AutoML: Gaussian Processes

Covariance Functions for GPs - Advanced

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

MS-Continuity and Differentiability I

We wish to describe a Gaussian process in terms of its smoothness. There are several notions of continuity for random variables. One is the continuity/differentiability in mean square (MS).

Definition

A Gaussian process $f(\mathbf{x})$ is said to be **MS continuous** at $\mathbf{x}_*$, if

$$\mathbb{E}[|f(\mathbf{x}^{(k)}) - f(\mathbf{x}_*)|^2] \xrightarrow{k\to\infty} 0 \text{ for all converging sequences } \mathbf{x}^{(k)} \xrightarrow{k\to\infty} \mathbf{x}_*.$$

A Gaussian process $f(\mathbf{x})$ is said to be **MS differentiable** along the i direction, if the following limit exists, with $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ being the unit vector along the i-th axis.

$$\lim_{h\to 0} \mathbb{E}[|\frac{f(\mathbf{x} + h\,\boldsymbol{e}_i) - f(\mathbf{x})}{h}|]$$

MS-Continuity and Differentiability II

- The MS continuity/differentiability do not necessarily lead to the continuity/differentiability of sampled functions!
- The MS continuity/differentiability of a Gaussian process can be derived from the smoothness properties of the kernel.
- The GP is continuous in MS iff the covariance function $k(\mathbf{x}, \mathbf{x}')$ is continuous.
- The MS derivative of a Gaussian process exists iff the second derivative $\frac{\partial^2 k(\mathbf{x},\mathbf{x}')}{\partial \mathbf{x} \partial \mathbf{x}'}$ exists.

Squared Exponential Covariance Function

The squared exponential function is one of the most commonly used covariance functions.

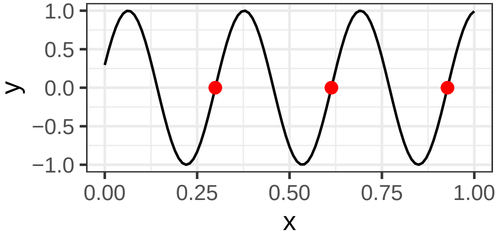
$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right).$$

Properties:

- It depends merely on the distance $r = \|\mathbf{x} - \mathbf{x}'\|$ $\rightarrow$ isotropic and stationary.
- Infinitely differentiable $\rightarrow$ the corresponding GP is too smooth.
- It utilizes strong smoothness assumptions $\rightarrow$ unrealistic for modeling most of the physical processes.

Upcrossing Rate and Characteristic Length-Scale I

- Another way to describe a Gaussian process is the expected number of up-crossings at level-0 on the unit interval, which we denote by N_0.



- For an isotropic covariance function $k(r)$, the expected number of up-crossings can be calculated explicitly:

$$\mathbb{E}[N_0] = \frac{1}{2\pi}\sqrt{\frac{-k''(0)}{k(0)}}.$$

Upcrossing Rate and Characteristic Length-Scale II

Example (squared exponential):

$$\begin{aligned} k(r) &= \exp\left(-\frac{r^2}{2\ell^2}\right) \\ k'(r) &= -k(r) \cdot \frac{r}{\ell^2} \\ k''(r) &= k(r) \cdot \frac{r^2}{\ell^4} - k(r) \cdot \frac{1}{\ell^2} \end{aligned}$$

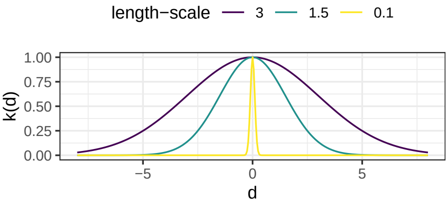
The expected number of upcrossings at level-0 is

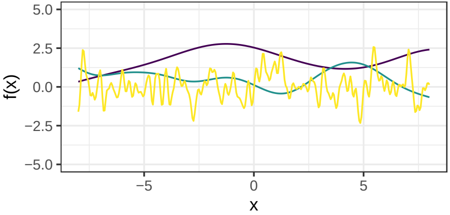
$$\mathbb{E}[N_0] = \frac{1}{2\pi}\sqrt{\frac{-k''(0)}{k(0)}} = \frac{1}{2\pi}\sqrt{\frac{1}{\ell^2}} = (2\pi\ell)^{-1}.$$

Upcrossing Rate and Characteristic Length-Scale III

ℓ is called **characteristic length-scale**. Loosely speaking, the characteristic length-scale describes how far you need to move in input space for the function values to become uncorrelated.

- Left plot: for higher ℓ the correlation between function values (for unchanged distance of input points) is also higher
- Right plot: a higher ℓ induces a smoother function and thus fewer level-0 upcrossings





Upcrossing Rate and Characteristic Length-Scale IV

For more than $p = 2$ dimensions, the squared exponential can be parameterized as follows:

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right)^\top \boldsymbol{M}\left(\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right)\right)$$

Possible choices for the matrix $\boldsymbol{M}$ include

$$\boldsymbol{M}_1 = \ell^{-2}\boldsymbol{I} \qquad \boldsymbol{M}_2 = \mathsf{diag}(\boldsymbol{\ell})^{-2} \qquad \boldsymbol{M}_3 = \Gamma\Gamma^\top + \mathsf{diag}(\boldsymbol{\ell})^{-2}$$

where $\boldsymbol{\ell}$ is a p-vector of positive values and Γ is a $p \times k$ matrix.

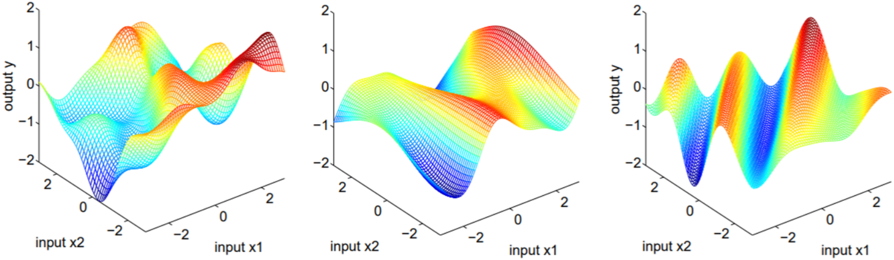
Here again, $\boldsymbol{\ell} = (\ell_1, \dots, \ell_p)$ are characteristic length-scales for each dimension.

Upcrossing Rate and Characteristic Length-Scale V

What is the benefit of learning an individual hyperparameter ℓ_i for each dimension?

- The $\ell_1, \ldots, \ell_p$ hyperparameters play the role of **characteristic length-scales**.
- Losely speaking, ℓ_i describes how far you need to move along axis i in input space for the function values to be uncorrelated.
- Such a covariance function implements **automatic relevance determination** (ARD), since the inverse of the length-scale ℓ_i determines the relevancy of input feature i to the regression.
- If ℓ_i is very large, the covariance will become almost independent of that input, effectively removing it from inference.
- If the features are on different scales, the data can be automatically **rescaled** by estimating $\ell_1, \ldots, \ell_p$

Upcrossing Rate and Characteristic Length-Scale VI



For the first plot, we have chosen $\boldsymbol{M} = \boldsymbol{I}$: the function varies the same in all directions. The second plot is for $\boldsymbol{M} = \text{diag}(\boldsymbol{\ell})^{-2}$ and $\boldsymbol{\ell} = (1, 3)$: The function varies less rapidly as a function of x_2 than x_1 as the length-scale for x_1 is less. In the third plot $\boldsymbol{M} = \Gamma\Gamma^T + \text{diag}(\boldsymbol{\ell})^{-2}$ for $\Gamma = (1, -1)^\top$ and $\boldsymbol{\ell} = (6, 6)^\top$. Here Γ gives the direction of the most rapid variation. (Image from Rasmussen & Williams, 2006)