Непараметрическое оценивание. Бутстреп

Центр биоэлектрических интерфейсов, 21 ноября 2018 г.

Денис Деркач, Влад Белавин

Оглавление

Непараметрическое оценивание

MSE, функция риска

Пусть в точке x_0 построена оценка $\hat{p}_n(x_0)$ плотности. Рассматривая квадратичную функцию потерь, приходим к следующему понятию.

Определение

Mean Square Error:

$$MSE(\hat{p}_n, p; x_0) = \mathbb{E}_p[(\hat{p}_n(x_0) - p(x_0))^2].$$

MISE

Если же построена оценка $\hat{p}_n(x)\, \forall x \in \mathbb{R}$, то

Определение

Mean Integrated Squared Error:

$$MISE(\hat{p}_n, p) = \mathbb{E}_p \left[\int_{\mathbb{R}} (\hat{p}_n(x) - p(x))^2 dx \right].$$

Bias

Определение

Смещение (bias)

$bias(x_0) = \mathbb{E}_p \hat{p}_n(x_0) - p(x_0)$

Разложение ошибки

Лемма

$$MSE(\hat{p}_n, p, x_0) = bias^2(x_0) + \mathrm{Var_p \hat{p}_n (x_0)}$$

Лемма

$$MISE(\hat{p}_n, p) = \int\limits_{\mathbb{R}} bias^2(x)dx + \int\limits_{\mathbb{R}} \mathrm{Var_p \hat{p}_n (x) dx}$$

Доверительная трубка для плотности

Пусть необходимо построить доверительные интервалы для p.
Для этого будем использовать гистограмму $\hat{p}_n(x)$, определенную ранее.
Определим

$$\overline{p_n}(x) = \mathbb{E}\hat{p}_n(x) = \frac{\int_{\Delta_j} p(u)du}{h} \text{ для } x \in \Delta_j.$$

По сути, $\overline{p_n}$ - "гистограммное" усреднение плотности p.

Определение

Пара функций $(p_-(x), p_+(x))$ является $1 - \alpha$ доверительной областью (трубкой), если для любого x:

$$\mathbb{P}_p\left(p_-(x) \le \overline{p_n}(x) \le p_+(x)\right) \ge 1 - \alpha$$

Доверительная трубка для плотности

Теорема

Пусть $M = M(n)$ - число ячеек в гистограмме $\hat{p}_n$, причем $M(n) \to \infty$ и $\frac{M(n)\log(n)}{n} \to \infty$ при $n \to \infty$.

Определим

$$p_-(x) = (max\{\sqrt{\hat{p}_n(x)} - C, 0\})^2, p_+(x) = (\sqrt{\hat{p}_n(x)} + C)^2,$$

где $C = \frac{1}{2} z_{\frac{\alpha}{2M}} \sqrt{\frac{M}{n(b-a)}}$

Тогда $(p_-(x), p_+(x))$ является $1-\alpha$ доверительным интервалом.

Доверительная трубка для плотности

Из центральной предельной теоремы

$$\frac{\nu_j}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{x_i \in \Delta_j\} \sim \mathcal{N}\left(\int\limits_{\Delta_j} p(x)dx, \frac{\int\limits_{\Delta_j} p(x)dx(1 - \int\limits_{\Delta_j} p(x)dx)}{n}\right)$$

Согласно дельта-методу $\sqrt{\frac{\nu_j}{n}} \sim \mathcal{N}\left(\sqrt{\int\limits_{\Delta_j} p(x)dx}, \frac{1}{4n}\right)$. Более того, можно показать, что $\sqrt{\frac{\nu_j}{n}}$ приблизительно независимы. Тогда $2\sqrt{n}\left(\sqrt{\frac{\nu_j}{n}} - \sqrt{\int\limits_{\Delta_j} p(x)dx}\right) \approx \xi_j$, где $\xi_0, \ldots, \xi_{M-1} \sim \mathcal{N}(0,1)$.

Доверительная трубка

$$
\begin{aligned}
A &= \{p_-(x) \leq \overline{p_n}(x) \leq p_+(x) \forall x\} = \\
&= \{\sqrt{p_-(x)} - c \leq \sqrt{\overline{p_n}(x)} \leq \sqrt{p_+(x)} + c \forall x\ \} = \\
&= \{\max_x |\sqrt{\hat{p}(x)} - \sqrt{\overline{p_n}(x)}| \leq c\}
\end{aligned}
$$

Доверительная трубка для плотности

Тогда $\mathbb{P}(A^c) = \mathbb{P}\{\max\limits_x |\sqrt{\hat{p}_n(x)} - \sqrt{\overline{p_n}(x)}| > c\} =$

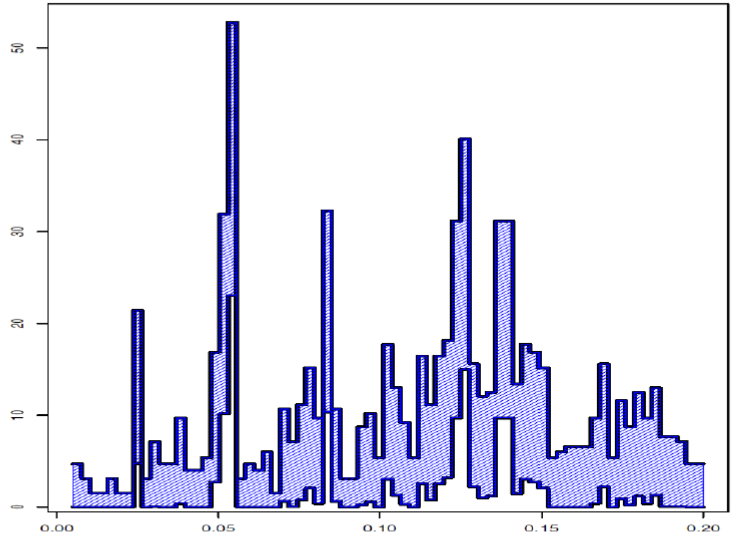
$$\mathbb{P}\left\{\max_{j=\overline{0,M-1}} \left| \sqrt{\frac{\nu_j}{nh}} - \sqrt{\frac{\int_{\Delta_j} p(x)dxh}{h}} \right| > c \right\} \approx$$

$$\mathbb{P}\left\{\max_{j=\overline{0,M-1}} \frac{|\xi_j|}{2\sqrt{nh}} > \frac{z_{\frac{\alpha}{2n}}}{2}\sqrt{\frac{M}{n(b-a)}}\right\} = \mathbb{P}\{\max_{j=\overline{0,M-1}} |\xi_j| > z_{\frac{\alpha}{2M}}\} \leq$$

$$\sum_{j=0}^{M-1} \mathbb{P}\{|\xi_j| > z_{\frac{\alpha}{2M}}\} = \sum_{j=0}^{M-1} \frac{\alpha}{M} = \alpha,$$

т.е. для предъявленных $p_-(x), p_+(x)$ выполнено определение доверительной трубки.

Доверительная трубка для плотности



Комментарии о доверительных трубках

- Важным условием для предыдущего вывода является наличие большого количества семплов n. В случае малого количества семплов ситуация может отличаться, в зависисости от использованного метода оценки.
- Разные отрасли используют разные определения ширины доверительной трубки (от 68% до 100%).

Ядерная оценка плотности

Позволяет получить более гладкие по сравнению с гистограммной оценки, быстрее сходящиеся к плотности.

Определение

Ядро - функция K такая, что

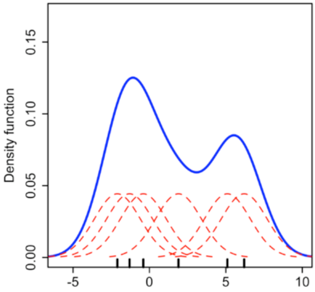
$$K(x) \geq 0, \int\limits_{\mathbb{R}} K(x)dx = 1, \int\limits_{\mathbb{R}} xK(x)dx = 0, \sigma_K^2 \equiv \int\limits_{\mathbb{R}} x^2 K(x)dx$$

Ядерная оценка плотности

Определение

Ядерная оценка плотности имеет вид:

$$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), h \text{ — \underline{ширина ядра}}$$



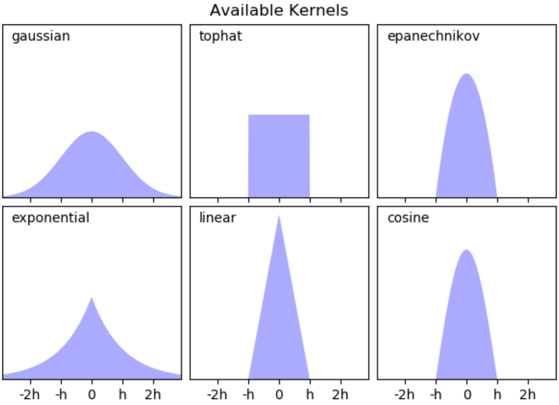
Виды ядер

Examples

- $K(x) = \frac{1}{2}\mathbb{I}\{|x| < 1\}$ — прямоугольное ядро
- $K(x) = (1 - |x|)\mathbb{I}\{|x| < 1\}$ — треугольное ядро
- $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ — Гауссовское ядро
- $K(x) = \frac{3}{4}(1 - x^2)\mathbb{I}\{|x| < 1\}$ — ядро Епанечникова

Далее мы будем рассматривать только гладкие ядра.

Примеры ядер



Вид ядерной функции K влияет на “качество” оценки не так сильно, как выбор ширины ядра h.

Picture credit:http://scikit-learn.org/stable/modules/density.html

Выбор ширины ядра

Теорема

$$MISE(\hat{p}_n, p) \approx \frac{1}{4}\sigma_K^4 h^4 \int\limits_{\mathbb{R}} (p''(x))^2 dx + \frac{1}{nh} \int\limits_{\mathbb{R}} (K(x))^2 dx$$

Минимум достигается при $h = h^*$:

$$h^* = \left(\frac{1}{n} \frac{\int\limits_{\mathbb{R}} (K(x))^2 dx}{\left(\int\limits_{\mathbb{R}} x^2 K(x) dx \right)^2 \left(\int\limits_{\mathbb{R}} p''(x))^2 dx \right)} \right)^{\frac{1}{5}}$$

При этом $MISE(\hat{p}_n, p) = O\left(n^{-\frac{4}{5}}\right)$

Выбор ширины ядра

Воспользуемся bias-variance decomposition:

$$bias(x) = \mathbb{E}_p \hat{p}_n(x) - p(x) =$$

$$\int\limits_{\mathbb{R}} \left(\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)\right) p(x_1) \ldots p(x_n) dx_1 \ldots dx_n -$$

$$\frac{1}{n} \sum_{i=1}^{n} \int\limits_{\mathbb{R}} K(z) p(x) dz \approx \int\limits_{\mathbb{R}} K(z) [-p'(x) zh + p''(x) \frac{(zh)^2}{2}] dz =$$

$$\frac{1}{2} \sigma_K^2 h^2 p''(x)$$

$$\int\limits_{\mathbb{R}} (bias(x))^2 dx = \frac{1}{4} \sigma_K^4 h^4 \int\limits_{\mathbb{R}} [p''(x)]^2 dx$$

Выбор ширины ядра

$$\int\limits_{\mathbb{R}} \mathbb{V}ar_p \hat{p}_n(x)dx = \int\limits_{\mathbb{R}} \mathbb{V}ar_p[\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})dx =$$

$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} \mathbb{V}ar_p K(\frac{x-x_i}{h})dx \leq \frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} \mathbb{E}_p K(\frac{x-x_i}{h})^2 dx =$$

$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}} K(\frac{x-x_i}{h})^2 p(x_i)dx_i dx =$$

$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} p(x_i)\int\limits_{\mathbb{R}} K(\frac{x-x_i}{h})^2 dx dx_i =$$

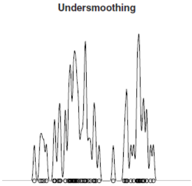
$$\frac{1}{(nh)^2}\sum_{i=1}^{n}\int\limits_{\mathbb{R}} p(x_i)dx_i h\int\limits_{\mathbb{R}} K^2(z)dz = \frac{1}{nh}\int\limits_{\mathbb{R}} K^2(z)dz$$

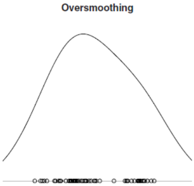
Минимум $MISE(\hat{p}_n, p)$ достигается в некотором h^*.

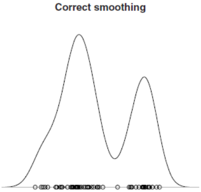
Выбор ширины ядра

Подставляя h^* в $\hat{p}_n$, получаем, что $MISE = O(n^{-\frac{4}{5}})$, т.е. сходимость ядерной оценки лучше, чем у гистограммы. Можно показать, что при достаточно общих условиях нельзя получить скорость лучше, чем $n^{-\frac{4}{5}}$.

Выбор ширины ядра







Как и в случае с гистограммной, при больших h имеет место oversmoothing, а при маленьких - undersmoothing из-за bias-variance tradeoff.

Доверительный интервал

Определим $\overline{p_n}(x) = \mathbb{E}\hat{p}_n(x) = \int\limits_{\mathbb{R}} \frac{1}{h}K(\frac{x-u}{h})p(u)du$. Допустим, что $supp(p) \subset (a, b)$.

Тогда определим $(1-\alpha)$ доверительную трубку.

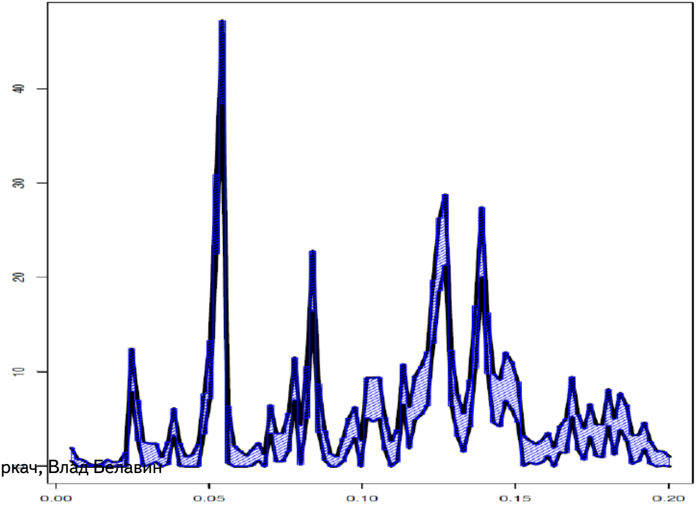
$$p_-(x) = \hat{p}_n(x) - \frac{z_\alpha}{\sqrt{n}}s(x),$$

$$p_+(x) = \hat{p}_n(x) + \frac{z_\alpha}{\sqrt{n}}s(x).$$

Где $s^2(x) = \frac{1}{n-1}\sum\limits_{i=1}^{n}[Y_i(x) - \overline{Y_n}(x)]^2, Y_i(x) = \frac{1}{h}K(\frac{x-X_i}{h})$,

$z_\alpha = \Phi^{-1}\left(\frac{1+(1-\alpha)^{\frac{w}{b-a}}}{2}\right)$, $\Phi-$ функция стандартного нормального распределения. w - эффективная ширина ядра.

Доверительный интервал для усредненной плотности



Ядерная оценка плотности: многомерный случай

Пусть теперь данные многомерные, то есть $i-$ое наблюдение - вектор размерности d:

$$X_i = [X_i^1, \dots X_i^d]^T.$$

Пусть $h = [h_1, \dots, h_d]^T$ - вектор ширины ядра вдоль каждого измерения.

Тогда:

$$\hat{p}_n(x) = \frac{1}{nh_1 \cdot \ldots \cdot h_d} \sum_{i=1}^{n} \left[\prod_{j=1}^{d} K\left(\frac{x_j - X_i^j}{h_j} \right) \right],$$

где $x = [x_1, \dots, x_d]^T$ — произвольная точка в $\mathbb{R}^d$

Ядерная оценка плотности: многомерный случай

Для такой оценки риск

$MISE(\hat{p}_n, p) \approx$

$$\frac{1}{4}\sigma_K^4 \left[\sum_{j=1}^{d} h_j^4 \int\limits_{\mathbb{R}^d} p_{jj}^2(x)dx + \sum_{j \neq k} h_j^2 h_k^2 \int\limits_{\mathbb{R}^d} p_{jj}(x) p_{kk}^2(x) dx\right] +$$

$$\frac{\left(\int\limits_{\mathbb{R}^d} K^2(x)dx\right)^d}{nh_1 \cdot \ldots \cdot h_d},$$

где $p_{jj}(x) = \frac{\partial^2 p(x)}{\partial x_j^2}$

Оптимальная ширина ядра $h_i^* \approx cn^{-\frac{1}{4+d}}$

При этом риск имеет порядок: $MISE(\hat{p}_n, p) = O(n^{-\frac{4}{4+d}})$.

Проклятие размерности

Оптимальный порядок риска $O(n^{-\frac{4}{4+d}})$, т.е. наблюдаем "проклятье размерности" - при росте d скорость сходимости к истинной плотности падает.

Рассмотрим таблицу объёмов выборки, необходимых для того, чтобы средний квадрат ошибки в нуле был меньше 0.1 в зависимости от размерности наблюдений в случае многомерной нормальной плотности и оптимальной ширины ядра:

d	1	2	3	4	5	6	7	8	9
n	4	19	67	223	768	2790	10700	43700	187000

где d — размерность данных, n — необходимый объём выборки.

Непараметрическая регрессия

Пусть имеется n наблюдений: $(X_1, Y_1) \dots (X_n, Y_n)$,
сгенерированных из совместной плотности $p(x, y)$.
Наблюдения связаны соотношением:
$Y_i = r(X_i) + \varepsilon_i, \varepsilon_i - i.i.d, \mathbb{E}\varepsilon_i = 0, \mathbb{V}ar\varepsilon_i = \sigma^2$
Необходимо оценить функцию регрессии:

$$r(x) = \mathbb{E}(Y|X = x) = \int\limits_{\mathbb{R}} yp(y|x)dy = \frac{\int\limits_{\mathbb{R}} yp(x,y)dy}{\int\limits_{\mathbb{R}} p(x,y)dy} = \frac{\int\limits_{\mathbb{R}} yp(x,y)dy}{p(x)}.$$

Непараметрическая регрессия

Определение

Пусть $\hat{p}_n(x)$ и $\hat{p}_n(x,y)-$ ядерные оценки плотностей по выборкам $\{X_1,\ldots,X_n\}$ и $\{(X_1,Y_1)\ldots,(X_n,Y_n)\}$ соответственно с ядром K. Тогда, если $\hat{p}_n(x)\neq 0$, то

$$\hat{r}_n(x) = \frac{\int\limits_{\mathbb{R}} y\hat{p}_n(x,y)dy}{\hat{p}_n(x)}$$

.

Непараметрическая регрессия

Для оценки $r(x)$ используется оценка Надарая-Ватсона:

Определение

$\hat{r}_n^{NW} = \sum\limits_{i=1}^{n} w_i(x) Y_i$, где $w_i = \frac{K(\frac{x-X_i}{h})}{\sum\limits_{j=1}^{n} K(\frac{x-X_j}{h})}$, K заданная ядерная функция

Таким образом,это взвешенная сумма Y_i, где точки близкие к x имеют больший вес.

NB: оценку Надарая-Ватсона можно применять и в случае, когда X_i — фиксированные и детерминированные числа (например, $X_i = \frac{i}{n}$).

Непараметрическая регрессия

Перейдём к риску и выбору ширины ядра.

Теорема

$$MISE(\hat{r}_n^{NW}, r) \approx$$
$$\approx \quad \frac{h^4}{4}(\int\limits_{\mathbb{R}} x^2K^2(x)dx)^4 \int (r''(x) \quad + \quad 2r'(x)\frac{p'(x)}{p(x)})^2 dx \quad +$$
$$\frac{1}{h}\int\limits_{\mathbb{R}} \frac{\sigma^2 \int\limits_{\mathbb{R}} K^2(x)dx}{np(x)}dx$$

Оптимальная ширина ядра: $h^* = cn^{-\frac{1}{5}}$

Порядок риска при этой ширине: $MISE(\hat{r}_n^{NW}, r) = O(n^{-\frac{4}{5}})$

Непараметрическая регрессия

Опять же, h^* нельзя выписать на практике, так как она зависит от неизвестных $r(x), p(x)$.

Поэтому минимизируют по h оценку риска

$$\hat{\mathcal{J}}(h) = \sum_{i=1}^{n} (Y_i - \hat{r}_{(-i)}^{NW}(X_i))^2,$$

где $\hat{r}_{(-i)}^{NW}$ - оценка Надарайя-Ватсона, построенная по выборке, из которой удалено наблюдение (X_i, Y_i)

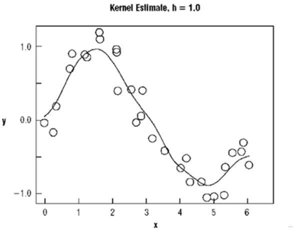
Теорема

$$\hat{\mathcal{J}}(h) = \sum_{i=1}^{n} \left(Y_i - \hat{r}_n^{NW}(X_i)\right)^2 \frac{1}{\left(1 - \frac{K(0)}{\sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)}\right)^2},$$

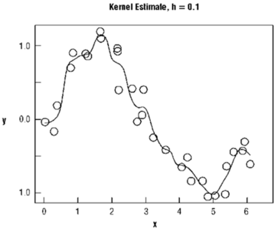
Непараметрическая регрессия

Как и в случае с гистограммной и ядерной оценкой плотности, наблюдается bias-variance tradeoff: при больших h имеет место oversmoothing - оценка слишком сглажена, а при маленьких h имеет место undersmoothing - оценка излишне подстроилась под данные.

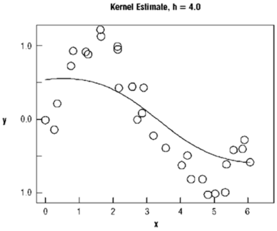
Непараметрическая регрессия



Correct smoothing



Undersmoothing



Oversmoothing

Непараметрическое оценивание

- Доверительная трубка для функции регрессии
- Многомерный случай

Непараметрический бутстреп

- Стандартная постановка задачи
- Идея бутстрепа
- Алгоритм

Параметрический бутстреп

Доверительное оценивание на основе бутстрепа

- Нормальный интервал
- Центральный интервал

Метод складного ножа

Доверительная трубка для функции регрессии

Построим доверительную область. Сначала оценим σ^2. Пусть X_i упорядочены по возрастанию. Предполагая, что $r(x)$ - гладкая функция, получаем $r(X_{i+1}) - r(X_i) \approx 0$.
Тогда:

$$Y_{i+1} - Y_i = [r(X_{i+1}) + \varepsilon_{i+1}] - [r(X_i) + \varepsilon_i] \approx \varepsilon_{i+1} - \varepsilon_i$$

$$\mathbb{V}ar(Y_{i+1} - Y_i) \approx \mathbb{V}ar(\varepsilon_{i+1} - \varepsilon_i) = \mathbb{V}ar\varepsilon_{i+1} + \mathbb{V}ar\varepsilon_i = 2\sigma^2$$

$$\Rightarrow \hat{\sigma}^2 = \tfrac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2$$

Будем строить доверительную область для сглаженной версии $\overline{r}_n(x) = \mathbb{E}(\hat{r}_n^{NW}(x))$ настоящей функции регрессии r.

Доверительная трубка для функции регрессии

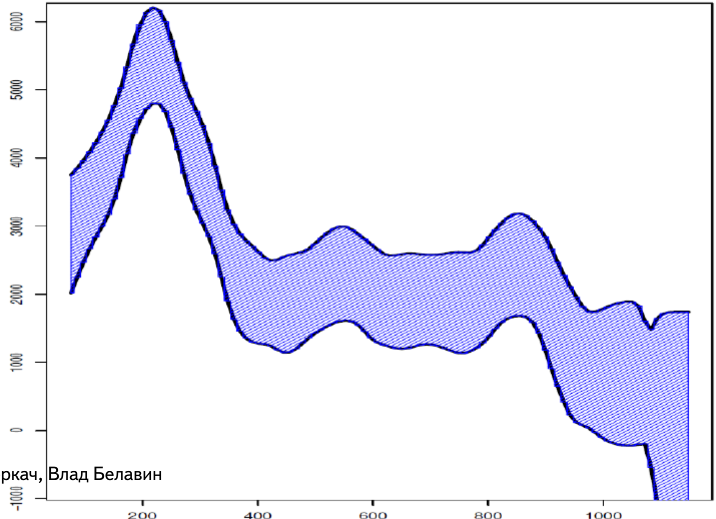
Приближенный $(1-\alpha)$ доверительный интвервал для $\overline{r}_n(X)$ имеет вид:

$$r_-(x) = \hat{r}_n^{NW}(x) - C$$
$$r_+(x) = \hat{r}_n^{NW}(x) + C,$$

Где $\hat{\sigma}, w_i$ - определены выше, $C = z_\alpha \hat{\sigma} \sqrt{\sum_{i=1}^{n} w_i^2(x)}$,

$z_\alpha = \Phi^{-1}\left(\frac{1+(1-\alpha)^{\frac{w}{b-a}}}{2}\right)$, Φ — функция стандартного нормального распределения, w - эффективная ширина ядра, $X_1, \ldots, X_n \in (a;b)$

Доверительная трубка для функции регрессии



Непараметрическое оценивание

- Доверительная трубка для функции регрессии
- Многомерный случай

Непараметрический бутстреп

- Стандартная постановка задачи
- Идея бутстрепа
- Алгоритм

Параметрический бутстреп

Доверительное оценивание на основе бутстрепа

- Нормальный интервал
- Центральный интервал

Метод складного ножа

Непараметрическая регрессия: многомерный случай

Если $X = [X_1, \ldots, X_n]^T$, то из-за проклятия размерности бесполезно обобщать оценку Надарайя-Ватсона аналогично способу, как мы делаем в ядерной оценке плотности.
Вместо этого можно рассмотреть аддитивную модель

- $Y = \sum_{j=1}^{d} r_j(X^j) + \alpha + \varepsilon$
  или
- $Y = \sum_{j=1}^{d} r_j(X^j) + \sum_{j<k} r_{jk}(X^j X^k) + \alpha + \varepsilon$

Непараметрическая регрессия: многомерный случай

Подготовка первой аддитивной модели:

Algorithm (Backfitting)

Инициализация : $\hat{\alpha} = \overline{Y_n}; \hat{r}_1, \dots \hat{r}_d$

Пока не стабилизируется $\hat{r}_1, \dots, \hat{r}_d$ повторять;

- Для всех $j = 1, \dots, d$:
  1. Вычислить $\widetilde{\varepsilon}_i = Y_i - \hat{\alpha} - \sum\limits_{k \neq j} \hat{r}_k(X_i^k), i = 1, \dots, n$
  2. Получить $\hat{r}_j(X^j)$ - функцию регрессии $\widetilde{\varepsilon}_i$ на j-ую компоненту X^j (то есть в качестве наблюдений имеем $\{(X_1^j, \widetilde{\varepsilon}_1), \dots, (X_n^j, \widetilde{\varepsilon}_n)\}$
  3. Положить $\hat{r}_j := \hat{r}_j - \frac{1}{n} \sum\limits_{i=1}^{n} \hat{r}_j(X_i^j)$

Непараметрический бутстреп

Стандартная постановка задачи

- Модель:
  - Имеем конечную простую выборку $\{X_i\}_{i=1}^n \subset \mathbb{R}$, порожденную распределением вероятности F.
  - Задана некоторая статистика $T_n = T_n(X_1, \dots, X_n)$.
- **Задача:** оценить дисперсию $\mathbb{V}ar_F(T_n)$, которая зависит от неизвестного распределения F.

Пример

Пусть $T_n = \overline{X}_n$.
Тогда $\mathbb{V}ar_F(T_n) = \sigma^2/n$, где $\sigma^2 = \int (x-\mu)^2 dF(x)$ и $\mu = \int x dF(x)$. Таким образом, дисперсия T_n есть функция F.

Эмирическая оценка функции распределения

Определение

Если есть выборка $X_i \sim F$ iid, то эмпирическая (выборочная) функция распределения: $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \theta(x - X_i)$

Лемма

Гливенко-Кантелли

$\lim_{n\to\infty} \sup_x (\widehat{F}_n(x) \to F(x)) = 0$ почти наверняка.

Идея бутстрепа

Шаг 1. Оценить $\mathbb{V}ar_F(T_n)$ с помощью $\mathbb{V}ar_{\widehat{F}_n}(T_n)$.

Шаг 2. Приблизить $\mathbb{V}ar_{\widehat{F}_n}(T_n)$ при помощи моделирования.

Пример

- Для $T_n = \overline{X}_n$, $\mathbb{V}ar_{\widehat{F}_n}(T_n) = \widehat{\sigma}^2/n$, где $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$.
- В данном случае шага 1 достаточно.
- Однако зачастую не удаётся выписать явно $\mathbb{V}ar_{\widehat{F}_n}(T_n)$. В таком случае прибегают к шагу 2.

Оценка дисперсии на основе бутстрепа

Предположим, что $Y_1, \ldots, Y_B$ — реализации i. i. d.случайных величин с функцией распределения G. Согласно закону больших чисел,

$$\overline{Y}_n = \frac{1}{B}\sum_{j=1}^{B} Y_j \xrightarrow{P} \int y dG(y) = \mathbb{E}Y, \quad B \to \infty.$$

Таким образом, мы можем использовать $\overline{Y}_n$ при достаточно больших B для приближения $\mathrm{E}\ Y$. Более того, для любой функции h с конечным мат. ожиданием имеем:

$$\frac{1}{B}\sum_{j=1}^{B} h(Y_j) \xrightarrow{P} \int h(y) dG(y) = \mathrm{E}\left(h(Y)\right), \quad B \to \infty.$$

Оценка дисперсии на основе бутстрепа

В частности, это означает, что мы можем моделировать и дисперсию:

$$\frac{1}{B}\sum_{j=1}^{B}(Y_j - \overline{Y}_n)^2 = \frac{1}{B}\sum_{j=1}^{B}(Y_j)^2 - \left(\frac{1}{B}\sum_{j=1}^{B}Y_j\right)^2 \xrightarrow{\mathrm{P}}$$

$$\to \mathrm{P} \int y^2 dG(y) - \left(\int y dG(y)\right)^2 = \mathbb{V}ar(Y), \quad B \to \infty.$$

Это означает, что мы можем использовать выборку для оценки дисперсии. Данная процедура позволяет нам находить $\mathbb{V}ar_{\widehat{F}_n}(T_n)$ — «дисперсию T_n при данных, распределённых по $\widehat{F}_n$».

Оценка дисперсии на основе бутстрепа

Теперь ситуация выглядит следующим образом.

- С точки зрения реальности:

$$F \Rightarrow X_1, \ldots, X_n \Rightarrow T_n = g(X_1, \ldots, X_n)$$

- С точки зрения бутстрепа:

$$\widehat{F}_n \Rightarrow X_1^*, \ldots, X_n^* \Rightarrow T_n^* = g(X_1^*, \ldots, X_n^*)$$

- **Проблема:** как получить $X_1^*, \ldots, X_n^*$ из $\widehat{F}_n$?
- **Решение:** при подсчёте мат. ожидания с помощью $\widehat{F}_n$ мы использовали одинаковую массу $\frac{1}{n}$. Это значит, что получение наблюдения из $\widehat{F}_n$ эквивалентно выбору случайной точки из исходной выборки.

Алгоритм: оценка дисперсии

Приведём **алгоритм** оценки дисперсии с помощью бутстрепа:

1. Выбираем $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$
2. Вычисляем $T_n^* = g(X_1^*, \ldots, X_n^*)$
3. Повторяем шаги 1 и 2 пока не получим $T_{n,1}^*, \ldots, T_{n,B}^*$
4. Положим

$$v_{boot} = \frac{1}{B}\sum_{b=1}^{B}(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*)^2$$

В итоге получаем:

$$\mathbb{V}ar_F(T_n) \approx \mathbb{V}ar_{\widehat{F}}(T_n) \approx v_{boot}$$

Количество выборок B

Количество B зависит от необходимой точности (и компьютерного времени). Скорее всего для оценки дисперсии необходимо достаточно будет поставить $50 < B < 200$. Более точная оценка:

Лемма

$\mathbb{V}ar_B T_n = (\mathbb{V}ar^2_{B\to\infty}\widehat{T}_n + \frac{E(\Delta)+2}{4B})^{\frac{1}{2}}$, где Δ зависит от распредения F и пропорциональная коэффициенту эксцесса.

Алгоритм: оценка смещения

Приведём **алгоритм** оценки смещения с помощью бутстрепа:

1. Выбираем $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$
2. Вычисляем $T_n^* = g(X_1^*, \ldots, X_n^*)$
3. Повторяем шаги 1 и 2 пока не получим $T_{n,1}^*, \ldots, T_{n,B}^*$
4. Положим оценку бустрепом смещения:

$$b_{boot} = \frac{1}{B} \sum_{r=1}^{B} T_{n,r}^* - \widehat{T}$$

Проблемы подхода

Для выборки $X_1, \dots X_n \sim U[0, \theta]$, оценим $\widehat{\theta} = \max\{x_1, \dots x_n\}$

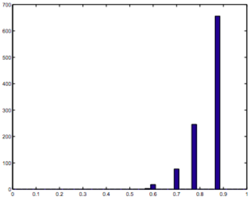


Figure: Histogram of the nonparametric bootstrap replications $\hat{\theta}^*$ with $n = 10$, $B = 1000$, $\hat{\theta} = 0.8722$. The maximum peak is at $\hat{\theta} = 0.8722$ with a probability of $\mathcal{P}(\hat{\theta} \in \mathbf{x}^*) = 0.6560 \approx 1 - (1 - 1/n)^n = 0.6513$.

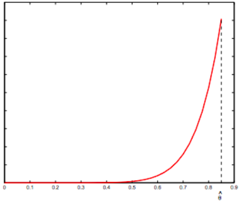


Figure: Theoretical results (extreme values) says that $\mathcal{P}(\hat{\theta}^*) = n\frac{(\hat{\theta}^*)^{n-1}}{\hat{\theta}^n}$.

Проблемы происходят из-за того, что $\widehat{F_n}$ неточно воспроизводит F: нужно сгладить.

Параметрический бутстреп

Параметрический бутстреп

- Предположим, что $F(x) \in \{F(x, \theta) \ : \ \theta \in \Theta \subset \mathbb{R}^d\}$.
- Тогда с помощью максимизации правдоподобия найдем параметр θ, а именно:

$$\theta = \arg\max_{\theta \in \Theta} \mathcal{L}(\vec{X}, \theta)$$

- Вместо ОМП можно использовать метод моментов.
- Далее действуем по описанной схеме непараметрического бутстрепа.

$(+)$ Непрерывная выборка, в маленькой выборке, как правило, недооценка разброса.

$(-)$ Произвольная модель и оценка параметров.

- Обычно выборки с менее чем 10 элементами считаются ненадежными для непараметрического бутстрепа.

Оценка смещения

1. Строим $\widehat{F_n}$ используя свои знания о системе
2. Выбираем $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$
3. Вычисляем $T_n^* = g(X_1^*, \ldots, X_n^*)$
4. Повторяем шаги 1 и 2 пока не получим $T_{n,1}^*, \ldots, T_{n,B}^*$
5. Положим оценку бустрепом смещения:

$$b_{boot} = \frac{1}{B} \sum_{r=1}^{B} T_{n,r}^* - \widehat{T}$$

Доверительное оценивание на основе бутстрепа

Нормальный интервал

- Если предположить, что данные распределены нормально, то имеет смысл рассмотреть следующий доверительный интервал:

$$(T_n - z_{\alpha/2}\widehat{se}_{boot}, T_n + z_{\alpha/2}\widehat{se}_{boot}),$$

- При этом $z_\alpha : F_{N(0,1)}(z_\alpha) = 1 - \alpha$, $\widehat{se}_{boot} = \sqrt{v_{boot}}$.

Центральный интервал

- Пусть $\theta = T(F)$ и $\widehat{\theta}_n = T(\widehat{F}_n)$.
- Пусть $\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,B}$ — получены итерированием шагов 1 и 2 алгоритма бутстрепа.
- Пусть θ^*_β — обозначает β-квантиль для $(\theta^*_{n,1}, \ldots, \theta^*_{n,B})$.
- Тогда центральный $(1-\alpha)$-доверительный интервал :

$$C_n = (2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2}, 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2}).$$

Центральный интервал

Теорема

При некоторых несильных условия на $T(F)$,

$$\mathrm{P}_F(T(F) \in C_n) \to 1 - \alpha, \qquad n \to \infty,$$

$$C_n = (2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2}, 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2})$$

Общие проблемы параметрического и непараметрического бутстрепа

- неполные данные;
- скоррелированные данные;
- большое количество выбросов.

Метод складного ножа

Мотивация

Непараметрический бутстреп производит семплирование из некоторого F^* построенного по результатам данных. Кроме того, в принципе,непараметрический бутстреп может занимать довольно много времени.

Метод складного ножа (jackknife)

- Пусть $T_n = (X_1, \ldots, X_n)$.
- Рассмотрим n подвыборок:
  $X_{(-i)} = \{X_1, \ldots X_{i-1}, X_{i+1} \ldots X_n\}$
- Пусть $\overline{T}_n = n^{-1} \sum_{i=1}^{n} T_{(-i)}$.
- Построим следующую оценку $\mathbb{V}ar(T_n)$:

$$v_{jack} = \left(\frac{n-1}{n} \sum_{i=1}^{n} (T_{(-i)} - \overline{T}_n)^2 \right)^{\frac{1}{2}}$$

- Аналогично:

$$b_{jack} = (n-1)(\sum_{i=1}^{n} T_{(-i)} - \overline{T}_n)$$

delete-d складной нож

- метод складного ножа может быть улучшен за счёт изымания большего количества примеров из выборки.
- в этом случае оценки будут выглядеть:

$$v_{delete-d} = \frac{r}{\binom{n}{d}} \sum_k (T_{(-k)} - \overline{T}_n)^2$$

где $X_{(-k)}$ - выборка без d элементов, $\overline{T}_n = \frac{\sum T_{(-k)}}{\binom{n}{d}}, n = r \cdot d$.

- Для состоятельности медианной оценки при $n \to \infty$, нужно выбирать $\sqrt{n} < d < n$.

Метод складного ножа

- delete-d складной нож — аппроксимация бутстрепа.
- Бутстреп — разные результаты, метод складного ножа — всегда одинаковые.
- Метод складного ножа проще применять для сложных схем семплирования.