Метод k ближайших соседей

М. Кореневская, гр. 23641/3

18 октября 2017 г.

Стандартные типы задач

- Обучение с учителем
  - Классификация
  - Регрессия
  - Ранжирование
  - Прогнозирование
- Обучение без учителя
- Частичное обучение

  и др.

Метод k ближайших соседей использует расстояния (метрики) в пространстве объектов.

Гипотезы компактности и непрерывности

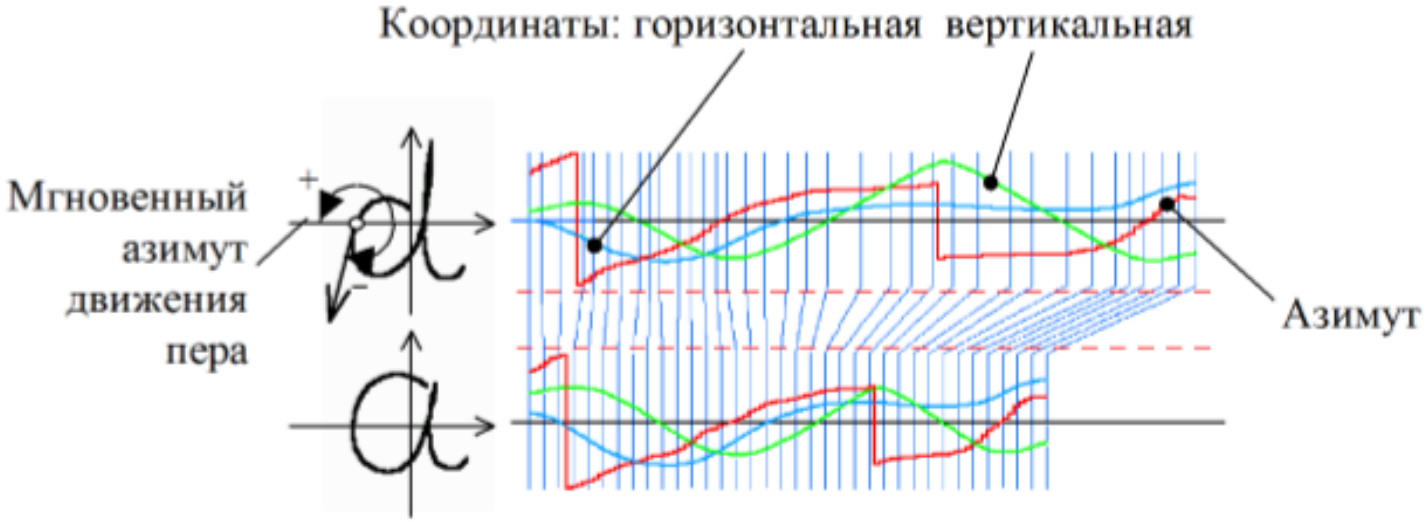
Гипотеза непрерывности (для регрессии):
близким объектам соответствуют близкие ответы
Гипотеза компактности (для классификации):
близкие объекты, как правило, лежат в одном классе
Формализация понятия "близости":
задана функция расстояния $\rho : X \times X \to [0, \infty)$.
Пример. Евклидово расстояние и его обобщение:

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

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

Другие примеры расстояний

- **между текстами** (редакционное расстояние Левенштейна):
  GCTAAAGGTCAGCC . .TTTAGAAA .GGGCCATTAGGAAATTGC
  GACTAA . . . . AGCCTATTTACAAATGGGCCATTAGG . . .TTGC
- **между сигналами** (энергия сжатий и растяжений):



Обобщенный метрический классификатор

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

$$\rho\left(x, x^{(1)}\right) \leq \rho\left(x, x^{(2)}\right) \leq \ldots \leq \rho\left(x, x^{(l)}\right),$$

$x^{(i)}$ — i-й сосед объекта x среди $x_1, \ldots, x_l$;
$y^{(i)}$ — ответ на i-м соседе объекта x.
Метрический алгоритм классификации:

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

$w\left(i, x\right)$ — **вес, оценка сходства объекта x с его i-м соседом, неотрицательная, не возрастающая по i.**
$\Gamma_y\left(x\right) = \sum_{i=1}^{l} \left[y^{(i)} = y\right] w\left(i, x\right)$ — **оценка близости объекта x к классу y.**

Метод k ближайших соседей (k nearest neighbors, kNN)

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

Преимущества:

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

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

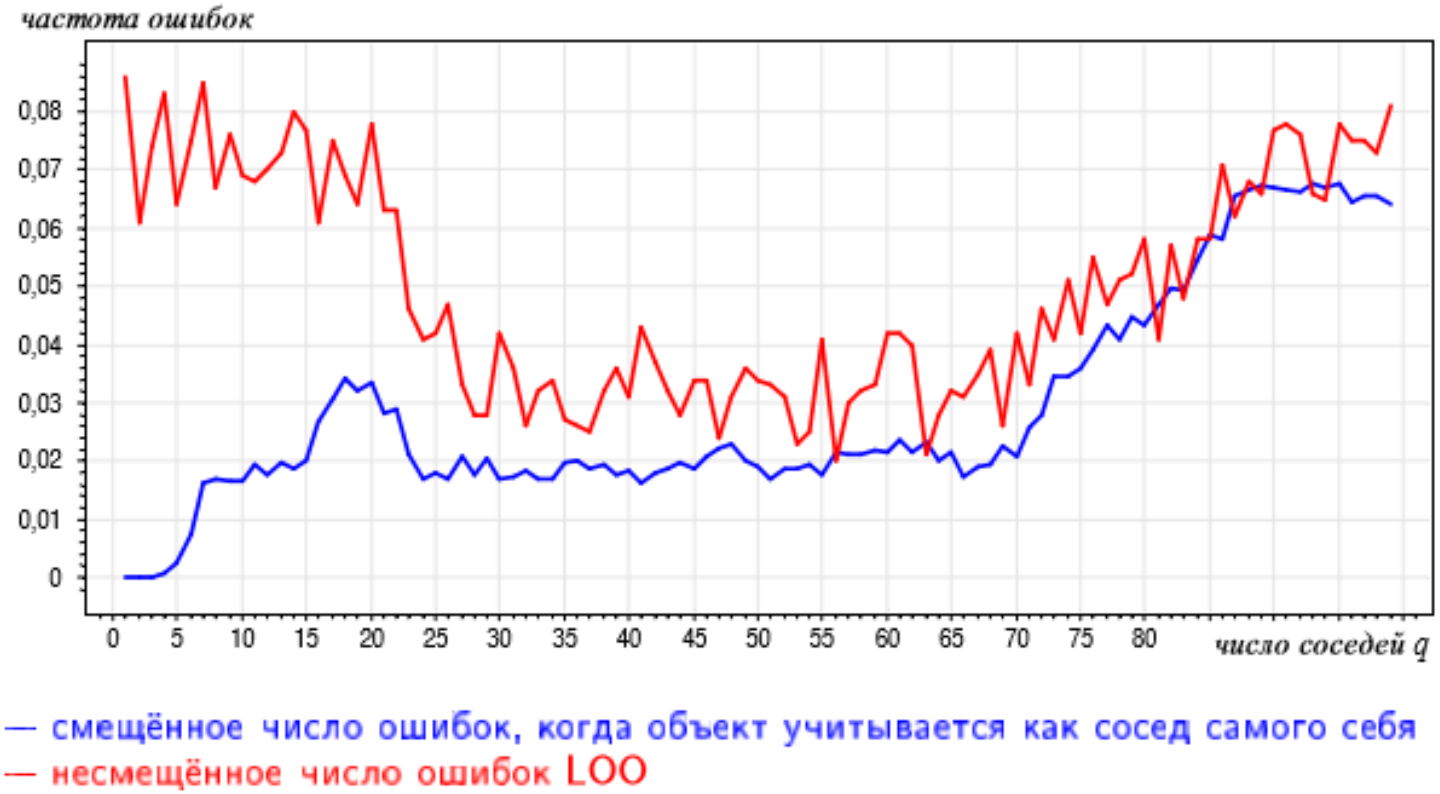
Проблемы:

- 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.



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