\ell = \{x_1, \ldots, x_\ell\}$; параметры ε и m;
Выход: разбиение выборки на кластеры и шумовые выбросы;
$U := X^\ell$ — непомеченные; $a := 0$;
пока в выборке есть непомеченные точки, $U \neq \varnothing$:
  взять случайную точку $x \in U$;
  если $|U_\varepsilon(x)| < m$ **то**
    пометить x как, возможно, шумовой;
  иначе
    создать новый кластер: $K := U_\varepsilon(x)$; $a := a + 1$;
    для всех $x' \in K$, не помеченных или шумовых
      если $|U_\varepsilon(x')| \geqslant m$ **то** $K := K \cup U_\varepsilon(x')$;
      иначе пометить x' как граничный кластера K;
    $a_i := a$ для всех $x_i \in K$;
    $U := U \setminus K$;

Преимущества алгоритма DBSCAN

- быстрая кластеризация больших данных:
  $O(\ell^2)$ в худшем случае,
  $O(\ell \ln \ell)$ при эффективной реализации $U_\varepsilon(x)$;
- кластеры произвольной формы (долой центры!);
- деление объектов на корневые, граничные, шумовые.



Агломеративная иерархическая кластеризация

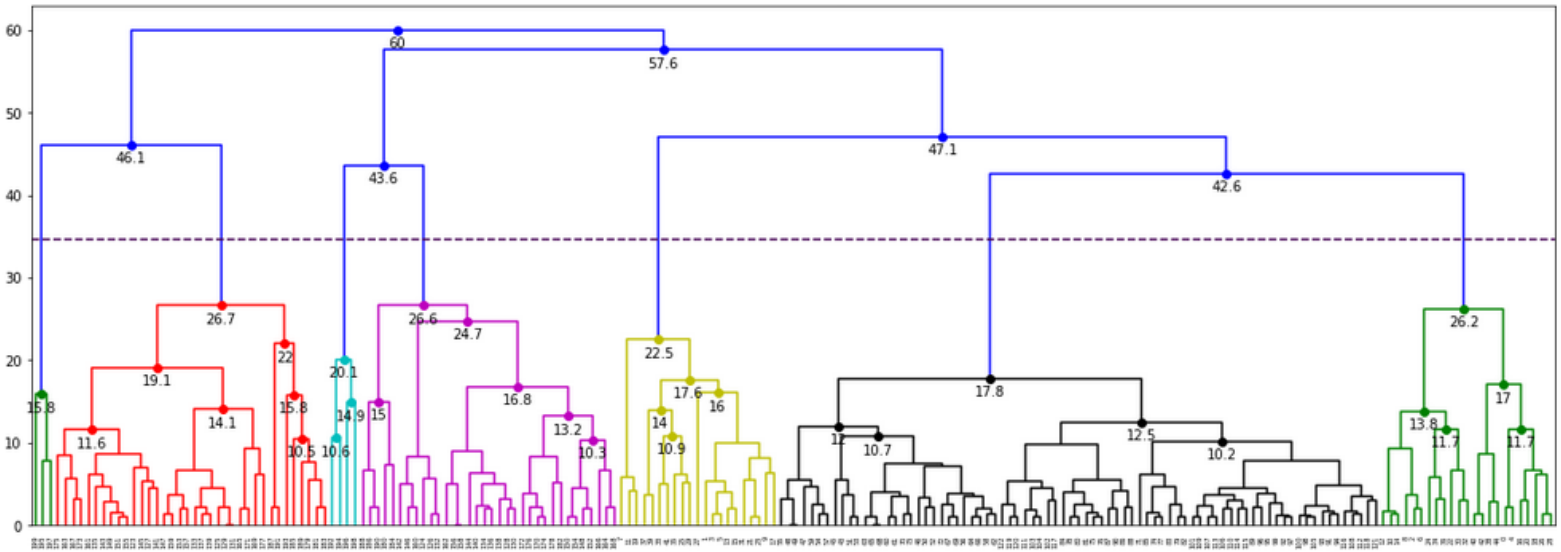
Алгоритм иерархической кластеризации (Ланс, Уильямс, 1967): итеративный пересчёт расстояний R_{UV} между кластерами U, V.

$C_1 := \{\{x_1\}, \ldots, \{x_\ell\}\}$ — все кластеры 1-элементные;
$R_{\{x_i\}\{x_j\}} := \rho(x_i, x_j)$ — расстояния между ними;
для всех $t = 2, \ldots, \ell$ (t — номер итерации):
  найти в C_{t-1} пару кластеров (U, V) с минимальным R_{UV};
  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_{WS} по формуле Ланса-Уильямса:
    $R_{WS} := \alpha_U R_{US} + \alpha_V R_{VS} + \beta R_{UV} + \gamma \left| R_{US} - R_{VS} \right|$;

G.N.Lance, W.T.Williams. A general theory of classificatory sorting strategies: II. Clustering systems. 1967

Дендрограмма — визуализация иерархической кластеризации

- Кластеры группируются вдоль горизонтальной оси
- По вертикальной оси откладываются расстояния R_t
- Расстояния возрастают, линии нигде не пересекаются
- Верхние уровни различимы лучше, чем нижние
- Уровень отсечения определяет число кластеров



Напоминание. Многомерное шкалирование (MDS)

Дано: $(i,j) \in E$ — выборка рёбер графа $\langle V, E \rangle$
R_{ij} — расстояния между вершинами ребра (i,j)
Например, R_{ij} — длина кратчайшего пути по графу (IsoMAP)

Найти: векторные представления вершин $z_i \in \mathbb{R}^d$ так, чтобы близкие вершины (в смысле малого R_{ij}) имели близкие z_i и z_j

Критерий стресса (stress):

$$\sum_{(i,j)\in E} R_{ij}^{\gamma} \big(\rho(z_i, z_j) - R_{ij}\big)^2 \to \min_Z, \quad Z \in \mathbb{R}^{V \times d},$$

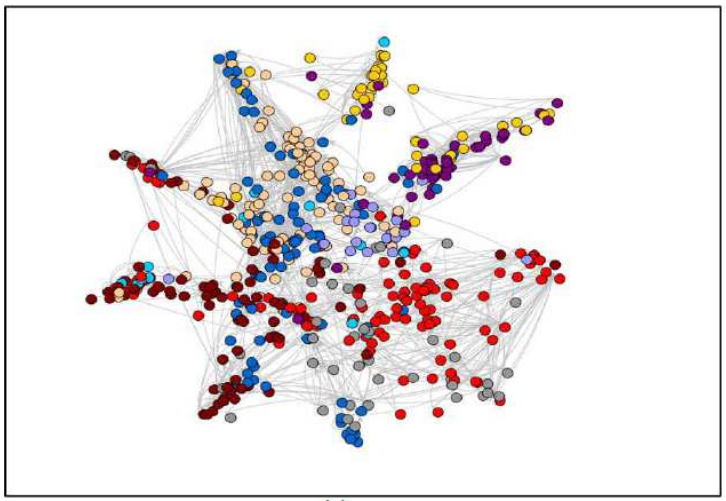
где $\rho(z_i, z_j) = \|z_i - z_j\|$ — обычно евклидово расстояние,
R_{ij}^{γ} — веса, важнее расстояния большие ($\gamma > 0$) или малые ($\gamma < 0$)

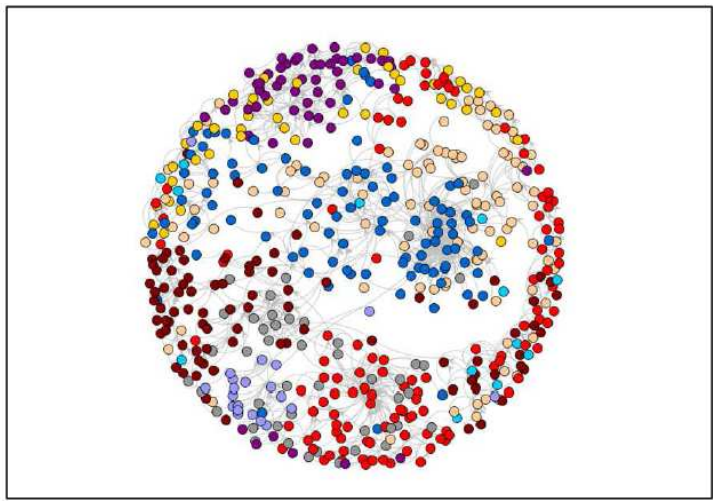
Обычно решается методом стохастического градиента (SG)

I.Chami et al. Machine learning on graphs: a model and comprehensive taxonomy. 2020.

Напоминание. MDS для визуализации данных

При $d = 2$ осуществляется проекция выборки на плоскость





- используется для визуализации кластерных структур
- форму облака точек можно настраивать весами и метрикой
- наиболее популярные методы — t-SNE, UMAP
- недостаток всех методов — искажения неизбежны

Laurens van der Maaten, Geoffrey Hinton. Visualizing data using t-SNE. 2008
Leland McInnes, John Healy, James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. 2020

Резюме

- Метрические методы — простейшие в машинном обучении, обучение сводится к запоминанию выборки (lazy learning)
- Усложняя метрические методы, можно обучать:
  — число ближайших соседей k или ширину окна h
  — веса (значимости, информативности) объектов
  — множество эталонов (prototype learning)
  — метрику (distance learning, similarity learning), в частности, веса признаков в метрике Минковского
- Метод потенциальных функций = линейный классификатор
  расстояние до опорного объекта = новый признак
- Качество обучения зависит от метрики и ширины окна, слабо зависит от вида ядра сглаживания
- Непараметрические методы обходятся без модели?
  Нет, моделируется функция сходства между объектами