Оценивание плотностей. Непараметрическая регрессия.

1 Рассмотрим задачу непараметрической регрессии $Y = m(X)+\varepsilon$, где $X = (X_1, \ldots, X_n)$ – вектор признаков. В рамках метода локальной линейной регрессии

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

получить оценку $\widehat{m}(x)$ регрессионной функции $m(x)$ в явном виде.

2 Пусть $(X_1, \ldots, X_n)$ – выборка из распределения с непрерывной плотностью $p(x)$. Рассмотрим гистограмму выборки

$$H_{n,\Delta}(x) = \sum_{j\in\mathbb{Z}} \#\{i : X_i \in [j\Delta, (j+1)\Delta)\} I\{x \in [j\Delta, (j+1)\Delta)\}.$$

Доказать, что $\forall x \in \mathbb{R}$ $\frac{1}{n\Delta} H_{n,\Delta}(x)$ стремится по вероятности к $p(x)$ при $\Delta \to 0$ и $n \to \infty$.

3 Выданы данные $\{(X_i, Y_i), i = 1, \ldots, n\}$. Рассмотрим задачу непараметрической регрессии $Y_i = m(X_i)+\varepsilon_i$. Построить непараметрическую регрессию с помощью оценки Надарая-Ватсона и методом сглаживающего сплайна (функция smooth.spline в R):

$$SS(h) = \sum_{i=1}^{n} (Y_i - m(X_i))^2 + h \int_{X_{(1)}}^{X_{(n)}} [m''(x)]^2 dx \longrightarrow \min_{m}$$

и вывести графики получившихся приближений. Зачем, на ваш взгляд, добавлять в выражение суммы квадратов интеграл от квадрата второй производной функции m? Какой выбор h, на ваш взгляд, является оптимальным? Где отличаются графики оценки Надарая-Ватсона и сглаживающего сплайна и как вы это объясните?

4 Выданы данные $\{(X_i, Y_i), i = 1, \ldots, n\}$. Построить по ним оценку функции $m(x)$ в модели непараметрической регрессии $Y = m(X) + \varepsilon$ методами Надарая-Ватсона и Гассера-Мюллера. Оптимизировать ширину окна пропускания предложенными в лекции методами. Задать меру качества приближения и выяснить, какая оценка получилась лучше.

5 Выданы данные $\{(y_i, x_{ij}), i = 1, \ldots, n+q,\ \ j = 1, \ldots, k\}$, причем $y_{n+1}, \ldots, y_{n+q}$ известны. Используя пройденные методы регрессионного анализа, в рамках линейной регрессионной модели и модели непараметрической регрессии предсказать значения откликов объектов с номерами $n+1, \ldots, n+q$. Описать и объяснить проделанные процедуры. Используя метрику RSS, определить, какое предсказание получилось лучше.

6 Пусть $X_1, \ldots, X_n$ – выборка из распределения с плотностью

$$p(x) = \frac{3}{8}(2|x| - x^2) I(|x| \le 2).$$

Исследовать поведение ядерных оценок $\widehat{p}_n(x)$ плотности $p(x)$, построенных по ядру Епанечникова и гауссовскому ядру, при возрастании n. Опишите метод выбора ширины окна пропускания $h_n(x)$. Использовать равномерную метрику $\rho(f, g) = \sup_{x\in[-2,2]} |f(x) - g(x)|$ как меру качества приближения и вывести на графике зависимость $\rho(\widehat{p}_n, p)$ для обеих оценок.