Universität Bielefeld
Tanja Pasurek
Ehsan Abedi WiSe 2024/25

Foundations of Statistics

Homework 8

Kernel density estimation (Chapter 2.4)

Exercise 1 (Moments of kernel density estimators).

Given a kernel K, a fixed bandwidth $b > 0$, and real-valued samples $x_1, ..., x_n$, the kernel density estimator $\hat{f}(x)$ is defined as (cf. Ch. 2.4, Def. (1))

$$\hat{f}(x) := \frac{1}{nb} \sum_{i=1}^{n} K(\frac{x - x_i}{b}), \quad \forall x \in \mathbb{R}.$$

As a function depending on samples, $\hat{f}$ is a *random* function, and so are its moments.

(a) Show that $\hat{f}$ is a probability density, that is

$$\int_{-\infty}^{+\infty} \hat{f}(x) \, \mathrm{d}x = 1.$$

(b) Show that the 1st moment of $\hat{f}$ is

$$m_1(\hat{f}) := \int_{-\infty}^{+\infty} x \hat{f}(x) \, \mathrm{d}x = \bar{x}_n,$$

i.e. the 1st moment of $\hat{f}$ is equal to the sample mean $\bar{x}_n := \frac{1}{n} \sum_{i=1}^{n} x_i$.
Hint: use the normalization and symmetry property of the kernel.

(c) Show that the 2nd moment of $\hat{f}$ is

$$m_2(\hat{f}) = \int_{-\infty}^{+\infty} x^2 \hat{f}(x) \, \mathrm{d}x = b^2 m_2(K) + \frac{1}{n} \sum\nolimits_{i=1}^{n} x_i^2,$$

where

$$m_2(K) := \int_{-\infty}^{+\infty} x^2 K(x) \, \mathrm{d}x$$

is the 2nd moment of the kernel K.

(d) Finally, show that the variance of the density $\hat{f}$ is given by

$$\mathrm{var}(\hat{f}) := m_2(\hat{f}) - [m_1(\hat{f})]^2 = b^2 m_2(K) + \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

Observe that the second term is equal to the sample standard deviation with denominator n.

(e) Compare the results of (b) and (d) and comment on your observation.

(f) Now assuming that the samples are iid from a certain density f with 1st moment (=mean) $\mu \in \mathbb{R}$ and 2nd moment $M \in \mathbb{R}_+$. Compute $\mathbb{E}[m_1(\hat{f})]$ and $\mathbb{E}[m_2(\hat{f})]$ and comment on your observation.

Exercise 2 (A property of the empirical CDF).

Let $X_1, ..., X_n$ be i.i.d. real-valued samples from a CDF $x \mapsto F(x)$, and let $x \mapsto \hat{F}_n(x)$ denote the ECDF:

$$\hat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x) \qquad \text{for all } x \in \mathbb{R}.$$

(a) Show that

$$\mathrm{Cov}[\hat{F}_n(x), \hat{F}_n(y)] = \frac{1}{n}[F(x \wedge y) - F(x)F(y)] \quad \text{for all } x, y \in \mathbb{R},$$

where $x \wedge y = \min(x, y)$.

(b) Conclude that $\hat{F}_n(x)$ and $\hat{F}_n(y)$ are positively correlated: If $\hat{F}_n(x)$ overshoots $F(x)$, then $\hat{F}_n(y)$ will tend to overshoot $F(y)$.

Exercise 3 (Towards an optimal bandwidth).

The goal of this exercise is to estimate the optimal bandwidth in order to apply kernel density estimation using the *Epanechnikov* and *Cosine* kernels.

(a) The *Cosine* kernel is defined as

$$K_C(u) := \begin{cases} \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) & -1 \leq u \leq 1, \\ 0 & \text{elsewhere.} \end{cases}$$

Show that this is indeed a kernel density function. Compute its relative efficiency (as defined in the lectures with respect to the Epanechnikov kernel).

(b) From the built-in `faithful` dataset, define `A=faithful$eruptions*60`, which is the eruption time in seconds. Plot the histogram of the eruption data with the command `hist(A,prob=T)` and observe that it does *not* look like a normal distribution.

(c) To estimate the density using a kernel K, we first need to determine a bandwidth b. Recall that the optimal bandwidth b_{opt}, which is obtained by minimizing asymptotic mean integrated squared error (cf. Eq.(28) in Ch. 2.4), is given by

$$b_{opt} = \left(\frac{j_2(K)}{n[k_2(K)]^2 \int_{-\infty}^{\infty} f''(x)^2 \mathrm{d}x} \right)^{1/5},$$

But, this depends on the true density f, which is unknown!

Therefore, as a first step, we assume that f can be approximated by the density f_G of a Gaussian distribution $\mathcal{N}(\mu, \sigma)$ with some mean μ and standard deviation $\sigma > 0$. Show that for the density f_G, we have

$$\left(\frac{1}{\int_{-\infty}^{\infty} f_G''(x)^2 \mathrm{d}x} \right)^{1/5} = \sigma \cdot \left(\frac{8}{3}\sqrt{\pi} \right)^{1/5}.$$

Hint: First, show the fundamental identity $f_G'(x)/f_G(x) = -\frac{x-\mu}{\sigma^2}$.Then, apply a change variable and use $\int_{-\infty}^{\infty} e^{-z^2}(z^2-1)^2 dz = 3\sqrt{\pi}/4$.

(d) Now, calculate the optimal bandwidth for the Epanechnikov and Cosine kernels for the eruption data, using the following approximation

$$\left(\frac{1}{\int_{-\infty}^{\infty} f''(x)^2 \mathrm{d}x} \right)^{1/5} \approx s \cdot \left(\frac{8}{3}\sqrt{\pi} \right)^{1/5}.$$

where s is the sample standard deviation of the data. Create a single plot in `R` showing the histogram of the data and the kernel density estimates for the Epanechnikov and Cosine kernels.

(e) Explain briefly how one can improve the approximation above to obtain a better bandwidth.

Exercise 4. The built-in-dataset `WWWusage` in the package `stats` contains a time series of the numbers of users connected to the Internet through a server every minute.

(a) Calculate the quantiles, maximum, minimum, mean, median, IQR and mode with `R`. (For mode, one needs to write a function.)

(b) A value x of the dataset is called an outlier if

$$x < x_{0.25} - 1.5 \times \text{IQR} \quad \text{or} \quad x > x_{0.75} + 1.5 \times \text{IQR}.$$

Here by x_α we mean the α-quantiles. Given this definition, are there outliers among this dataset?

(c) Draw a boxplot of the data using the command `boxplot(WWWusage)`. Add the median and mode as colored horizontal lines on the boxplot. Describe the other lines of the boxplot by inserting them manually.

(d) With the command `histo <- hist(WWWusage)` draw a default histogram (with automatically chosen bins and absolute frequencies). Add the command `rug(WWWusage)`.
What is the chosen bin size? Get numerical information about this histogram by using the command `str(histo)`.

(e) We want to estimate the probability that a sample lies in the bin $(100, 110]$. To this end, adjust the `breaks` manually and use the output of `hist` function. Check the result by direct computation.

(f) With the commands `density` and `lines` add a kernel density plot to your histogram. Try Gaussian, Epanechnikov, rectangular, and triangular kernels and vary the bandwidth. Describe the results.