AutoML: Gaussian Processes

Covariance Functions for GPs

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

Covariance function of a GP I

The marginalization property of the Gaussian process implies that for any finite set of input values, the corresponding vector of function values is Gaussian:

$$\boldsymbol{f} = \left[f\left(\mathbf{x}^{(1)}\right), \ldots, f\left(\mathbf{x}^{(n)}\right)\right] \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{K}\right).$$

- The covariance matrix $\boldsymbol{K}$ is constructed according to the chosen inputs $\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right\}$.
- Each entry $\boldsymbol{K}_{ij}$ is computed by $k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)$.
- Technically, to be a valid covariance matrix, $\boldsymbol{K}$ needs to be positive semi-definite for **every** choice of inputs $\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right\}$.
- A function $k(\cdot, \cdot)$ that satisfies this condition is called **positive definite**.

Covariance function of a GP II

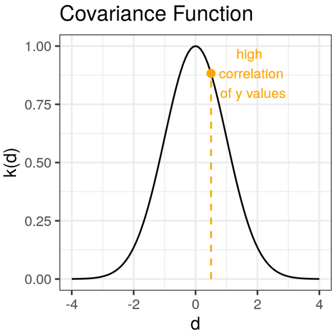
- Recall that the purpose of the covariance function is to control to which degree the following condition is fulfilled:

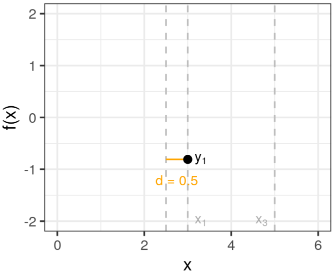
  If $\mathbf{x}^{(i)}$ *and* $\mathbf{x}^{(j)}$ *are close in the* $\mathcal{X}$*-space, their function values* $f(\mathbf{x}^{(i)})$ *and* $f(\mathbf{x}^{(j)})$ *should be close in* $\mathcal{Y}$*-space.*

Closeness of $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ in the input space $\mathcal{X}$ is measured by $\boldsymbol{d} = \mathbf{x}^{(i)} - \mathbf{x}^{(j)}$.

Covariance function of a GP: Example I

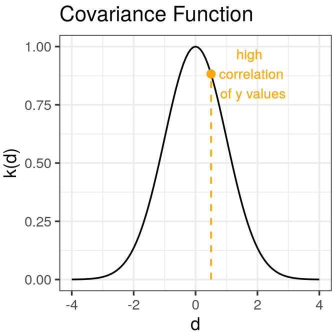
- Let $f(\mathbf{x})$ be a GP with $k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2}\|\boldsymbol{d}\|^2)$ where $\boldsymbol{d} = \mathbf{x} - \mathbf{x}'$.
- Consider two points $\mathbf{x}^{(1)} = 3$ and $\mathbf{x}^{(2)} = 2.5$. To investigate how correlated their function values are, compute their correlation!

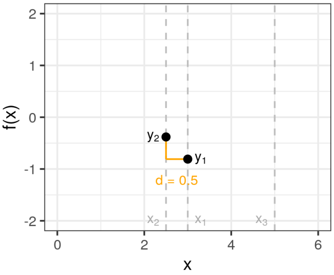




Covariance function of a GP: Example II

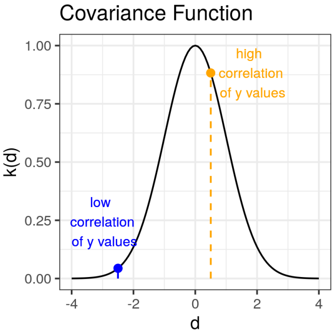
- Assume that we observe a value of $y^{(1)} = -0.8$. Under the said assumption for the Gaussian process, the value of $y^{(2)}$ should be close to $y^{(1)}$.

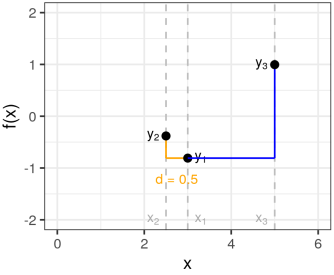




Covariance function of a GP: Example III

- Now, let us take a new point $\mathbf{x}^{(3)}$ which is not too close to $\mathbf{x}^{(1)}$.
- Their function values should not be so correlated. That is, $y^{(1)}$ and $y^{(3)}$ are probably far away from each other.





Covariance Functions

Three types of properties are commonly used in covariance functions:

- k is **stationary** if it depends only on $\boldsymbol{d} = \mathbf{x} - \mathbf{x}'$ and is denoted by $k(\boldsymbol{d})$.
- k is **isotropic** if it depends only on $r = \|\mathbf{x} - \mathbf{x}'\|$and is denoted by $k(r)$.
- k is a **dot product** if it depends only on $\mathbf{x}^T\mathbf{x}'$.

- Isotropy implies stationarity.
- Isotropic functions are rotationally invariant.
- Stationary functions are translationally invariant:

$$k(\mathbf{x}, \mathbf{x} + \boldsymbol{d}) = k(\mathbf{0}, \boldsymbol{d}) = k(\boldsymbol{d})$$

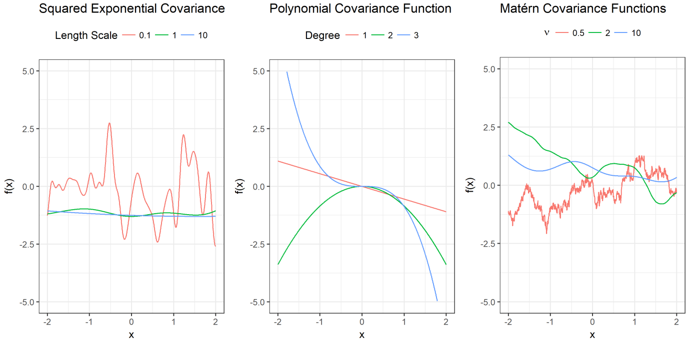
Commonly Used Covariance Functions I

Name	$k(\mathbf{x}, \mathbf{x}')$
constant	σ_0^2
linear	$\sigma_0^2 + \mathbf{x}^T\mathbf{x}'$
polynomial	$(\sigma_0^2 + \mathbf{x}^T\mathbf{x}')^p$
squared exponential	$\exp(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\ell^2})$
Matérn	$\frac{1}{2^\nu\Gamma(\nu)}\left(\frac{\sqrt{2\nu}}{\ell}\|\mathbf{x}-\mathbf{x}'\|\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}}{\ell}\|\mathbf{x}-\mathbf{x}'\|\right)$
exponential	$\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|}{\ell}\right)$

$K_\nu(\cdot)$ is the modified Bessel function of the second kind.

Commonly Used Covariance Functions II

- Some random functions drawn from Gaussian processes with a Squared Exponential Kernel (left), Polynomial Kernel (middle), and a Matérn Kernel (right, $\ell = 1$).
- The length-scale hyperparameter determines the "wiggliness" of the function.
- For Matérn, the ν parameter determines how differentiable the process is.



Squared Exponential Covariance Function

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

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

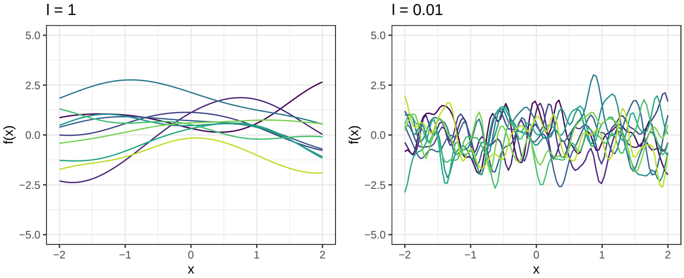
Properties:

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

Characteristic Length-Scale I

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\ell^2}\|\mathbf{x} - \mathbf{x}'\|^2\right)$$

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. Higher ℓ induces smoother functions, lower ℓ induces more wiggly functions.



Characteristic Length-Scale II

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.

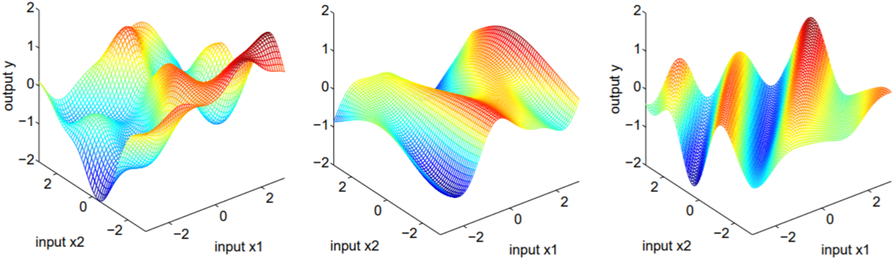
Here again, $\boldsymbol{\ell} = (\ell_1, \dots, \ell_p)$ are characteristic length-scales for each dimension.

Characteristic Length-Scale III

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

- The $\ell_1, \ldots, \ell_p$ hyperparameters play the role of **characteristic length-scales**.
- Loosely 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 $\ell_1, \ldots, \ell_p$

Characteristic Length-Scale IV



For the first plot, we have chosen $\boldsymbol{M} = \boldsymbol{I}$: the function varies the same in all directions. The second plot is for $\boldsymbol{M} = \mathrm{diag}(\boldsymbol{\ell})^{-2}$ 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 $\boldsymbol{M} = \Gamma\Gamma^T + \mathrm{diag}(\boldsymbol{\ell})^{-2}$ for $\Gamma = (1, -1)^\top$ and $\boldsymbol{\ell} = (6, 6)^\top$. Here Γ gives the direction of the most rapid variation.

[Rasmussen and Williams. 2006]