Introduction to Gaussian Processes

Miguel López-Pérez

University of Granada

November 22, 2019

Overview

1. Gaussian distribution
2. Gaussian Processes
3. Bayesian inference
4. Gaussian Processes for regression
5. Conclusions

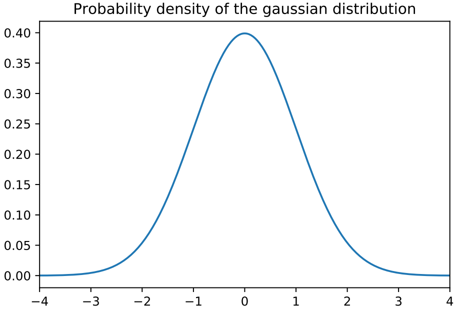
Section 1

Gaussian distribution

Univariate Gaussian distribution

- It has several good properties: easy computations, central limit theorem, ... It will be the central tool of the gaussian processes.
- Knowing the parameters for the mean μ and the variance σ^2, the point density function is given by

$$\mathrm{p}(f|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{f-\mu}{\sigma}\right)^2} \tag{1}$$



Multivariate Gaussian distribution

- Let $\mathbf{f}$ be multivariate, i.e. $\mathbf{f} = (f_1, \ldots, f_n)^T$, we can extend the notion of gaussian distribution.
- Knowing the vector of means $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ the point density function is given by:

$$\mathrm{p}(\mathbf{f}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{f} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{f} - \boldsymbol{\mu})\right) \qquad (2)$$

- Note: $\boldsymbol{\Sigma}$ must be a (semi)definite positive matrix.

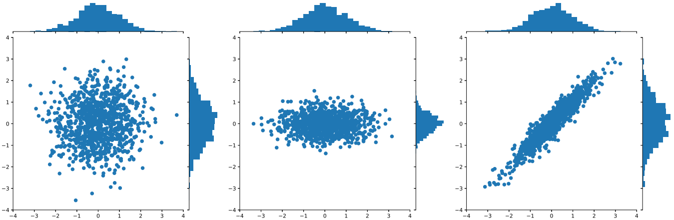
Some nice properties of the multivariate gaussian distribution

- The marginals are also gaussian distributed, i.e, $\mathrm{p}(f_i), \mathrm{p}(f_i, f_j), \ldots$ are gaussian.
- The conditional distributions are also gaussian distributed, i.e. $\mathrm{p}(f_i|f_j), \mathrm{p}(f_i, f_j|f_k), \mathrm{p}(f_i|f_j, f_k), \ldots$ are gaussian.

Example of multivariate gaussian distribution in $\mathbb{R}^2$

- $\mathbf{f} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The mean in all examples is $\boldsymbol{\mu} = (0,0)^T$ while the covariance matrix changes:

$$\underset{(a)}{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}} \quad \underset{(b)}{\begin{pmatrix} 1 & 0 \\ 0 & 0.2 \end{pmatrix}} \quad \underset{(c)}{\begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}}$$



(a) (b) (c)

Figure: 1k samples from multivariate gaussian distributions.

Example of multivariate gaussian distribution in $\mathbb{R}^3$

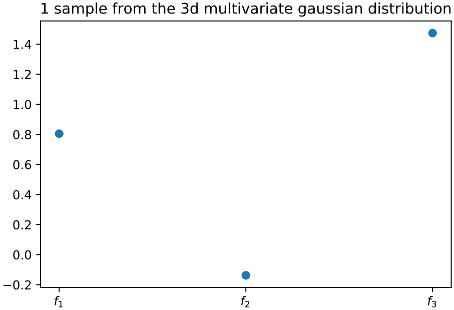
- The 3-dimensional multivariate gaussian distribution is more difficult to observe. So we are going to plot the samples in an axis.
- $\mathbf{f} = (f_1, f_2, f_3) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. With mean $\boldsymbol{\mu} = (0, 0, 0)^T$ and covariance matrix:

$$\begin{pmatrix} 1 & 0.1 & 0.9 \\ 0.1 & 1 & 0.1 \\ 0.9 & 0.1 & 1 \end{pmatrix}$$

Example of multivariate gaussian distribution in $\mathbb{R}^3$

- The 3-dimensional multivariate gaussian distribution is more difficult to observe. So we are going to plot the samples in an axis.
- $\mathbf{f} = (f_1, f_2, f_3) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. With mean $\boldsymbol{\mu} = (0, 0, 0)^T$ and covariance matrix:

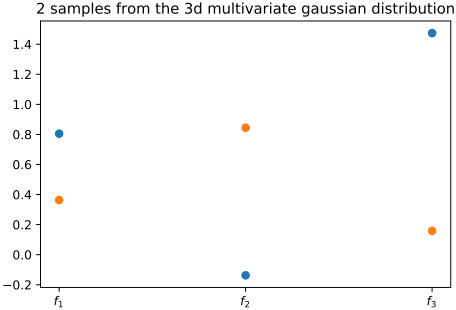
$$\begin{pmatrix} 1 & 0.1 & 0.9 \\ 0.1 & 1 & 0.1 \\ 0.9 & 0.1 & 1 \end{pmatrix}$$



Example of multivariate gaussian distribution in $\mathbb{R}^3$

- The 3-dimensional multivariate gaussian distribution is more difficult to observe. So we are going to plot the samples in an axis.
- $\mathbf{f} = (f_1, f_2, f_3) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. With mean $\boldsymbol{\mu} = (0, 0, 0)^T$ and covariance matrix:

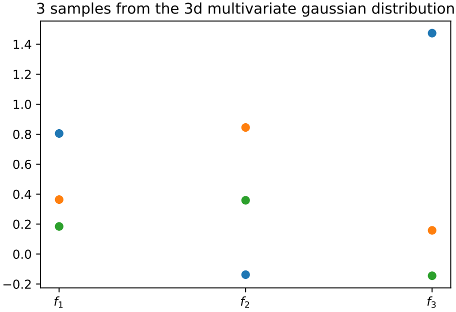
$$\begin{pmatrix} 1 & 0.1 & 0.9 \\ 0.1 & 1 & 0.1 \\ 0.9 & 0.1 & 1 \end{pmatrix}$$



Example of multivariate gaussian distribution in $\mathbb{R}^3$

- The 3-dimensional multivariate gaussian distribution is more difficult to observe. So we are going to plot the samples in an axis.
- $\mathbf{f} = (f_1, f_2, f_3) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. With mean $\boldsymbol{\mu} = (0, 0, 0)^T$ and covariance matrix:

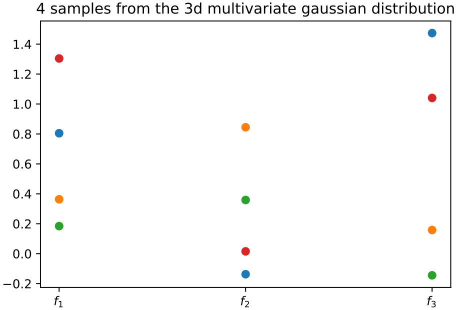
$$\begin{pmatrix} 1 & 0.1 & 0.9 \\ 0.1 & 1 & 0.1 \\ 0.9 & 0.1 & 1 \end{pmatrix}$$



Example of multivariate gaussian distribution in $\mathbb{R}^3$

- The 3-dimensional multivariate gaussian distribution is more difficult to observe. So we are going to plot the samples in an axis.
- $\mathbf{f} = (f_1, f_2, f_3) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. With mean $\boldsymbol{\mu} = (0, 0, 0)^T$ and covariance matrix:

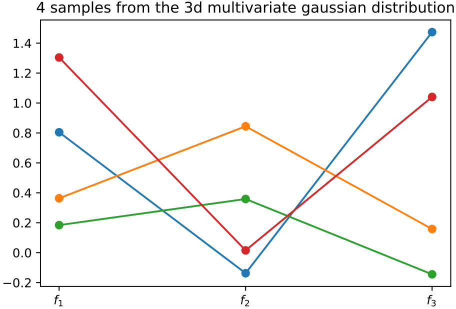
$$\begin{pmatrix} 1 & 0.1 & 0.9 \\ 0.1 & 1 & 0.1 \\ 0.9 & 0.1 & 1 \end{pmatrix}$$



Example of multivariate gaussian distribution in $\mathbb{R}^3$

- The 3-dimensional multivariate gaussian distribution is more difficult to observe. So we are going to plot the samples in an axis.
- $\mathbf{f} = (f_1, f_2, f_3) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. With mean $\boldsymbol{\mu} = (0, 0, 0)^T$ and covariance matrix:

$$\begin{pmatrix} 1 & 0.1 & 0.9 \\ 0.1 & 1 & 0.1 \\ 0.9 & 0.1 & 1 \end{pmatrix}$$



Section 2

Gaussian Processes

Gaussian Process definition

Definition

A **Gaussian Process** is a collection of random variables such that every finite collection of those random variables has a multivariate normal distribution. A **Gaussian process** is fully specified by a mean function $\boldsymbol{\mu}(\cdot)$ and kernel (covariance) function $k(\cdot, \cdot)$

- We say that

$$\mathbf{f} \sim GP\left(m(\cdot), k(\cdot, \cdot)\right).$$

  So $\mathbf{f} = \{f(\mathbf{x}) : x \in \mathcal{X}\}$ and for every finite combination of indexes $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_n \Rightarrow f(\mathbf{x}_1) \ldots, f(\mathbf{x}_n) \sim \mathcal{N}(m(\mathbf{X}), k(\mathbf{X}, \mathbf{X}))$.
- When the set of indexes $\mathcal{X}$ is finite, it is a multivariate gaussian distribution. But it is interesting the case of $\mathcal{X}$ being infinite, e.g., one continue subset of $\mathbb{R}^d$.
- We can also see the gaussian process as a distribution over functions.

Gaussian Process example

Let $\mathcal{X} = [0, 1]$ be the space of indexes and $\mathbf{f} \sim GP((m(\cdot), k(\cdot, \cdot))$. We define the following mean and kernel (covariance) functions:

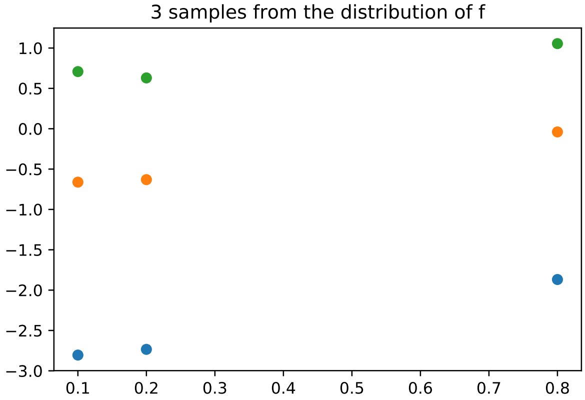
$$m(\mathbf{x}) = \mathbf{0}, \quad k(\mathbf{x}, \mathbf{x}') = \exp\left(||\mathbf{x} - \mathbf{x}'||^2\right). \tag{3}$$

If we choose any finite set of indexes we will obtain a multivariate gaussian distribution, e.g., we have the following sample $\mathbf{X} = \{0.1, 0.2, 0.8\}$.
The resulting gaussian distribution is $\mathbf{f}(\mathbf{X}) \sim \mathcal{N}(\mu, \mathbf{\Sigma})$:

$$\mu = 0, \quad \mathbf{\Sigma} = \begin{pmatrix} 1 & 0.9900 & 0.6126 \\ 0.9900 & 1 & 0.6976 \\ 0.6126 & 0.6976 & 1 \end{pmatrix} \tag{4}$$

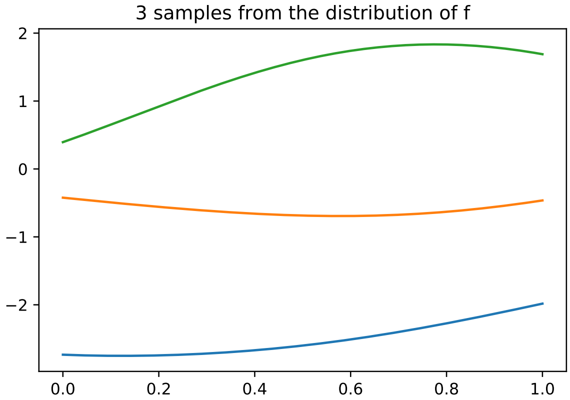
Gaussian Process example

We can take several sample from this normal:



Gaussian Process example

We can also increase the number of indexes used leading to **functions**:



We are sampling **functions** and all of them seem to have similar properties.

How to define a Gaussian Process?

- As it is said before a GP is completely defined by its mean and kernel (covariance) functions.
- Usually, the mean function is fixed to zero: $m(\mathbf{x}) = 0$ without losing generality.
- The main issue will be how to define the kernel function $k(\cdot, \cdot)$. This kernel function will define the desirable properties of the functions. **Notice:** $k(\cdot, \cdot)$ must define a semidefinite positive matrix.

Example of kernel functions: RBF

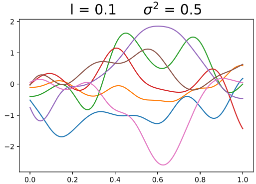
The Radial Basis Function (RBF) is the most used because it has a great power of represention.

Radial Basis Function kernel (RBF)

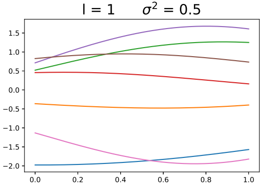
$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(\frac{-||\mathbf{x} - \mathbf{x}'||^2}{2l^2}\right)$$

It has the hyperparameters l and σ^2 that control the properties of the function.

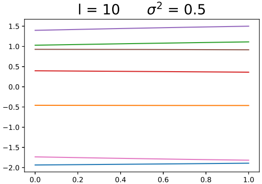
Example of kernel functions: RBF



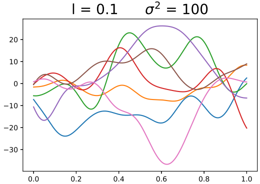
(a)



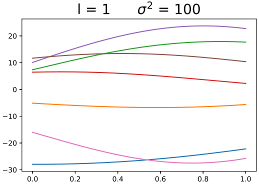
(b)



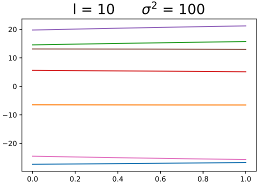
(c)



(d)



(e)



(f)

Example of kernel functions: RBF

- It is called a **stationary** kernel because it depends of the distance of two points, i.e., $||x' - x||$.
- It is clear that σ controls the **amplitude** of the values of f (look at the values of the y-axis!!).
- As we could see the RBF kernel imposes **smoothness** per se. We can control the amount of smoothness tuning the parameter l. The higher the parameter the higher the smoothness.
- This property of **smoothness** is desirable in many scenarios, in addition, it is very **flexible** and it has a great power of **representation** which leads to be the most used.

Example of kernel functions: Matern

The Matern functions are a family of kernels:

Matern12

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

Matern32

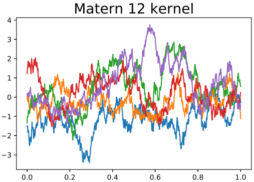
$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \left(1 + \sqrt{3}\frac{||\mathbf{x} - \mathbf{x}'||}{2l}\right) \exp\left(-\sqrt{3}\frac{||\mathbf{x} - \mathbf{x}'||}{2l}\right)$$

Matern52

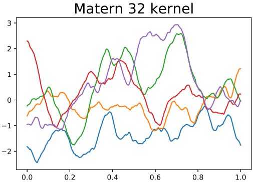
$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \left(1 + \sqrt{5}\frac{||\mathbf{x} - \mathbf{x}'||}{2l} + \frac{5}{3}\frac{||\mathbf{x} - \mathbf{x}'||^2}{2l^2}\right) \exp\left(-\sqrt{5}\frac{||\mathbf{x} - \mathbf{x}'||}{2l}\right)$$

It is also a stationary kernel and it is controlled by the variance σ^2 and lengthscale l parameters.

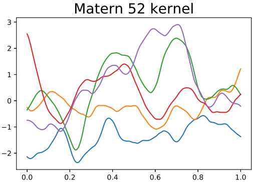
Example of kernel functions: Matern



(a)



(b)



(c)

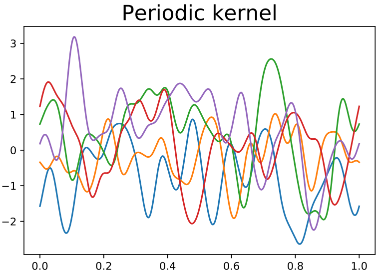
Example of kernel functions: Periodic

The periodic kernel is used for periodic data:

Periodic

$$k(x, x') = \sigma^2 \exp\left(-\frac{\sin(\pi \| x - x' \|^2 / p)}{l^2}\right)$$

It is also a stationary kernel and it has three parameters: variance σ^2 for the amplitude, lengthscale l for the smoothness and phase p for the periodicity parameters.



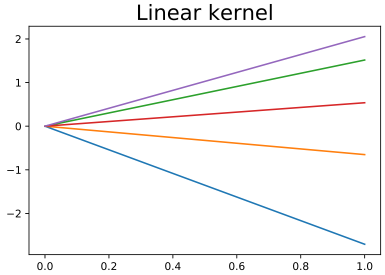
Example of kernel functions: Linear

The linear kernel is used for linear data:

Linear

$$k(x, x') = \sigma^2 x \cdot x'$$

It has the parameter σ^2 which controls the slope of the lines.



Example of kernel functions: White noise

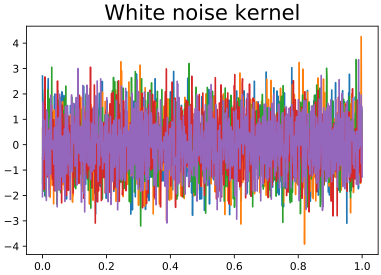
The white noise kernel is used for gaussian noisy data:

White noise

$$k(x, x') = \sigma^2 \delta_{xx'}$$

where $\delta_{xx'}$ is the kronecker delta.

All the points are independent and the σ^2 parameter control the amplitude of this noise.

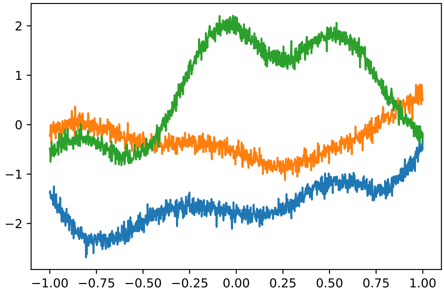


Combining kernels: Sum

We can also combine kernels by summing them. Look that it also defines a semidefinite positive matrix!

$$k(x, x') = k_1(x, x') + k_2(x, x')$$

It acts like a OR operator if one of them is high it will high:

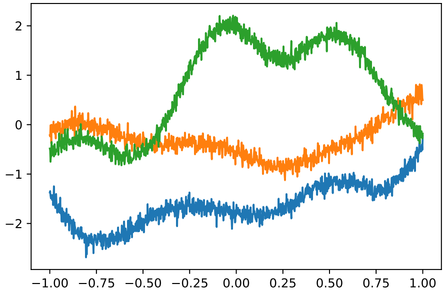


Combining kernels: Sum

We can also combine kernels by summing them. Look that it also defines a semidefinite positive matrix!

$$k(x, x') = k_1(x, x') + k_2(x, x')$$

It acts like a OR operator if one of them is high it will high:



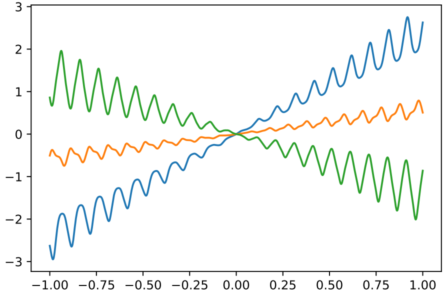
RBF kernel + White noise kernel

Combining kernels: Product

We can also combine kernels by summing them. Look that it also defines a semidefinite positive matrix!

$$k(x, x') = k_1(x, x') \times k_2(x, x')$$

It acts like an AND operator both of them must be high for high values:

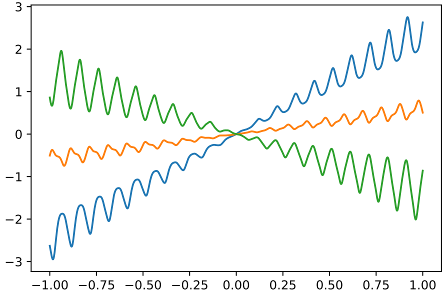


Combining kernels: Product

We can also combine kernels by summing them. Look that it also defines a semidefinite positive matrix!

$$k(x, x') = k_1(x, x') \times k_2(x, x')$$

It acts like an AND operator both of them must be high for high values:



Linear kernel x Periodic kernel

Section 3

Bayesian inference

Problem to solve

- We have observed the following data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.
- Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$
- We model the regression with an unknown function corrupted by gaussian noise:
$$\mathbf{y} = \mathbf{f}(\mathbf{X}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \tag{5}$$
- Once we learn this function we can infer the distribution on unseen data:
$$\mathrm{p}(\mathbf{y}_* | \mathbf{y}, \mathbf{X}, \mathbf{X}_*) = \int \mathrm{p}(\mathbf{y}_* | \mathbf{f}_*) \mathrm{p}(\mathbf{f}_* | \mathbf{f}) \mathrm{p}(\mathbf{f} | \mathbf{y}) d\mathbf{f} d\mathbf{f}_* \tag{6}$$
- So we want to calculate the distribution of $\mathrm{p}(\mathbf{f}|\mathbf{y})$ and then the distribution of $\mathrm{p}(\mathbf{y}_*|\mathbf{y}, \mathbf{X}, \mathbf{X}_*)$. For calculating these posterior distributions we use the Bayes's Rule.

Bayes's Rule explained

Bayes's Rule

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})} \tag{7}$$

- $p(\mathbf{f}|\mathbf{y})$ is the posterior distribution. That is the distribution of $\mathbf{f}$ knowing that we have observed $\mathbf{y}$.
- $p(\mathbf{y}|\mathbf{f})$ is the likelihood. How probable is the seen data for a value of the latent function $\mathbf{f}$.
- $p(\mathbf{f})$ is the prior distribution. It is the distribution of $\mathbf{f}$ before we have seen anything. This distribution imposes prior knowledge or properties to the desired posterior distribution. It acts like a regularizer.
- $p(\mathbf{y})$ is the evidence. This is how probable is our observation.

Section 4

Gaussian Processes for regression

Regression problem with GP prior

- We have the observed the following data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.
- Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$
- We model the regression with an unknown function corrupted by gaussian noise:

$$\mathbf{y} = \mathbf{f}(\mathbf{X}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \tag{8}$$

- We can impose that the latent function $\mathbf{f}$ follows a GP prior, i.e., $\mathbf{f} \sim GP(0, k(\cdot, \cdot))$.
- The joint distribution is:

$$\mathrm{p}(\mathbf{y}, \mathbf{f}) = \underbrace{\mathrm{p}(\mathbf{y}|\mathbf{f})}_{likelihood} \underbrace{\mathrm{p}(\mathbf{f}|\mathbf{X})}_{GPprior} \tag{9}$$

  - Likelihood Gaussian: $p(\mathbf{y}|\mathbf{f}) \sim \mathcal{N}\left(\mathbf{f}, \sigma^2 \mathbf{I}\right)$
  - Gaussian prior: $\mathrm{p}(\mathbf{f}|\mathbf{X}) \sim \mathcal{N}(0, K(\mathbf{X}, \mathbf{X}))$

Noise-free predictions

- We have observed the following noise-free data $(\mathbf{x}_1, f_1), \ldots, (\mathbf{x}_n, f_n)$.
- If we have unseen values $\mathbf{X}_*$, which are the values of the latent function $\mathbf{f}_*$?
- We know that $\mathbf{f}$ and $\mathbf{f}_*$ follows jointly the following gaussian distribution because of the GP prior:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right) \tag{10}$$

- Using the rules of conditioning in a gaussian multivariate distribution we can calculate the posterior distribution:

$$\begin{aligned} &\mathrm{p}(\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{f}) \sim \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \\ &\boldsymbol{\mu}_* = K(\mathbf{X}_*, \mathbf{X}) K(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f} \\ &\boldsymbol{\Sigma}_* = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X}) K(\mathbf{X}, \mathbf{X})^{-1} K(\mathbf{X}, \mathbf{X}_*) \end{aligned} \tag{11}$$

Noisy predictions

- We have the observed following noisy data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.
- If we have unseen values $\mathbf{X}_*$, which are the values of the latent function $\mathbf{y}_*$?
- Note that $\mathrm{p}(\mathbf{y}|\mathbf{X}) = \int \mathrm{p}(\mathbf{y}|\mathbf{f})\mathrm{p}(\mathbf{f}|\mathbf{X})d\mathbf{f} = \mathcal{N}(0, K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I})$.
- We know that $\mathbf{y}$ and $\mathbf{f}_*$ follows jointly the following gaussian distribution because of the GP prior:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I} & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) + \sigma^2\mathbf{I} \end{bmatrix}\right). \tag{12}$$

- Using the rules of conditioning in a gaussian multivariate distribution we can calculate the posterior distribution:

$$\begin{aligned} &\mathrm{p}(\mathbf{y}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{y}) \sim \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \\ &\boldsymbol{\mu}_* = K(\mathbf{X}_*, \mathbf{X})\left(K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y} \\ &\boldsymbol{\Sigma}_* = K(\mathbf{X}_*, \mathbf{X}_*) + \sigma^2\mathbf{I} - K(\mathbf{X}_*, \mathbf{X})\left(K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I}\right)^{-1}K(\mathbf{X}, \mathbf{X}_*) \end{aligned} \tag{13}$$

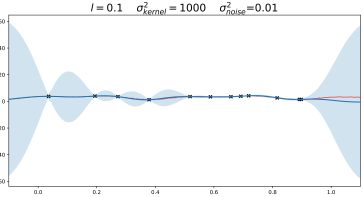
Marginal likelihood

- We want to compute marginal likelihood of the model, i.e., how probable is the observation of the model given the data.
- The **marginal likelihood** of the model is given by:

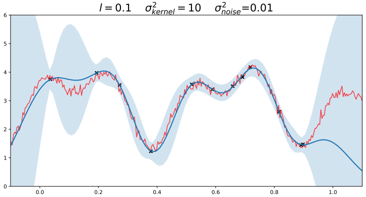
$$
\begin{aligned}
\log p(\mathbf{y}|\mathbf{X}) &= \log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\right) && (14)\\
&= -\frac{1}{2}\mathbf{y}^T\left(K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y} && (15)\\
&\quad -\frac{1}{2}\log\left|K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\right| && (16)\\
&\quad -\frac{n}{2}\log(2\pi) && (17)
\end{aligned}
$$

- The parameters of the kernel are computed by maximizing the marginal likelihood. Notice that $K(\mathbf{X},\mathbf{X})$ depends on the chosen kernel and its hyperparameters.

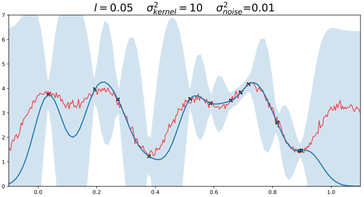
Example of the RBF kernel for regression



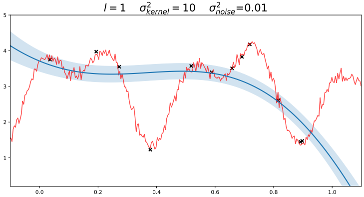
(a)



(b)

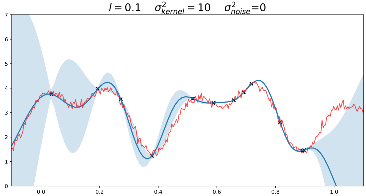


(c)

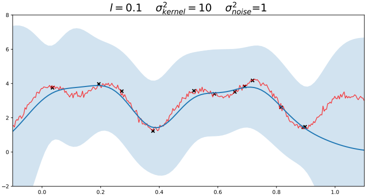


(d)

Example of the RBF kernel for regression

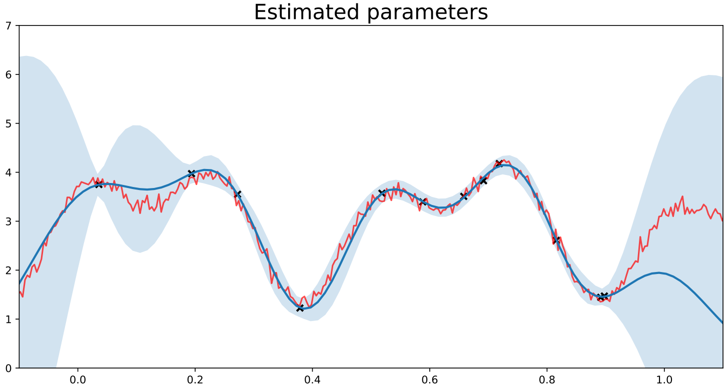


(e)



(f)

Example of the RBF kernel for regression



Section 5

Conclusions

Conclusions

- Gaussian processes are amazing. We were doing bayesian linear regression with infinite basis functions!!
- Gaussian processes are useful when:
  - little data is provided.
  - we know prior information about data.
  - we desire uncertainty in the predictions.
- Main drawbacks:
  - Scalability. It is $\mathcal{O}(n^3)$. This is solved by using Sparse Gaussian Processes.
  - Inference. Although inference is easy in the regression case is more difficult with non-gaussian likelihood, e.g., in classification. The state of the art is the variational inference and the MCMC.
  - Engineering of the kernel. Deep Gaussian Processes offer much more complex model without engineering complex kernels.

Useful Resources

David Duvenaud. *Kernel Cookbook*.
https://www.cs.toronto.edu/~duvenaud/cookbook/

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for machine learning*. The MIT Press. 2006.
http://www.gaussianprocess.org/gpml/.

GPflow: Gaussian processes in TensorFlow.
https://gpflow.readthedocs.io/en/latest/index.html