Gaussian Process Regression

Mario Teixeira Parente (m.parente@fz-juelich.de)

ISIS LENS Machine Learning School

February 19, 2021

MLZ is a cooperation between

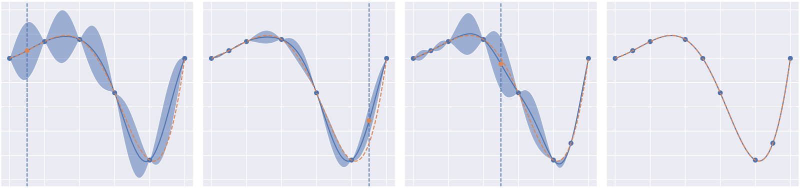
JÜLICH FORSCHUNGSZENTRUM

Helmholtz-Zentrum Geesthacht
Zentrum für Material- und Küstenforschung

Technische Universität München

Introduction

- With **Gaussian Process Regression** (GPR), you can
  - *approximate functions* while at the same time you
  - *quantify uncertainty* about the approximation itself.



- It is *flexible* and based on "not too many assumptions".
- However, it can become *computationally intractable* with functions defined on higher-dimensional domains, i. e., when the number of dimensions exceeds a few dozens.

Bayesian statistics

- In Bayesian statistics, we estimate a parameter θ of a distribution μ with observational data $Y = (Y_1, \ldots, Y_N)^\top$.

- More exactly, we estimate a distribution for θ, the so-called **posterior distribution**.

- We start by including knowledge about θ *a priori*, i. e., before making any observations. This prior information is encoded in a **prior distribution** $\rho(\theta)$.

Bayesian statistics – II

- After observing data $Y_i \sim \mu$, we incorporate this information and *update* the prior to get the posterior information.
- We use *Bayes' theorem* to get

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}, \tag{1}$$

i. e.,

$$\rho(\theta|Y) = \frac{\rho(Y|\theta) \times \rho(\theta)}{\rho(Y)}. \tag{2}$$

Gaussian Process Regression

- In GPR, we do not estimate a distribution for finite-dimensional parameters θ, but for *functions*.
- All distributions involved, i. e., priors and posteriors, are *Gaussian*.
- A **Gaussian Process** f is such a distribution over functions.

Gaussian Process Regression – II

- It is uniquely determined by its **mean function** $m(\mathbf{x})$ and **kernel function** $k(\mathbf{x}, \mathbf{x}')$, i. e.,

$$f \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \tag{3}$$

- For each $\mathbf{x}$, $f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}))$.

- For $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we have that

$$\mathbf{f} := (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^\top \sim \mathcal{N}(\mathbf{m}, K) \tag{4}$$

  with $\mathbf{m} = (m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n))^\top$ and $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

- W.l.o.g., we set $m(\mathbf{x}) \equiv 0$.

Kernel function

- The kernel function k is the most important quantity in GPR.

- It describes the *correlation* between two random function values $f(\mathbf{x})$, $f(\mathbf{x}')$ and hence the "shape" of the function,

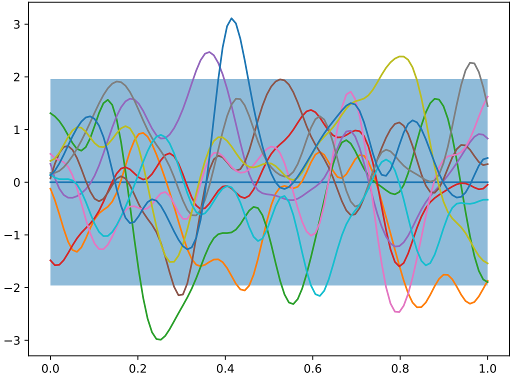
$$k(\mathbf{x}, \mathbf{x}') = \mathbf{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \tag{5}$$

- There are kernel functions available for many contexts.

Kernel function – II

Example: RBF kernel (**R**adial **B**asis **F**unction) a.k.a. Gaussian kernel,

$$k_\ell(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2\ell^2}\right), \quad \ell = 0.05 \ (\textit{length scale}). \qquad (6)$$



Regression

- *Regression* is done when we observe function values at $\mathbf{x}_1, \dots, \mathbf{x}_n$ and would like to predict values in between.
- Let us assume that we observe $y_i = f(\mathbf{x}_i) + \eta$, $\eta \sim \mathcal{N}(0, \sigma^2)$, at locations $X = (\mathbf{x}_i)_i$ and we want to predict $f(\mathbf{x}^*)$ at some point $\mathbf{x}^*$.
- For $\mathbf{y} = (y_i)_i$, using Gaussian identities we get that

$$f(\mathbf{x}^*)|\mathbf{y} \sim \mathcal{N}(\hat{f}(\mathbf{x}^*), \hat{\sigma}^2(\mathbf{x}^*)), \tag{7}$$

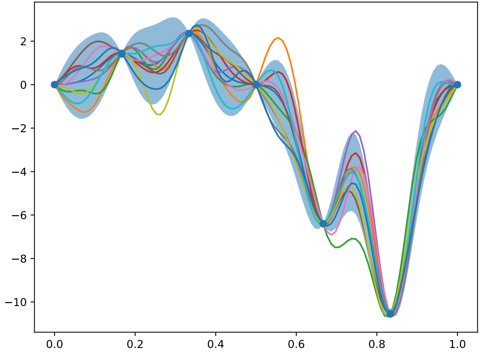
where

$$\hat{f}(\mathbf{x}^*) = k(\mathbf{x}^*, X)(k(X, X) + \sigma^2 I)^{-1}\mathbf{y}, \tag{8}$$

$$\hat{\sigma}^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, X)(k(X, X) + \sigma^2 I)^{-1} k(X, \mathbf{x}^*). \tag{9}$$

Regression – II

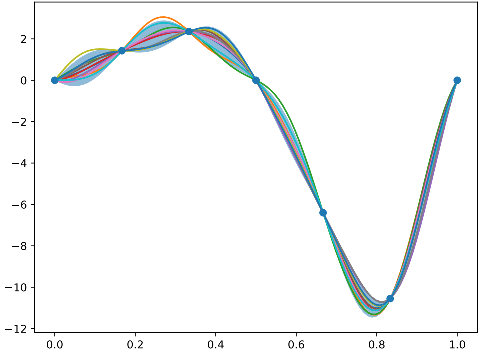
- **Example:** We observe at equidistant locations $\mathbf{x}_i \in [0, 1]$.



- **Problem:** Length scale ℓ does not fit to the data.

Regression – III

- **Solution:** Optimize for length scale ℓ. ($\rightarrow \ell = 0.109$)



Time for the tutorial!

More **details** on GPR can be found in
C. Williams and C. Rasmussen. *Gaussian Processes for Machine Learning*. MIT Press. 2006.
(`http://www.gaussianprocess.org/gpml/chapters/RW.pdf`).