# Домашнее задание № 2

*Тема: Непараметрическое оценивание плотности.*

*Каждое задание оценивается в 2 балла.*

*Крайний срок сдачи: 29 октября 2021 г., 18:00*

## 1

N1 (i) Смоделируйте выборку размера $n = 1000$ из гамма-распределения с фиксированными параметрами shape=3 и rate=4.

(ii) Оцените качество ядерных оценки с ядром Епанечникова, перебирая значения параметра bandwidth от 0.1 до 5 с шагом 0.1 (используйте встроенные функции - например, в языке R используйте функцию density). Для этого найдите значение ядерных оценок $\hat{p}_n$ в точках $x_1, ..., x_M$, выбранных по равномерной решётке на отрезке $[0, 2]$ с шагом 0.01; если это сделать невозможно, то найдите значение в ближайших точках, для которых значение оценки плотности известно. Выберете параметр bandwidth, при котором минимальна ошибка MISE, оценённая по формуле

$$\widehat{MISE}(x) = \frac{1}{n} \sum_{k=1}^{n} \big(\hat{p}_n(x_k) - p(x_k)\big)^2, \tag{1}$$

где $p(\cdot)$ - истинная плотность.

(iii) Проделайте тоже самое для ядерной оценки плотности $\hat{p}_n$ с ядром

$$K(x) = \sum_{k=0}^{2} \varphi_k(0)\varphi_k(x)\mathbb{I}\big\{|x| \le 1\big\}, \tag{2}$$

где

$$\varphi_0(x) = \frac{1}{\sqrt{2}}, \qquad \varphi_1(x) = \sqrt{\frac{3}{2}}x, \qquad \varphi_2(x) = \frac{\sqrt{5}}{2\sqrt{2}}\left(3x^2 - 1\right)$$

- это первые 3 многочлена Лежандра на отрезке $[-1, 1]$. Графически сравните оценки MISE, полученные для ядерных оценок плотности с ядром Епанечникова и с ядром (2), в зависимости от значения параметра bandwidth.

T1 Найдите наименьшее значение параметра bandwidth, при котором ядерная оценка с ядром (2) лучше любой ядерной оценки с ядром Епанечникова ("любой" = "с любым значением параметра bandwidth") в смысле асимптотического поведения MISE. Предполагается, что как и в задаче N1, истинное распределение выборки есть гамма-распределение с параметрами shape=3 и rate=4.

*Подсказка:* используйте Proposition 1.7 из книги A.Tsybakov "Introduction to nonparametric estimation". Про это утверждение я рассказывал на лекции.

# 2

N2 Плотность распределения "Bart Simpson" равна

$$p_{BS}(x) = \frac{1}{2}p_{(0,1)}(x) + \frac{1}{10}\sum_{j=0}^{4} p_{((j/2)-1,1/10)}(x), \qquad x \in \mathbb{R},$$

где $p_{(\mu,\sigma)}$ - плотность нормального распределения со средним $\mu$ и дисперсией $\sigma^2$. Промоделируйте выборку с такой плотностью (см. семинар).

(i) Перебирая различные значения количества компонент (от 2 до 10), постройте параметрические оценки плотности как смеси нормальных распределений (EM-алгоритм). Найдите оценку с наибольшим значением логарифма функции правдоподобия.

(ii) Постройте ядерные оценки плотности, используя методы выбора параметра bandwidth, связанные с процедурой кросс- проверки. Если доступно несколько таких методов, то используйте тот, который приводит к построению оценки плотности, наиболее близкой (по виду графика) к истинной.

Среди оценок, построенных на предыдущем шаге, выберете оценку, наиболее близкую к построенной ядерной оценке плотности. В качестве меры близости используйте выражение

$$\frac{1}{J}\sum_{j=1}^{J}\big(\hat{p}_n^{EM}(x_j)-\hat{p}_n^{K}(x_j)\big)^2,$$

где $\hat{p}_n^{EM}$ - оценка, полученная при помощи EM-алгоритма, $\hat{p}_n^{K}$ - ядерная оценка плотности, $x_1, ...x_J$ - набор точек, для которых известно значение $\hat{p}_n^{K}$.

T2 Как известно, для любой оценки плотности $\hat{p}_n(x)$ $MISE$ определяется как

$$\begin{aligned} MISE(\hat{p}_n) &= \mathbb{E}\int\Big(\hat{p}_n(x)-p(x)\Big)^2dx \\ &= \mathbb{E}\Big(\int\hat{p}_n^2(x)dx-2\int\hat{p}_np(x)dx\Big)+\int p^2(x)dx, \end{aligned}$$

и поиск оптимального параметра $h$ оценки $\hat{p}_n(x)$ (например, оптимального параметра bandwidth $h$ для ядерной оценки плотности) сводится к минимизации функции

$$G(h)=\mathbb{E}\Big(\int\hat{p}_n^2(x)dx-2\int\hat{p}_np(x)dx\Big).$$

Оценка кросс-валидации параметра $h$ определяется как точка минимума функции

$$\hat{G}(h)=\int\hat{p}_n^2(x)dx-\frac{2}{n}\sum_{i=1}^{n}\hat{p}_{(-i)}(x_i),$$

где $\hat{p}_{(-i)}$ - оценка, построенная по всем значениям, кроме $i-$го (leave-one-out estimate). Докажите, что в случае ядерных оценок плотности $\hat{p}_n, \hat{p}_{(-i)}$ оценка $\hat{G}(h)$ является несмещённой оценкой $G(h)$.

# 3

N3 Рассмотрим переменную eruptions из базы данных faithful (длительность извержений Old Faithful Geyser, см. `https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/faithful.html`). Предположим, что эта выборка имеет распределение смеси двух нормальных распределений,

$$p(x) = \pi p_{(\mu_1,\sigma_1)}(x) + (1-\pi)p_{(\mu_2,\sigma_2)}(x),$$

где $p_{(\mu_i,\sigma_i)}(x)$ - это плотность нормального распределения со средним $\mu_i$ и стандартным отклонением $\sigma_i, i = 1, 2$. Для оценки параметров $\pi \in (0,1), \mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}, \sigma_1 > 0, \sigma_2 > 0$ используются два метода:

1) "профессиональный" - EM - алгоритм;
2) "кустарный" :
   i. строится ядерная оценка плотности;
   ii. средние значения $\mu_1, \mu_2$ оцениваются x-координатами "горбов";
   iii. находится точка минимума оценки плотности, лежащая между "горбами" (далее будем называть эту точку разделительной);
   iv. стандартные отклонения $\sigma_1, \sigma_2$ оцениваются стандартными отклонениями подвыборок со значениями слева и справа от разделительной точки;
   v. наконец, $\pi$ оценивается как доля наблюдений слева от разделительной точки.

Найдите оценки параметров первым и вторым методом. Определите какой из этих методов лучше описывает переменную eruptions, применив критерий хи-квадрат при делении области значений переменной на 20 интервалов.

# 4

T3* Напомним, что эффективностью ядра $K : \mathbb{R} \to \mathbb{R}_+$ называется функционал

$$J(K) = \Big(\int_{\mathbb{R}} K^2(x)dx\Big)^{4/5}\Big(\int_{\mathbb{R}} x^2 K(x)dx\Big)^{2/5}.$$

Докажите, что минимальное значение этого функционала для чётных функций $K$, обладающих свойством $\int_{\mathbb{R}} K(x)dx = 1$, достигается на ядре Епанечникова

$$K(x) = \frac{3}{4}\big(1 - x^2\big) \cdot \mathbb{I}\big\{|x| \leq 1\big\}.$$