Методы современной прикладной статистики
10. Нелинейные модели регрессии.

Родионов Игорь Владимирович
vecsell@gmail.com

Весна, 2018

Ядерная оценка плотности

Для произвольной функции распределения её оценкой является эмпирическая функция распределения $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$. Хочется получить оценку и для плотности.

Знаем, что оценкой для плотности является гистограмма, но хочется чего-то более точного, например, аналога $\frac{\partial}{\partial x}\widehat{F}_n(x)$.

Но $\widehat{F}_n(x)$ – дискретная функция распределения, поэтому сгладим её с помощью свёртки с непрерывной функцией распределения.

Ядерная оценка плотности

Рассмотрим случайную величину

$$Z_n = X + h_n Y,$$

где X – случайная величина с ф.р. F, плотность $p(x)$ которой мы хотим оценить, Y независима с X и имеет известную плотность $q(y)$, h_n – последовательность, стремящаяся к 0, n – размер выборки.

Плотность $h_n Y$ равна $\frac{1}{h_n} q(y/h_n)$, тогда по формуле свертки

$$p_{Z_n}(z) = \frac{1}{h_n} \int_{\mathbb{R}} q\left(\frac{z-x}{h_n}\right) dF(x) = \frac{1}{h_n} \int_{\mathbb{R}} q\left(\frac{z-x}{h_n}\right) p(x)dx. \tag{1}$$

Оценка Розенблатта-Парзена

Поскольку $Z_n \xrightarrow{P} X$ при $h_n \to 0$, то (при некоторых дополнительных условиях) $p_{Z_n} \to p(x)$.

Заменим теперь в формуле (1) F на эмпирическую функцию распределения $\widehat{F}_n$, получая тем самым **ядерную оценку** для $p(x)$:

$$\widehat{p}_n(z) = \frac{1}{h_n} \int_{\mathbb{R}} q\left(\frac{z-x}{h_n}\right) \widehat{F}_n(dx) = \frac{1}{nh_n} \sum_{i=1}^{n} q\left(\frac{z-X_i}{h_n}\right),$$

а функцию q будем называть **ядром**.

Theorem

Пусть выполнены следующие условия

1. $q(y)$ – *непрерывна и ограничена;*
2. $\alpha = \int q^2(y)dy < +\infty$;
3. $h_n \to 0$ *и* $h_n n \to +\infty$ *при* $n \to +\infty$.

Тогда

$$\widehat{p}_n(z) = p_{Z_n}(z) + \frac{\xi_n(z)}{\sqrt{nh_n}},$$

где $p_{Z_n}(z) \to p(z)$ *почти всюду, а*

$$\xi_n(z) \xrightarrow{d} \xi(z) \sim N(0, \alpha p(z)).$$

Замечания

Наилучшая скорость сходимости в теореме достигается при $h_n = Cn^{-1/5}$, таким образом, скорость сходимости в теореме равна $n^{2/5}$. Это неплохо, потому что скорость сходимости в ЦПТ равна $\sqrt{n}$, и выше вряд ли могло бы быть.

Однако на практике h_n выбирается эмпирически, исходя из сгруппированности данных и их количества.

На ядре Епанечникова $q^*(y) = \frac{3}{4}(1 - y^2)I(|y| \leq 1)$ в теореме достигается наилучшая скорость сходимости (при некоторых условиях, налагаемых на ядро), но существует много ядер, на которых скорость сходимости близка к наилучшей.

Виды ядер

1) Ядро Епанечникова

$$q^*(y) = \frac{3}{4}(1 - y^2)I(|y| \leq 1).$$

Если предположить, что у ядра носитель A, который является конечным интервалом, что оно дважды непрерывно дифференцируемо с условием $\int_A [q''(y)]^2 dy = \gamma < +\infty$ и что второй момент $\beta = \int y^2 q(y) dy < +\infty$, то ядро Епанечникова среди них наилучшее.

2) Квартическое ядро

$$q(z) = \frac{15}{16}(1 - y^2)^2 I(|y| \leq 1).$$

В отличие от ядра Епанечникова, дифференцируемо в точках 1 и -1.

Виды ядер

3) Треугольное ядро $q(y) = (1 - |y|)I(|y| \leq 1)$.
Позволяет быстро пересчитывать $\widehat{p}_n(z)$ в случае увеличения z с фиксированным шагом Δz.

4) Ядро Гаусса

$$q(y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Бесконечно дифференцируемо, но оценка $\widehat{p}_n(z)$ высчитывается медленно из-за многократного подсчета значений экспонент.

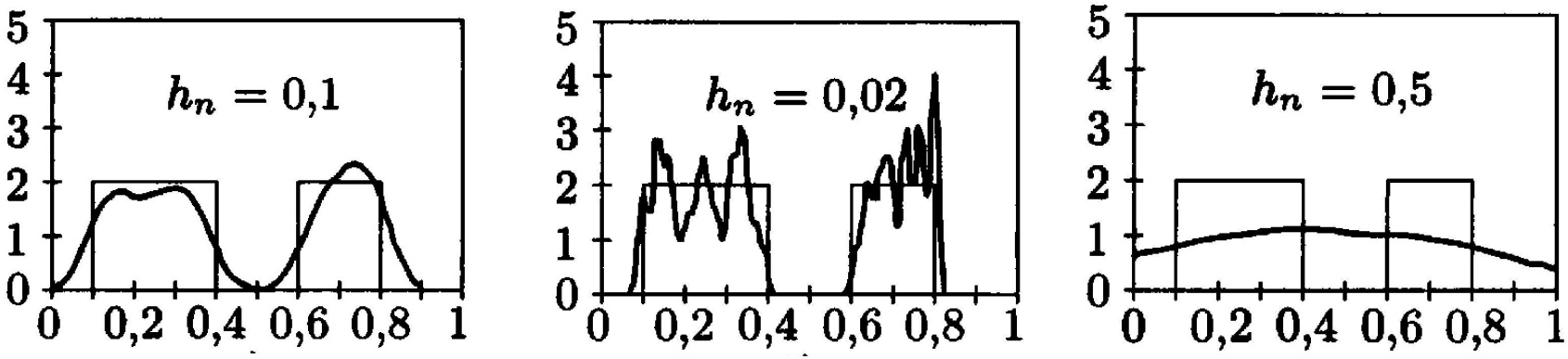
5) Прямоугольное ядро

$$q(y) = \frac{1}{2} I(|y| \leq 1).$$

По сути, ядром не является, поскольку имеются разрывы в точках 1 и -1, но его можно приблизить трапециями. Его основным преимуществом является очень простой вид.

Выбор h_n

Как на практике выбирать h_n?



На рисунке – выборка размера 100 с плотностью $p(x) = 2(I(z \in [0.1, 0.4]) + I(z \in [0.6, 0.8]))$, вычислены $\widehat{p}_n(z)$ для $h_n = 0.1, 0.02, 0.5$.

Выбор h_n

Видим, что выбор малого h_n ведет к быстро меняющейся, неустойчивой оценке, так как $\widehat{p}_n(z)$ опирается лишь на небольшое количество наблюдений в узкой окрестности z, а слишком большие значения h_n ведут к чрезмерному сглаживанию плотности.

Можно выбирать h_n следующим образом:

1. $h_n(x) = \inf\{h : \sharp\{X_i : |X_i - x| < h\} = K\}$, т.е. чтобы оценка в любой точке всегда строилась по K элементам выборки.
2. $h_n = \frac{1}{n}\sum_{i=1}^{n} d_{iK}$, где d_{iK} – расстояние от X_i до K-того ближайшего соседа.

Непараметрическая регрессия

Рассмотрим модель

$$Y_i = m(X_i) + \varepsilon_i,$$

где X_i – значение признака на i-том объекте (будем пока считать, что признак один), $\{\varepsilon_i\}$ – н.о.р. случайные ошибки, $E\varepsilon_i = 0$, $D\varepsilon_i = \sigma^2$.

Наша задача – оценить функцию m. На предыдущих двух лекциях решалась задача, когда $m(x)$ являлась линейной функцией от x, теперь же решим её в общем случае.

Непараметрическая регрессия

С помощью метода локального усреднения, имеем

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} \omega_i(x) Y_i}{\sum_{i=1}^{n} \omega_i(x)},$$

где веса $\omega_i(x)$ велики для X_i, близких к x, и малы для остальных X_i.

Определим $\omega_i(x)$ с помощью применения ядерных оценок. Пусть $q(y)$ – ядро, выберем

$$\omega_i(x) = \frac{1}{h} q\left(\frac{x - X_i}{h}\right),$$

где $h = h_n$ – окно пропускания (bandwidth).

Оценка Надарая-Ватсона

Подставляя значения весов в формулу локального усреднения, получаем оценку Надарая-Ватсона

$$\widehat{m}(x) = \frac{\sum\limits_{i=1}^{n} q\left(\frac{x-X_i}{h}\right) Y_i}{\sum\limits_{i=1}^{n} q\left(\frac{x-X_i}{h}\right)},$$

в знаменателе которой, как легко заметить, стоит ядерная оценка плотности выборки X_i, умноженная на nh.

Оценка Надарая-Ватсона

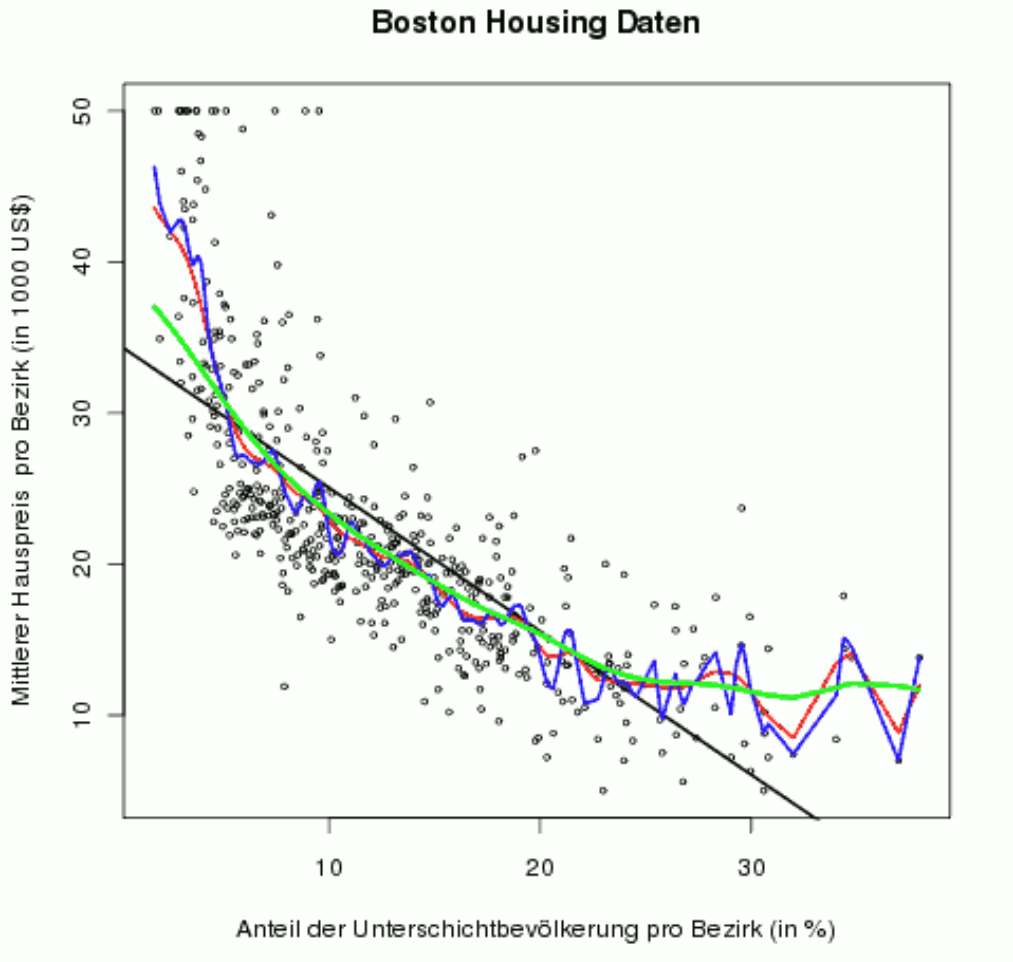


Рис.: График зависимости между процентом бедного населения и средней ценой квартиры в кварталах Бостона.

Другие методы определения весов

1) Менее популярным методом задания весов является метод Гассера-Мюллера. Определим

$$\widetilde{\omega}_i(x) = \frac{1}{h} \int_{X_{(i-1)}}^{X_{(i)}} q\left(\frac{x-y}{h}\right) dy,$$

где $-\infty = X_{(0)} < X_{(1)} < \ldots < X_{(n)} < X_{(n+1)} = +\infty$ – вариационный ряд значений признака X.

2) В случае, если признаков в регрессионной модели – k, можно взять

$$\omega_i(x) = \frac{1}{h^k} \prod_{j=1}^{k} q\left(\frac{x_j - X_{ij}}{h}\right) \text{ или } \omega_i(x) = \frac{1}{h} q\left(\frac{\|x - X_i\|}{h}\right).$$

Состоятельность оценки Надарая-Ватсона

Theorem

Пусть выполнены следующие условия:
1) $\{(X_i, Y_i)\}$ *– н.о.р. случайные векторы (размерности 2);*
2) ядро q *таково, что* $\int_{\mathbb{R}} |q(y)|dy < \infty$ *и* $yq(y) \to 0$ *при* $y \to \infty$;
3) $E(Y^2|X = x) < \infty \ \forall x$;
4) последовательность h_n *такова, что* $h_n \to 0$ *и* $h_n n \to +\infty$ *при* $n \to +\infty$.
Тогда в точках непрерывности $p(x)$

$$\widehat{m}(x) \xrightarrow{P} m(x) = E(Y|X = x).$$

Оптимизация bandwidth

1) LOO-метод (leave-one-out):

$$h_n = arg \min_h \sum_{i=1}^{n} (\widetilde{m}_i(X_i) - Y_i)^2,$$

где оценка $\widetilde{m}_i(x)$ получена по набору объясняющей переменной X, откуда мы удалили наблюдение X_i.

2) KNN-метод. Можно воспользоваться одним из методов со слайда 10, например, таким:

$$h_n(x) = \inf\{h : \sharp\{X_i : |X_i - x| < h\} = K\},$$

а дальше найти оптимальное K с помощью LOO или кросс-валидации.

Проблема краевых эффектов

В задаче непараметрической регрессии часто наблюдается значительное смещение $\widehat{m}(x)$ от истинной зависимости $m(x)$ вблизи минимального и максимального значений $\{X_i\}$.

Смещение возникает, поскольку объекты выборки $\{X_i\}$ располагаются "по одну сторону" от x – т.е. отсутствие объектов "по другую сторону" никак не учитывается.

В этом случае вместо аппроксимации регрессионной зависимости в окрестности x числом будем аппроксимировать зависимость линейной функцией (или многочленом).

Проблема краевых эффектов

Итак, оценим $m(x)$ вблизи краевых точек как $\hat{a}(x) + \hat{b}(x)x$, где

$$\hat{a}(x), \hat{b}(x) =$$

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

Название метода – local linear regression model. Если положить $b(x) = 0$, то получим оценку Надарая-Ватсона.

Проблема краевых эффектов

Если в методе локальной линейной регрессии заменить выражение $Y_i - a(x) - b(x)(X_i - x)$ на

$$Y_i - a(x) - \sum_{j=1}^{d} b_j(x)(X_i - x)^j,$$

то мы получим метод локальных полиномов (local polynomial regression model), который точнее предыдущего, но не всегда его применение осмысленно.

В многомерном случае метод локальной регрессии требует решать задачу линейной регрессии $Y = a(x) + B(X - x)$ в каждой точке x, что требует значительных вычислительных затрат.

Проблема выбросов

Оценка Надарая-Ватсона крайне чувствительна к большим одиночным выбросам. Это проблема решается с помощью алгоритма LOWESS (локально взвешенное сглаживание). Обозначим $q_h(x) = \frac{1}{h} q(x/h)$, где q – ядро (возможно, многомерное).

1) Положим $\gamma_i = 1$ для всех $i \in \{1, \ldots, n\}$;

$2N$) Вычисляем оценки скользящего контроля на каждом объекте,

$$a_i = \widehat{m}(X_i, X \backslash X_i) = \frac{\sum_{j \neq i} Y_j \gamma_j q_{h(X_i)}(X_j - X_i)}{\sum_{j \neq i} \gamma_j q_{h(X_i)}(X_j - X_i)};$$

$2N + 1$) Вычисляем коэффициенты $\gamma_i = \overline{q}(a_i - Y_i)$, где $\overline{q}$ – какое-нибудь другое ядро.

Проблема выбросов

Алгоритм продолжает свою работу до тех пор, пока γ_i не стабилизируются. Сходится алгоритм достаточно быстро. Результатом алгоритма является следующая оценка регрессии

$$\widehat{m}(x) = \frac{\sum_{j=1}^{n} Y_j \gamma_j q_{h(x)}(X_j - x)}{\sum_{j=1}^{n} \gamma_j q_{h(x)}(X_j - x)}.$$

Варианты выбора ядра $\overline{q}$:

1) на $2N+1$ шаге строится вариационный ряд ошибок $\varepsilon_i = |a_i - Y_i| : \varepsilon_{(1)} \leq \ldots \leq \varepsilon_{(n)}$, тогда $\overline{q}(\varepsilon) = I(\varepsilon \leq \varepsilon_{(n-k)})$ (жесткая фильтрация);

2) $\overline{q}(\varepsilon) = q_Q\left(\frac{\varepsilon}{6med\{\varepsilon_i\}}\right)$, где q_Q – квартическое ядро (мягкая фильтрация).

Нелинейные обобщения линейной регрессии

Предположение о том, что модель регрессии линейна по параметрам, удобно для построения численных методов, но не всегда хорошо согласуется со знаниями о предметной области.

Пусть задана нелинейная модель регрессии $Y = f(X, \alpha) + \varepsilon$, хотим минимизировать функционал качества

$$Q(\alpha) = \sum_{i=1}^{n} (f(x_i, \alpha) - y_i)^2$$

по вектору параметров $\alpha \in \mathbb{R}^p$, чтобы найти оценку отклика.

Сведем решение нелинейной задачи к решению последовательности линейных задач.

Метод Ньютона-Рафсона

Для минимизации $Q(\alpha)$ воспользуемся методом Ньютона–Рафсона. Выберем начальное приближение $\alpha^0 = (\alpha_1^0, \ldots, \alpha_p^0)$ и организуем итерационный процесс

$$\alpha^{t+1} := \alpha_t - h_t (Q''(\alpha^t))^{-1} Q'(\alpha^t),$$

где Q' – градиент Q по α, Q'' – матрица вторых производных (гессиан) Q по α, h_t – величина шага, в самом простом варианте метода равная 1.

Обращать матрицу Q'', а тем более много раз, не очень хочется, поэтому желательно придумать метод, упрощающий это действие.

Метод Ньютона-Рафсона

Выпишем компоненты градиента

$$\frac{\partial}{\partial \alpha_j} Q(\alpha) = 2 \sum_{i=1}^{n} (f(x_i, \alpha) - y_i) \frac{\partial f}{\partial \alpha_j}(x_i, \alpha).$$

Выпишем компоненты гессиана (вместо $f(x_i, \alpha)$ пишем $f(x_i)$)

$$\frac{\partial^2}{\partial \alpha_j \partial \alpha_k} Q(\alpha) = 2 \sum_{i=1}^{n} \frac{\partial f(x_i)}{\partial \alpha_j} \frac{\partial f(x_i)}{\partial \alpha_k} - 2 \sum_{i=1}^{n} (f(x_i) - y_i) \frac{\partial^2 f(x_i)}{\partial \alpha_j \partial \alpha_k}.$$

Было бы неплохо, если бы второе слагаемое обратилось в 0, тогда гессиан равнялся бы произведению двух матриц и обращать его стало бы гораздо проще.

Метод Ньютона-Гаусса

Предположим, что f дважды непрерывно дифференцируема, тогда её можно линеаризовать в окрестности α^t,

$$f(x_i, \alpha^t) = f(x_i, \alpha) + \sum_{j=1}^{p} \frac{\partial f(x_i)}{\partial \alpha_j}(x_i, \alpha)(\alpha_j^t - \alpha_j).$$

Тогда если заменить в гессиане функцию f на её линеаризацию, то вторые производные от f обратятся в 0.

Данный метод называется методом Ньютона-Гаусса, в остальном он ничем не отличается от метода Ньютона-Рафсона. Скорость сходимости у него тоже практически такая же, как у метода Ньютона-Рафсона.

Метод Ньютона-Гаусса

Обозначим $F_t = (\frac{\partial f}{\partial \alpha_j}(x_i, \alpha^t))$, $1 \le i \le n$, $1 \le j \le p$, – матрица первых производных на t-й итерации, $f_t = (f(x_i, a^t))_{i=1}^n$ – вектор значений аппроксимирующей функции на t-й итерации.

Тогда формула t-й итерации метода Ньютона–Гаусса в матричной записи примет вид

$$\alpha^{t+1} = \alpha_t - h_t(F_t^T F_t)^{-1} F_t^T (f_t - Y).$$

Легко заметить, что в правой части записано решение линейной регрессионной задачи $\|(f_t - Y) - F_t\delta\|^2 \to \min_\delta$, т.е. мы свели решение задачи нелинейной регрессии к последовательности линейных регрессионных задач.

Нелинейные преобразования признаков

На практике встречаются ситуации, когда линейная модель регрессии представляется необоснованной, но предложить адекватную нелинейную модель $f(x, \alpha)$ также не удаётся. Тогда в качестве компромисса строится модель вида

$$f(X, \alpha) = \sum_{j=1}^{k} \varphi_j(X_j),$$

где $\varphi_j : \mathbb{R} \to \mathbb{R}$ – некоторые преобразования исходных признаков, в общем случае нелинейные. Будем подбирать оптимальные $\{\varphi_j\}$ с точки зрения минимизации квадратичной функции потерь.

Алгоритм настройки с возвращением

На первом шаге полагаем $\varphi_j(x) = \alpha_j x$, коэффициенты α_j находятся с помощью линейной регрессии.

На каждом последующем шаге выбирается одна из функций φ_j, все остальные фиксируются, и выбранная функция строится заново. Для этого решается стандартная задача наименьших квадратов

$$\varphi_j = arg \min_{\varphi} \sum_{i=1}^{n} \left(\varphi(x_{ij}) - z_i\right)^2,$$

где $z_i = y_i - \sum_{k \neq j} \varphi_k(x_{ik})$ – не зависит от φ_j. К решению этой задачи приводит использование, например, оценки Надарая-Ватсона, можно также приближать φ_j полиномами или рядом Фурье.

Finita!