МСПС | Оценивание плотности и непараметрич. регрессия |

Для произвольной ф.р. $F(x)$ её оценкой является $F_n(x)=\frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$ — эмпирич. ф.р. А как оценить плотность? Можно гистограммой. Но хочется чего-то более точного, например, аналога $\frac{\partial}{\partial x}F_n(x)$.

Идея состоит в том, чтобы сгладить $F_n(*)$ за счёт её свертки с абс. непрерывным распределением: рассм. сл.в. $Z_n = X + h_n Y$, где Y имеет извест. плот. $q(y)$ и $Y \perp\!\!\!\perp X$, $h_n \to 0$, а $X \sim F_n$ с плотн. $p(x)$.

Плотность $h_n Y$ — $\frac{1}{h_n} q\left(\frac{y}{h_n}\right) \Rightarrow p_{Z_n}(z) = \frac{1}{h_n}\int_{\mathbb{R}} q\left(\frac{z-x}{h_n}\right)F(dx) \overset{(1)}{=} \frac{1}{h_n}\int_{\mathbb{R}} q\left(\frac{z-x}{h_n}\right)p(x)dx$

При $h_n \to 0$ $Z_n \xrightarrow{P} X$, а значит, $p_{Z_n}(z) \to p(z)$ (при неких доп. условиях)

Заменим в ф-ле (1) $F(x)$ на $F_n(x)$ — получим оценку для $p(z)$:

$\hat{p}_n(z) = \frac{1}{h_n}\int_{\mathbb{R}} q\left(\frac{z-x}{h_n}\right)F_n(dx) = \frac{1}{nh_n}\sum_{i=1}^{n} q\left(\frac{z-X_i}{h_n}\right)$ — оценка Розенблатта–Парзена, или ядерная оценка плотности (плотность $q(z)$ называется ядром)

<u>Теорема</u> Пусть вып. следующие условия:

1) $q(y)$ непр. и огр., причём $\alpha = \int q^2(y)dy < +\infty$; 2) $h_n \to 0$ и $h_n \cdot n \to +\infty$ при $n \to +\infty$.

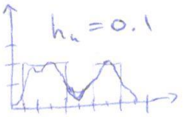
Тогда $\hat{p}_n(z) = p_{Z_n}(z) + \frac{\xi_n(z)}{\sqrt{nh_n}}$, где $p_{Z_n}(z) \to p(z)$ при почти всех z, а $\xi_n(z) \xrightarrow{d} \xi(z) \sim N(0, \alpha p(z))$.

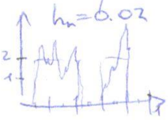
Оказывается, что оптим. скорость сходимости в этой теореме достигается при $h_n = C \cdot n^{-\frac{1}{5}}$ (т.е. скорость сходимости — $n^{2/5}$, это неплохо, потому что выше $\sqrt{n}$ быть не может), а наилучшее с точки зрения сходимости — т.н. ядро Епанечникова $q^*(y) = \frac{3}{4}(1-y^2)\cdot I(|y|\le 1)$

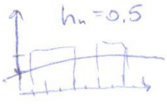
Ядро	$q(y)$	Достоинства и недостатки
Епанечникова	$\frac{3}{4}(1-y^2)I(\lvert y\rvert\le 1)$	Требования ко всем ядрам: носитель $q(y)$ — конеч. интервал, и на нём ядро 2 раза диффер.
Квадратич.	$\frac{15}{16}(1-y^2)^2 I(\lvert y\rvert\le 1)$	Дифф. в точках −1 и 1, в отличие от ядра Епанеч.
Треугол.	$(1-\lvert y\rvert)I(\lvert y\rvert\le 1)$	Можно быстро производить пересчёт $\hat{p}_n(z)$ при увеличении z.
Гаусса	$\frac{1}{\sqrt{2\pi}}e^{-y^2/2}$	Беск. дифф., но $\hat{p}_n(z)$ вычисл. медленно из-за подсчёта знач. экспонент
Прямоуг.	$\frac{1}{2}I(\lvert y\rvert\le 1)$	По сути, не ядро, но можно приближ. трапециями. От. простой выр.

По скорости сходимости все ядра примерно одинаковы (как и по дисперсии $\hat{p}_n(z)$). Но как всё-таки на практике h_n выбирать?

<u>Пример</u> Выборка размера $n=100$ с плотн. $p(z) = 2(I[0.1; 0.4] + I[0.6; 0.8])$, вычислены $\hat{p}_n(z)$ для $h_n = 0.02;\ 0.1;\ 0.5$







Выводы: выбор малого h_n ведёт к быстро мен., неустойчивой оценке, т.к. $\hat{p}_n(z)$ опирается лишь на небольшое кол-во наблюдений в узкой окрест. z, а слишком большое знач. h_n влечёт чрезмерное сглаживание плотности.

Непараметрич. регрессия

Рассм. модель $Y_i = m(X_i) + \varepsilon_i$. Пусть пока X_i – числа, а не строки, ε_i – н.о.р., $E\varepsilon_i = 0$, $D\varepsilon_i = \sigma^2$. Задача – оценить m.

С помощью метода лок. усреднения имеем:
$\hat{m}(x) = \left(\sum_{i=1}^{n} \omega_i(x) \cdot Y_i\right) / \left(\sum_{i=1}^{n} \omega_i(x)\right)$, где веса $\omega_i(x)$ велики для X_i, близких к т. x и малы для остальных X_i.

Определим $\omega_i(x)$ с помощью применения ядерных оценок. Пусть $q(y)$ – ядро, тогда $\omega_i(x) = q_h(x - X_i)$, где $q_h(y) = \frac{1}{h} q\left(\frac{y}{h}\right)$ и $h = h_n$ – окно пропускания (bandwidth).

В случае, когда X_i – многомерный, можно взять $\omega_i(x_1 \dots x_n) = \prod_{j=1}^{n} q_h(x_j - X_{ij})$ или $\omega_i(x) = q_h(\|X - X_i\|)$.

Менее популяр. метод задания весов – метод Гассера–Мюллера:
$\tilde{\omega}_i(x) = \int_{X_{(i-1)}}^{X_{(i)}} q_h(x-y)\,dy$, $\quad -\infty = X_{(0)} < X_{(1)} \le \dots \le X_{(n)} < X_{(n+1)} = +\infty$ – вар. ряд.

Если же задаём веса самым первым методом, то получается след.:

$$\hat{m}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} q_h(x - X_i) \cdot Y_i}{\frac{1}{n}\sum_{i=1}^{n} q_h(x - X_i)}$$

– оценка Надарая–Ватсона, а снизу, как легко видеть, стоит ядерная оц. плотности перем. X.

Теорема Пусть X – одномер. и вып. след. условия: 0) ~~выб.~~ (x_i, y_i) – выборка
1) $\int |q(y)| < +\infty$; 2) $y\,q(y) \to 0$ при $y \to +\infty$; ~~3) $E\,Y^2 < +\infty$;~~ 3) $n \to \infty$, $h_n \to 0$, $n h_n \to +\infty$.
4) $E(Y^2 | X = x) < +\infty \ \forall x$

Тогда $\hat{m}(x) \xrightarrow{P} m(x)$ в $\forall$ точке непр. функций $m(x)$, $p(x)$ и $D(Y|X=x)$.

Если говорить о скорости сходимости $\hat{m}(x)$ к $m(x)$, то наилучшая (по $R(x,h) = E[\hat{m}(x) - m(x)]^2$) скорость сход. достигается на $h_n = \frac{c}{n^{1/5}}$, как и для оценок плотности.

Оптимизация ширины окна пропускания

Дело в том, что X_i могут быть распределение неравномерно, т.е. где-то густо, а где-то пусто. Чтобы оценка регрессии была устойчивой, применяется адаптивное ядро: $\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i\, q_{h(x)}(x - X_i)}{\sum_{i=1}^{n} q_{h(x)}(x - X_i)}$, где $h(x)$ можно по-разному выбирать. Популярный способ: $h(x) = \inf\{h : \#\{X_i : |X_i - x| < h\} = k\}$, т.е. фактич. метод k ближайших соседей.

Можно ещё делать так (метод LOO – leave-one-out)
$LOO(h, X) = \sum_{i=1}^{n} \left(\hat{m}(x_i, X \setminus \{x_i\}) - y_i\right)^2 \to \min_h$, где оценка $\hat{m}(x_i, X \setminus \{x_i\})$ – оценка регрессии в т. x_i, построенная по набору объясняющей переменной X, откуда мы исключили значение x_i.

Проблема краевых эффектов

В одномер. случае (т.е. когда X_i — числа) часто наблюдается значит. смещение $\hat{m}(x)$ от истинной зависимости $m(x)$ вблизи миним. и макс. значений X_i (смещение возникает, когда объекты выборки X_i располагаются по одну сторону объекта x — и в многомер. случае такая ситуация возникает чаще).

Решается эта задача так: вместо аппроксимации зависимости в окрест. x числом аппроксимируем её лин. ф-цией, т.е.

$$\sum_{i=1}^{n} q_{h(x)}(x-X_i)\left(Y_i - a(x) - b(x)(X_i - x)\right)^2 \to \min_{a(x),\, b(x)}$$

Если $b(x)=0$, то получим оц. Надарая-Ватсона, а этот метод назыв. local linear regression model. Если минимизируем ф-л

$$\sum_{i=1}^{n} q_{h(x)}(x-X_i)\left(Y_i - a(x) - \sum_{j=1}^{d} b_j(x)(X_i-x)^j\right)^2 \to \min_{a(x),\, b_j(x)}$$

, то такой метод назыв. методом локальных полиномов (local polynomial regression model).

В многомерном случае нужно решать задачу лин. регрессии $Y = \vec{a}(x) + B(X-x)$ в каждой точке x, что сопряжено с большими выч. затратами.

Проблема выбросов.

Оценка Надарая-Ватсона крайне чувствит. к большим одиночным выбросам.

Решается эта проблема так: алгоритм LOWESS (локально взвешенное сглаживание)

1) положим $\gamma_i = 1\ \forall i = 1 \dots n$

2N) вычисляем оц. скользящего контроля (LOO) на каждом объекте:

$$a_i := \hat{m}(x_i, X \setminus \{x_i\}) = \frac{\sum_{j=1, j\neq i}^{n} Y_j \gamma_j q_{h(x_i)}(X_j - X_i)}{\sum_{j=1, j\neq i}^{n} \gamma_j q_{h(x_i)}(X_j - X_i)}$$

, q — возможно, многомер. ядро.

2N+1) вычисляем коэфф.

$\gamma_i := \bar{q}(a_i - Y_i)$, $i = 1 \dots n$, где $\bar{q}$ — какое-то другое ядро.

Пока коэфф. γ_i не стабилизируются. Говорят, сходится довольно быстро.

Варианты для ядра $\bar{q}(y)$:

1) Жёсткая фильтрация: строится на 2N+1 шаге вариац. ряд ошибок $\varepsilon_{(1)} \le \dots \le \varepsilon_{(n)}$, где $\varepsilon_i = |a_i - Y_i|$, тогда $\gamma_{(i)} := I\{i \le n-k\}$, т.е. ядро $\bar{q}(\varepsilon) = I\{\varepsilon \le \varepsilon_{(n-k)}\}$.

2) Мягкая фильтрация: $\bar{q}(\varepsilon) = q_Q\left(\frac{\varepsilon}{6\,\mathrm{med}\{\varepsilon_i\}}\right)$, где q_Q — квадратич. ядро.