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:

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

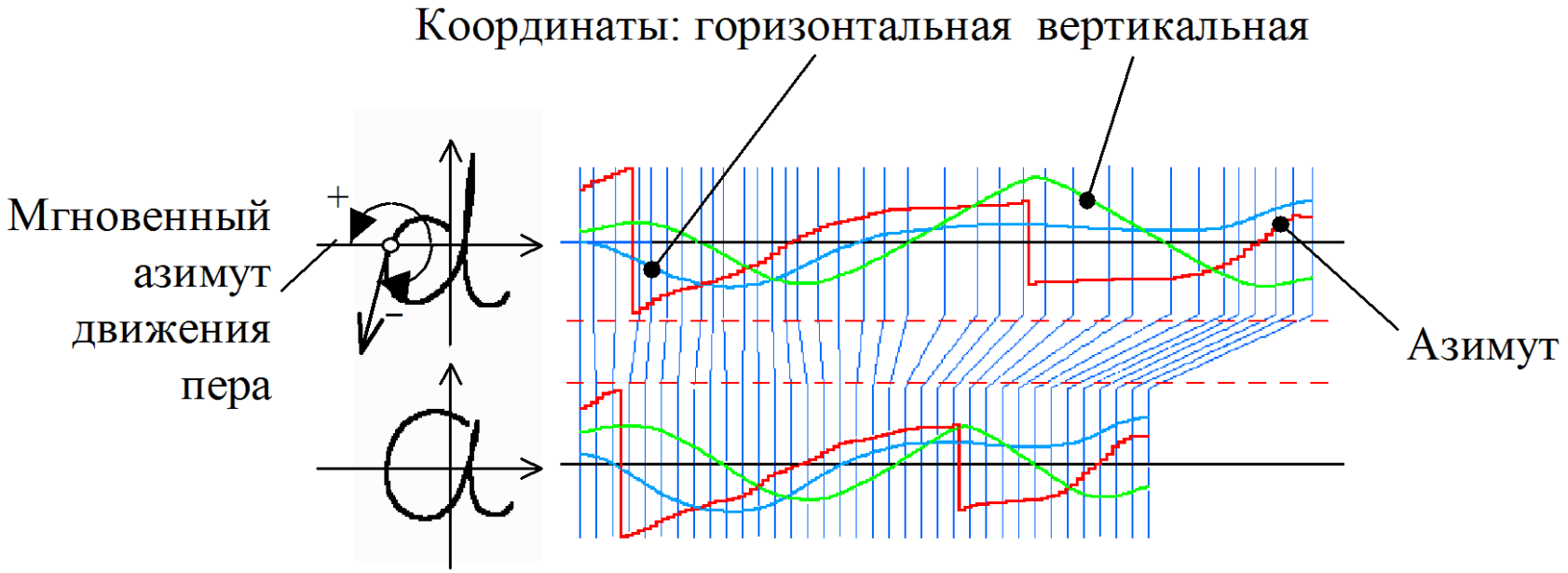
$x = (x^1, \ldots, x^n)$ — вектор признаков объекта x,
$x_i = (x_i^1, \ldots, x_i^n)$ — вектор признаков объекта x_i.

Ещё примеры расстояний:

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

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

— **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.

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

Преимущества:

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

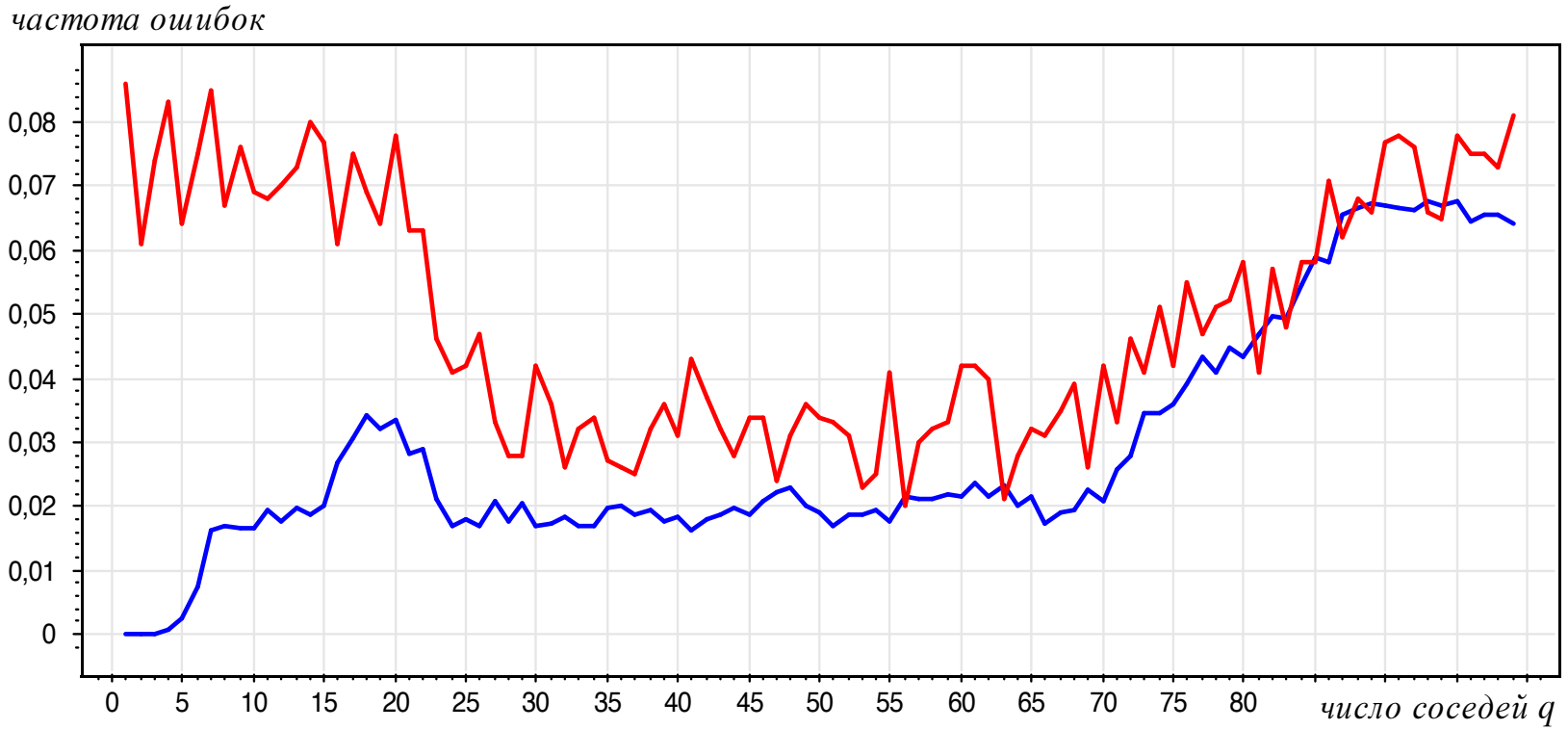
$$\mathrm{LOO}(k, X^\ell) = \sum_{i=1}^{\ell} \Big[a\big(x_i; X^\ell \setminus \{x_i\}, k\big) \neq y_i \Big] \to \min_k .$$

Проблемы:

- возможны ситуации, когда классификация не однозначна: $\Gamma_y(x) = \Gamma_s(x)$ для пары классов $y \neq s$
- учитываются не значения расстояний, а только их ранги

Пример зависимости LOO от числа соседей

Пример. Задача Iris.



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

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