Forecasting new COVID-19 hospitalizations in Belgium using Gaussian Processes Regression

Project description

The aim of this project is to perform a Gaussian process (GP) regression on a simple toy example and on the prediction of the number of new hospital admissions in Belgium due to COVID-19. In particular, the goal is to assess the performance of Gaussian processes, which are known to be extremely flexible machine learning models, to predict future new hospital admissions.

A Gaussian process is a generalization of a multivariate Gaussian distribution to *infinite* dimensions. It essentially defines a probability measure on a function space. When we say that $f(\cdot)$ is a GP, we mean that it is a random variable that is actually a function. Mathematically, we write:

$$f(\cdot) \sim \text{GP}\left(m(\cdot), k(\cdot,\cdot)\right), \tag{1}$$

where $m : \mathbb{R}^d \to \mathbb{R}$ is the *mean function* and $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the *covariance function*. So, compared to a multivariate normal we have:

- A random function $f(\cdot)$ instead of a random vector $\mathbf{x}$.
- A mean function $m(\cdot)$ instead of a mean vector $\boldsymbol{\mu}$.
- A covariance function $k(\cdot,\cdot)$ instead of a covariance matrix $\boldsymbol{\Sigma}$.

More specifically, let $\mathbf{x}_{1:n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be n points in $\mathbb{R}^d$ and $\mathbf{f} \in \mathbb{R}^n$ be the outputs of $f(\cdot)$ on each one of the elements of $\mathbf{x}_{1:n}$, i.e.,

$$\mathbf{f} = \begin{pmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \end{pmatrix}.$$

The fact that $f(\cdot)$ is a GP with mean and covariance function $m(\cdot)$ and $k(\cdot,\cdot)$, respectively, means that the vector of outputs $\mathbf{f}$ at the arbitrary inputs in $\mathbf{X}$ is the following multivariate-normal:

$$\mathbf{f}|\mathbf{x}_{1:n}, m(\cdot), k(\cdot,\cdot) \sim \mathcal{N}\left(\mathbf{m}(\mathbf{x}_{1:n}), \mathbf{K}(\mathbf{x}_{1:n}, \mathbf{x}_{1:n})\right),$$

with mean vector:

$$\mathbf{m}(\mathbf{x}_{1:n}) = \begin{pmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_n) \end{pmatrix},$$

and covariance matrix:

$$\mathbf{K}(\mathbf{x}_{1:n}, \mathbf{x}_{1:n}) = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \ldots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \ldots & k(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix}.$$

A more detailed description of GP can be found [1].

Tasks

In this project, you will complete the following tasks:

Task 1: Build a Gaussian process regression in which the observed data is generated by a deterministic process but the observations are stochasticaly perturbed by random noise. More specifically, consider the following toy example with a gaussian heteroscedastic noise:

$$g(x) = \log(10(|x - 0.03| + 0.03))\ \sin(\pi(|x - 0.03| + 0.03)) + \delta(x) \tag{2}$$

where $\delta(x) = \frac{\epsilon}{\exp(2(|x-0.03|+0.03))}$, and $\epsilon \sim N(0, \sigma^2)$.

Here, assume access to N training pairs $\{x_i, g(x_i)\}, i = 1, ..., N$ randomly sampled in the interval $x \in [-2, 2]$.

The steps to solve the first exercise are:

- Build a GP regression using a Radial Basis Function (RBF) kernel to predict the $g(x)$ using the noisy x measurements as input. Train the model by gradually increasing the number of samples from $N = 10$ to $N = 100$. Note that as the number of samples increases it doesn't really matter what your prior knowledge of the data. Describe how the number of train points affects the epistemic uncertainty of the model.
- Experiment with different measurement noises, for instance $\sigma^2 = [0.1, 0.5, 1.0]$. What do you observe?
- Experiment with difference likelihood variances. What happens for very big variances? What happens for very small variances?
- Experiment with different kernel variances. This the s^2 parameter of the squared exponential covariance function. It specifies our prior variance about the function values. What is its effect?
- Try some other kernels. How can you pick the right kernel? How would you pick the correct values for the hyperparameters specifying the kernel?
- Finally, evaluate if the model constructed is good. Create a validation dataset and compute some accuracy measurements, such as mean squared error (MSE) and the coefficient of determination (R2-score). Also, provide some statistical diagnostics comparing the predictive distribution to the distribution of the validation dataset.

Note: For this specific task, there is a Jupyter notebook provided in the GitHub repository, which will help you start up your work.

Task 2: COVID-19 affects different people in different ways. Most infected people will develop mild to moderate illness and recover without hospitalization. However, those who need to be hospitalized often need large hospital resources. That is the case for severely ill patients and those who are at risk of severe disease in which more advanced respiratory support is needed, such as oxygen and ventilation. Therefore, hospitals must construct models able to correctly predict possible future hospital admissions in order to have the necessary hospital resources available to save lives.

Machine learning techniques have been used to build predictive models that support decision-making in the health area. Build a GP regression model to predict future daily hospital admissions in Belgium due to COVID-19. The dataset can be found in the GitHub repository. More specifically, the dataset consists of hospital admissions from 15 March 2020 to 04 April 2022. An overview of the dataset can be seen in Fig. 1. More details about the dataset you can found here. Note that to construct the GP regression model you can use the kernel function that you prefer. However, the choice must be justified.

Now consider that you only have information only one day of the week, how does that affect the predictability of the model? And just one day a month? Also, if now you add

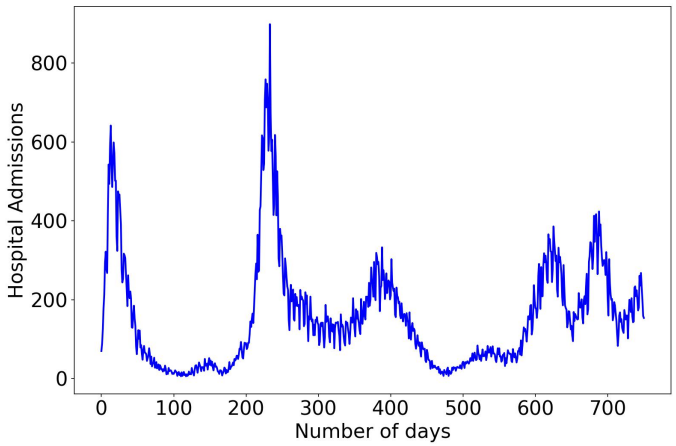

Figure 1: Hospital admissions due COVID-19 in Belgium.

some noise to the dataset. How does this affect the model? Finally, compare the results provided by the GP with the real data provide after 4 April 2022, which can be found here.

Written report

Your written report should contain the following:

Section 1: A high-level introduction of the mathematical formulation of the GP regression. In particular, describe the interpretation of the mean function and the interpretation of the covariance function. Describe the main properties of the covariance function and how to encode prior beliefs in the covariance function.

Section 2: Results corresponding to **Tasks 1**, compared to the analytical solution. The result includes relevant figures and a detailed discussion of the results.

Section 3: Results corresponding to **Tasks 2**, including relevant figures and a detailed discussion of the results.

Section 4: Final discussion and conclusions, discuss the main findings of using GP for regression tasks. Don't forget to cite the relevant literature to strengthen your conclusions.

Resources

GPy is a Gaussian Process (GP) framework written in Python. It includes support for basic GP regression, multiple output GPs (using coregionalization), various noise models, sparse GPs, non-parametric regression and latent variables.

You can find a more in-depth discussion of GP in Carl Edward Rasmussen and Chris Williams's textbook. The book deals with the supervised-learning problem for both regression and classification and includes detailed algorithms. The book is available for download in electronic format here.

In case of any doubts regarding this assignment please contact Rodolfo Freitas on TEAMS or at rodolfo.da.silva.machado.de.freit@ulb.be.

References

[1] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.