M156 – Lab assignment #6: Nonlinear regression using Gaussian processes

The objective of this lab is to use Gaussian processes to perform nonlinear regression on a real dataset: the target values are real, and the training point are real (one dimensional) as well.

1 Theory

1. Recall the definition of a Gaussian process.

2. We are going to assume that the function $y(x)$ is a Gaussian process. Then, we are assuming the following regression model:
$$y(x) = \mathbf{w}^\top \boldsymbol{\phi}(x)$$
where $\boldsymbol{\phi}(x) = [\phi_0(x), \phi_1(x), ..., \phi_{M-1}(x)]^\top \in \mathbb{R}^M$, with $\phi_0 \equiv 1$.

   For N training data points $\{t_n, x_n\}$, we assume that $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M)$. Derive the expression of the pdf of $\mathbf{y} = [y_1, y_2, ..., y_n]^\top = \boldsymbol{\Phi}\mathbf{w}$, where $(\boldsymbol{\Phi})_{ij} = \phi_j(x_i)$ (we already know $\mathbf{y}$ is going to be Gaussian, as a linear combination of Gaussian random variables).

3. For any test point x_{N+1}, recall (no proof is required) the expression of the predictive distribution $p(t_{N+1}|\mathbf{t})$, where $\mathbf{t} = [t_1, t_2, ..., t_3]^\top$. Bear in mind that this distribution is implicitly conditioned on the values $x_1, x_2, ..., x_N$.

2 Experimental Part

The dataset we are using (`temperature_data.mat`) is a time series of average temperature measurements (variable `t`) over the planet. There is one data point for each year between 1880 and 2016 (variable `x`). Those average temperatures are expressed as the difference between the temperatures that were actually measured and the mean of those values between 1951 and 1980 (in degrees Celsius). The dataset was taken from NASA's website.

The goal of this lab is to perform regression analysis on the data and to try to extrapolate it to make predictions of the average temperatures on Earth for the years to come.

We are going to assume we want to recover a "trend" in the data, and will consider the distance of any point to this trend as noise. Then our regression model is:

$$t_n = y(x_n) + \epsilon_n$$

where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise.

1. Fit a linear regression model $y(x) = wx + b$ to the dataset (that is, without using nonlinear basis functions). Plot the data and the obtained regression on the same graph. Is this model appropriate for regression on this data? Why?

2. Extrapolate the data values until year 2100. Do you think the obtained values from year 2017 on are accurate predictions (in terms of the general trend you can visually see in the data)?

3. We want to perform a nonlinear regression on this dataset. We are going to use Gaussian Processes (GP) with the Gaussian kernel:
$$k_{\tau^2}(x, x') = \exp\left(-\frac{(x - x')^2}{2\tau^2}\right)$$

It is strongly recommended to code a function computing the kernel values as a function of x,x', and τ^2, since it is going to be extensively used.

4. One parameter we need to know to apply GP regression is the variance of the noise σ^2. We are going to use the empirical value of $\sigma^2 = 0.01$. For now, set the value of the width of the Gaussian kernel to $\tau^2 = 10$. Sample a few realizations of the corresponding GP prior, for values of $x \in [1880, 2100]$, with 1000 points in this interval.

5. Compute, for values of $x \in [1880, 2100]$, with 1000 points in this interval, the mean and variance of the predictive distribution. Plot them on the same graph as the data. You can represent the variance of each point by plotting the mean plus/minus two standard deviations. Do you think the obtained distribution provides a good estimate of the trend in the data, as well as predicted values for years to come?

6. We also need to tune the width of the Gaussian kernel τ^2. What is the role of this parameter on the GP (show this with plots of some realizations of the GP) and on the mean of the predictive distribution? How do the extrapolated (predicted) values behave when τ^2 is decreased/increased? To help you, comment on the previous points, using the values $\tau^2 = 10, 1000, 10000, 100000$.

7. Explain theoretically why it is going to be hard to extrapolate the data for values of $|x - x'| \geq \tau$. Then tune τ^2 accordingly to obtain a sufficiently good recovery of the trend of the data, and to get predicted values until year 2100 which make sense. Once you have settled for a value, comment on the uncertainty on the predictions, depending on the values of x.

8. Conclude: Which value of τ^2 would you use to extrapolate the data until year 2200? Year 2100? Which one would you use to capture more localized changes in the data (i.e. changes occuring over a few years)?