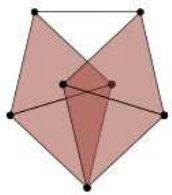
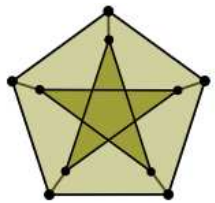
Теория и практика машинного обучения
• Лекция 2 •
Метрические алгоритмы классификации

Воронцов Константин Вячеславович
МФТИ • МГУ • ВШЭ • ВЦ РАН • Яндекс • FORECSYS



Комбинаторика и алгоритмы для школьников



• Летняя школа — 2015 •
19 августа 2015

Содержание

Напоминания

Задача восстановления зависимости $y^*: X \to Y$
по точкам *обучающей выборки* (x_i, y_i), $i = 1, \ldots, \ell$.

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

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

В задачах классификации Y — конечное множество.

В некоторых задачах трудно придумать хорошие признаки,
но легко оценивать степень сходства объектов друг с другом.

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 $\rho\colon X \times X \to [0, \infty)$.

Пример. Евклидово расстояние и его обобщение:

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

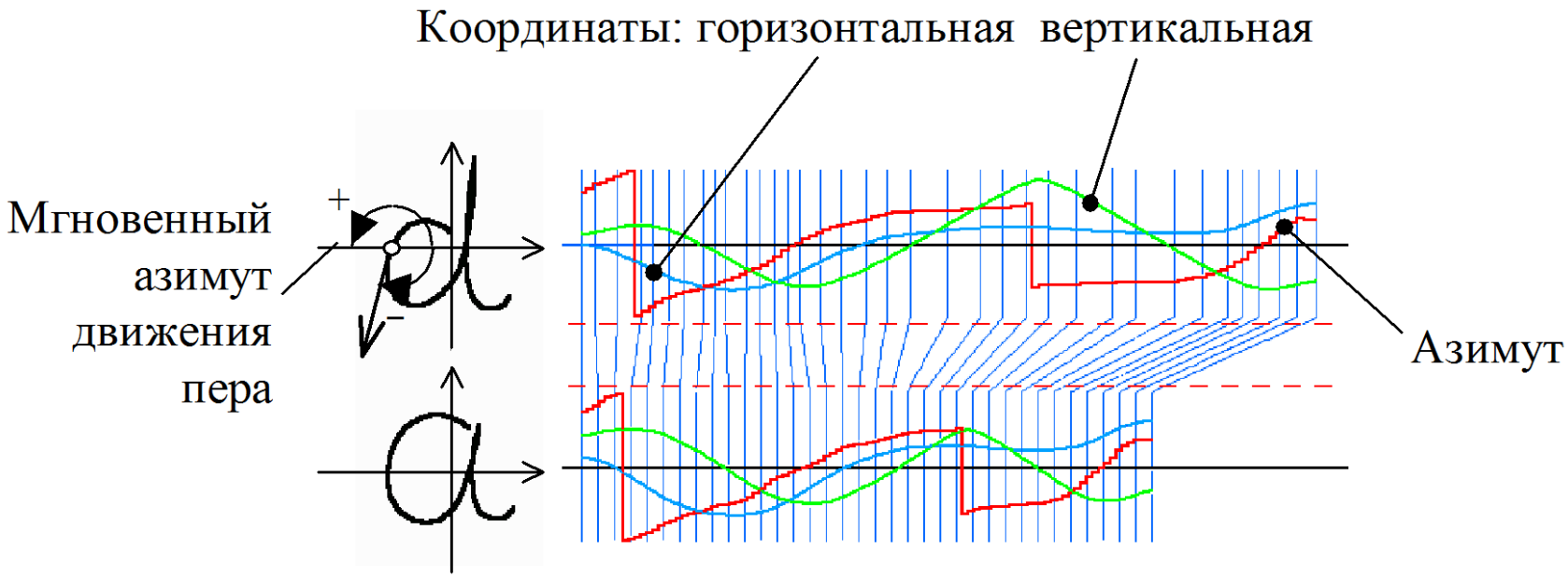
$x = (x^1, \ldots, x^n)$ — вектор признаков объекта x,
$x_i = (x_i^1, \ldots, x_i^n)$ — вектор признаков объекта x_i.

Ещё примеры расстояний:

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

```
CTGGGCTAAAAGGTCCCTTAGCC..TTTAGAAAAA.GGGCCATTAGGAAATTGCAA
CTGGGACTAAA....CCTTAGCCTATTTACAAAAATGGGCCATTAGG...TTGCAA
```

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



Обобщённый метрический классификатор

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:

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

$w(i, x)$ — 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 q ближайших соседей (k nearest neighbors, kNN)

$w(i,x) = [i \leqslant q]$.
$w(i,x) = [i \leqslant 1]$ — метод ближайшего соседа.

Преимущества:

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

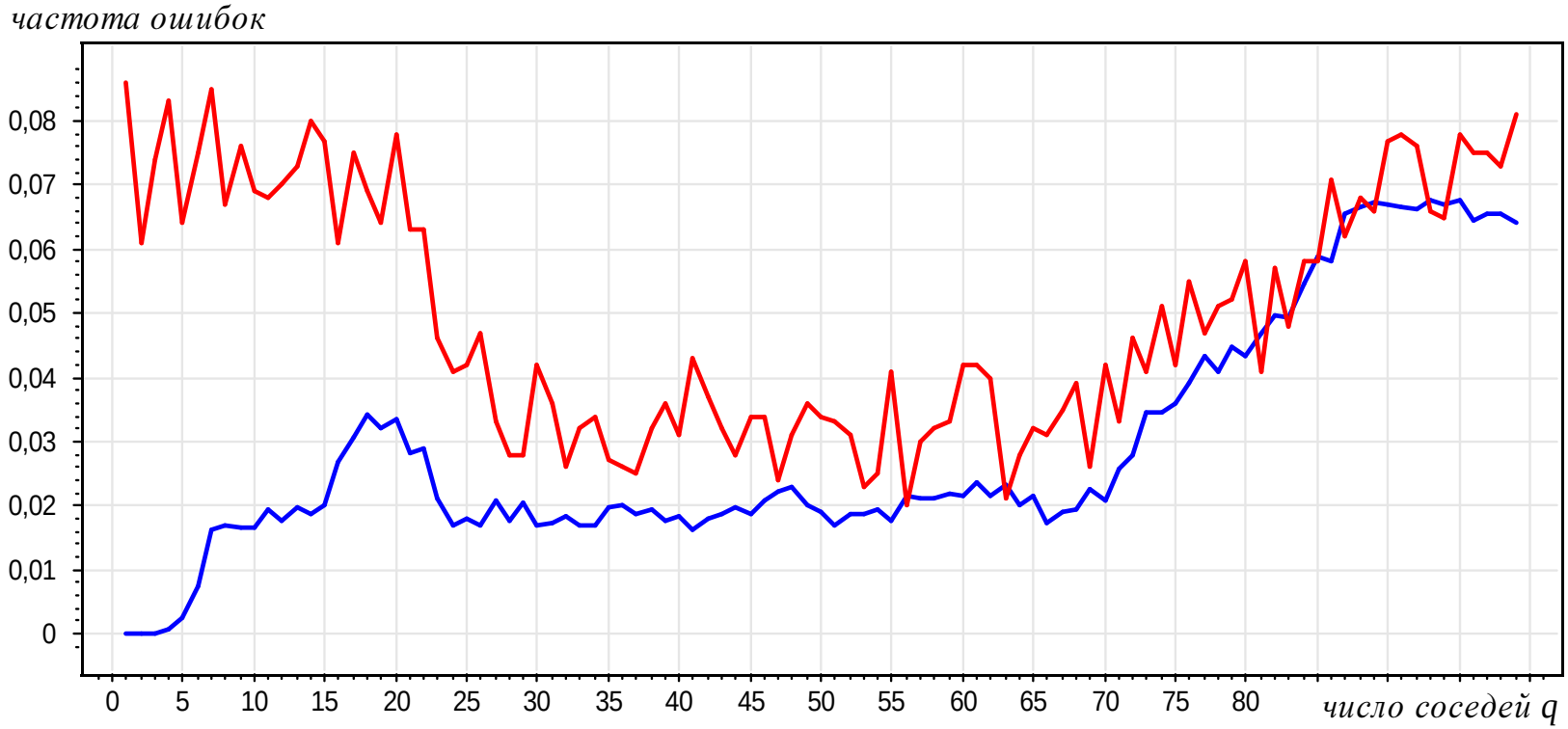
$$\mathrm{LOO}(q, X^\ell) = \sum_{i=1}^{\ell} \Big[a\big(x_i; X^\ell \setminus \{x_i\}, q\big) \neq y_i \Big] \to \min_q .$$

Проблема:

- неоднозначность классификации при $\Gamma_y(x) = \Gamma_s(x), \;\; y \neq s$.

Пример зависимости LOO от числа соседей

Пример. Задача Iris, усреднение по 50 случайным разбиениям



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

В реальных задачах минимум редко бывает при $q = 1$.

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 *фиксированной ширины:*

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

Метод парзеновского окна *переменной ширины:*

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

Оптимизация параметров — по критерию LOO:

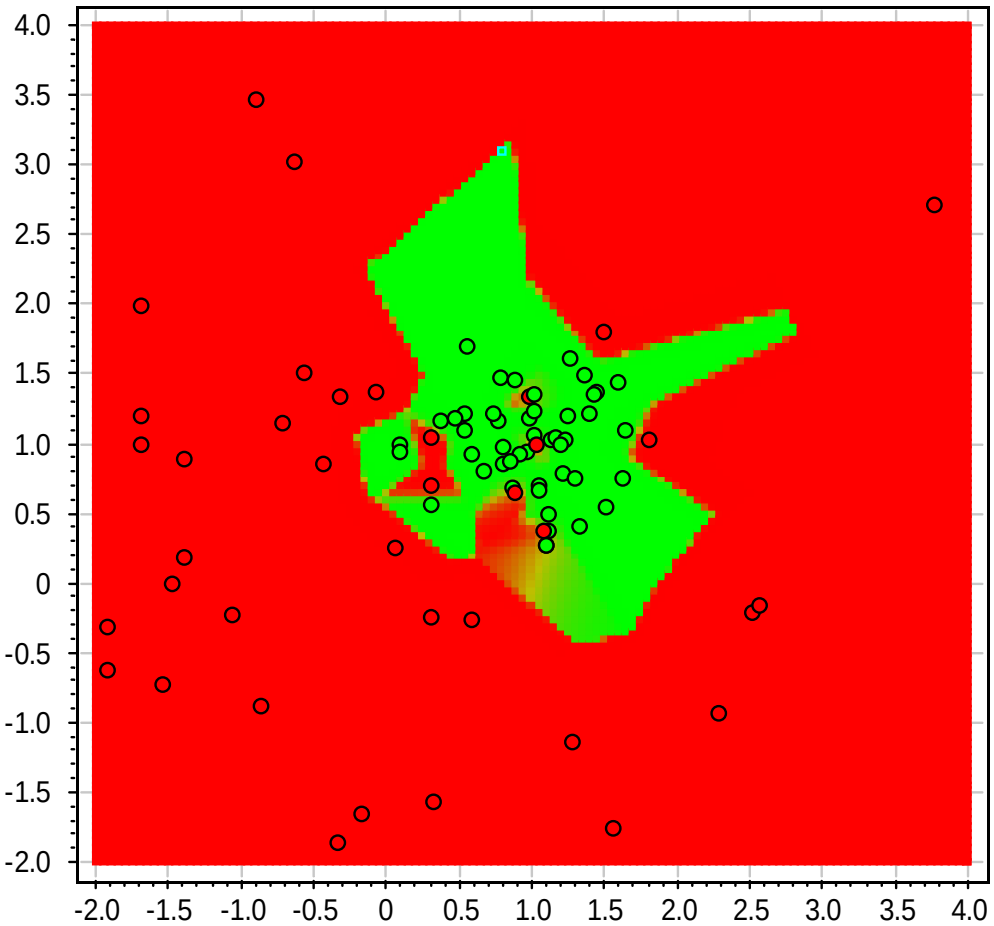
- выбор ширины окна h или числа соседей q
- выбор ядра K слабо влияет на качество классификации

Парзеновское окно фиксированной ширины h

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

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

$h = 0.05$

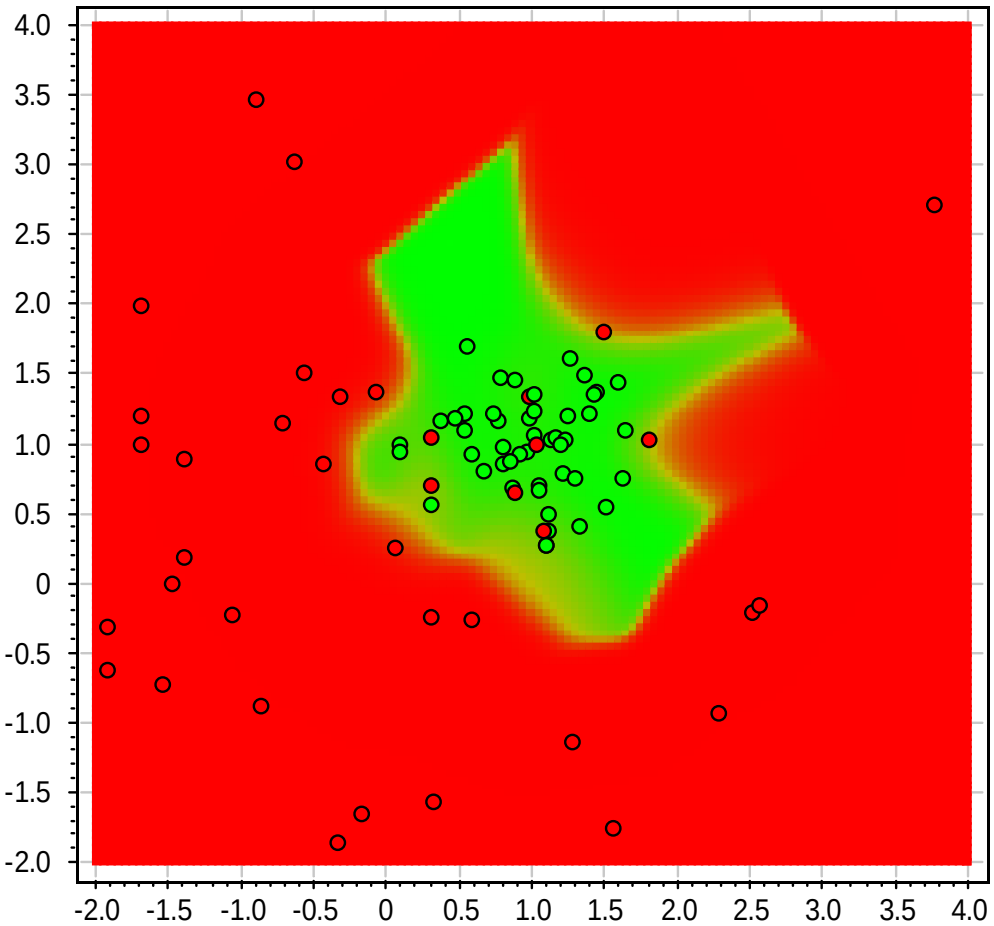


Парзеновское окно фиксированной ширины h

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

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

$h = 0.2$

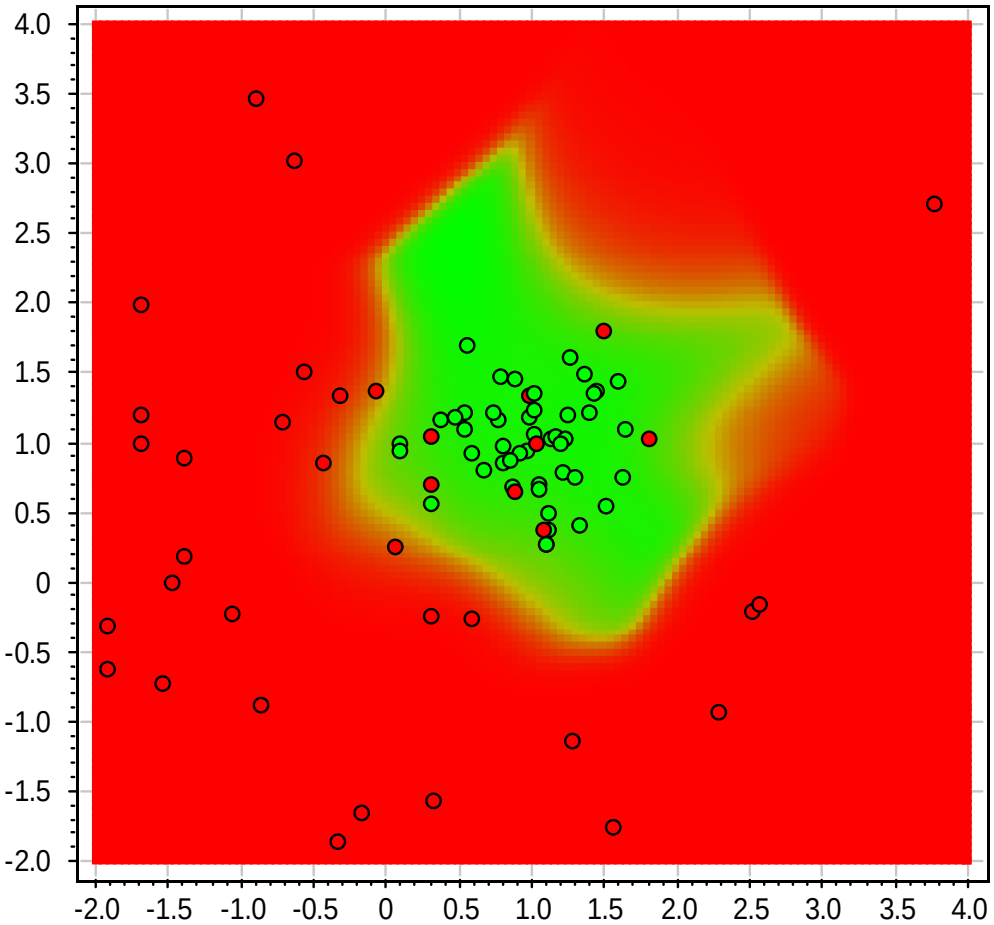


Парзеновское окно фиксированной ширины h

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

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

$h = 0.3$

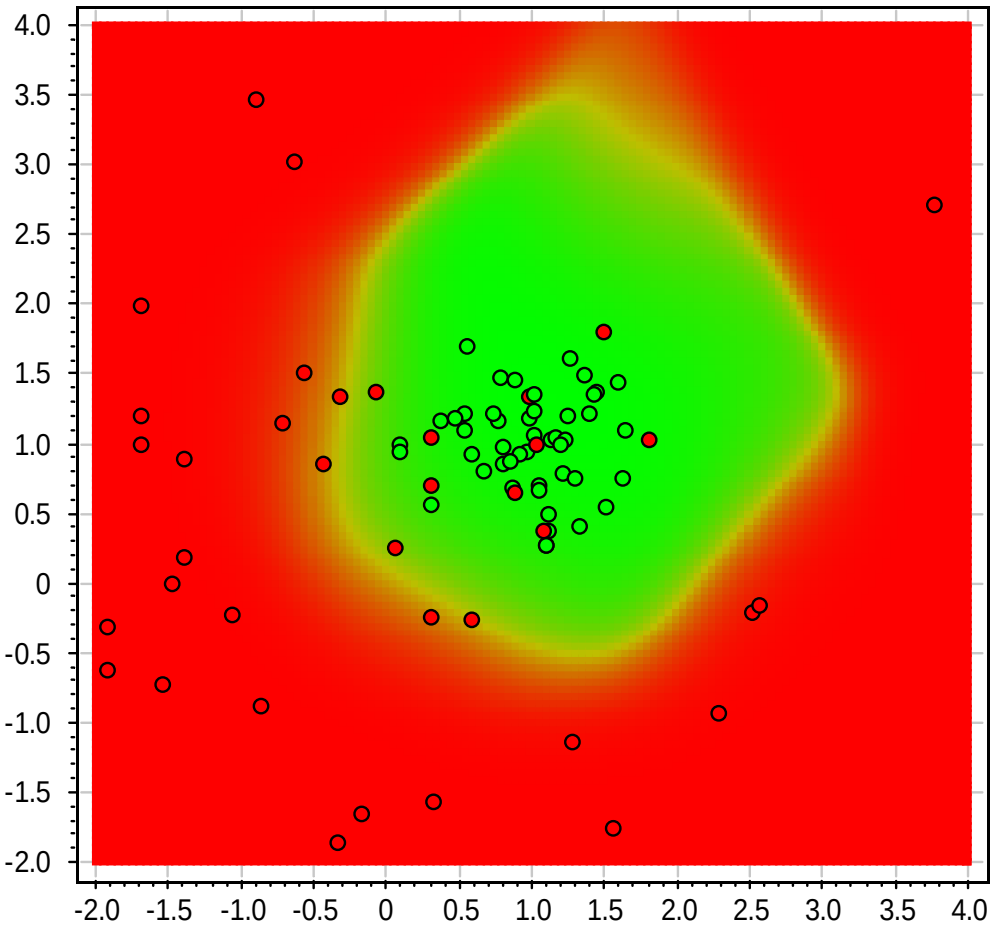


Парзеновское окно фиксированной ширины h

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

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

$h = 0.5$

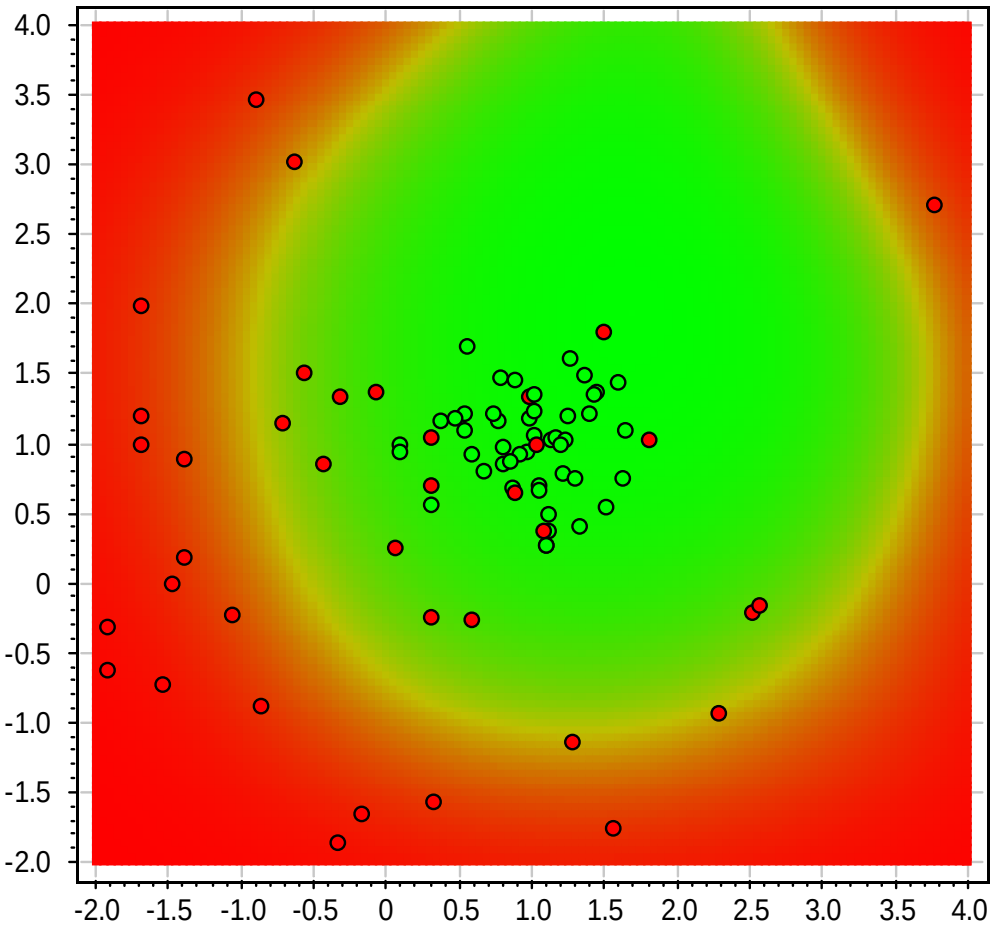


Парзеновское окно фиксированной ширины h

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

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

$h = 1.0$

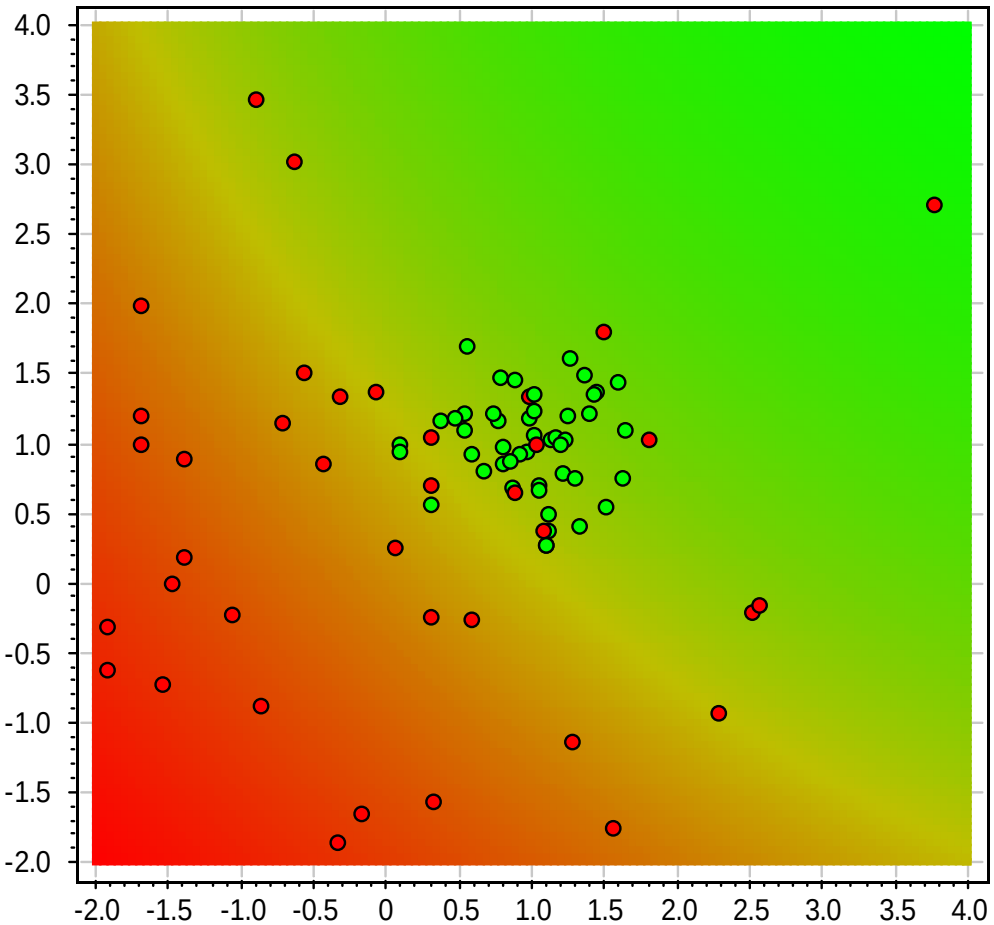


Парзеновское окно фиксированной ширины h

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

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

$h = 5.0$



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

Определения и обозначения

$\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечное *генеральное множество* объектов.

Пусть все C_L^ℓ разбиений $X \sqcup \bar{X} = \mathbb{X}$ равновероятны, где
X — обучающая выборка объёма ℓ,
$\bar{X}$ — контрольная выборка объёма k, $L = \ell + k$.

a_X — классификатор, полученный в результате обучения по X,
$\nu(a_X, \bar{X})$ — частота его ошибок на контрольной выборке $\bar{X}$:

$$\nu(a_X, \bar{X}) = \frac{1}{k} \sum_{x_i \in \bar{X}} \big[a_X(x_i) \neq y_i\big].$$

Критерий обобщающей способности Complete Cross-Validation:

$$\mathrm{CCV} = \frac{1}{C_L^\ell} \sum_{X \subset \mathbb{X}} \nu\big(a_X, \bar{X}\big) = \mathsf{E}\, \nu\big(a_X, \bar{X}\big).$$

Классификатор ближайшего соседа (NN, nearest neighbor)

Определение

Профиль компактности $K(m)$ — это доля объектов x_i, у которых m-й сосед $x_i^{(m)}$ принадлежит другому классу:

$$K(m) = \frac{1}{L}\sum_{i=1}^{L}\left[y_i^{(m)} \neq y_i\right]; \quad m = 1, \ldots, L-1,$$

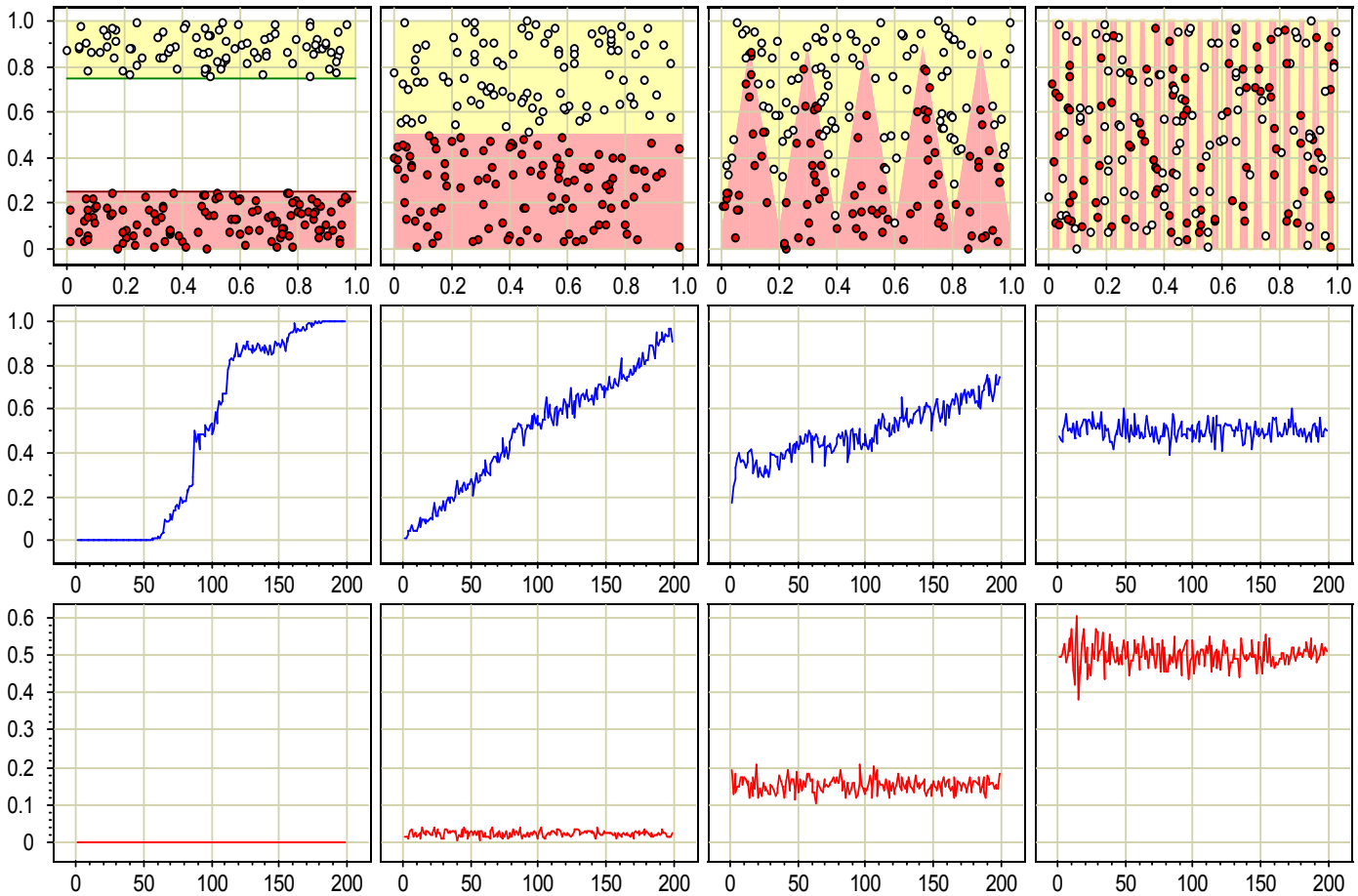
Теорема

Для классификатора ближайшего соседа

$$\text{CCV} = \sum_{m=1}^{k} K(m)\frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^{\ell}}.$$

Профили компактности для серии модельных задач

средний ряд: профили компактности,
нижний ряд: зависимость CCV от длины контроля $k = |\bar{X}|$.



Доказательство

Основная идея — просуммировать по разбиениям аналитически:

$$\mathrm{CCV} = \mathsf{E}\nu(a_X, \bar{X}) = \frac{1}{C_L^\ell} \sum_{X \subset \mathbb{X}} \frac{1}{k} \sum_{i=1}^{L} \big[x_i \in \bar{X}\big] \big[a_X(x_i) \neq y_i\big] =$$

$$= \frac{1}{C_L^\ell} \sum_{X \subset \mathbb{X}} \frac{1}{k} \sum_{i=1}^{L} \sum_{m=1}^{k} \big[y_i^{(m)} \neq y_i\big] \big[x_i^{(m)} \in X\big] \big[x_i, x_i^{(1)}, \ldots, x_i^{(m-1)} \in \bar{X}\big] =$$

$$= \sum_{m=1}^{k} \sum_{i=1}^{L} \frac{\big[y_i^{(m)} \neq y_i\big]}{k C_L^\ell} \sum_{X \subset \mathbb{X}} \big[x_i^{(m)} \in X\big] \big[x_i, x_i^{(1)}, \ldots, x_i^{(m-1)} \in \bar{X}\big] =$$

$$= \sum_{m=1}^{k} \sum_{i=1}^{L} \frac{\big[y_i^{(m)} \neq y_i\big]}{k C_L^\ell} C_{L-1-m}^{\ell-1} = \sum_{m=1}^{k} \underbrace{\frac{1}{L} \sum_{i=1}^{L} \big[y_i^{(m)} \neq y_i\big]}_{K(m)} \frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^{\ell}}.$$

■

Некоторые свойства профиля компактности и оценки CCV

- $\gamma_m = \frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^{\ell}} \to 0$ быстрее геометрической прогрессии:
$$\frac{\gamma_{m+1}}{\gamma_m} = 1 - \frac{\ell - 1}{L - 1 - m} < \frac{k}{L-1}.$$
- CCV тем меньше, чем меньше $K(m)$ при малых m, то есть чем чаще близкие объекты лежат в одном классе.
- При малых k

$$k = 1: \quad \text{CCV} = K(1) = \text{LOO};$$
$$k = 2: \quad \text{CCV} = K(1)\tfrac{\ell}{\ell+1} + K(2)\tfrac{1}{\ell+1};$$
$$k = 3: \quad \text{CCV} = K(1)\tfrac{\ell}{\ell+2} + K(2)\tfrac{2\ell}{(\ell+1)(\ell+2)} + K(3)\tfrac{2}{(\ell+1)(\ell+2)}.$$

- Профиль компактности вычисляется за $O(\ell^2)$ операций.
- Минимизация CCV позволяет отбирать эталонные объекты.

Задача отбора эталонов $\Omega \subseteq \mathbb{X}$ (prototype learning)

$a_{X|\Omega}$ — классификатор ближайшего соседа, которому разрешено использовать только объекты из $\Omega \subseteq \mathbb{X}$.
Требуется найти оптимальное подмножество эталонов Ω.

Определение (*профиль компактности* относительно Ω)

$$K(m,\Omega) = \frac{1}{L}\sum_{i=1}^{L}\left[y_i^{(m|\Omega)} \neq y_i\right]; \quad m = 1,\ldots,|\Omega|.$$

где $x_i^{(m|\Omega)}$ — m-й сосед объекта x_i из множества Ω;

Теорема

$$\mathrm{CCV}(\Omega) = \sum_{i=1}^{L} \underbrace{\sum_{m=1}^{k}\left[y_i^{(m|\Omega)} \neq y_i\right]\frac{C_{L-1-m}^{\ell-1}}{LC_{L-1}^{\ell}}}_{T(i,\Omega)\text{ — вклад объекта } x_i \text{ в CCV}}.$$

Жадные алгоритмы отбора эталонов

Задача: найти Ω: $\mathrm{CCV}(\Omega) \to \min$.

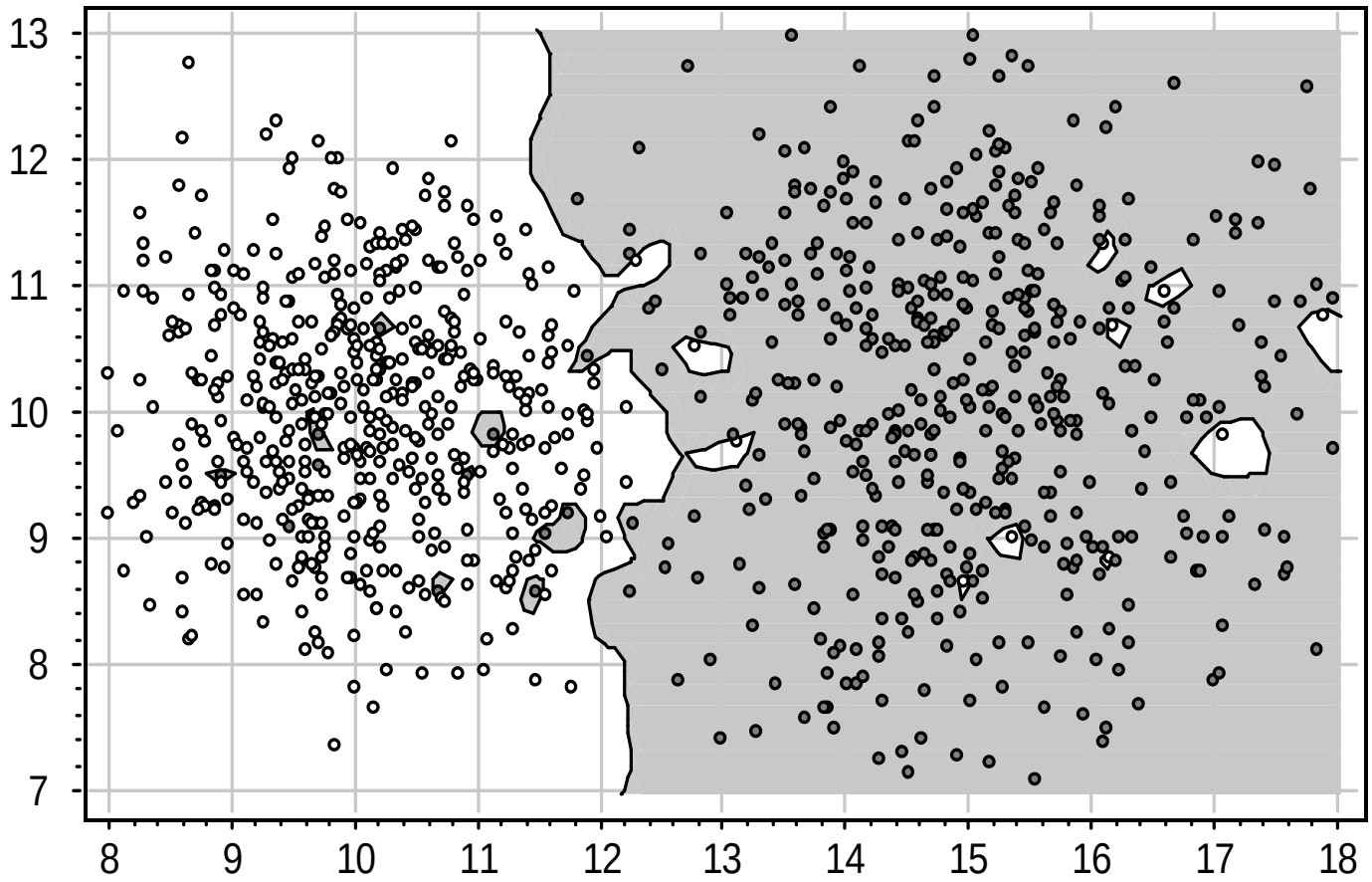
Жадный алгоритм удаления не-эталонов:

$\Omega := \mathbb{X}$;
повторять
  найти $x \in \Omega$: $\mathrm{CCV}(\Omega \setminus \{x\}) \to \min$;
  $\Omega := \Omega \setminus \{x\}$;
пока CCV уменьшается или почти не увеличивается;

Жадный алгоритм добавления эталонов:

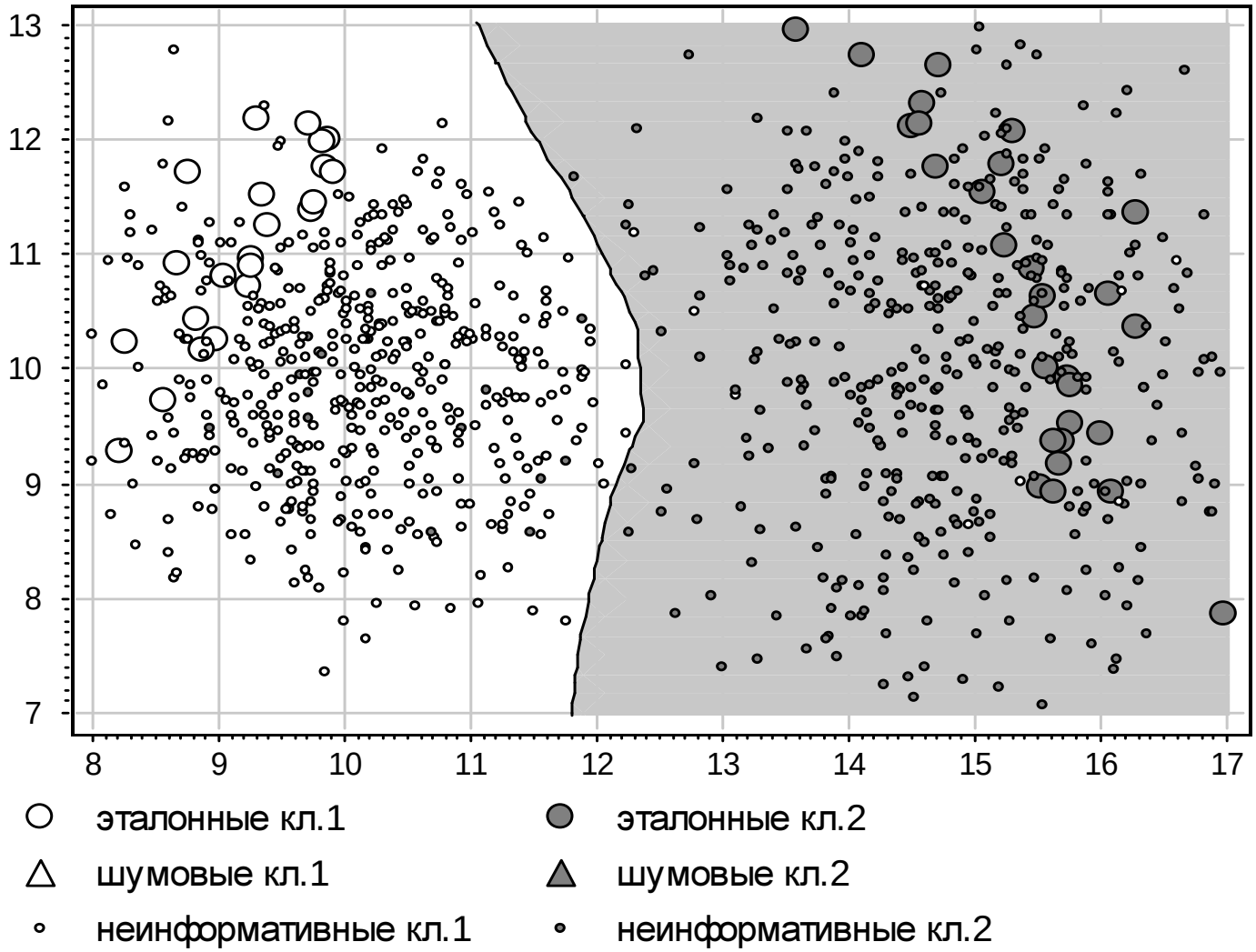
$\Omega :=$ {по одному объекту от каждого класса};
повторять
  найти $x \in \mathbb{X} \setminus \Omega$: $\mathrm{CCV}(\Omega \cup \{x\}) \to \min$;
  $\Omega := \Omega \cup \{x\}$;
пока CCV уменьшается;

Модельные данные

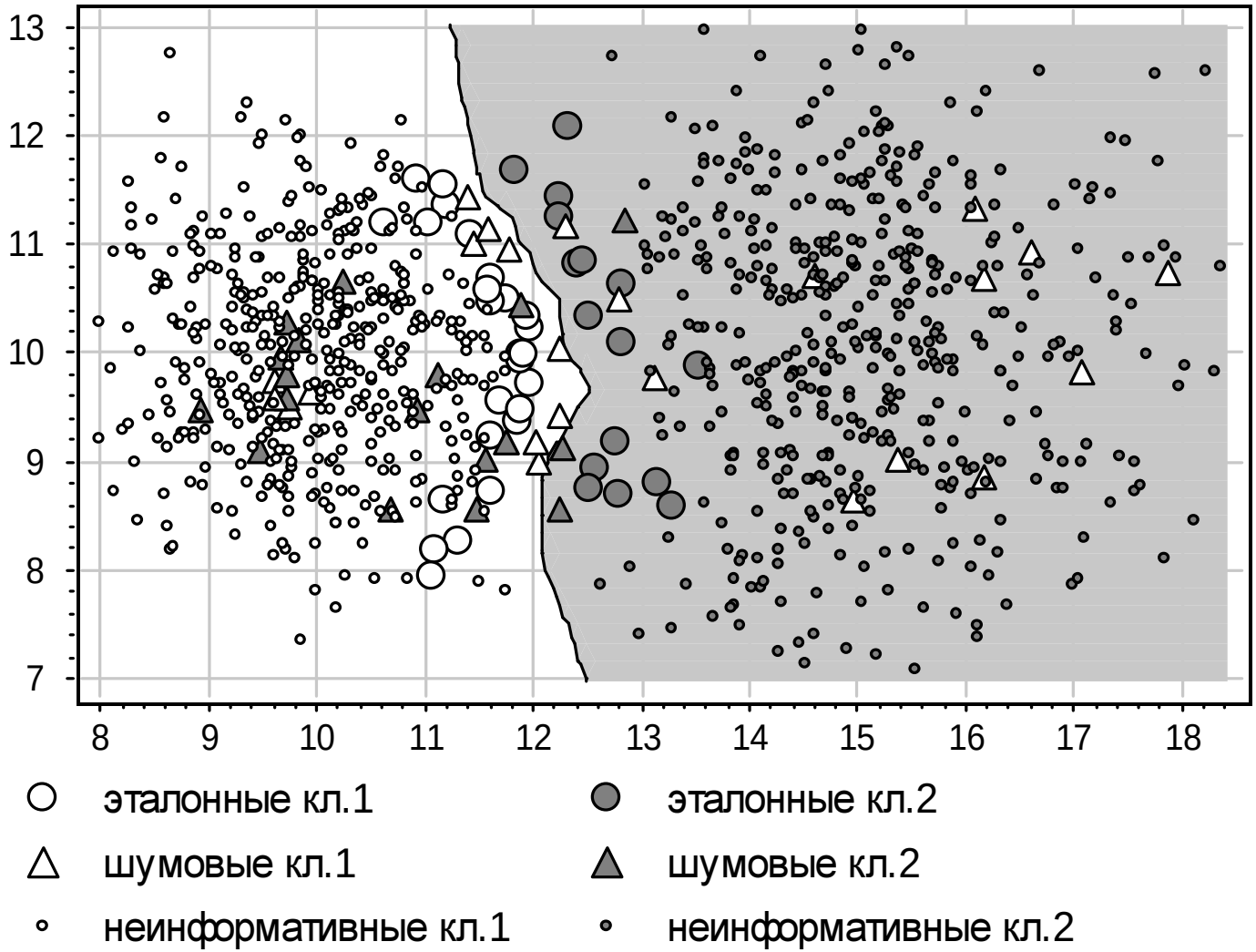


Модельная задача классификации: 1000 объектов, метод NN.

Жадное добавление эталонных объектов

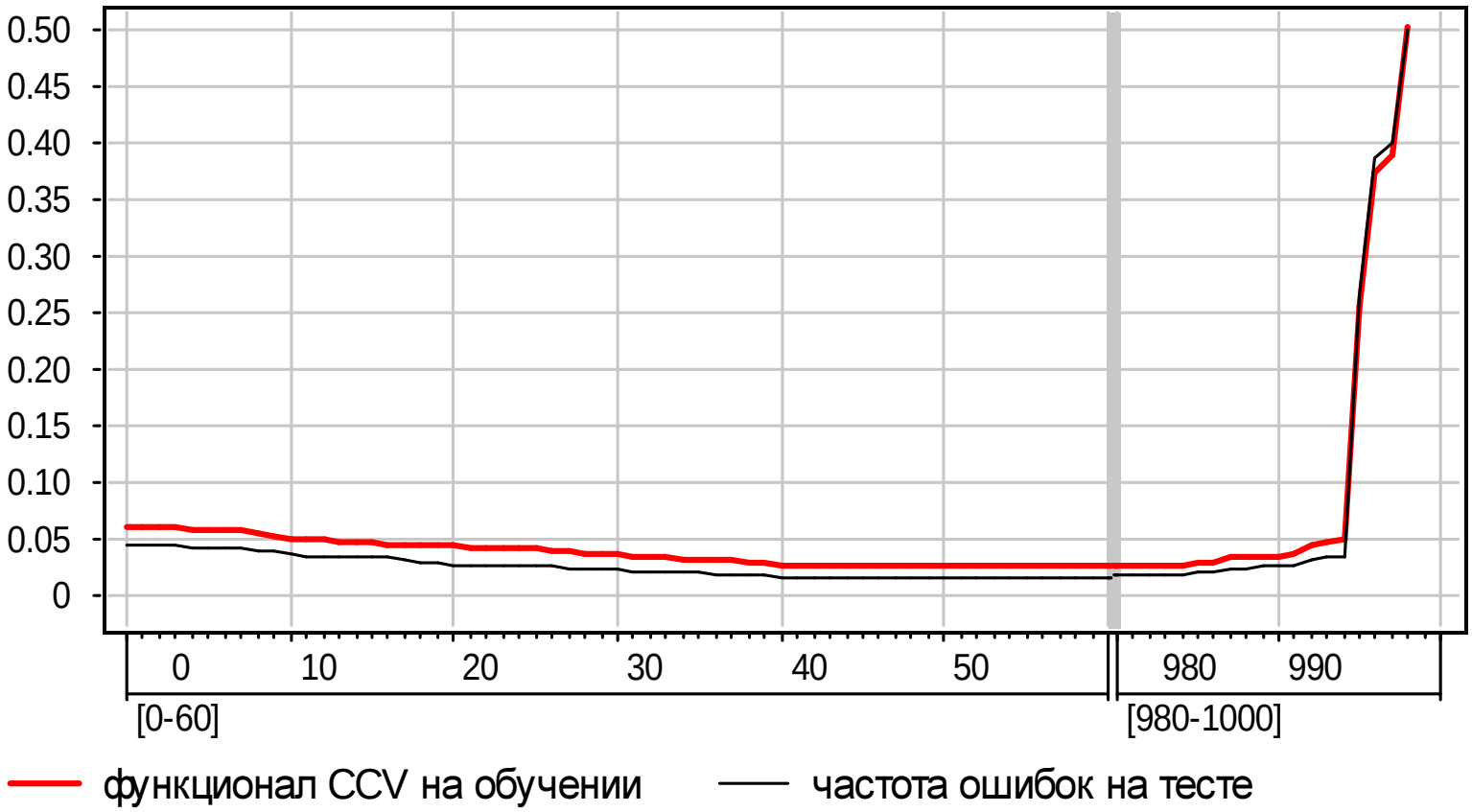


Жадное удаление не-эталонных объектов



Жадное удаление не-эталонных объектов

Зависимость CCV от числа удалённых неэталонных объектов.



Вывод: при отборе эталонов переобучения нет.

Резюме

- Метрические модели классификации определяются функцией расстояния $\rho(x, x_i)$ и функцией близости $w(i, x)$.
- Комбинаторная оценка CCV формализует гипотезу компактности, которая долгое время считалась эвристикой.
- Имеется обобщение оценки CCV для kNN.
- Обобщения для произвольной $w(i, x)$ пока нет.
- Оценку CCV можно использовать для отбора эталонов и для жадной оптимизации весов признаков в метрике.

Воронцов Константин Вячеславович
voron@forecsys.ru

www.MachineLearning.ru • Участник:Vokov

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