UNIVERSITE DE GENEVE
Faculté des sciences

Section de Mathématiques

Statistical Machine Learning
TP 4

1D smoother, cross validation

1. Program the running mean

$$\hat{\mu}_\lambda^{\mathrm{RM}}(x) = \frac{\sum_{j=1}^n y_j 1_{|x_j - x| \le \lambda/2}}{\sum_{j=1}^n 1_{|x_j - x| \le \lambda/2}},$$

and running median estimators for a window width $\lambda > 0$ to estimate the underlying function μ at any point $x \in \mathbb{R}$ based on a realisation $(y_1, \ldots, y_n)$ of

$$Y_i = \mu(x_i) + \epsilon_i \quad \text{with} \quad \epsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2). \tag{1}$$

Note that the x_i's are not necessarily equispaced. To estimate $\mu(x)$ at a point x outside the range of the data $[x_{(1)}, x_{(n)}]$, we define

$$\hat{\mu}_\lambda^{\mathrm{RM}}(x) = \begin{cases} \hat{\mu}_\lambda^{\mathrm{RM}}(x_{(1)}) & x \le x_{(1)} \\ \hat{\mu}_\lambda^{\mathrm{RM}}(x_{(n)}) & x \ge x_{(n)} \end{cases}.$$

2. To test your function, choose a function μ, a sample size n, locations $x_1, \ldots, x_n$, a variance σ^2 to simulate data according to (1). Then choose a value of λ to plot $\hat{\mu}_\lambda^{\mathrm{RM}}(x)$ on a grid of x's on $[.8\ x_{(1)}, 1.2\ x_{(n)}]$.

3. For the choice of λ:

   (a) Choose your best λ by eye.

   (b) Since you are oracle, choose λ by minimizing w.r.t. λ the ℓ_2 distance between $\hat{\mu}_\lambda^{\mathrm{RM}}$ and μ on a fine grid on $[x_{(1)}, x_{(n)}]$:

$$\ell_2^2(\lambda) = \sum_{x_k \in \mathrm{grid}} (\hat{\mu}_\lambda^{\mathrm{RM}}(x_k) - \mu(x_k))^2.$$

   You may want to plot the ℓ_2 distance as a function of λ.

   (c) Simulate a validation set according to (1) and optimize its prediction as a function of λ.

   (d) Choose λ by 2-fold cross-validation (odd versus even indexes). You may want to plot $\mathrm{CV}(\lambda)$ on the same plot as $\ell_2^2(\lambda)$.

4. Instead of Gaussian errors in (1), you may want to use Sudent distribution with 3 degrees of freedom and observe which of the running mean and running median is more robust to heavy tails.