ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 21 **МАТЕМАТИКИ** **Выпуск 5**
2014

А. Н. Тырсин, **О. В. Ворфоломеева** (Челябинск, ЮУрГУ).
Алгоритм непараметрического оценивания индекса детерминации регрессионных зависимостей.

Форма регрессионной модели во многих случаях не известна. Неправильный выбор формы может привести к ухудшению статистических характеристик уравнения, а также делает невозможным использование построенной модели за пределами выборки. Для оценки адекватности модели ее целесообразно сравнить с эмпирической регрессией. Это можно сделать с помощью сравнения индексов детерминации этих зависимостей.

Индекс детерминации показывает долю дисперсии результативной переменной Y, объясненной вариацией факторных переменных $X_1, X_2, \ldots, X_m$, включенных в нелинейную модель регрессии

$$R^2_{Y/X_1\ldots X_m} = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}, \tag{1}$$

где $\overline{y}$ — среднее значение результативной переменной, $\widehat{y}_i$ — значения регрессии, y_i — фактические значения переменной Y.

У эмпирической регрессии функциональная форма не известна, поэтому значения $\widehat{y}_i$ определяют без построения уравнения регрессии. В [1] рассмотрен ряд непараметрических алгоритмов построения регрессии. Они основаны на сглаживании значений y_i. Основным недостатком данного подхода является проблема выбора размера окрестности усреднения (апертуры сглаживающего фильтра). Описанный ниже алгоритм устраняет данный недостаток.

Пусть имеем многомерную выборку $(x_{i1}, \ldots, x_{im}, y_i)$, $i = 1, \ldots, n$. Формируем матрицу расстояний $\mathbf{R}$, элементы которой $r_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$, $i, j = 1, \ldots, n$.

Для каждого i-го наблюдения подбираем оптимальную выборку из L_i ближайших соседей согласно матрице $\mathbf{R}$, так чтобы построенное уравнение линейной регрессии

$$\widetilde{y}_l = a_{i0} + \sum_{k=1}^m a_{ik} x_{lk}, \qquad l \in L_i = \{l_1, l_2, \ldots, l_{L_i}\} \tag{2}$$

имело минимальную дисперсию ошибок регрессии

$$s^2_{L_i} = \frac{\sum_{j \in L_i} (y_j - \widetilde{y}_j)^2}{L_i - m - 1} \to \min_{3(m+1) < L_i \leqslant n}.$$

Для оптимальной локальной выборки из L_i ближайших соседей по формуле (2) формируем значение $\widehat{y}_i$ непараметрической регрессии

$$\widehat{y}_i = a_{i0} + \sum_{k=1}^m a_{ik} x_{ik}.$$

© Редакция журнала «ОПиПМ», 2014 г.

По сформированному таким образом множеству значений $\widehat{y}_i,\ \ i = 1, \ldots, n,$ по (1) найдем оценку индекса детерминации множественной регрессии.

Очевидно, что оптимальные размеры локальных выборок L_i в общем случае будут различными. Это позволит учесть изменение градиента теоретической функции регрессии в зависимости от значений факторных переменных и дисперсии случайной компоненты. Так как не использовалось никаких предположений о виде функции регрессии, то описанный алгоритм является непараметрическим.

Отметим, что оценку коэффициентов регрессии в (2) выполняют по-разному, в зависимости от особенностей исходных данных, например, можно использовать метод наименьших квадратов, метод наименьших модулей, а также робастные методы.

СПИСОК ЛИТЕРАТУРЫ

1. *Хардле В.* Прикладная непараметрическая регрессия. М.: Мир, 1993, 349 с.