Р О С С И Й С К А Я   А К А Д Е М И Я   Н А У К

СИБИРСКОЕ ОТДЕЛЕНИЕ

А В Т О М Е Т Р И Я

2005, том 41, № 3

УДК 519.234

В. Г. Алексеев

(Звенигород Московской обл.)

О ДОПУСТИМЫХ НЕПАРАМЕТРИЧЕСКИХ ОЦЕНКАХ ПЛОТНОСТИ ВЕРОЯТНОСТИ

Формулируются рекомендации по построению непараметрических оценок плотности вероятности, минимизирующих интегральную среднеквадратичную ошибку оценивания.

Данную работу следует рассматривать как расширенный комментарий к работе [1], в которой непараметрическая оценка неизвестной плотности вероятности используется в задачах восстановления стохастических зависимостей и распознавания образов. Величина $W_2(\bar{p}(x))$, определенная в [1, с. 56] и названная среднеквадратичным отклонением оценки $\bar{p}(x)$ плотности вероятности $p(x)$, представляет собой не что иное, как интегральную среднеквадратичную ошибку (ИСКО), т. е. среднюю (по ансамблю реализаций) ошибку, измеряемую в метрике пространства $L_2(\quad,\quad)$:

$$W_2(\bar{p}(x)) \quad \left\langle \quad [\bar{p}(x) \quad p(x)]^2 dx \right\rangle. \tag{1}$$

Далее будут приведены некоторые соображения, касающиеся непараметрического оценивания плотности вероятности $p(x)$. Учитывая критерий (1) качества оценки плотности вероятности [1], сформулируем рекомендации, направленные, во-первых, на уменьшение ошибки оценивания функции $p(x)$, измеряемой в метрике пространства $L_2(\quad,\quad)$, и, во-вторых, на сокращение объема вычислений, ведущих к искомой статистической оценке, а также коротко обсудим некоторые вопросы, касающиеся сходимости оценок плотности вероятности в метрике пространства непрерывных функций $C(\quad,\quad)$.

Формулируемые рекомендации по существу сводятся к использованию так называемых допустимых (в известном смысле неулучшаемых) оценок плотности вероятности. Что же касается оценок, не являющихся допустимыми, то их применение неизбежно приводит к завышенным значениям ИСКО, поэтому с самого начала их следует исключить из рассмотрения, поскольку избранным нами критерием качества оценки плотности вероятности является величина (1). Обозначения оценок плотности вероятности и других математических объектов в данной работе не будут совпадать с обозначениями в работе [1].

Итак, пусть $X_1, X_2, \ldots, X_n$ – выборка из n независимых наблюдений случайной величины X с неизвестной плотностью вероятности $p(x)$. Как и в работах [2–5], ограничимся рассмотрением ядерной оценки функции $p(x)$, которая определяется соотношением

$$p_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{2}$$

где $h = h(n)$ – некоторая последовательность положительных чисел, стремящаяся к нулю (но не слишком быстро) при $n \to \infty$, а $K(x)$ – некоторая четная ограниченная функция, удовлетворяющая условиям $\int K(x)dx = 1$ и $K(x) \in L_2(-\infty, \infty)$ (т. е. $\int K^2(x)dx < \infty$). Здесь (и далее) интеграл без указания пределов обозначает интегрирование в пределах от $-\infty$ до ∞.

Понятие допустимости весовой функции $K(x)$ впервые введено в [6]. Весовая функция $K(x)$ (а вместе с ней и оценка плотности вероятности (2)) называется допустимой, если ее преобразование Фурье $\varphi(t) = \int \exp(itx)K(x)dx$ неотрицательно и не превосходит 1 для всех $t \in \mathbf{R}$ (вещественность функции $\varphi(t)$ следует из четности преобразуемой функции $K(x)$). Если функция $K(x)$ допустима в смысле [6], то ИСКО получаемой с ее помощью оценки (2) не может быть уменьшена одновременно (с помощью одной и той же весовой функции $K_*(x)$) для всех плотностей вероятности $p(x)$ и для всех n и h из области их значений. Если же весовая функция $K(x)$ не является допустимой, то всегда есть возможность уменьшить ИСКО получаемой с ее помощью оценки (2) одновременно для всех плотностей вероятности $p(x)$ и для всех натуральных n.

Далее понадобится понятие порядка весовой функции $K(x)$. Как и в работах [4, 5], порядком весовой функции $K(x)$ будем называть наименьшее четное число $r \geq 2$, для которого $\int x^r K(x)dx \neq 0$. Очевидно, функция $K(x)$ знакопеременна, если $r > 2$. Хорошо известно, что при достаточно больших n (а в ряде случаев и при не очень больших n) применение весовых функций высших порядков (т. е. порядков $r > 2$) позволяет существенно уменьшить ошибку оценивания (измеряемую во многих совершенно разных метриках), если только оцениваемая плотность вероятности $p(x)$ является достаточно гладкой (многократно дифференцируемой) функцией. С учетом этого обстоятельства предложим весовые функции $K(x)$ не только минимального (второго) порядка, но и порядков $r > 2$.

Допустимыми весовыми функциями второго порядка являются функции из [3, 7] соответственно:

$$K(x) = \begin{cases} (1\ \ )(1 - |x|\ \ )/2, & |x| \leq 1; \\ 0, & |x| > 1, \end{cases} \tag{3}$$

где 0   1, и

$$K(x) \qquad \begin{array}{ll} (3/2)[1 \quad x^2(10 \quad 20|x| \quad 15x^2 \quad 4|x|^3)], & |x| \quad 1; \\ 0, & |x| \quad 1. \end{array} \qquad (4)$$

Что же касается допустимых весовых функций высших порядков, то они могут быть найдены в работах [4, разд. 1] и [5, разд. 1]. Все предложенные в них весовые функции $K(x)$ являются линейными комбинациями B-сплайнов Шенберга, представляющих значительный интерес не только для непараметрической статистики, но и для теории интерполирования [8–10]. При этом все весовые функции $K(x)$ из [5] обладают еще одним благоприятным свойством: они дифференцируемы. В этом случае с помощью весовой функции $K(x)$ может быть построена оценка, которая при n сходится к оцениваемой плотности вероятности $p(x)$ не только в метрике $L_2(\quad,\quad)$, но и в метрике пространства непрерывных функций $C(\quad,\quad)$ (если только оцениваемая плотность вероятности $p(x)$ непрерывна). Отметим, что весовая функция второго порядка (4) также дифференцируема.

Замечание 1. Следует иметь в виду, что базовая весовая функция $W_6(x)$, входящая в некоторые из функций $K(x)$ в качестве одного из слагаемых, приведена в работе [5] с опечаткой: в одной из шести формул, определяющих функцию $W_6(x)$ для каждого из шести смежных интервалов единичной длины, значение нормирующего делителя, следующего за знаком «/» после круглой скобки, несколько завышено. Правильное значение нормирующего делителя во всех шести формулах, определяющих функцию $W_6(x)$, одно и то же: 39916800.

Замечание 2. Алгоритмы в работах [4, 5] позволяют строить допустимые (неулучшаемые в метрике $L_2(\quad,\quad)$) оценки не только самой плотности вероятности $p(x)$, но и ее производных до третьего порядка включительно. Необходимость статистического оценивания производных плотности вероятности возникает при решении многих задач обработки результатов наблюдений [11, разд. 3.6].

Наконец, заметим, что носителем каждой из предлагаемых нами весовых функций $K(x)$ является конечный интервал. Это последнее обстоятельство существенно ускоряет вычисление оценки (2) плотности вероятности: при достаточно больших n правая часть формулы (2) будет реально зависеть лишь от небольшой доли исходных наблюдений X_i.

СПИСОК ЛИТЕРАТУРЫ

1. Лапко А. В., Лапко В. А. Непараметрические методики анализа множеств случайных величин // Автометрия. 2003. 39, № 1. С. 54.

2. Алексеев В. Г. Об оценке плотности вероятности и ее производных // Мат. заметки. 1972. 12, № 5. С. 621.

3. Алексеев В. Г. О допустимых непараметрических оценках плотности вероятности и ее производных // Проблемы передачи информации. 1994. 30, № 2. С. 36.

4. Алексеев В. Г. Новые допустимые оценки плотности вероятности и ее производных. Ч. I // Мат. заметки ЯГУ. 2001. 8, № 2. С. 3.

5. Алексеев В. Г. Новые допустимые оценки плотности вероятности и ее производных. Ч. II // Мат. заметки ЯГУ. 2003. 10, № 1. С. 7.
6. Cline D. B. H. Admissible kernel estimators of a multivariate density // Ann. Statist. 1988. 16, N 3. P. 1421.
7. Алексеев В. Г. К построению оценок спектральных плотностей периодически коррелированного случайного процесса // Проблемы передачи информации. 1990. 26, № 3. С. 106.
8. Unser M. Sampling – 50 years after Shannon // Proc. IEEE. 2000. 88, N 4. P. 569.
9. Meijering E. A chronology of interpolation: from ancient astronomy to modern signal and image processing // Proc. IEEE. 2002. 90, N 3. P. 319.
10. Алексеев В. Г. B-сплайны Шенберга и их применения в радиотехнике и в смежных с ней дисциплинах // Радиотехника. 2003. № 12. С. 21.
11. Шапиро Е. И. Непараметрические оценки плотности вероятности в задачах обработки результатов наблюдений // Зарубеж. радиоэлектрон. 1976. № 2. С. 3.

Институт физики атмосферы РАН

Поступило в редакцию
18 мая 2004 г.