January 31, 2018

Course "Multivariate analysis and nonparametric statistics". Home Assignment 1.

Topic: Density estimation.

Deadline: February 13, 2018, before the classes.

Main rules:

1. *Please do the home assignment individually.*
2. *The solutions to the numerical part should contain the programming code and several pictures (at least 1 picture for each item), which serve as an evidence that the solution is correct. All codes and pictures should be "pasted" into a single pdf file. You can write your solutions to the theoretical tasks by hand, and then take pictures of your solutions, but all photos should be also "pasted" into a single PDF file.*
3. *The report (1 PDF file) should be submitted by email to vpanov@hse.ru. The deadline is strict.*
4. *It is not obligatory to make implementations in R - any other programming language can be used.*

Numerical part

N1 Consider the database "president" , containing quarterly approval rating for the President of the United States from the first quarter of 1945 to the last quarter of 1974, see `https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/presidents.html`. The general aim is to construct several estimators for the probability density of the rating.

(i) Construct the histogram estimator with amount of bins selected by the Sturges rule.

(ii) Construct the kernel estimators with various kernels (apply all kernels available in the R language). The bandwidth can be chosen by default.

(iii) Construct the kernel estimators under various choices of bandwidth (apply all rules for bandwidth selection, which are implemented in the R language). The kernel can be chosen by default.

(iv) Among the kernel estimators obtained on steps (ii) and (iii), find an estimator which is closest to the histogram estimator obtained in (i). For the measure of closeness between a kernel estimator $f_n^{(K)}$ and the histogram $f_n^{(H)}$, use

$$\frac{1}{n}\sum_{i=1}^{n}\left(f_n^{(K)}(x_i) - f_n^{(H)}(x_i)\right)^2,$$

where $x_1, ..., x_n$ are the points, for which the values of $f_n^{(K)}$ are known.

N2 (i) Simulate a sample of length $N = 1000$ having the distribution with density

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

where ϕ^N is the density of the standard normal distubution (with zero mean and variance equal to 1), ϕ^E is the density of the standard exponential distribution (with rate 1), equal to

$$\phi^E(x) = \begin{cases} e^{-x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Hint: use "rexp" for modelling from exponential distribution. Random variables with densities $\phi^E(x+1), \phi^E(-x+1)$ can be modelled by $(rexp(1) - 1)$ and $(-rexp(1) + 1)$ resp.

(ii) Construct the histogram estimator $\hat{p}_n(x)$ with amount of bins chosen according to the Sturges rule. Calculate the empirical analogue of MISE, namely

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

where $x_1, ..., x_Q$ form the equidistant grid on $[-3, 3]$, and $Q = 10000$.

(iii) Estimate MISE more precisely: namely, repeat the steps 1 and 2 many times (say, $J = 20$ times), get the estimates $\hat{p}_n^{(1)}(x), ...\hat{p}_n^{(J)}(x)$ and afterwards estimate MISE by

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

(iv) Repeat steps (i)-(iii), but using other methods for bandwidth selection on step (ii) (Freedman-Diaconis, Scott's rules). Which choice of the method is better in this situation, i.e., which choice leads to smaller values of (3)?

(v) Consider the values of the bandwidth taking from an equidistant grid on $(0, 1)$ with step 0.01. For each value of bandwidth, construct the kernel estimator with Epanechnikov kernel. Estimate the MISE and plot the graph, which illustrates the dependence between h and estimated MISE. Under which choice of h the MISE for the kernel estimator is minimal?

(vi) On one same graph, display

- the plot of the best histogram estimator, that is, the histogram estimator with best choice of bandwidth, see item (iv);

- the plot of the best kernel estimator, see (v);

- the plot of the true density function.

Theoretical part

T1 Let $p(x)$ be a function defined by (1).

(i) Explain why $p(x)$ is the probability density function.

(ii) Calculate mathematical expectation and variance of a random variable with this distribution.

T2 Assuming that the data follow the normal distribution with mean 0 and variance σ^2.

(i) Calculate the optimal value of bandwidth of the histogram estimator, that is, calculate the value which minimizes the AMISE of the histogram estimator.

(ii) With the optimal choice of bandwidth, analyse how

1. the part of AMISE corresponding to the bias depends on σ;
2. the part of AMISE corresponding to the variance depends on σ.

T3 Calculate (without using a computer) the theoretical efficiencies of

(i) the triangular kernel;

(ii) the Gaussian kernel.

T4 Assume that the data have standard exponential distribution (rate=1).

(i) Calculate the value of bandwidth, which minimizes the AMISE of the kernel estimator constructed with Epanechnikov kernel.

(ii) Propose a method for estimation of bandwidth, taking into account the result of the previous item.