Processus Gaussiens en Apprentissage Machine

1. Définition du processus gaussien

Soit une fonction $f : \mathbb{R}^d \to \mathbb{R}$ distribuée selon un processus gaussien :

$$f \sim \mathcal{GP}(0, k_\lambda(x, x'))$$

où $k_\lambda(x, x')$ est un noyau de covariance, ici le noyau exponentiel classique (RBF) :

$$k_\lambda(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\lambda^2}\right)$$

2. Distribution conjointe en 2 ou 3 points

Supposons que l'on observe les valeurs de la fonction aux points x_1, x_2, x_3, on note :

$$\mathbf{f} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$$

où la matrice de covariance est :

$$\mathbf{K} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & k(x_1, x_3) \\ k(x_2, x_1) & k(x_2, x_2) & k(x_2, x_3) \\ k(x_3, x_1) & k(x_3, x_2) & k(x_3, x_3) \end{bmatrix}$$

3. Distribution conditionnelle

Soit x_* un nouveau point pour lequel on souhaite prédire $f_* = f(x_*)$, en connaissant les valeurs de f aux points $x_1, \ldots, x_n$.

On note :

- $\mathbf{f}_{\text{train}} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$
- $f_* = f(x_*)$

La distribution conjointe est :

$$\begin{bmatrix} \mathbf{f}_{\text{train}} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{\text{train}} & \mathbf{k}_* \\ \mathbf{k}_*^\top & k_{**} \end{bmatrix} \right)$$

avec :

- $\mathbf{K}_{\text{train}} \in \mathbb{R}^{n \times n}$, la matrice de covariance entre les points d'entraînement.
- $\mathbf{k}_* = \begin{bmatrix} k(x_*, x_1) \\ \vdots \\ k(x_*, x_n) \end{bmatrix}$
- $k_{**} = k(x_*, x_*)$

Alors la loi conditionnelle de f_* sachant $\mathbf{f}_{\text{train}}$ est :

$$f_* \mid \mathbf{f}_{\text{train}} \sim \mathcal{N}(\mu_*, \sigma_*^2)$$

avec :

$$\mu_* = \mathbf{k}_*^\top \mathbf{K}_{\text{train}}^{-1} \mathbf{f}_{\text{train}}$$

$$\sigma_*^2 = k_{**} - \mathbf{k}_*^\top \mathbf{K}_{\text{train}}^{-1} \mathbf{k}_*$$

4. Exemple numérique (configuration)

Soient :

$$\begin{aligned} x_1 &= 0, \quad f_1 = f(0) \\ x_2 &= 1, \quad f_2 = f(1) \\ x_* &= 0.5 \end{aligned}$$

Alors :

- $\mathbf{f}_{\text{train}} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$
- $\mathbf{K}_{\text{train}} = \begin{bmatrix} k(0,0) & k(0,1) \\ k(1,0) & k(1,1) \end{bmatrix}$
- $\mathbf{k}_* = \begin{bmatrix} k(0.5, 0) \\ k(0.5, 1) \end{bmatrix}$
- $k_{**} = k(0.5, 0.5) = 1$

section*5. Variance conditionnelle en un point observé

Supposons que l'on observe la valeur de la fonction f en un point x_i, c'est-à-dire que $f(x_i)$ fait partie de $\mathbf{f}_{\text{train}}$, et que l'on cherche à prédire $f_* = f(x_i)$, soit le même point.

Dans ce cas, la variance conditionnelle du processus gaussien est donnée par :

$$\text{Var}[f(x_i) \mid \mathbf{f}_{\text{train}}] = k(x_i, x_i) - \mathbf{k}_i^\top \mathbf{K}_{\text{train}}^{-1} \mathbf{k}_i$$

Mais ici :

- $\mathbf{k}_i$ est la i-ème colonne (ou ligne) de $\mathbf{K}_{\text{train}}$
- Donc $\mathbf{k}_i = \mathbf{K}_{\text{train}} \mathbf{e}_i$, où $\mathbf{e}_i$ est le vecteur unité de la base canonique
- Ainsi :

$$\mathbf{k}_i^\top \mathbf{K}_{\text{train}}^{-1} \mathbf{k}_i = (\mathbf{e}_i^\top \mathbf{K}_{\text{train}}^\top) \mathbf{K}_{\text{train}}^{-1} (\mathbf{K}_{\text{train}} \mathbf{e}_i) = \mathbf{e}_i^\top \mathbf{e}_i = 1$$

  (puisque $\mathbf{K}_{\text{train}}^{-1} \mathbf{K}_{\text{train}} = I$)

Donc :

$$\text{Var}[f(x_i) \mid \mathbf{f}_{\text{train}}] = k(x_i, x_i) - 1 = 0$$

Conclusion : la variance conditionnelle est nulle si l'on prédit la valeur de f en un point x_i déjà observé. Cela reflète le fait que le processus gaussien interpole parfaitement les données d'entraînement (en l'absence de bruit).