

О СОСТОЯТЕЛЬНОСТИ ПОСЛЕДОВАТЕЛЬНОСТИ НЕПАРАМЕТРИЧЕСКИХ ОЦЕНОК РЕГРЕССИИ**Мангалова Екатерина Сергеевна**

аспирант, Сибирский государственный аэрокосмический университет им. ак. М.Ф. Решетнева, РФ, г. Красноярск

Consistency of nonparametric estimators sequence**Ekaterina Mangalova***postgraduate student, Siberian State Aerospace University, Russia, Krasnoyarsk*

Аннотация. Приводится формализация алгоритма последовательного построения непараметрических оценок Надарая-Ватсона. Доказывается состоятельность последовательности оценок с уменьшающимися параметрами размытости.

Abstract. Algorithm based on nonparametric Nadaraya-Watson estimator sequence is described. In case of decreasing bandwidths, theorem about consistency of nonparametric estimators sequence is proved.

Ключевые слова: непараметрическая регрессия; состоятельность; ансамбль.

Keywords: nonparametric estimator; consistency; ensemble.

Постановка задачи восстановления регрессии. Задача восстановления регрессии может быть записана следующим образом [1]. Имеется множество наблюдений:

$G = \{g_1, g_2, \dots, g_n\}$, где каждое наблюдение представимо набором

$g_i = \{x_i^1, x_i^2, \dots, x_i^m, y_i\}$, где $x^1, x^2, \dots, x^m$ – независимые переменные, значения которых известны и на основании которых определяется значение переменной y . Требуется восстановить зависимость между входными переменными $x^1, x^2, \dots, x^m$ и зависимой переменной y .

В работе [2] предложена процедура решения поставленной задачи при помощи последовательности оценок регрессии Надарая-Ватсона.

Алгоритм последовательного построения непараметрических оценок Надарая-Ватсона. Идея построения последовательности непараметрических оценок состоит в улучшении некоторой базовой оценки регрессии за счет последовательного добавления непараметрических оценок невязок.

Последовательностью нулевого уровня $\hat{H}_0(x|G)$ является непараметрическая оценка

$$\hat{H}_0(x|G) = \frac{\sum_{i=1}^n \prod_{j=1}^m K\left(\frac{x^{(j)} - x_i^{(j)}}{c_0^{(j)}}\right) y_i}{\sum_{i=1}^n \prod_{j=1}^m K\left(\frac{x^{(j)} - x_i^{(j)}}{c_0^{(j)}}\right)},$$

где $c_0 = (c_0^{(1)}, c_0^{(2)}, \dots, c_0^{(m)})$ – параметры размытости, которые выбираются достаточно большими, чтобы оценка существовала во всем признаковом пространстве.

Каждая следующая оценка дополняет текущую последовательность $\hat{H}_{q-1}(x|G)$,

основываясь на невязке между выходом объекта y и $\hat{H}_{q-1}(x|G)$:

$$\hat{H}_q(x|G) = \hat{H}_{q-1}(x|G) + \begin{cases} \frac{\sum_{i=1}^n \prod_{j=1}^m K\left(\frac{x^{(j)} - x_i^{(j)}}{c_q^{(j)}}\right) y_i}{\sum_{i=1}^n \prod_{j=1}^m K\left(\frac{x^{(j)} - x_i^{(j)}}{c_q^{(j)}}\right)}, & \sum_{i=1}^n \prod_{j=1}^m K\left(\frac{x^{(j)} - x_i^{(j)}}{c_q^{(j)}}\right) > 0, \\ 0, & \sum_{i=1}^n \prod_{j=1}^m K\left(\frac{x^{(j)} - x_i^{(j)}}{c_q^{(j)}}\right) = 0. \end{cases}$$

На параметры размытости $c_0, c_1, \dots, c_Q$ накладывается условие: как минимум одна компонента c_{q-1} должна быть меньше соответствующей компоненты c_q .

Построение последовательности происходит до тех пор, пока уменьшается значение среднеквадратического отклонения выхода модели от выхода объекта.

Состоятельность последовательности непараметрических оценок. Рассмотрим задачу оценивания $y = \varphi(x) \forall x \in R^1$ по выборке независимых синхронных измерений случайных величин $(x_i, y_i, i = 1, 2, \dots, n)$ при известной плотности вероятности $p(x)$.

Теорема. Пусть:

а) $\varphi(x), p(x)$ в области определения $y = \varphi(x)$ ограничены и непрерывны вместе со

своими производными до второго порядка включительно, а также

$$\int \left[\varphi^{(m)}(x) \right]^2 dx < \infty, \int \left[p^{(m)}(x) \right]^2 dx < \infty, m \leq 2.$$

б) ядерная функция K удовлетворяет следующим условиям:

$$0 \leq K(u) < \infty, K(u) = K(-u), \int K(u) du = 1, \int u^2 K(u) du < \infty.$$

в) на последовательность коэффициентов размытости накладывается условие:

$$\lim_{n \rightarrow \infty} c_{\varrho}(n) = 0, \lim_{n \rightarrow \infty} n c_{\varrho}(n) = \infty.$$

Тогда ансамбль непараметрических оценок $\hat{y} = \hat{\varphi}(x) = \hat{H}_{\varrho}(x) = \hat{H}_{\varrho-1}(x) + \hat{h}_{\varrho}(x)$

обладает свойством состоятельности: $D(\hat{\varphi}(x)) = 0$.

Доказательство. 1. Рассмотрим ансамбль, состоящий из двух моделей:

$$\hat{H}_{\varrho}(x) = \hat{H}_0(x) + \hat{h}_{\varrho}(x) = \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{c_0}\right) y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{c_0}\right)} + \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{c_{\varrho}}\right) (y_i - \hat{H}_0(x_i))}{\sum_{i=1}^n K\left(\frac{x-x_i}{c_{\varrho}}\right)},$$

где $c_0 \rightarrow \infty$, $\hat{H}_0(x)$ - эквивалентно выборочному среднему.

Представим дисперсию в следующем виде:

$$D(\hat{\varphi}(x)) = M\left((\varphi(x) - \hat{\varphi}(x))^2\right) + \left(M(\varphi(x) - \hat{\varphi}(x))\right)^2 = M\left((\varphi(x) - \hat{\varphi}(x))^2\right)$$

Вычислим среднеквадратическое отклонение:

$$M\left((\varphi(x) - \hat{\varphi}(x))^2\right) = M(\hat{\varphi}^2(x)) - \varphi^2(x) = M\left(\hat{H}_0(x) + 2\hat{H}_0(x)\hat{h}_{\varrho}(x) + \hat{h}_{\varrho}^2(x)\right) - \hat{H}_0(x) - 2H_0(x)h_{\varrho}(x) - h_{\varrho}^2(x) = M\left(\hat{h}_{\varrho}^2(x)\right) - h_{\varrho}^2(x)$$

Найдем $M(\hat{h}_{\varrho}^2(x))$:

$$M\left(\hat{h}_{\varrho}^2(x)\right)=\frac{1}{nc_{\varrho}^2p^2(x)}\int\left(h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)\right)^2dt+\frac{n-1}{nc_{\varrho}^2p^2(x)}\left(\int h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)dt\right)^2$$

Сделаем замену переменных $u=(x-t)/c_{\varrho}, t=x-c_{\varrho}u, du=-c_{\varrho}dt, dt=du/c_{\varrho}$;

$$M\left(\hat{h}_{\varrho}^2(x)\right)=\frac{1}{nc_{\varrho}p^2(x)}\int\left(h_{\varrho}(x-c_{\varrho}u)p(x-c_{\varrho}u)K(u)\right)^2du+\\ +\frac{n-1}{np^2(x)}\left(\int h_{\varrho}(x-c_{\varrho}u)p(x-c_{\varrho}u)K(u)du\right)^2$$

$$\lim_{n \rightarrow \infty} nc_{\varrho}(n) = \infty$$

Т.к.

$$M\left(\hat{h}_{\varrho}^2(x)\right)=\lim_{n \rightarrow \infty} \frac{n-1}{np^2(x)}\left(\int h_{\varrho}(x-c_{\varrho}u)p(x-c_{\varrho}u)K(u)du\right)^2=h_{\varrho}^2(x)+\lim_{n \rightarrow \infty} \frac{1}{n}h_{\varrho}^2(x)=h_{\varrho}^2(x)$$

2. Покажем, что свойство состоятельности не нарушится при добавлении в

последовательность промежуточной модели с параметром размытости $c_1 > c_{\varrho}$;

$$\hat{H}_{\varrho}^+(x)=\hat{H}_{\varrho}(x)+\hat{z}(x)$$

$$\hat{z}(x)=\frac{\sum_{i=1}^n K\left(\frac{x-x_i}{c_1}\right)\left(y_i-\hat{H}_0(x_i)\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{c_1}\right)}-\frac{\sum_{i=1}^n K\left(\frac{x-x_i}{c_{\varrho}}\right)\left(\frac{\sum_{j=1}^n K\left(\frac{x_i-x_j}{c_1}\right)\left(y_j-\hat{H}_0(x_j)\right)}{\sum_{j=1}^n K\left(\frac{x_i-x_j}{c_1}\right)}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{c_{\varrho}}\right)}$$

Вычислим среднеквадратическое отклонение:

$$M\left(\left(\varphi(x)-\hat{\varphi}(x)\right)^2\right)=2M\left(\hat{H}_{\varrho}(x)\hat{z}(x)\right)+M\left(\hat{z}^2(x)\right)$$

Докажем, что оба слагаемых равны нулю:

$$\begin{aligned}
M\left(\hat{H}_{\varrho}(x)\hat{z}(x)\right) &= H_0(x)M\left(\hat{z}(x)\right)+M\left(\hat{h}_{\varrho}(x)\hat{z}(x)\right)= \\
&= \frac{1}{nc_1c_{\varrho}p^2(x)}\int\left(h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_1}\right)\right)\left(h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)\right)dt+ \\
&+ \frac{n-1}{nc_1c_{\varrho}p^2(x)}\int\left(h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_1}\right)\right)dt\int\left(h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)\right)dt \\
&\frac{1}{nc_{\varrho}^2p^2(x)}\int\left(a(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)\right)\left(h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)\right)dt+ \\
&+ \frac{n-1}{nc_{\varrho}^2p^2(x)}\int\left(a(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)\right)dt\int\left(h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)\right)dt \\
&a(x)=\frac{1}{c_1p(x)}\int h_1(t)p(t)K\left(\frac{x-t}{c_1}\right)dt
\end{aligned}$$

где

Проведя замену переменных и применяя преобразования, аналогичные использованным в п. 1, получаем:

$$M\left(\hat{H}_{\varrho}(x)\hat{z}(x)\right)=\frac{(n-1)h_{\varrho}(x)}{nc_1p(x)}\int\left(h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_1}\right)\right)dt-\frac{n-1}{n}a(x)h_{\varrho}(x)=0$$

Аналогично и для второго слагаемого:

$$\begin{aligned}
M\left(\hat{z}^2(x)\right) &= \frac{1}{nc_1^2p^2(x)}\int\left(h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_1}\right)\right)^2dt+\frac{n-1}{nc_1^2p(x)}\left(\int h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_1}\right)dt\right)^2- \\
&- \frac{2}{nc_1c_{\varrho}p^2(x)}\int\left(h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_1}\right)\right)\left(a(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)\right)dt- \\
&- \frac{2(n-1)}{nc_1c_{\varrho}p^2(x)}\int\left(h_{\varrho}(t)p(t)K\left(\frac{x-t}{c_1}\right)\right)dt\int\left(a(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)\right)dt+ \\
&+ \frac{1}{nc_{\varrho}^2p^2(x)}\int\left(a(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)\right)^2dt+\frac{n-1}{nc_{\varrho}^2p^2(x)}\left(\int a(t)p(t)K\left(\frac{x-t}{c_{\varrho}}\right)dt\right)^2
\end{aligned}$$

Проведя замену переменных и применяя преобразования, аналогичные использованным в п. 1, получаем:

$$M(\hat{z}^2(x)) = \frac{1}{nc_1^2 p^2(x)} \int \left(h_{\varrho}(t) p(t) K\left(\frac{x-t}{c_1}\right) \right)^2 dt + \frac{n-1}{nc_1^2 p^2(x)} \left(\int h_{\varrho}(t) p(t) K\left(\frac{x-t}{c_1}\right) dt \right)^2 -$$

$$- 2 \frac{(n-1)a(x)}{nc_1 p(x)} \int h_{\varrho}(t) p(t) K\left(\frac{x-t}{c_1}\right) dt + \frac{n-1}{n} a^2(x) = 0$$

Таким же образом может быть показано, что добавление любого количества промежуточных моделей не нарушает свойства состоятельности.

3. Свойство состоятельности не нарушится при добавлении в последовательность оценки с параметром размытости $c_{\varrho+1} < c_{\varrho}$:

$$\hat{H}_{\varrho+1}(x) = \hat{H}_{\varrho}(x) + \hat{h}_{\varrho+1}(x).$$

Далее доказательство аналогично приведенному в п.1.

Таким образом, последовательность непараметрических оценок, как и непараметрическая оценка Надарая-Ватсона, является состоятельной.

Список литературы:

1. Барсегян А.А., Куприянов М.С., Холод И.И., Тесс М.Д., Елизаров С.И. Анализ данных и процессов 3-е изд. — СПб.: БХВ-Петербург, 2009. — 512 с.
2. Мангалова, Е. С., Шестернева О.В. О последовательном построении оценок регрессии // Вестник Сибирского государственного аэрокосмического университета им. академ. М. Ф. Решетнева. – 2015. – Т. 16. – С. 604-610.