Введение в машинное обучение

Национальный исследовательский университет "Высшая школа экономики"
Yandex School of Data Analysis

Неофициальный конспект по курсу.

9 апреля 2021 г.

1 Метрические методы

Следующие методы, которые мы рассмотрим, будут **метрические методы**, которые используют функции *расстояния* или *метрики* в пространстве объектов.

Исходная идея заключается в предположении, что в практических задачах часто встречаются зависимости, которые непрерывны (хотя это справедливо скорее для задач регрессии, уточнения описаны ниже):

- **Гипотеза непрерывности** (для регрессии):

  близким объектам соответствуют близкие объекты.

- **Гипотеза компактности** (для классификации):

  близкие объекты, как правило, лежат в одном классе.

Выглядит логично, однако, как формализовать эту «близость»?

- **Формализация понятия «близости»:**

  Задана функция расстояния $\rho : X \times X \to [0, \infty)$.

По сути эта функция от пары объектов, которая паре ставит соответствие — неотрицательное число.

Часто также накладывают требование, чтобы это была метрика в пространстве объектов, то есть чтобы она была и симметричной, и выполнялось неравенство треугольника. Однако формально в методах нигде не используется предположение о том, что это метрика, поэтому, в общем-то, и функции расстояния, не являющиеся метриками, нам тоже подходят.

Самым известным примером расстояния, наверное, является евклидово расстояние в признаковом пространстве:

- **Пример.** Евклидово пространствов и его обобщение:

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

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

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