Лабораторная работа № 2 Ядерные оценки плотности

Предположим, что распределение наблюдаемой случайной величины ξ абсолютно непрерывно, обозначим $p(x) = F'(x)$ плотность распределения, где $F(x) = \mathbf{P}\{\xi \le x\}$ – функция распределения ξ. Если число N наблюдений достаточно велико, то можно оценить значения плотности $p(x)$ распределения по выборке $X_1, X_2, \ldots, X_N$.

В лабораторной работе 2 для получения оценок функции плотности $p(x)$ мы будем использовать гистограмму, полигон частот и ядерные оценки – современный метод оценивания плотности.

Опишем кратко последний метод. Ядерные оценки были введены Розенблаттом (Rosenblatt, 1956) и Парзеном (Parzen, 1962) и получили широкое распрстранение в различных задачах статистики. Это связано с тем, что использование ядерных оценок может обеспечить существенно более высокую степень точности по сравнению с гистограммами для плотностей распределения, удовлетворяющих более сильным условиям гладкости.

Состоятельной и несмещенной оценкой неизвестной функции распределения $F(x)$ является эмпирическая функция распределения $\widehat{F}_N(x) = N^{-1}\sum_{i=1}^{N} \mathbb{I}_{\{X_i \le x\}}$, где $\mathbb{I}_{\{X_i \le x\}}$ – индикатор, равный единице, если событие $\{X_i \le x\}$ выполняется, и нулю в противном случае. Так как $\widehat{F}_N(x)$ – кусочно-постоянная функция, ее производная равна нулю всюду, за исключением точек скачков $\widehat{F}_N(x)$ (т. е. точек наблюдений X_i), и не годится в качестве оценки для $p(x) = F'(x)$.

Основная идея при получении ядерных оценок состоит в предварительном *сглаживании* эмпирического распределения за счет его свертки с распределениями, имеющими плотности и сходящимися к сосредоточенному в точке 0 распределению. Точнее: рассмотрим случайную величину $\zeta_N = \xi + \delta_N\eta$, где η – случайная величина, не зависящая от ξ и имеющая плотность $K(y)$, а $\delta_N > 0$ – последовательность чисел, $\delta_N \to 0$ при $N \to \infty$. Согласно известным формулам преобразования плотности, плотностью случайной величины $\delta_N\eta$ служит ${\delta_N}^{-1}K(y/\delta_N)$. Поскольку случайные величины ξ и η независимы, для плотности $q_N(z)$ случайной величины ζ_N с учетом формулы свертки получаем:

$$q_N(z) = \frac{1}{\delta_N}\int\limits_{-\infty}^{\infty} K\left(\frac{z-x}{\delta_N}\right) dF(x) = \frac{1}{\delta_N}\int\limits_{-\infty}^{\infty} K\left(\frac{z-x}{\delta_N}\right) p(x)\,dx. \tag{2.1}$$

Очевидно, что при $\delta_N \to 0$ случайные величины ζ_N будут сходиться к ξ по вероятности, плотности $q_N(z)$ – к плотности $p(z)$ при почти всех z, если, например, плотность $K(y)$ непрерывна, ограничена и $\int_{-\infty}^{\infty} K^2(y)\,dy < \infty$[1]. Заменяя в формуле (2.1) функцию распределения $F(x)$ ее оценкой $\widehat{F}_N(x)$, получаем оценку для плотности $p(z)$:

$$\widehat{p}_N(z) = \frac{1}{\delta_N}\int\limits_{-\infty}^{\infty} K\left(\frac{z-x}{\delta_N}\right) d\widehat{F}_N(x) = \frac{1}{N\,\delta_N}\sum_{i=1}^{N} K\left(\frac{z-X_i}{\delta_N}\right), \tag{2.2}$$

которая называется *оценкой Розенблатта – Парзена.*

Ответ на вопрос о статистических свойствах этой оценки дает следующая теорема.

[1]На самом деле достаточно (см., например, Лагутин М.Б. "Наглядная математическая статистика М.: БИНОМ. Лаборатория знаний, 2007, с.389) ограниченности плотности К(у) и выполнения при некоторых $C > 0$ и $\varepsilon > 0$ неравенства $K(y) \le C/|y|^{1+\varepsilon}$.

Теорема 1. *Пусть выполнены следующие условия:*

1) *плотность* $K(y)$ *непрерывна и ограничена, причем:*
$\alpha = \int K^2(y)\,dy < \infty$

2) $\delta_N \to 0$ *при* $N \to \infty$ *так, что* $N\delta_N \to \infty$.

Тогда

$$\widehat{p}_N(z) = q_N(z) + \xi_N(z)/\sqrt{N\delta_N},$$

где $q_N(z) \to p(z)$ *при почти всех* z, *а случайные величины* $\xi_N(z)$ *асимптотически нормальны:* $\xi_N(z) \Rightarrow \xi(z) \sim \mathcal{N}(0, \alpha p(z))$.

Ответ на вопрос о наилучшем выборе ядра $K(y)$ и выборе скорости сходимости к нулю *ширины окна* δ_N зависит от свойств гладкости оцениваемой плотности $p(x)$. Можно доказать (см., например, Лагутин М.Б. "Наглядная математическая статистика М.: БИНОМ. Лаборатория знаний, 2007, с.389), что если носитель распределения (множество $\{x : p(x) > 0\}$) – конечный интервал, а плотность $p(x)$ дважды непрерывно дифференцируема, причем $\int_{-\infty}^{\infty}[p''(x)]^2\,dx < \infty$, то среди непрерывных ограниченных ядерных функций $K(y)$, удовлетворяющих условиям $K(-y) = K(y)$ (четность) и $\int_{-\infty}^{\infty} K^2(y)\,dy < \infty$, оптимальным по точности получаемой оценки является ядро Епанечникова (Epanechnikov, 1969))[2] (см. табл. 2.1), рекомендуемая при этом ширина ядра δ_N – величина порядка $N^{-1/5}$, достигаемый при этом выборе порядок точности оценки составляет $N^{-2/5}$.

В общей ситуации конкуренцию ядру Епанечникова могут составить и другие ядерные функции. Например, перечисленные в следующей таблице.

Таблица 2.1

№	Ядро	$K(y)$
1	Епанечникова (Бартлетта)	$(3/4)\big(1-y^2\big)\mathbb{I}_{\{\lvert y\rvert\le 1\}}$
2	Квадратическое	$(15/16)\big(1-y^2\big)\mathbb{I}_{\{\lvert y\rvert\le 1\}}$
3	Треугольное (Симпсона)	$\big(1-\lvert y\rvert\big)\mathbb{I}_{\{\lvert y\rvert\le 1\}}$
4	Нормальное (Гаусса)	$(1/\sqrt{2\,\pi})\exp(-y^2/2)$
5	Прямоугольное (равномерное)	$(1/2)\,\mathbb{I}_{\{\lvert y\rvert\le 1\}}$
6	Двойное экспоненциальное (Лапласа)	$(1/2)\,\exp(-\lvert y\rvert)$

Лабораторная работа № 2

Методы оценки плотности распределения, ядерные оценки

Задание

1. В качестве выборки использовать результаты моделирования смеси распределений, полученные в лабораторной работе №1, согласно своему варианту.

2. Построить графики следующих функций в одном графическом окне с наложением:

а) график плотности $p(x)$;
б) график гистограммы $\widehat{g}_N(x)$;

[2]Известно, что это ядро было введено несколько раньше Бартлеттом (Bartlett, 1963).

в) график полигона частот $\widehat{l}_N(x)$;
г) графики ядерных оценок плотности $\widehat{p}_N(x)$ при различных ядерных функциях[3].

3. Сравнить визуально точность применяемых методов оценки плотности, изменяя тип ядра, ширину ядра δ_N, объем выборки.

Примеры функций, реализующих вычисление значений ядер (в MATLAB)

function $z = kernel4(y)$ % *Ядро Гаусса* (№ 4 в табл. 2.1):
$z = 1/sqrt(2 * pi) * exp(-y.\hat{}2/2);$

function $z = kernel5(y)$ % *Прямоугольное ядро* (№ 5 в табл. 2.1):
$z = (abs(y) <= 1/2);$

function $z = kernel3(y)$ % *Ядро Симпсона* (№ 3 в табл. 2.1):
$z = (y + 1). * (-1 <= y \ \& \ y < 0) + (1 - y). * (0 <= y \ \& \ y <= 1);$

Пример выполнения задания для показательного распределения (в MATLAB)

clc; % Очистка командного окна
$N = 1000;;$ % Задание объема выборки
clf % Инициализация графического окна
$L = 1.5;$ % Параметр показательного распределения
$X = -log(rand(1, N))/L;$ % Моделирование выборки из показательного
% распределения объема $N = 1000$
$Xs = sort(X);$ % Сортировка (упорядочение) выборки
$a = 0;$ % Задание точками a и b границ диапазона реализовавшихся значений
$b = Xs(N) + 1;$ % случайной величины, на котором будут строится графики
$h = 0.01;$ % Задание величины шага
$x = a : h : b;$ % Точки разбиения отрезка $[a, b]$ с шагом h (задание сетки)
$f = L * exp(-L * x);$ % вычисление вектора значений плотности показатель-
% ного распределения с параметром L в точках x
$plot(x, f);$ % построение графика функции плотности
hold on % Включение режима наложения графиков
$m =$*hist1*; % Вызов функции, строящей графики гистограммы и полигона
% частот, приведенной на с. **??**.
$h1 = 0.001;$ % Задание шага для построения ядерных оценок,
$dN = N\hat{}(-1/5);$ % Задание ширины ядра
$x1 = a : h1 : b;$ % Создание сетки
$l = length(x1);$ % Определение числа элементов вектора $x1$ (узлов сетки)
$for\ i = 1 : l$ % Вычисление значений ядерных оценок в l точках $x(i)$
$f4(i) = 1/(N * dN) * sum(kernel4((Xs - x1(i))/dN));$
$f5(i) = 1/(N * dN) * sum(kernel5((Xs - x1(i))/dN));$
$f3(i) = 1/(N * dN) * sum(kernel3((Xs - x1(i))/dN));$
end

[3]Вычисление значения ядерной функции удобно оформлять как функцию , которую можно записывать в отдельном скрипт-файле (с расширением r) в текущей папке и вызывать из основной программы, или оформлять как функцию в теле основной программы (в этом случае описание должно предшествовать вызову функции), формат функции в $\mathbb{R}$ см. help(function)

```
plot(x1, f4, 'r');   % Построение графиков
plot(x1, f5, 'g');   % ядерных оценок (для трех видов ядер)
plot(x1, f3, 'm');   % плотности распределения
hold off             % Отключение режима наложения графиков
```