

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 $\rho: X \times X \rightarrow [0, \infty)$.



Пример. Евклидово расстояние и его обобщение:

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

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

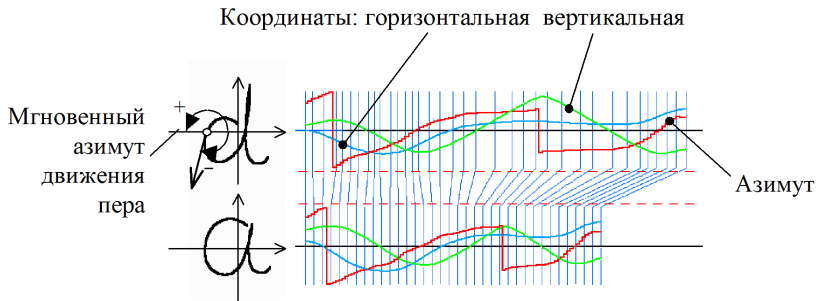
$x_i = (x_i^1, \dots, x_i^n)$ — вектор признаков объекта x_i .

Ещё примеры расстояний:

— между текстами (редакторское расстояние Левенштейна):

CTGGGCTAAAAGGTCCTTAGCC..TTTAGAAAAA.GGGCCATTAGGAAATTGC
CTGGGACTAAA....CCTTAGCCATTTTACAAAAATGGGCCATTAGG...TTGC

— между сигналами (энергия сжатий и растяжений):



Для произвольного $x \in X$ отранжируем объекты x_1, \dots, x_ℓ :

$$\rho(x, x^{(1)}) \leq \rho(x, x^{(2)}) \leq \dots \leq \rho(x, x^{(\ell)}),$$

$x^{(i)}$ — i -й сосед объекта x среди x_1, \dots, x_ℓ ;

$y^{(i)}$ — ответ на i -м соседе объекта x .

Метрический алгоритм классификации:

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

$w(i, x)$ — 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 .

$$w(i, x) = [i \leq k].$$

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

Преимущества:

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

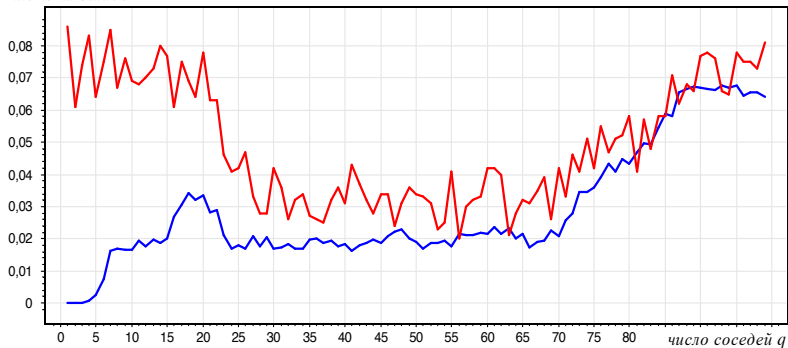
$$\text{LOO}(k, X^\ell) = \sum_{i=1}^{\ell} [a(x_i; X^\ell \setminus \{x_i\}, k) \neq y_i] \rightarrow \min_k.$$

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. Задача Iris.

частота ошибок



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

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