УДК 004.02

ИССЛЕДОВАНИЕ ПРИМЕНЕНИЯ ЯДЕРНОЙ ОЦЕНКИ ПЛОТНОСТИ ПРИ РЕШЕНИИ ЗАДАЧИ АППРОКСИМАЦИИ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ

© Олешко Р.С., Лёзина И.В.

e-mail: zonaRostATY63@yandex.ru

Самарский национальный исследовательский университет имени академика С.П. Королёва г. Самара, Российская Федерация

При моделировании непрерывного процесса известными данными является выборка некоторых значений функции, описывающей этот процесс, в случайные моменты времени. В таком случае, элементы выборки считаются независимыми случайными величинами, имеющими одинаковое распределение. Основной задачей, в данном случае, является оценка плотности этого распределения.

Стандартным подходом и наиболее простым решением для оценки плотности являются гистограммы. Использование сглаженных гистограмм способно обеспечить высокое качество оценок плотности вероятности, однако, лишь в случае последовательностей большой длины.

Наряду с гистограммами для оценки плотности используются также оценки вида:

$$f(x) = (n * h)^{-1} * \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \quad (1)$$

где h – малый параметр, обозначающий диапазон, параметр сглаживания; значение $\frac{x-x_i}{h}$ обозначается как некоторый параметр u; K(u) – функция ядра, удовлетворяющая условиям: K(u) ≥ 0, K(u) = K(-u), $\int K(u)\,du = 1$, K(u) → 0, u→∞ [1].

Такой способ вычисления плотности распределения вероятности называется ядерной оценкой плотности (kernel density estimation). В отличие от построения гистограмм, функции ядерной оценки плотности не зависят от выбора положения разрядов и, вследствие этого, легко обобщаются на многомерный случай.

Ядро (K(u)) – это непрерывная ограниченная симметрическая вещественная функция с единичным интегралом. Результирующая оценка плотности так же будет являться непрерывной и дифференцируемой, как и функция-ядро.

В основе разрабатываемого приложения лежит исследование оценки плотности распределения вероятности различными ядерными функциями и исследование их эффективности относительно друг друга.

Разработанное приложение использует следующие виды ядер:

1) ядро Гаусса (K(u) = $\frac{1}{\sqrt{2\pi}} * exp\left(\frac{-u^2}{2}\right)$) – ядро с неограниченным носителем (определено при изменении его параметра от минус до плюс бесконечности). Является примером радиально базисной – функции ядра;

2) ядро Епанечникова (K(u) = $\frac{3}{4} * max\{1 - u^2, 0\}$) – характеризуется наилучшими результатами при работе с полиномами первой степени [2]. Так же, ядро является ядром с конечным носителем, все значения вне диапазона не существуют, т.е. = 0;

3) треугольное ядро (K(u) = max{1-abs(u), 0}) – так же называемое неупорядоченным ядром, является примером только условно положительно определенного ядра;

4) экспоненциальное ядро, тесно связанное с ядром Гаусса, имеет ограничения на области определения [3];

5) косинусное ядро (K(u) = $\frac{\pi}{4} * cos\left(\frac{\pi * u}{2}\right) * I(|u| \leq 1)$).

Выбор ядра будет влиять на гладкость и дифференцируемость итогового распределения. Однако значительным параметром метода оценки плотности распределения вероятности при помощи ядерной оценки плотности является ширина полосы пропускания (h).Она определяет поведение оценки в конечных выборках.

Так же, следует отметить, что при малом значении полосы пропускания получится тот же эффект, что и при использовании гистограммы значений. В обратном случае, плотность распределения может выродиться в константу.

В алгоритме разработанного приложения, вычисление ширины полосы пропускания производится по наилучшему выявленному критерию. Например, вычисление h для экспоненциального ядра и ядра Гаусса осуществляются по эмпирическому правилу Сильвермана, где ширина окна рассчитывается по формуле:

$$h = 0.9 * A * n^{\frac{-1}{5}} \quad (2)$$

где A – вычисляется как минимальное из значений стандартного отклонения последовательности и интерквартильного диапазона последовательности, разделенного в последствии на 1.34; n – длина исходного массива данных.

Во многих случаях, при отклонении формы оцениваемой плотности от нормальной, учет интерквартильного диапазона позволяет скорректировать в меньшую сторону значение величины h, что делает предложенную оценку достаточно универсальной [4].

Так же следует отметить, что чем более плотно расположены значения объектов выборки, тем меньшее значение параметра ширины полосы пропускания будет получено.

Библиографический список

1. Айвазян, С. А. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное изд. [Текст]/ С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М.: Финансы и статистика, 1983. – 471 с.

2. Christopher G. Small D.L. McLeish, Hilbert Space Methods in Probability and Statistical Inference [Текст]/, Copyright © 1994 John Wiley & Sons, Inc. – 228 p.

3. Воронцов, К. В. Математические методы обучения по прецедентам [Текст]/. 2012.

4. Simon J. Sheather, Density Estimation [Текст]/, Statistical Science 2004, Vol. 19, No. 4, 588–597.DOI: 10.1214/088342304000000297 © Institute of Mathematical Statistics, 20.