Самостоятельная работа 3

«Исследование непараметрической оценки прямой регрессии»

Задание: Реализовать три имитационных модели объекта: линейный объект, квадратичный объект и любой нелинейный объект по выбору студента. Для каждого из объектов построить оценку прямой регрессии с использованием четырех видов ядер (прямоугольное ядро, треугольно ядро, параболическое ядро и кубическое ядро). Объем выборки задаются студентом. Коэффициента размытости ядра автоматически подбирается в программе либо задается студентом. Критерий оптимальности подбора коэффициента размытости выбирается студентом. Необходимо построить графики для полученных оценок прямой регрессии. На графике должны быть изображены (для визуальной оценки качества подученной оценки регрессии): исходный объект без помехи, точки выборки, полученная оценка регрессии. Сравнить между собой полученные оценки с разными ядрами, выделить лучшее и худшее ядро. Построить графики оценки регрессии при граничных значениях коэффициента размытости.

Методические указания

Регрессией называют первый начальный условный момент:

$$M\{Y \mid x\} = \int_{-\infty}^{\infty} y f(y \mid x) dy = \eta(x)$$

Оценка $\eta_n(x)$ регрессии строится на основе серии измерений выхода и входа объекта: $x_i, y_i,\ i = \overline{1,n}$:

$$\widehat{M}\{Y \mid x\} \equiv \eta_n(x) = \sum_{i=1}^{n} K_N\left(\frac{x - x_i}{h}\right) y_i ,$$

$$K_N\left(\frac{x - x_i}{h}\right) = \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)}.$$

Колоколообразная функция $K_N\left(\frac{x - x_i}{h}\right)$ по форме повторяет ядро $K(\cdot)$ и отличается от него на нормирующий множитель $1/\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)$. За счет

этого $\sum_{i=1}^{n} K_N\left(\frac{x-x_i}{h}\right)=1$, т. е. ядро $K_N\left(\frac{x-x_i}{h}\right)$ нормировано на 1 на системе экспериментальных точек.

Нормированность ядер $K_N\left(\frac{x-x_i}{h}\right)$ приводит к условию:

$$\min\{y_i,\ i=\overline{1,n}\}\le \eta_n(x)\le \max\{y_i,\ i=\overline{1,n}\},$$

которое говорит о существовании полосы, за пределы которой не выходит непараметрическая оценка регрессии.

Усечённость нормированных ядер $K_N(\cdot)$ (в силу усечённости ядра $K(\cdot)$) позволяет при построении оценки $\eta_n(x)$ в каждой фиксированной точке x учитывать только несколько близлежащих значений x_i и не "перелопачивать" всю выборку.

Основное влияние на оценку регрессииоказывает положительная константа c, но зависимость от c при возрастании n ослабевает. Форма ядра усеченная параболическая. Константа c, определяющая коэффициент размытости, вычисляется по выборке путём минимизации эмпирических показателей (характеризующих наилучшее сглаживание экспериментальных данных).

Считаем, что выборке $(x_i, y_i), i=\overline{1,n}$ измерениях входа находятся на равных расстояниях друг от друга $\Delta = x_{i+1}-x_i (i=\overline{1,n-1})$, а объем выборки n фиксирован. Перейдем от размерного параметра c (его размерность, обратная размерности x) к безразмерному β:

$$\beta = c^{-1}\Delta n^{1/5}\ (0\le\beta\le 1).$$

Оценка регрессии приобретает вид:

$$\hat{M}\{Y\mid x\}\equiv \eta_n(x)=\sum_{i=1}^{n} K_N\left(\beta\frac{x-x_i}{\Delta}\right)y_i$$

$$K_N\left(\beta\frac{x-x_i}{\Delta}\right)=\frac{K\left(\beta\frac{x-x_i}{\Delta}\right)}{\sum_{j=1}^{n} K\left(\beta\frac{x-x_j}{\Delta}\right)}$$

При $\beta=0$ оценка регрессии $\eta_n(x)$ не зависит от x. Такой вариант, хотя и редко, но возможен. Выбранный вход объекта не оказывает влияния на выход объекта.

При $\beta = 1$ оценка регрессии $\eta_n(x)$ точно проходит через экспериментальные точки, т. е. оценка не осуществляет сглаживания экспериментальных дынных. Такой вариант тоже возможен, если сигнальная часть выхода объекта не зашумлена помехой.

При наличии помех в выходе объекта оценка должна сглаживать экспериментальные дынные. Этот наиболее распространённый вариант соответствует параметру β, находящемуся внутри интервала [0; 1]. Для его вычисления необходимо строить критерии оптимальности.

Критерий оптимальности

1. Метод "скользящего экзамена"

$$I_{1n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{\eta_n}(x_i))^2 = \min_{\beta},$$

$$\overline{\eta_n}(x_i) = \sum_{\substack{k=1 \\ k\neq i}}^{n} K_N\left(\beta\frac{x_i - x_k}{\Delta}\right)y_k ,$$

Выборка x_i, y_i $(i = \overline{1,n})$ при этом своеобразно разбивается на две части: одна используется для построения модернизированной модели $\overline{\eta}_n(x)$, вторая – для ее проверки (по вышеуказанному критерию). Первое слагаемое в I_{1n} (т. е. при $i = 1$) равно квадрату невязки между выходом объекта y_1 и выходом модели $\overline{\eta}_n(x_1)$ в первой экзаменующей точке (x_1, y_1). Эта экзаменующая точка не участвует в построении (в обучении) модели $\overline{\eta}_n(x_1)$. Затем берется вторая экзаменующая точка (x_2, y_2) и в ней вычисляется квадрат невязки между выходами объекта y_2 и модели $\overline{\eta}_n(x_2)$, где модель $\overline{\eta}_n(x_2)$ построена по всей выборке кроме точки (x_2, y_2), и т. д.

2. Разбиение выборки на две части: $x_i, y_i,\ i \in M_{n_1}$, $x_i, y_i,\ i \in M_{n_2}$

По одной из них (объема n_1) строится оценка регрессии, по второй (объема n_2) – показатель качества:

$$I_{3n_2} = \frac{1}{n_2}\sum_{l\in M_{n_2}}\left(y_i - \eta_{n_1}(x_l)\right)^2 = \min_{0\le\beta\le 1}$$

$$\eta_{n_1}(x) = \sum_{i\in M_{n_1}} K_N\left(\beta\frac{x - x_i}{\Delta}\right)y_i .$$

При наличии одного входа разбиение на множества M_{n_1}, M_{n_2} можно выполнить сравнительно легко. Надо упорядочить выборку x_i, y_i, $i = \overline{1,n}$ по x, т. е. сделать в выборке все $x_i < x_{i+1}$. Затем выборочные точки с нечетными номерами отнести к первой группе, с четными номерами – ко второй группе. После настройки параметра β оценка регрессии $\eta_n(x)$ (при дальнейшем её использовании) строится по всей выборке.

3. Условие минимального квадратичного уклонения

$$I = \left(\hat{\sigma}_y^2 - I_n\right)^2 = \min_{\beta};\ I_n = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \eta_n(x_i)\right)^2$$

Для поиска параметра β в одномерном случае хорошо подходят: метод деления отрезка пополам и метод золотого сечения.

Имитация объекта

1. Формирование сигнальной части объекта. Точки выборки находятся на равных расстояниях друг от друга $x_{ji} = x_i^0 + j\Delta, i = 1,2;\ j = \overline{0,n}$

$\Delta = x_{i+1} - x_i (i = \overline{1,n-1})$.

2. Формирование аддитивной помехи (метод полярных координат)

Процедура расчета оценок

1. $tmp_i = \sum_{j=1}^{n} K\left(\beta\frac{x_i - x_j}{\Delta}\right)$

2. $\eta_n(x_i) = \frac{\sum_{j=1}^{n} K\left(\beta\frac{x_i - x_j}{\Delta}\right) y_j}{tmp_i}$

Расчет оптимального β методом "скользящего экзамена"

1 Способ: Перебор

1. Формирование базы значений β: $\beta_j = j * 0.01, j = \overline{0,100}$
2. $tmp_i = \sum_{k=1}^{n} K\left(\beta_j\frac{x_i - x_k}{\Delta}\right), k \neq i$
3. $\eta_n(x_i) = \frac{\sum_{k=1}^{n} K\left(\beta_j\frac{x_i - x_k}{\Delta}\right) y_j}{tmp_i}, k \neq i$
4. $I_{1n}(\beta_j) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \eta_n(x_i)\right)^2 = min$

2 Способ: С использованием метода деления отрезка пополам

0. $a = 0, b = 1, \varepsilon = 0.0001$

1. $x^m = \frac{a+b}{2}, L = b - a, I_{1n}(x^m)$

2. $x_1 = a + \frac{L}{4}, x_2 = b - \frac{L}{4}, I_{1n}(x_1), I(x_2)$

3. $I_{1n}(x_1) < I_{1n}(x^m)$: исключить $(x^m, b], b = x^m, x^m = x_1$. Перейти к шагу 5

4. $I_{1n}(x^m) > I_{1n}(x_2)$: исключить $[a, x^m), a = x^m, x^m = x_2$. Перейти к шагу 5

Если $I_{1n}(x^m) \leq I_{1n}(x_2)$: $b = x_2, a = x_1$

5. $L = b - a, |L| < \varepsilon$. Иначе возврат к шагу 2