Непараметрическое оценивание плотности распределения вероятностей

Кирилл Андреев

k.andreev@skoltech.ru

Сколковский институт науки и технологий (Сколтех)

Современные методы теории информации, оптимизации и управления

Сочи, Россия

2–23 августа, 2020

Содержание

1. Введение
2. Критерии эффективности. Дилемма смещения–дисперсии
3. Выбор ширины и формы окна
4. Применение KDE в анализе нейронных сетей

Введение

Мотивация

- Оценка энтропии (а в общем случае любых интегральных функционалов) сложных многомерных распределений
- Определить условия, при которых эти оценки будут сходиться, а также определить скорость сходимости

M.P. Wand, M.C. Jones
"Kernel smoothing" // Monographs on Statistics and Applied Probability, 1995

Ядерная оценка плотности

- Будем строить оценки плотности распределений в виде

$$p(x) = \sum_{i=1}^{N} K(x - x_i)$$

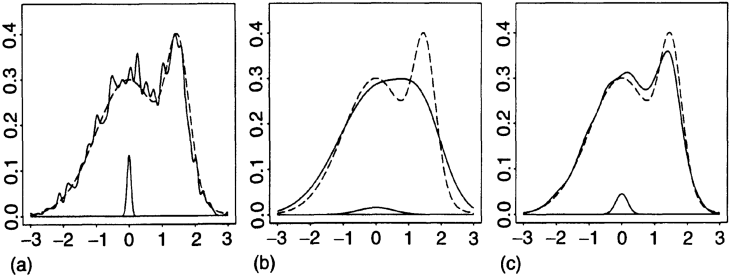
- $x_i \in \mathbb{R}^d$ – элемент располагаемой выборки
- N – объем выборки
- d – размерность пространства

Z. Goldfeld, E. Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, Y. Polyanskiy
Estimating Information Flow in Deep Neural Networks // arXiv: 1810.05728

Z. Goldfeld, K. Greenewald, J. Weed and Y. Polyanskiy
Optimality of the Plug-in Estimator for Differential Entropy Estimation under Gaussian Convolutions // 2019 IEEE Int. Symp. on Inf.n Theory (ISIT), Paris, France, 2019

Yanjun Han, Jiantao Jiao, Tsachy Weissman, Yihong Wu
Optimal rates of entropy estimation over Lipschitz balls // arXiv: 1711.02141

Основные задачи в одномерном случае



- Каковы критерии оптимальности ядерной оценки плотности?
- При какой ширине окна и форме оконной функции достигается оптимальная оценка?
- Какова скорость сходимости ядерной оценки плотности и при каких условиях?

$$\hat{f}(h,x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

Содержание

1. Введение
2. Критерии эффективности. Дилемма смещения–дисперсии
3. Выбор ширины и формы окна
4. Применение KDE в анализе нейронных сетей

Эффективность параметрических оценок

$$MSE\left(\hat{\theta}\right) = \mathbb{E}\left(\hat{\theta} - \theta\right)^2 = Var\left(\hat{\theta}\right) + \left(\mathbb{E}\hat{\theta} - \theta\right)^2$$

- Аналогичным образом рассмотрим ошибку оценивания плотности в точке x_0

$$\mathbb{E}\hat{f}\left(x_0, h\right) = \mathbb{E}K_h\left(x_0 - x\right) = \int K_h\left(x_0 - y\right) f\left(y\right) dy = \left(K_h * f\right)\left(x_0\right) - f\left(x_0\right)$$

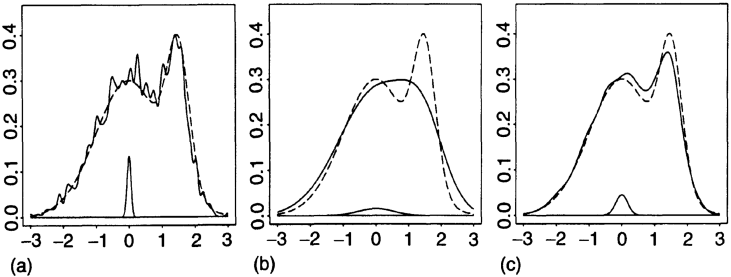
- x – реализация случайной величины с плотностью распределения вероятностей $f(x)$

$$Var\left\{\hat{f}\left(x_0, h\right)\right\} = \frac{1}{N}\left(\left(K_h^2 * f\right)\left(x_0\right) - \left(K_h * f\right)^2\left(x_0\right)\right)$$

$$MSE\left\{\hat{f}\left(x_0, h\right)\right\} = \frac{1}{N}\left(\left(K_h^2 * f\right)\left(x_0\right) - \left(K_h * f\right)^2\left(x_0\right)\right) + \left\{\left(K_h * f\right)\left(x_0\right) - f\left(x_0\right)\right\}^2$$

Дилемма смещения-дисперсии

$$MSE\left\{\hat{f}\left(x_0, h\right)\right\} = \frac{1}{N}\left(\left(K_h^2 * f\right)\left(x_0\right) - \left(K_h * f\right)^2\left(x_0\right)\right) + \left\{\left(K_h * f\right)\left(x_0\right) - f\left(x_0\right)\right\}^2$$



Интегральные критерии непараметрических оценок плотности

$$MSE\left\{\hat{f}\left(x_0,h\right)\right\} = \frac{1}{N}\left(\left(K_h^2 * f\right)\left(x_0\right) - \left(K_h * f\right)^2\left(x_0\right)\right) + \left\{\left(K_h * f\right)\left(x_0\right) - f\left(x_0\right)\right\}^2$$

$$MISE(h) = \int MSE\left\{\hat{f}\left(x,h\right)\right\}dx = \int \mathbb{E}\left\{\hat{f}(x,h) - f(x)\right\}^2 dx$$

Ограничения на функцию K

$$\int K(z)dz = 1, \quad \int zK(z)dz = 0, \quad \int z^2K(z)dz = \mu_2(K) < +\infty$$

$$\mathbb{E}\hat{f}(x,h) = \int K(z)f(x-hz)dz$$

$$\mathbb{E}\hat{f}(x,h) = f(x) + \frac{1}{2}h^2f''(x)\int z^2K(z)dx + o(h^2)$$

Интегральные критерии непараметрических оценок плотности

$$Var\left\{\hat{f}\left(x_0,h\right)\right\} = \frac{1}{N}\left(\left(K_h^2 * f\right)(x_0) - \left(K_h * f\right)^2(x_0)\right)$$

$$Var\left\{\hat{f}\left(x_0,h\right)\right\} = \frac{1}{Nh}\int K^2(z)f(x-hz)dz - \frac{1}{N}\left\{\mathbb{E}\hat{f}(x,h)\right\}^2$$

$$Var\left\{\hat{f}\left(x_0,h\right)\right\} = \frac{1}{Nh}\int K^2(z)\left(f(x)+o(1)\right)dz - \frac{1}{N}\left\{f(x)+o(1)\right\}^2$$

$$Var\left\{\hat{f}\left(x_0,h\right)\right\} = \frac{1}{Nh}R(K)\left(f(x)+o(1)\right)dz + o\left(\frac{1}{Nh}\right), \text{ где } \quad R(K) = \int f^2(x)dx$$

Асимптотические интегральные оценки

$$MISE(h, N, f) = \frac{1}{Nh} R(K) f(x) + \frac{1}{4} h^2 \mu_2^2(K) f''(x)^2 + o\left(\frac{1}{Nh} + h^4\right)$$

$$MISE(h, N, f) = AMISE(h, N) + o\left(\frac{1}{Nh} + h^4\right)$$

Асимптотическая формула по h и N:

$$AMISE(h, N, f) = \frac{1}{Nh} R(K) + \frac{1}{4} h^2 \mu_2^2(K) R(f'')$$

- $\mu_2(\nu)$ – второй момент $\int \nu^2 f(\nu) dx$
- $R(\nu)$ – интеграл от квадрата $\int \nu^2 d\nu$

Содержание

1. Введение
2. Критерии эффективности. Дилемма смещения–дисперсии
3. Выбор ширины и формы окна
4. Применение KDE в анализе нейронных сетей

Оптимальная ширина окна

$$AMISE(h, N, f) = \frac{1}{Nh} R(K) + \frac{1}{4} h^2 \mu_2^2(K) R(f'')$$

- Минимальное значение $AMISE$ доставляется следующей шириной окна:

$$h_{AMISE} = \left\{ \frac{R(K)}{\mu_2(K)^2 R(f'') n} \right\}^{1/5}$$

- Асимптотическое значение интегральной MSE

$$\inf_{h>0} AMISE = \frac{5}{4} \left\{ \mu_2(K)^2 R(K)^4 R(f'') \right\}^{1/5} \frac{1}{n^{4/5}}$$

Задача выбора оптимального окна

$$AMISE(h, N, f) = \frac{1}{Nh} R(K) + \frac{1}{4} h^2 \mu_2^2(K) R(f'')$$

- Если $K_\delta(x) = K\left(\frac{x}{\delta}\right)/\delta$, то задача оптимизации AMISE факторизуется
- Данное условие выполнено, когда $R(K_\delta) = \mu_2(K_\delta)^2 \Rightarrow \delta_0 = \left\{\frac{R(K)}{\mu_2(K)^2}\right\}^{1/5}$
- Минимум интегральной ошибки достигается для окна Епанечникова

Kernel	$\{C(K^*)/C(K)\}^{5/4}$
Epanechnikov	1.000
Biweight	0.994
Triweight	0.987
Normal	0.951
Triangular	0.986
Uniform	0.930

Комментарии

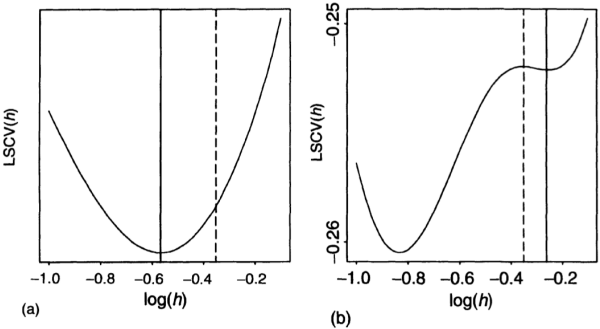
- Возможно применение переменной ширины окна

$$h_{AMSE} = \left\{ \frac{R(K) f(x)}{\mu_2(K)^2 f''(x)^2 n} \right\}^{1/5}$$

- Выбор формы окна не оказывает существенного эффекта на скорость сходимости

Least squares cross validation

$$LSCV(h) = \int \hat{h}(x)^2 dx - \frac{2}{N}\sum_{i=1}^{N} \hat{f}_{-i}(X_i, h), \quad \hat{f}_{-i}(X_i, h) = \frac{1}{N-1}\sum_{j=1, j\neq i}^{N} K_h(x - x_j)$$ [1]



[1] $LSCV$ – несмещенная оценка $MISE(h) - \int f^2(x)dx$

Многомерный случай – проклятие размерностей

$$\hat{f}(x, \mathbf{H}) = \frac{1}{N}\sum_{i=1}^{N} K_{\mathbf{H}}(x - x_i), \quad K_{\mathbf{H}}(x) = |H|^{-1/2} K\left(\mathbf{H}^{1/2}x\right), \int K(x)dx = 1$$

- Предположим, что $\mathbf{H} = h^2\mathbf{I}$
- Аналогично одномерному случаю, определим $AMISE$, после чего определим оптимальную ширину окна

$$h_{AMISE} = \left\{\frac{d \cdot R(K)}{N\mu_2(K)^2 \int \nabla^2 f(x)^2 dx}\right\}^{\frac{1}{d+4}} \sim C_0 \frac{1}{N^{1/(d+4)}}$$

$$\inf_{h>0} AMISE \sim C_1 \left(\frac{1}{N^4}\right)^{1/(d+4)}$$

- Ядерная оценка плотности требует экспоненциального объема выборки
- Асимптотика существенно ухудшается с ростом d

Содержание

1. Введение
2. Критерии эффективности. Дилемма смещения–дисперсии
3. Выбор ширины и формы окна
4. Применение KDE в анализе нейронных сетей

Skoltech
Skolkovo Institute of Science and Technology

Оценка интегральных функционалов

- Использовать разложение фон-Мизеса[2]
- Использовать “plug-in” метод оценивания
  - В общем случае «встраиваемое» оценивание может оказаться неэффективным[3]
  - Добавление Гауссового шума к выходу нейронно сети эквивалентно свертки плотности распределения вероятностей исходного распределения с плотностью нормального распределения
  - Свертка с нормальным распределением позволяет эффективно использовать встраиваемый метод оценивания энтропии и взаимной информации[4]

---

[2] “Nonparametric von Mises Estimators for Entropies, Divergences and Mutual Informations”, `https://www.cs.cmu.edu/~bapoczos/articles/kandasamy15NIPS.pdf`

[3] “Optimal rates of entropy estimation over Lipschitz balls”, `https://arxiv.org/pdf/1711.02141.pdf`

[4] “Optimality of the Plug-in Estimator for Differential Entropy Estimation under Gaussian Convolutions”, `http://people.lids.mit.edu/yp/homepage/data/isit2019_optimalit.pdf`

Спасибо за внимание!

СИРИУС

ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР

Skoltech
Skolkovo Institute of Science and Technology