TD3: Non-parametric regression (Cubic splines)

1 Context

Consider a pair of random variables $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$. The regression function of Y on X is given by:

$$f^\star(x) = \mathbb{E}(Y|X = x)$$

The goal in nonparametric regression is to construct an estimation $\widehat{f}_n$ of $f^\star$ using i.i.d. samples $(x_1, y_1), \dots (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ with no particular assumptions on $\widehat{f}$ nor on the distribution of the data except that f is twice differentiable over some segment $[a, b]$ and that we can write:

$$y_i = f(x_i) + \varepsilon_i \ ,$$

with mean zero random errors $\varepsilon_1, \dots, \varepsilon_n$.

Minimizing the data fitting term $\sum_{i=1}^n |y_i - f(x_i)|^2$ over the span of $\mathcal{C}^2$ functions leads to *overfitting*: $\widehat{f}_n$ would be too complex and learn the noise ε. To avoid learning functions that are too "wiggly", a good compromise is to minimize instead the regularized problem:

$$\min_{f \in \mathcal{C}^2} \frac{1}{n} \sum_{i=1}^n |y_i - f(x_i)|^2 + \alpha \int_a^b f''(x)^2 \mathrm{d}x \ , \tag{1}$$

for some fixed hyperparameter $\alpha > 0$.

Theorem

Problem (1) has is a unique solution given by piecewise cubic polynomials defined over the windows $[a, x_0], [x_0, x_1], \dots, [x_n, b]$ known as *natural cubic splines* with knots $(x_0, \dots, x_n)$.

Definition: Natural Cubic Spline (NCS)

A natural cubic spline (NCS) g with knots (i.e nodes) $(x_1, \dots, x_n)$ is a twice differentiable function such that:

1. g is a cubic polynomial on each of $[x_1, x_2], \dots [x_{n-1}, x_n]$,
2. g is a linear function outside $[x_1, x_n]$,
3. g is continuous and has continuous first and second derivatives at its knots $x_1, \dots, x_n$.

Let's prove this theorem.

2 Exercise 1: infinite dimensional problem

Let $n \geq 2$ and g is the NCS interpolant of the pairs (x_i, y_i) where $a < x_1 \cdots < x_n < b$. Let $\widetilde{g}$ be any other differentiable function on $[a, b]$ that interpolates the n pairs.

1. Let $h = g - \widetilde{g}$. Using integration by parts and the definition of a NCS, show that:

$$\int_a^b g''(x)h''(x)\mathrm{d}x = 0$$

2. Deduce that:
$$\int_a^b g''(x)^2 \mathrm{d}x \leq \int_a^b \widetilde{g}''(x)^2 \mathrm{d}x,$$
and that equality only holds if h is exactly 0 on [a, b].

3. Show that the solution of the problem (1) must be a natural cubic spline.

3 Characterizing the solution

Since the solution must be a NCS, problem (1) is equivalent to a finite dimensional problem: all we need to do is find the coefficients of the cubic polynomials that define it. Let g be a NCS with knots $x_1, \ldots, x_n$ and and define the quantities

$$g_i = g(x_i) \text{ and } \gamma_i = g''(x_i).$$

Notice that since g is linear outside $[x_1, x_n]$ (i.e on $[a, x_1]$ and $[x_n, b]$), it follows that $\gamma_1 = \gamma_n = 0$. Now define $\mathbf{g} = (g_1, \ldots, g_n)^\top \in \mathbb{R}^n$ and $\boldsymbol{\gamma} = (\gamma_2, \ldots, \gamma_{n-1}) \in \mathbb{R}^{n-2}$. Finally, let $h_i = x_{i+1} - x_i$ for $i = 1 \ldots n-1$ and the define the symmetric matrix $R \in \mathbb{R}^{n-2,n-2}$ and the rectangular matrix $Q \in \mathbb{R}^{n,n-2}$ with column indexing $j = 2 \ldots n-1$ same as the indexing of γ (the top left element of Q is thus Q_{12}) such that:

$$\begin{cases} Q_{j-1,j} = \frac{1}{h_{j-1}} \\ Q_{jj} = -\frac{1}{h_{j-1}} - \frac{1}{h_j} \\ Q_{j+1,j} = \frac{1}{h_j} \\ Q_{ij} = 0 \text{ if } |i-j| \geq 2 \end{cases} \tag{2}$$

$$\begin{cases} R_{ii} = \frac{1}{3}(h_{i-1} + h_i) \text{ for } i = 2, \ldots, n-1 \\ R_{i,i+1} = R_{i+1,i} = \frac{h_i}{6} \text{ for } i = 2, \ldots, n-2 \\ R_{ij} = 0 \text{ for } |i-j| \geq 2 \end{cases} \tag{3}$$

Sparring you the technical enforcing of the NCS conditions, it can be shown that:

Theorem

R is positive definite and $\mathbf{g}$ and $\boldsymbol{\gamma}$ specify a NCS g if and only if:

$$Q^\top \mathbf{g} = R\boldsymbol{\gamma},$$

and in that case:

$$\int_a^b g''(x)^2 \mathrm{d}x = \boldsymbol{\gamma}^\top R \boldsymbol{\gamma} = \mathbf{g}^\top \mathbf{K} \mathbf{g},$$

where $\mathbf{K} = QR^{-1}Q^\top$.

4 Exercise 2: solving (1), handling duplicated data, and choosing α

1. Using EX 1 and the theorem above, write the finite dimensional equivalent problem of (1) and show that the solution is characterized by $(\mathbf{g}, \boldsymbol{\gamma})$ and can be written:
$$\mathbf{g} = (I + \alpha K)^{-1}\mathbf{Y}, \qquad \boldsymbol{\gamma} = R^{-1}Q^\top \mathbf{g}$$

2. Denote $g_{-i}(x; \alpha)$ the solution of the problem after removing the i-th observation (x_i, y_i) i.e formally:
$$g_{-i}(.; \alpha) = \text{argmin}_g \sum_{j \neq i} |y_j - g(x_j)|^2 + \int_a^b g"(x)^2 \mathrm{d}x$$
In practice, we would like to set α that provides the smallest error on *new unseen data*. Thus, we minimize in α the cross validation score:
$$CV(\alpha) = \sum_{i=1}^n (Y_i - g_{-i}(x_i; \alpha))^2$$
Consider α and i fixed and define $\mathbf{Y}' \in \mathbb{R}^n$ such that $\mathbf{Y}'_j = \mathbf{Y}_j$ if $j \neq i$ and $\mathbf{Y}'i = g_{-i}(x_i, \alpha)$. Show that $\mathbf{g}_{-i} = A(\alpha)\mathbf{Y}'$ where $\mathbf{A}(\alpha) = (I + \alpha K)^{-1}$.

3. Deduce the formula for $CV(\alpha)$:

$$CV(\alpha) = \sum_{i=1}^{n} \left(\frac{Y_i - g(x_i)}{1 - A(\alpha)_{ii}} \right)^2,$$

where g is the NCS solution on the entire data $(x_1, y_1)..(x_n, y_n)$.

4. Computing CV(α) requires the diagonal elements of $A(\alpha)$. To reduce the computational cost, a simple idea is to replace the denominators by their average value trace$(A(\alpha))n^{-1}$. This is known as the generalized cross validation score:

$$\text{GCV}(\alpha) = \frac{\sum_{i=1}^{n}(Y_i - g(x_i))^2}{(1 - n^{-1}\text{tr}A(\alpha))^2}$$

Propose an algorithm to compute GCV for a grid of hyperparameters $\alpha_1, \ldots, \alpha_M$.

5. In practice, it is rather likely to have several observations $y_{i,1}, \ldots, y_{i,m}$ for the same observation x_i. Show that the problem (with duplicated x_i):

$$\min_g \sum_{i=1}^{n} \sum_{j=1}^{m} |y_{ij} - g(x_i)|^2 + \alpha \int_a^b g"(x)^2 \mathrm{d}x$$

is equivalent to the weighted regression problem:

$$\min_g \sum_{i=1}^{n} w_i |y'_i - g(x_i)|^2 + \alpha \int_a^b g"(x)^2 \mathrm{d}x$$

where w_i and y' must be determined.

6. Solve the problem and deduce a formula for the modified $A_w(\alpha)$ as well as $\text{GCV}_w(\alpha)$.

5 Exercise 3: Practice

See the jupyter notebook "TP - Splines".