СРАВНЕНИЕ ДВУХ АЛГОРИТМОВ НАСТРОЙКИ ДЛИНЫ РЯДА ДЛЯ ПРОЕКЦИОННОЙ ОЦЕНКИ ПЛОТНОСТИ ВЕРОЯТНОСТИ

Браништи В.В. ©
Старший преподаватель кафедры высшей математики,
Сибирский государственный аэрокосмический университет
имени академика М. Ф. Решетнёва

Аннотация

В работе рассматривается проекционная оценка функции плотности вероятности случайной величины. Рассматриваются два подхода к оцениванию длины ряда проекционной оценки, основанные на двух разных способах оценивания функционала качества. С помощью численных экспериментов показано, что подход, основанный на построении несмещённой оценки функционала качества, оказывается более эффективным.

Ключевые слова: функция плотности вероятности, статистическое оценивание, ортогональные системы, пространство $L_{2,w}$, проекционная оценка.

Keywords: probability density function, statistical estimation, orthogonal systems, $L_{2,w}$ space, projective estimate.

Оценивание функции плотности вероятности случайной величины является центральной задачей математической статистики [1, 5]. Большинство современных алгоритмов классификации, распознавания образов, восстановления стохастических зависимостей используют те или иные алгоритмы восстановления неизвестной плотности. При этом перспективным направлением является применение так называемых непараметрических методов восстановления, т.е. методов, не использующих информацию о виде закона распределения. К непараметрическим методам оценивания плотности относятся гистограммные оценки, оценки ядерного типа [2, 1065; 3, 23] и проекционные оценки плотности [4, 45].

При построении проекционной оценки предполагается, что истинная функция плотности $f(x)$ исследуемой случайной величины ξ принадлежит функциональному гильбертову пространству $L_{2,w}(\Omega)$. В работе [5, 21] показано, что для любой непрерывной случайной величины существует содержащее его пространство $L_{2,w}(\Omega)$. В этом случае функция $f(x)$ представима в виде ряда:

$$f(x) = \sum_{j=0}^{\infty} \alpha_j \varphi_j(x),$$

где

$$\alpha_j = \left(f, \varphi_j \right)_w = \int_{\Omega} f(x)\varphi_j(x)w(x)dx, \qquad (1)$$

$\{\varphi_1, \varphi_2, \ldots\}$ – ортонормированный базис пространства $L_{2,w}(\Omega)$, $\Omega \subseteq (-\infty; +\infty)$ – множество, на котором восстанавливается функция плотности.

Проекционной оценкой функции плотности вероятности называется проекция функции $f(x)$ на конечномерное подпространство пространства $L_{2,w}$:

$$f_l(x) = \sum_{j=0}^{l} \alpha_j \varphi_j(x),$$

где l – длина ряда.

© Браништи В.В., 2016 г.

Оптимальные коэффициенты α_j неизвестны, следовательно, подлежат оцениванию. Соответствующие оценки коэффициентов α_j обозначаются как a_j. Кроме того, оцениванию подлежит длина ряда l. Проекционная оценка функции плотности, в которой используются оценки a_j оптимальных коэффициентов α_j обозначается через $\hat{f}_l(x)$, а в которой, кроме того, используется оценка длины ряда l – через $\hat{f}(x)$. Критерием качества оценки плотности является математическое ожидание квадрата отклонения от истинной плотности $f(x)$ в пространстве $L_{2,w}$:

$$Q\left\{\hat{f}\right\} = M\left\{\left\| f - \hat{f} \right\|_w^2\right\}. \qquad (2)$$

Пусть имеется независимая выборка x_1, x_2, …, x_n. Стандартный приём к оцениванию коэффициентов α_j приведён в работе [4, 45] и состоит в следующем. Выражение (1) рассматривается как математическое ожидание случайной величины $\varphi_j(\xi)w(\xi)$:

$$\int_{-\infty}^{+\infty} f(x)\varphi_j(x)w(x)dx = M\left\{\varphi_j(\xi)w(\xi)\right\},$$

которое оценивается с помощью выборочного среднего:

$$\hat{M}\left\{\varphi_j(\xi)w(\xi)\right\} = \frac{1}{n}\sum_{i=1}^{n}\varphi_j(x_i)w(x_i).$$

Отсюда

$$a_j = \frac{1}{n}\sum_{i=1}^{n}\varphi_j(x_i)w(x_i). \qquad (3)$$

В работе [6, 10] показано, что при использовании оценок (3) существует оптимальное в смысле критерия (2) конечное значение длины ряда l.

Оценка $\hat{l}$ длины ряда l строится путём минимизации функционала (2), который преобразуется следующим образом:

$$Q\left\{\hat{f}_l\right\} = M\left\{\left(f - \hat{f}_l, f - \hat{f}_l\right)_w\right\} = M\left\{(f,f)_w - 2\left(f, \hat{f}_l\right)_w + \left(\hat{f}_l, \hat{f}_l\right)_w\right\} =$$

$$= M\left\{(f,f)_w - 2\left(f, \sum_{j=0}^{l} a_j\varphi_j\right)_w + \left(\sum_{j=0}^{l} a_j\varphi_j, \sum_{j=0}^{l} a_j\varphi_j\right)_w\right\} =$$

$$= M\left\{(f,f)_w - 2\sum_{j=0}^{l} a_j\left(f,\varphi_j\right)_w + \sum_{j_1=0}^{l}\sum_{j_2=0}^{l} a_{j_1}a_{j_2}\left(\varphi_{j_1},\varphi_{j_2}\right)_w\right\} =$$

$$= M\left\{(f,f)_w - 2\sum_{j=0}^{l} a_j\left(f,\varphi_j\right)_w + \sum_{j=0}^{l} a_j^2\right\} = (f,f)_w - M\left\{2\sum_{j=0}^{l} a_j\left(f,\varphi_j\right)_w - \sum_{j=0}^{l} a_j^2\right\}.$$

Так как при любом законе распределения выражение $(f,f)_w$ не зависит от l, то минимизация функционала (2) эквивалентна максимизации функционала

$$W\left\{\hat{f}_l\right\} = M\left\{2\sum_{j=0}^{l} a_j\left(f,\varphi_j\right)_w - \sum_{j=0}^{l} a_j^2\right\}. \qquad (4)$$

Функционал (4) также использует вид истинной плотности $f(x)$, которая считается неизвестной. Поэтому оценку $\hat{l}$ строят путём максимизации оценки функционала (4).

В работе [6, 24] строится смещённая оценка функционала (4), смещение которой пропорционально случайной составляющей ошибки приближения:

$$\hat{W}_l = \sum_{j=0}^{l}\left(\frac{n+k+1}{n-1}a_j^2 - \frac{k+2}{n(n-1)}\sum_{i=1}^{n}\varphi_j^2(x_i)w^2(x_i)\right), \qquad (5)$$

где k – коэффициент пропорциональности, рассчитываемый по формуле:

$$k = \frac{\sum_{j=1}^{l}\left(\frac{n+3}{n}a_j^2\sum_{i=1}^{n}\varphi_j^2(x_i)w^2(x_i) - \frac{2}{n^2}\left(\sum_{i=1}^{n}\varphi_j^2(x_i)w^2(x_i)\right)^2 - (n+1)a_j^4\right)}{\sum_{j=1}^{l}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi_j^2(x_i)w^2(x_i) - a_j^2\right)^2}.$$

Проекционную оценку плотности вероятности, в которой коэффициенты находятся по формулам (3), а длина ряда оценивается путём максимизации значения (5), будем называть оценкой (А) и обозначать через $\hat{f}_{\text{А}}(x)$.

При $k = 0$ получаем несмещённую оценку функционала (4):

$$\hat{W}_l = \sum_{j=0}^{l}\left(\frac{n+1}{n-1}a_j^2 - \frac{2}{n(n-1)}\sum_{i=1}^{n}\left(\varphi_j(x_i)w(x_i)\right)^2\right). \qquad (6)$$

Проекционную оценку плотности вероятности, в которой коэффициенты находятся по формулам (3), а длина ряда оценивается путём максимизации значения (6), будем называть оценкой (Б) и обозначать через $\hat{f}_{\text{Б}}(x)$.

В настоящей работе предлагается сравнение качества оценок (А) и (Б) в смысле критерия (2). В качестве тестовых восстанавливаемых плотностей были взяты следующие:

1) $f(x) = \begin{cases} 2, & x \in \left[0; \frac{1}{2}\right] \\ 0, & \text{иначе} \end{cases}$ – равномерное распределение на отрезке $\left[0; \frac{1}{2}\right]$;

2) $f(x) = \begin{cases} \frac{1}{2} - \frac{1}{4}|x-1|, & x \in [-1; 3] \\ 0, & \text{иначе} \end{cases}$ – треугольное распределение на отрезке [−1; 3];

3) $f(x) = \begin{cases} 2\left(1+4|x|\right)\left(1-2|x|\right)^2, & x \in \left[-\frac{1}{2}; \frac{1}{2}\right] \\ 0, & \text{иначе} \end{cases}$ – кубическое распределение на отрезке $\left[-\frac{1}{2}; \frac{1}{2}\right]$;

4) $f(x) = \begin{cases} e^{-x}, & x \in [0; +\infty) \\ 0, & \text{иначе} \end{cases}$ – показательное распределение, λ = 1;

5) $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}}$ – нормальное распределение, μ = 1, σ = 1.

Все восстанавливаемые плотности принадлежат пространству $L_{2,w}(\Omega)$ при $w(x) \equiv 1$. Для восстановления плотности в этом пространстве используются следующие полные ортонормированные системы [7, 447]:

1) $\varphi_m = \frac{1}{m!2^m}\sqrt{\frac{2m+1}{2}}\frac{d^m}{dx^m}(x^2-1)^m$ – базис Лежандра, Ω = [–1; 1];

2) $\varphi_m = \begin{cases} \frac{1}{\sqrt{2\pi}}, & m = 0 \\ \frac{1}{\sqrt{\pi}}\sin mx, & m = 1, 3, 5, \ldots \\ \frac{1}{\sqrt{\pi}}\cos mx, & m = 2, 4, 6, \ldots \end{cases}$ – базис Фурье, Ω = [–π; π];

3) $\varphi_m = \frac{e^{x/2}}{m!}\frac{d^m}{dx^m}(x^m e^{-x})$ – базис Лагерра, Ω = [0; +∞);

4) $\varphi_m = \frac{(-1)^m e^{x^2/2}}{\sqrt{\sqrt{\pi}2^m m!}}\frac{d^m}{dx^m}e^{-x^2}$ – базис Эрмита, Ω = (–∞; +∞).

Соответствие между восстанавливаемой плотностью, используемой в работе ортонормированной системой, множеством Ω, а также некоторыми свойствами восстанавливаемой плотности приведено в таблице 1.

Таблица 1

Свойства восстанавливаемых функций плотности вероятности

№	Распределение	Базис	Ω	Непрерывна	Дифференцируема
1	равномерное	Лежандра	[–1; 1]	нет	нет
2	треугольное	Фурье	[–π; π]	да	нет
3	кубическое	Лежандра	[–1; 1]	да	да
4	показательное	Лагерра	[0; +∞)	нет	нет
5	нормальное	Эрмита	(–∞; +∞)	да	да

Для каждой восстанавливаемой плотности строились оценки (А) и (Б), для которых находилось значение функционала (2). Так как расчёт теоретического значения функционала (2) затруднён, то для него находилось численное значение методом, предложенным в [8, 15]. Результаты расчётов при разных объёмах выборки *n* занесены в таблицу 2.

Таблица 2

Приближённые значения функционала (2) для оценок (А) и (Б) при восстановлении различных законов распределения

Равномерное распределение					
	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
$Q\{\hat{f}_А\}$	0,799±0,078	0,466±0,033	0,362±0,022	0,313±0,016	0,277±0,013
$Q\{\hat{f}_Б\}$	0,68±0,066	0,46±0,034	0,36±0,022	0,308±0,016	0,274±0,012
Треугольное распределение					
	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
$Q\{\hat{f}_А\}$	0,288±0,025	0,228±0,013	0,208±0,009	0,199±0,007	0,186±0,005
$Q\{\hat{f}_Б\}$	0,268±0,022	0,224±0,013	0,201±0,008	0,196±0,007	0,189±0,006
Кубическое распределение					
	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
$Q\{\hat{f}_А\}$	0,531±0,073	0,225±0,032	0,149±0,02	0,118±0,016	0,093±0,013

$Q\left\{\hat{f}_{\text{Б}}\right\}$	0,44±0,066	0,202±0,028	0,151±0,021	0,115±0,016	0,089±0,012
Показательное распределение					
	n = 10	n = 20	n = 30	n = 40	n = 50
$Q\left\{\hat{f}_{\text{А}}\right\}$	0,136±0,022	0,078±0,013	0,049±0,007	0,037±0,005	0,030±0,004
$Q\left\{\hat{f}_{\text{Б}}\right\}$	0,142±0,026	0,071±0,01	0,050±0,008	0,040±0,006	0,030±0,004
Нормальное распределение					
	n = 10	n = 20	n = 30	n = 40	n = 50
$Q\left\{\hat{f}_{\text{А}}\right\}$	0,102±0,013	0,051±0,006	0,034±0,004	0,027±0,003	0,021±0,002
$Q\left\{\hat{f}_{\text{Б}}\right\}$	0,102±0,013	0,047±0,005	0,031±0,004	0,025±0,003	0,020±0,002

Как видно из таблицы 2, не зависимо от объёма выборки и вида восстанавливаемого закона распределения метод настройки длины ряда l, основанный на максимизации значения (6), показывает близкие или лучшие результаты по сравнению с методом, основанным на максимизации значения (5). Кроме того, построение оценки (Б) требует меньше вычислительных затрат. Полученные результаты позволяют сделать вывод о том, что при отсутствии априорной информации о виде закона распределения в проекционной оценке плотности вероятности длину ряда l целесообразно оценивать методом максимизации несмещённой оценки функционала (4).

Литература

1. Деврой Л., Дьёрфи Л. Непараметрическое оценивание плотности. – М.: Мир, 1988. – 408 с.
2. Parzen E. – On estimation of a probability density function and mode // The Annals of Mathematical Statistics. – 1962. – Vol. 35, 3. – P. 1065–1076.
3. Лапко А. В., Лапко В. А. Непараметрические модели и алгоритмы обработки информации. – Красноярск: Изд-во СибГАУ, 2010. – 220 с.
4. Ченцов Н. Н. – Оценка неизвестной плотности распределения по наблюдениям // Доклады АН СССР. – 1962. – 147, 1. – С. 45–48.
5. Браништи В. В. – Введение пространства $L_{2,w}$ при построении проекционной оценки плотности вероятности // Вестник СибГАУ. – 2016. – №1. – С. 19–26.
6. Новоселов А. А. Об оптимальном выборе структуры функции плотности вероятности и регрессии: препринт – Красноярск: ВЦ СО АН СССР, 1979. – 31 с.
7. Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа – 6-е изд., испр. – М.: Наука, 1989. – 624 с.
8. Браништи В. В. – О параметрическом оценивании функции плотности вероятности // Научно-технический вестник Поволжья. – 2014. – №1. – С. 13–16.