Курс «Введение в машинное обучение»

Метрические методы машинного обучения

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

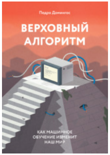
Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Введение в машинное обучение (курс лекций, К.В.Воронцов)»

МФТИ.ФПМИ.ИС.ИАД • 3 апреля 2025

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

1. **Введение расстояний между объектами**
   - Гипотезы компактности или непрерывности
   - Функции расстояния между векторами признаков
   - Беспризнаковые способы вычисления расстояний
2. **Метрические методы обучения с учителем**
   - Классификация
   - Непараметрическая регрессия
   - Задача отбора эталонов
3. **Метрические методы обучения без учителя**
   - Непараметрическое оценивание плотности
   - Кластеризация
   - Многомерное шкалирование

Гипотезы непрерывности и компактности

Задачи классификации и регрессии:

X — объекты, Y — ответы;

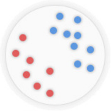
$X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка;

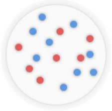
Гипотеза непрерывности (для регрессии):
близким объектам соответствуют близкие ответы.

выполнена: 

не выполнена: 

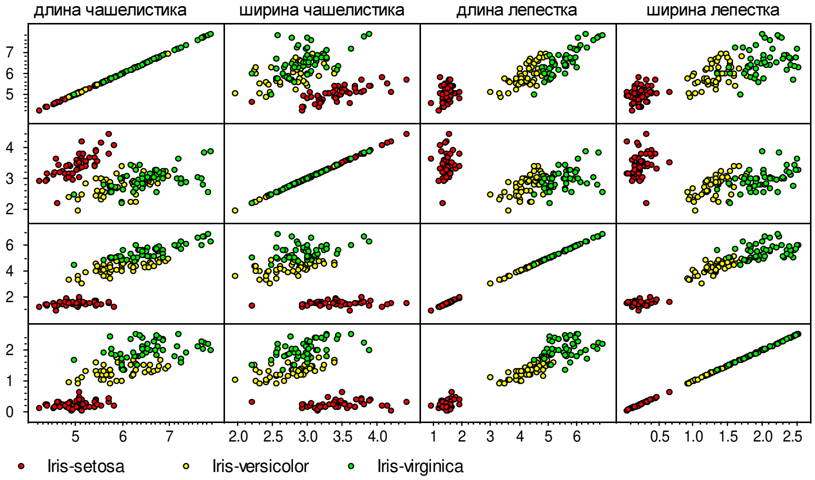
Гипотеза «компактности» (для классификации):
близкие объекты, как правило, лежат в одном классе.

выполнена: 

не выполнена: 

Пример: задача классификации цветков ириса [Фишер, 1936]

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



Формализация понятия «расстояние» (distance)

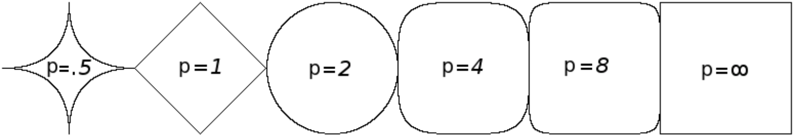
Евклидова метрика и *обобщённая метрика Минковского*:

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

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

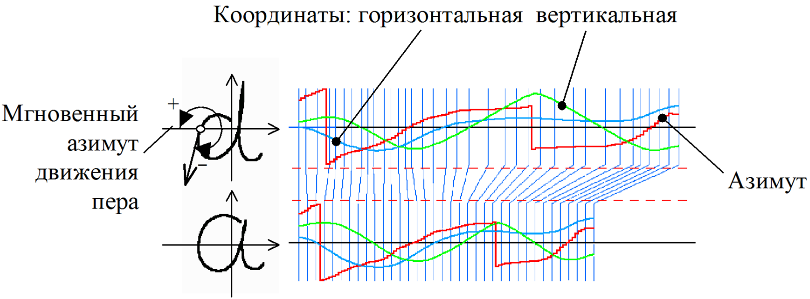


Расстояния между строками / сигналами

Для строк — редакторское расстояние Левенштейна:

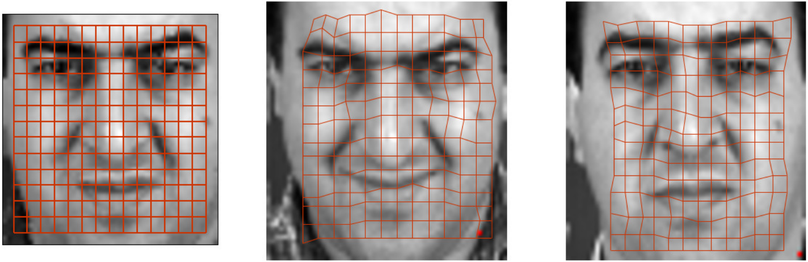
```
CTGGGCTAAAAGGTCCCTTAGCC..TTTAGAAAAA.GGGCCATTAGGAAATTGC
CTGGGACTAAA....CCTTAGCCTATTTACAAAAATGGGCCATTAGG...TTGC
```

Для сигналов — энергия сжатий и растяжений:



Расстояния между изображениями

Расстояние **между изображениями** на основе выравнивания:



Оценивается энергия растяжения прямоугольной сетки

Общая формула метрического классификатора

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

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

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

Метрический алгоритм классификации относит объект x к тому классу, которому принадлежат его ближайшие соседи:

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

$w(i, x)$ — *функция близости* к объекту x его i-го соседа, неотрицательная, не возрастает по i,
$\Gamma_y(x)$ — *оценка близости* объекта x к классу y.

Эвристические варианты метрического классификатора

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

$w(i,x) = K\bigl(\frac{1}{h}\rho(x, x^{(i)})\bigr)$ — метод окна Парзена, где
h — фиксированная *ширина окна* (bandwidth),
$K(r)$ — *ядро* (kernel), не возрастает и положительно на $[0,1]$

$w(i,x) = K\bigl(\frac{1}{h(x)}\rho(x, x^{(i)})\bigr)$ — метод окна Парзена, где
$h(x) = \rho(x, x^{(k+1)})$ — переменная ширина окна по k соседям

$w(i,x) = \alpha^{(i)} K\bigl(\frac{1}{h^{(i)}}\rho(x, x^{(i)})\bigr)$ — метод потенциальных функций,
$\alpha^{(i)}$ — обучаемый параметр «заряд потенциала»,
$h^{(i)}$ — «радиус действия потенциала» с центром в точке $x^{(i)}$

Метод потенциальных функций

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

Более простая запись (ведь можно не ранжировать объекты):

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

Физическая аналогия из электростатики:

$\alpha_i \geqslant 0$ — веса объектов, величина «заряда» в точке x_i

$h_i > 0$ — «радиус действия» потенциала с центром в точке x_i

α_i, h_i — обучаемые параметры (однородный потенциал, $h_i = h$)

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

$K(r) = \frac{1}{r^2}$ или $\frac{1}{r}$ или $\frac{1}{r+a}$ — вид потенциальной функции

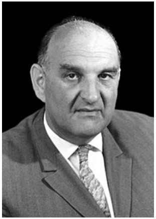
В задачах классификации нет ограничений на вид K и на $|Y|$

М.А.Айзерман, Э.М.Браверман, Л.И.Розоноэр. Метод потенциальных функций в теории обучения машин. М.: Наука, 1970.

Научная школа М. А. Айзермана

- *Гипотеза компактности*: схожие объекты, как правило, находятся в одном классе
- Идея *метода потенциальных функций* заимствуется из физики
- Линейная модель классификации: взвешенное голосование функций сходства $f_i(x) = K(x, x_i)$ между x и x_i:

$$a(x) = \arg\max_{y \in Y} \sum_{i\colon y_i = y} \alpha_{yi} K(x, x_i)$$



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.

Метод потенциальных функций = линейный классификатор

Два класса: $Y = \{-1, +1\}$.

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

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

Сравним с линейной моделью классификации:

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

- $f_j(x) = y_j K\bigl(\frac{1}{h_j}\rho(x, x_j)\bigr)$ — новые признаки объекта x, близость (сходство) объекта x и обучающего объекта x_j
- α_j — обучаемые веса линейного классификатора
- $n = \ell$ — число признаков равно числу объектов обучения

Выбор ядра K и ширины окна h

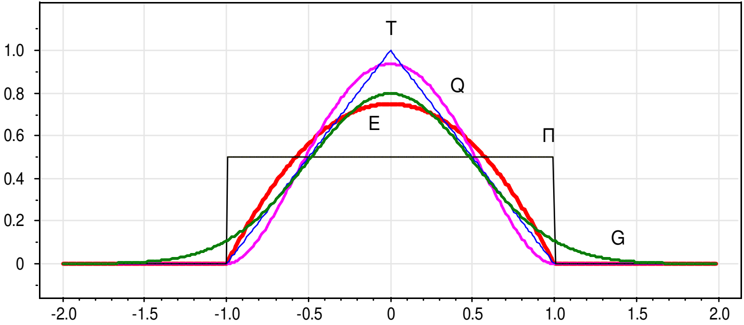
- Ядро $K(r)$
  — влияет на гладкость разделяющей поверхности
  — почти не влияет на качество классификации
- Ширина окна h
  — существенно влияет на качество классификации
- Переменная ширина окна по k ближайшим соседям:

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

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

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

Часто используемые ядра $K(r)$



$\Pi(r) = \bigl[|r| \leqslant 1\bigr]$ — прямоугольное

$T(r) = \bigl(1 - |r|\bigr)\bigl[|r| \leqslant 1\bigr]$ — треугольное

$E(r) = (1 - r^2)\bigl[|r| \leqslant 1\bigr]$ — квадратичное (Епанечникова)

$Q(r) = (1 - r^2)^2\bigl[|r| \leqslant 1\bigr]$ — квартическое

$G(r) = \exp\bigl(-2r^2\bigr)$ — гауссовское

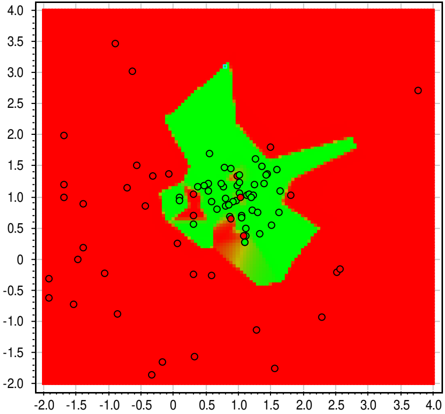
Влияние ширины окна h на качество классификации

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

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

ядро:
гауссовское

ширина окна:
$h = 0.05$



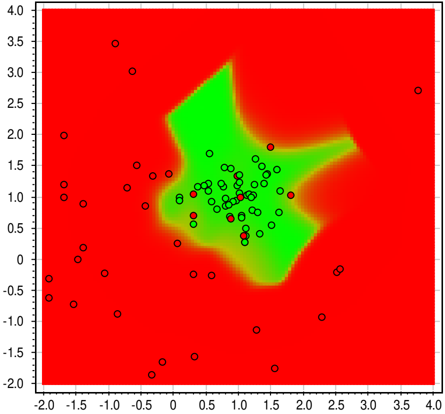
Влияние ширины окна h на качество классификации

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

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

ядро:
гауссовское

ширина окна:
$h = 0.2$



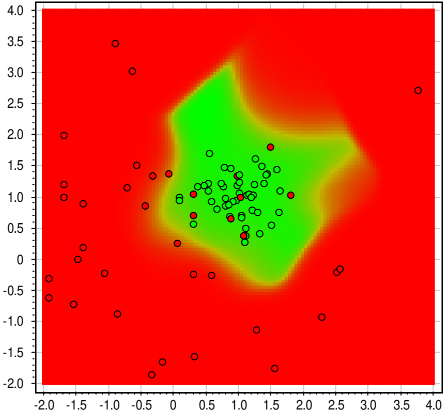
Влияние ширины окна h на качество классификации

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

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

ядро:
гауссовское

ширина окна:
$h = 0.3$



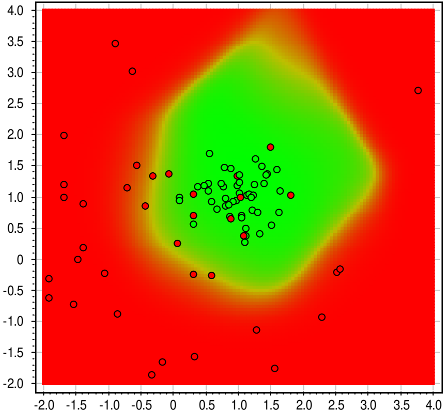
Влияние ширины окна h на качество классификации

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

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

ядро:
гауссовское

ширина окна:
$h = 0.5$



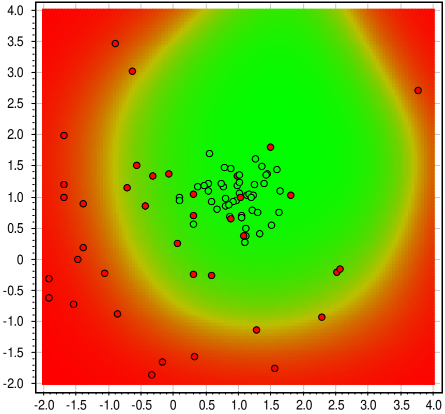
Влияние ширины окна h на качество классификации

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

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

ядро:
гауссовское

ширина окна:
$h = 1.0$



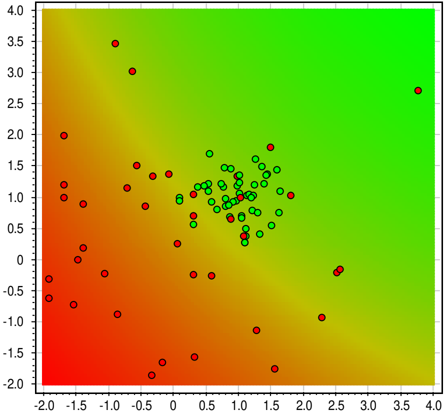
Влияние ширины окна h на качество классификации

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

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

ядро:
гауссовское

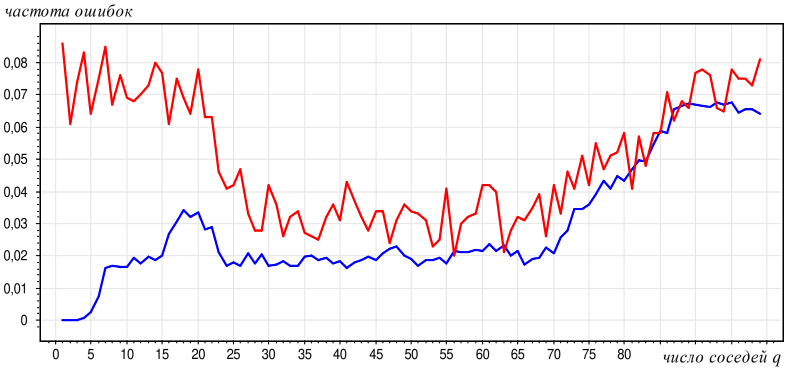
ширина окна:
$h = 5.0$



Зависимость LOO от числа соседей

Пример.
Задача UCI: Iris.

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



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

— несмещённое число ошибок LOO

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:

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

где w_i — весовой коэффициент, степень важности объекта x_i

Мотивация перехода к непараметрическим моделям:
нет теорий для создания «физичной» модели $f(x, \theta)$

Непараметрическая регрессия, формула Надарая–Ватсона

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

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

где $w_i(x) = K\left(\frac{\rho(x,x_i)}{h}\right)$ — веса объектов x_i относительно x;
$K(r)$ — *ядро* (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:*
$\lim\limits_{\ell\to\infty} h_\ell = 0$, $\lim\limits_{\ell\to\infty} \ell h_\ell = \infty$.

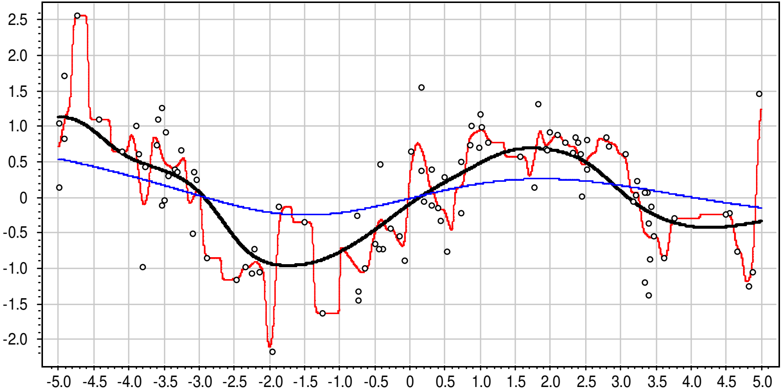
Тогда имеет место сходимость по вероятности:

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

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

Выбор ядра K и ширины окна h

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

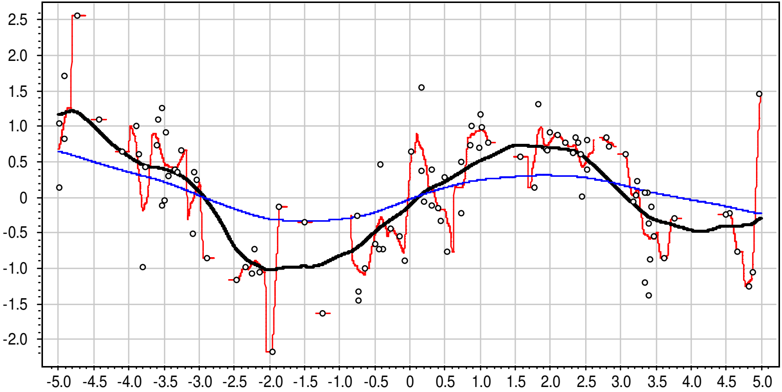


Гауссовское ядро $\Rightarrow$ гладкая аппроксимация

Ширина окна существенно влияет на точность аппроксимации

Выбор ядра K и ширины окна h

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

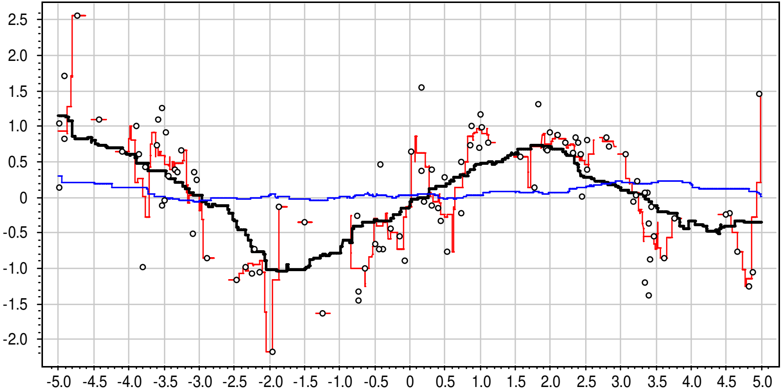


Треугольное ядро $\Rightarrow$ кусочно-линейная аппроксимация

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 $K(r)$
  — 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:

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

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 — неухудшение качества модели:

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

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

$z_j(x) = K\bigl(\frac{1}{h_j}\rho(x, x_j)\bigr)$, $j = 1, \ldots, \ell$ — признаки объекта x

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

$$a(x, w) = \operatorname{sign} \sum_{i=1}^{\ell} w_i y_i\, K\bigl(\tfrac{1}{h_i}\rho(x, x_i)\bigr) = \operatorname{sign}\langle w, y \otimes z(x)\rangle;$$

$$\sum_{i=1}^{\ell} L\bigl(y_i \langle w, y \otimes z(x_i)\rangle\bigr) + \tau \sum_{i=1}^{\ell} |w_i| \to \min_w;$$

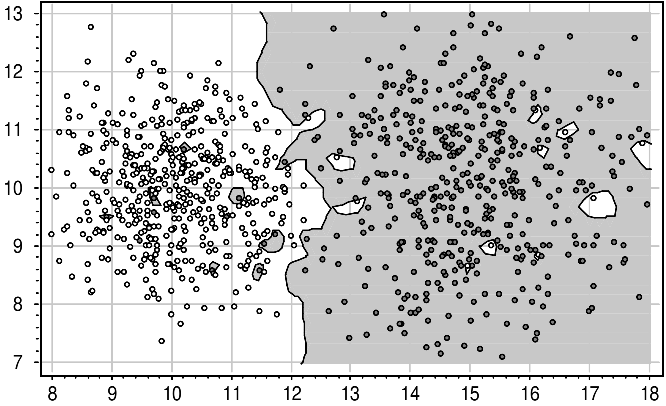
Линейная модель регрессии, $Y = \mathbb{R}$, обучение методом НК:

$$a(x, w) = \sum_{i=1}^{\ell} w_i y_i\, K\bigl(\tfrac{1}{h_i}\rho(x, x_i)\bigr) = \langle w, y \otimes z(x)\rangle$$

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

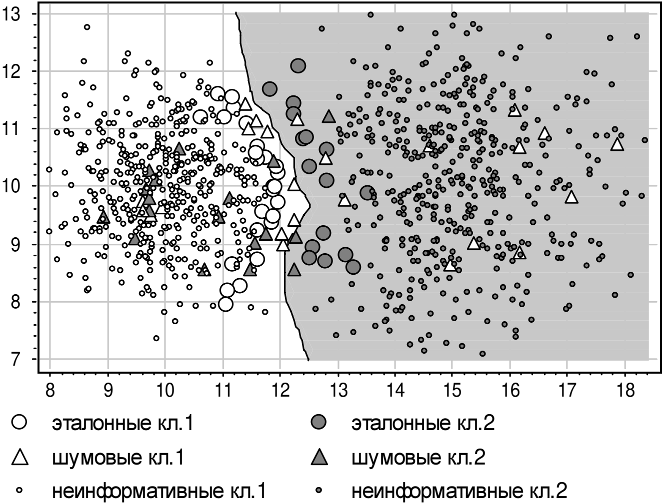
Чем больше τ, тем меньше остаётся эталонов

Пример. Отбор эталонов в задаче бинарной классификации



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

Пример. Отбор эталонов в задаче бинарной классификации



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):

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

Локальная непараметрическая оценка Парзена-Розенблатта

Эмпирическая оценка плотности по окну ширины h:

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

Обобщение: *оценка Парзена-Розенблатта* по окну ширины h:

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

где $K(r)$ — *ядро*, удовлетворяющее требованиям:
— чётная функция;
— нормированная функция: $\int K(r)\, dr = 1$;
— невозрастающая при $r > 0$, неотрицательная функция.

В частности, при $K(r) = \frac{1}{2}\big[|r| < 1\big]$ имеем эмпирическую оценку.

Обоснование оценки Парзена-Розенблатта

Другое название — Kernel Density Estimate (KDE)

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

Пусть *выполнены следующие условия:*

1) X^ℓ — простая выборка из распределения $p(x)$;
2) ядро $K(z)$ непрерывно и ограничено: $\int_X K^2(z)\,dz < \infty$;
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 $O(\ell^{-2/5})$.

А как быть в многомерном случае, когда $x_i \in \mathbb{R}^n$?

Два варианта обобщения на многомерный случай

1. Если объекты описываются n признаками $f_j\colon X \to \mathbb{R}$:

$$\hat{p}_{h_1\ldots h_n}(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \prod_{j=1}^{n} \frac{1}{h_j} K\left(\frac{f_j(x) - f_j(x_i)}{h_j}\right)$$

2. Если на X задана функция расстояния $\rho(x, x')$:

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

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

Сферическое гауссовское ядро — частный случай обоих:

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

Выбор ядра почти не влияет на качество восстановления

Функционал качества восстановления плотности:

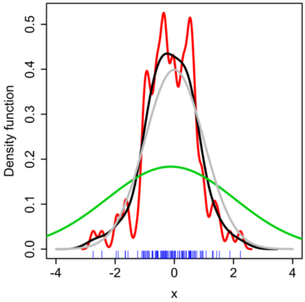
$$J(K) = \int_{-\infty}^{+\infty} \mathsf{E}\big(\hat{p}_h(x) - p(x)\big)^2\, dx.$$

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

Зависимость оценки плотности от ширины окна

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

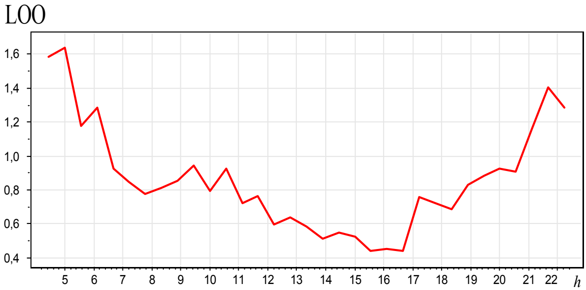
- Качество восстановления плотности существенно зависит от ширины окна h, но слабо зависит от вида ядра K
- При неоднородности локальных сгущений плотности можно задавать $h_k(x) = \rho\bigl(x, x^{(k+1)}\bigr)$, где k — число соседей

Выбор ширины окна

Скользящий контроль *Leave One Out* для оценки плотности:

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

Типичный вид зависимости $\mathrm{LOO}(h)$ или $\mathrm{LOO}(k)$:



Ретроспектива: (непара)метрические методы анализа данных

Классификация. Метод парзеновского окна:

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

Регрессия. Метод *ядерного сглаживания* Надарая–Ватсона:

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

Восстановление плотности. Метод Парзена–Розенблатта:

$$\hat{p}_h(x; X^\ell) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\Bigl(\frac{\rho(x, x_i)}{h}\Bigr)$$

Постановка задачи кластеризации

Дано:

X — пространство объектов;
$X^\ell = \{x_1, \dots, x_\ell\}$ — обучающая выборка;
$\rho\colon X \times X \to [0, \infty)$ — 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.

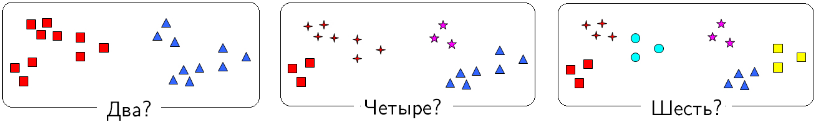
Это задача *обучения без учителя* (unsupervised learning).

Некорректность задачи кластеризации

Решение задачи кластеризации принципиально неоднозначно:

- точной постановки задачи кластеризации нет;
- существует много критериев качества кластеризации;
- существует много эвристических методов кластеризации;
- число кластеров $|Y|$, как правило, заранее не известно;
- результат кластеризации сильно зависит от метрики ρ, выбор которой также является эвристикой.

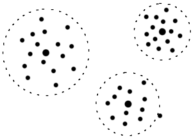
Пример: сколько здесь кластеров?



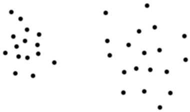
Цели кластеризации

- **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 представителю 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



внутрикластерные расстояния меньше межкластерных

ленточные кластеры

Типы кластерных структур

перемычки между кластерами

разреженный фон
из нетипичных объектов

перекрывающиеся кластеры

Типы кластерных структур

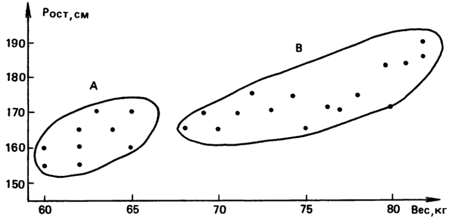
кластеры могут вообще отсутствовать

а это вообще не кластеры
и на практике такое не встречается

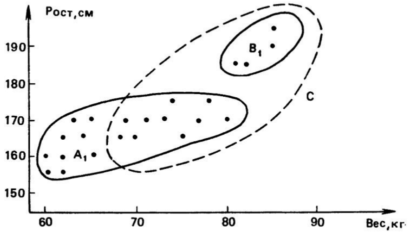
- Каждый метод кластеризации имеет свои ограничения и выделяет кластеры лишь некоторых типов.
- Понятие «тип кластерной структуры» зависит от метода и также не имеет формального определения.

Проблема чувствительности к выбору метрики

Результат зависит от нормировки признаков:



A — студентки,
B — студенты



после перенормировки
(сжали ось «вес» вдвое)

Задача кластеризации (clustering)

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

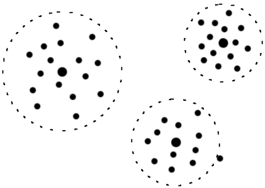
Найти:

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

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

Критерий: минимум суммы внутрикластерных расстояний

$$\sum_{i=1}^{\ell} \|x_i - \mu_{a_i}\|^2 \to \min_{\{a_i\}, \{\mu_j\}}$$



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

$$\|x - \mu_j\|^2 = \sum_{d=1}^{n} \bigl(f_d(x) - \mu_{jd}\bigr)^2$$

Метод K-средних (K-means) для кластеризации

Минимизация суммы квадратов внутрикластерных расстояний:

$$\sum_{i=1}^{\ell} \|x_i - \mu_{a_i}\|^2 \to \min_{\{a_i\},\,\{\mu_a\}}, \qquad \|x_i - \mu_a\|^2 = \sum_{j=1}^{n} \bigl(f_j(x_i) - \mu_{aj}\bigr)^2$$

Алгоритм Ллойда

вход: X^ℓ, $K = |Y|$; **выход:** центры кластеров μ_a, $a \in Y$;

$\mu_a :=$ начальное приближение центров, для всех $a \in Y$;

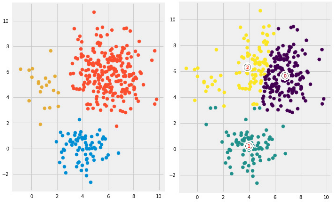
повторять

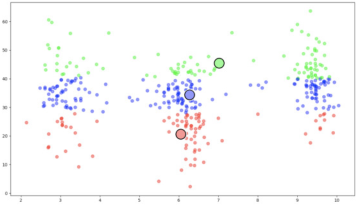
> отнести каждый x_i к ближайшему центру:
>
> $a_i := \arg\min\limits_{a \in Y} \|x_i - \mu_a\|$, $i = 1, \ldots, \ell$;
>
> вычислить новые положения центров:
>
> $$\mu_a := \frac{\sum_{i=1}^{\ell} [a_i = a] x_i}{\sum_{i=1}^{\ell} [a_i = a]}, \quad a \in Y;$$

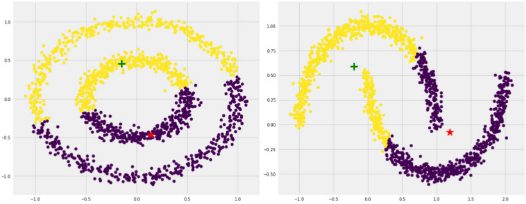
пока a_i не перестанут изменяться;

Примеры неудачной кластеризации k-means

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





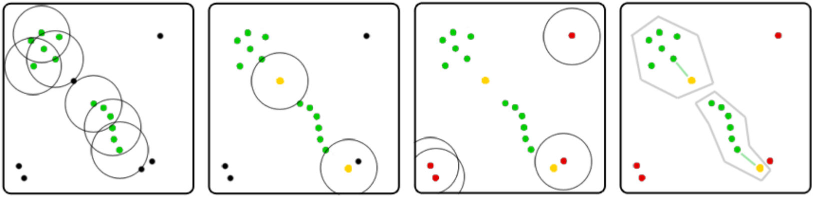


Алгоритм кластеризации DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

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

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



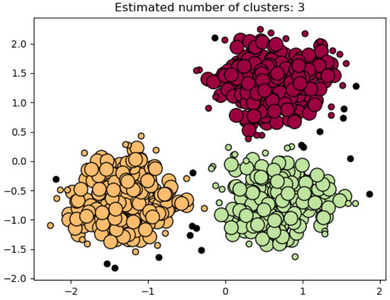
Ester, Kriegel, Sander, Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD-1996.

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.



Многомерное шкалирование (multidimensional scaling, MDS)

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

Найти: векторные представления вершин $z_i \in \mathbb{R}^d$, так, чтобы близкие (по графу) вершины имели близкие векторы.

Критерий стресса (stress):

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

где $\rho(z_i, z_j) = \|z_i - z_j\|$ — обычно евклидово расстояние,

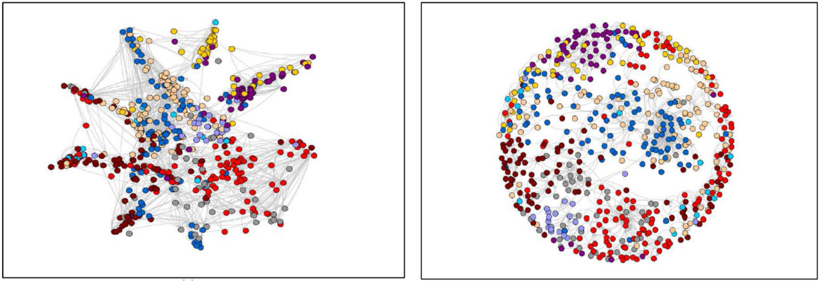
Обычно решается методом стохастического градиента (SG).

Вопрос: как лучше задавать веса $w(R_{ij}) = 1$? R_{ij}? $\frac{1}{R_{ij}}$?

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

Многомерное шкалирование для визуализации данных

При $d = 2$ осуществляется проекция выборки на плоскость



- Используется для визуализации кластерных структур
- Форму облака точек можно настраивать весами и метрикой
- Недостаток — искажения неизбежны
- Наиболее популярная разновидность метода — t-SNE

Laurens van der Maaten, Geoffrey Hinton. Visualizing data using t-SNE. 2008

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
  — набор 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, модельные предположения закладываются в метрику