

550.640 Project 1; kernel density estimators.

The solution must be written like you would write a report, with explanations and commented results. A list of numbers and figures is not sufficient.

Providing the program sources is optional. They are not graded (so no direct credit for them), but they can help understanding why results are not correct to decide whether partial credit can be given.

Due on Monday Feb. 25th.

(A) Write a program which, given an (M, d) matrix U , an (N, d) matrix X and a scalar parameter h , computes the kernel density estimator

$$f_h(u_i) = \frac{1}{Nh^d} \sum_{k=1}^N K((u_i - x_k)/h), \quad i = 1, \dots, M$$

where the x_k 's are the rows of X and the u_i are the rows of U . Provide the output in the form of an M vector. Use the Gaussian kernel,

$$K(x) = \exp(-|x|^2/2)/(2\pi)^{d/2}.$$

Illustrate the program with the result of the following 1D experiments in which you will plot $f(u)$ and the true density of X in function of u on the same figure.

- (i) $d = 1$, $N = 20$, $h = .1$, X is a random sample of the uniform distribution on $[0, 1]$, and U provides an even discretization of the interval $[-0.5, 1.5]$ in 100 points.
- (ii) Same as (i) with $h = .01$.
- (iii) Same as (i) with $N = 100$.
- (iv) Same as (i) with $h = .01$, $N = 100$.
- (v) Same as (i), but X is a sample of a standard gaussian distribution.
- (vi) Same as (v), with $N = 100$.

(B) We consider the following application of cross-validation to estimate an optimal parameter h in the previous estimator, based on the training set $T = \{x_1, \dots, x_N\}$. First separate T in q subsets of equal sizes T_1, \dots, T_q and let $T^j = T \setminus T_j$. Let f_h^j be the density estimator at scale h based on T^j . Define, for each $h > 0$, the number

$$L_q(h) = \sum_{j=1}^q \sum_{x \in T_j} \log f_h^j(x).$$

Let the final estimator be associated to

$$h^* = \operatorname{argmax} L_q.$$

(B-1) Explain why this procedure makes sense.

(B-2) Assume that no x_k is duplicated in T . Compute the limits of $L_q(h)$ when $h \rightarrow 0$ and when $h \rightarrow \infty$. Why does this show that a maximizer of L_q always exists?

(B-3) Define, for $x \in T_j$ and $y \in T^j$,

$$\pi_h^j(y|x) = \frac{\exp(-|x - y|^2/2h^2)}{\sum_{z \in T^j} \exp(-|x - z|^2/2h^2)}.$$

Prove that the optimal h satisfies

$$h^2 = \frac{1}{Nd} \sum_{j=1}^q \sum_{x \in T_j} \sum_{y \in T^j} |x - y|^2 \pi_h^j(y|x).$$

Simplify this in the case $q = N$. (This expression is not used in the next questions.)

(B-4) Based on the program written in (a), write a program which, given a training set X , plots $L_q(h)$ as a function of h for a given value of q . Use this to compute the optimal h^* with leave-one-out cross-validation ($q = N$), and plot the best density in the 6 cases given in Question (a) (for each case, provide the function L_q , the optimal h , the estimated density compared with the true one.)

(B-5) Plot $-\log h$ vs. $\log N$ where h is the optimal parameter computed with 10-fold and leave-one-out cross-validation for a 1D model in which the true density is standard Gaussian. Use discrete values of N between 50 and 1000 (not all of them!). Provide a commented result.

Hint: Use the previous obtained value of h to decide of the initial range for the next search.

(C) Provide a 2D image plot of the estimated density computed with the dataset *project1.dat* available on the class web site. This ASCII dataset that contains a list of 1000 points in 2 dimensions, and is directly readable using matlab load command.