Density estimation

Bandwidht choice by leave-one-out maximum likelihood

Pedro Delicado

20/set./2023

Histogram

1. At the slides we have seen the following relationship

$$\hat{f}_{h,(-i)}(x_i) = \frac{n}{n-1}\left(\hat{f}_h(x_i) - \frac{K(0)}{nh}\right)$$

   between the leave-one-out kernel density estimator $\hat{f}_{h,(-i)}(x)$ and the kernel density estimator using all the observations $\hat{f}_h(x)$, when both are evaluated at x_i, one of the observed data. Find a similar relationship between the histogram estimator of the density function $\hat{f}_{\text{hist}}(x)$ and its leave-one-out version, $\hat{f}_{\text{hist},(-i)}(x)$, when both are evaluated at x_i.

2. Read the CD rate data set and call `x` the first column. Then define

```
A <- min(x)-.05*diff(range(x))
Z <- max(x)+.05*diff(range(x))
nbr <- 7
```

and plot the histogram of `x` as

```
hx <- hist(x,breaks=seq(A,Z,length=nbr+1),freq=F)
```

The following sentence converts this histogram into a function that can be evaluated at any point of $\mathbb{R}$, or at a vector of real numbers:

```
hx_f <- stepfun(hx$breaks,c(0,hx$density,0))
```

Use `hx_f` to evaluate the histogram at the vector of observed data x. Then add the points $(x_i, \hat{f}_{\text{hist}}(x_i))$, $i = 1, \dots, n$, to the histogram you have plotted before.

3. Use the formula you have found before relating $\hat{f}_{\text{hist}}(x_i)$ and $\hat{f}_{\text{hist},(-i)}(x_i)$ to compute $\hat{f}_{\text{hist},(-i)}(x_i)$, $i = 1, \dots, n$. Then add the points $(x_i, \hat{f}_{\text{hist},(-i)}(x_i))$, $i = 1, \dots, n$, to the previous plot.

4. Compute the leave-one-out log-likelihood function corresponding to the previous histogram, at which `nbr=7` has been used.

5. **Choosing `nbr` by leave-one-out Cross Validation (looCV)**. Consider now the set `seq(1,15)` as possible values for `nbr`, the number of intervals of the histogram. For each of them compute the leave-one-out log-likelihood function (`looCV_log_lik`) for the corresponding histogram. Then plot the values of `looCV_log_lik` against the values of `nbr` and select the optimal value of `nbr` as that at which `looCV_log_lik` takes its maximum. Finally, plot the histogram of x using the optimal value of `nbr`.

6. **Choosing `b` by looCV**. Let `b` be the common width of the bins of a histogram. Consider the set

```
seq((Z-A)/15,(Z-A)/1,length=30)
```

as possible values for `b`. Select the value of `b` maximizing the leave-one-out log-likelihood function, and plot the corresponding histogram. *NOTE:* To avoid errors, use the following sintax for computing a histogram with bin width `b`

```
hx <- hist(x,breaks=seq(A,Z+b,by=b), plot=F)
```

and this sentence to plot it:

```
plot(hx,freq = FALSE)
```

7. Recycle the functions `graph.mixt` and `sim.mixt` defined at `density_estimation.Rmd` to generate $n = 100$ data from

$$f(x) = (3/4)N(x; m = 0, s = 1) + (1/4)N(x; m = 3/2, s = 1/3)$$

   Let `b` be the bin width of a histogram estimator of $f(x)$ using the generated data. Select the value of `b` maximizing the leave-one-out log-likelihood function, and plot the corresponding histogram. Compare with the results obtained using the Scott's formula:

$$b_{\text{Scott}} = 3.49\,\text{St.Dev}(X)n^{-1/3}.$$

Kernel density estimator

8. Consider the vector `x` of data you have generated before from the mixture of two normals. Use the relationship

$$\hat{f}_{h,(-i)}(x_i) = \frac{n}{n-1}\left(\hat{f}_h(x_i) - \frac{K(0)}{nh}\right)$$

   to select the value of `h` maximizing the leave-one-out log-likelihood function, and plot the corresponding kernel density estimator. *NOTE:* The following sentences converts the kernel density estimator obtained with the function `density` into a function that can be evaluated at any point of $\mathbb{R}$ or at a vector of real numbers:

```
kx <- density(x)
kx_f <- approxfun(x=kx$x, y=kx$y, method='linear', rule=2)
```