Домашнее задание № 2

Тема: Непараметрическое оценивание плотности

Максимальные баллы за задания:
- [T1]-[T4] - 1.25 балла;
- [N1], [N3] - 1.5 балла;
- [N2] - 2 балла;
- [T5] - 2 балла.*
Итого: за обязательную часть максимум равен 10 (5 - теоретические задачи, 5 - практические); дополнительные 2 балла можно набрать на бонусной задаче

Крайний срок сдачи: 25 октября 2020 г. (до конца дня).

1

N1 Рассмотрим базу данных "LakeHuron" , содержащую данные об уровне воды в озере Гурон (в футах) в 1875–1972 годах, см. `https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/LakeHuron.html`. Для студентов, работающих в Python: txt- файл в этими данными выложен в телеграм-канал.

Целью данного упражнения является оценка функции плотности уровня воды в озере.

(i) Постройте гистограмму с количеством столбиков, выбранном в соответствии с правилом Стёржеса. На том же графике отобразите график функции плотности нормального распределения с оценённым средним и дисперсией.

(ii) Среди всех гистограмм с количеством столбиков от 5 до 30, найдите оценку, наиболее близкую к плотности нормального распределения (см. предыдущий пункт). Проиллюстрируйте дилемму между смещением и дисперсией (bias-variance tradeoff) в данной ситуации.

(iii) Постройте ядерные оценки с различными ядрами (примените все ядра, доступные в языке R / Python, параметр bandwidth может быть выбран по умолчанию). Постройте ядерные оценки, построенные при помощи разных методов выбора bandwidth (примените все методы, доступные в R / Python, ядро может быть выбрано произвольным образом). Среди всех построенных ядерных оценок выберите ту, которая ближе всего к нормальному распределению.

T1 Вычислите (без использования компьютера) теоретические эффективности

(i) ядра boxcar

$$K(x) = \mathbb{I}\{x \in [-1/2, 1/2]\};$$

(ii) гауссовского ядра

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

2

N2 (i) Симулируйте выборку длины $N = 1000$ с распределением, имеющим плотность

$$p(x) = \frac{1}{2}\phi^N(x) + \frac{1}{4}\phi^E(x+2) + \frac{1}{4}\phi^E(-x+2), \qquad x \in \mathbb{R}, \quad (1)$$

где ϕ^N - плотность стандартного нормального распределения (с нулевым средним и единичной дисперсией), ϕ^E - плотность стандартного экспоненциального распределения (с параметром равным 1).

(ii) Постройте гистограмму $\hat{p}_n(x)$ с количеством стоблцов, выбранных по правилу Фридмана-Дьякони. Вычислите эмпирический аналог MISE, а именно

$$\widehat{MISE}(\hat{p}_n) = \frac{6}{Q} \sum_{q=1}^{Q} \left(\hat{p}_n(x_q) - p(x_q)\right)^2, \quad (2)$$

где точки $x_1, ..., x_Q$ выбраны по равномерной решётке на $[-3, 3]$, и $Q = 10000$.

(iii) Оцените MISE более точно: повторите шаги (i) и (ii) несколько раз (например, $J = 20$ раз), и на основе полученных оценок $\hat{p}_n^{(1)}(x), ...\hat{p}_n^{(J)}(x)$ оцените MISE как

$$\frac{1}{J}\sum_{j=1}^{J}\widehat{MISE}(\hat{p}_n^{(j)}). \tag{3}$$

(iv) Повторите шаги (i)-(iii), но используя другие методы для выбора параметра на шаге (ii) (правила Стёржеса, Скотта, или другие). Какой метод даёт наилучшие результаты в этой ситуации, т.е. при каком методе значение величины (3) меньше?

(v) Рассмотрим значения параметра bandwidth на интервале $(0, 1)$ с шагом 0.01. Для каждого значения параметра, постройте ядерную оценку с ядром Епанечникова. Оцените MISE и постройте график зависимости MISE от параметра h. Какое значение h минимизирует MISE в данной ситуации?

(vi) На одном и том же рисунке, отобразите

- график лучшей гистограммы, то есть гистограммы с оптимальным выбором количества столбиков, см. (iv);

- график лучшей ядерной оценки, см. (v);

- график истинной функции плотности.

T2 Вычислите математическое ожидание, дисперсию и характеристическую функцию случайной величины с плотностью (1).

T3 Допустим, что дана выборка из нормального распределения с нулевым средним и дисперсией σ^2.

(i) Вычислите значение параметра bandwidth, минимизирующее AMISE (asymptotic mean integrated squared error) ядерной оценки плотности, построенной на основе гауссовского ядра

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

(ii) При найдённом в предыдущем пункте значения параметра bandwidth, вычислите значение параметра σ, при котором часть AMISE, относящаяся к смещению, равна части AMISE, относящейся к дисперсии.

3

N3 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 μ и дисперсией σ^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

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

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 $\hat{p}_n^{K}$.

T4 Дана выборка $x_1, ..., x_n$ из смеси K нормальных распределений

$$p(x) = \sum_{k=1}^{K} \pi_k p_{(\mu_k,\sigma_k)}(x),$$

где $\pi_1, ..., \pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$, $p_{(\mu_k,\sigma_k)}(x)$ - плотность нормального распределения со средним 0 и дисперсией σ_k^2. Для применения EM-алгоритма вводится латентная случайная величина Y, принимающая значения 1,..,K и представляющая собой номер компоненты смеси. Обозначим соответствующий набор i.i.d. случайных величин через $Y_1, ..., Y_n$, реализацию набора - через $y_1, .., y_n$. Найдите значения вектора

$$\theta = (\mu_1, ..., \mu_K, \sigma_1, ...\sigma_K, \pi_1, ..., \pi_K),$$

являющиеся решением оптимизационной задачи

$$\arg\max_{\theta} \mathbb{E}_Y \Big[\log L(x_1, .., x_n, Y_1, ..., Y_n) \Big],$$

где L - совместная функция правдоподобия $x_1, .., x_n, y_1, ..., y_n$.

Комментарий. Ответ может содержать величины

$$e_{i,j} = \mathbb{P}\big\{ Y_i = j \big\}, \qquad i = 1..n, \quad j = 1..K.$$

4

T5* 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

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