Gaussian Processes Demystified

PGO

Ivan Pogrebnyak

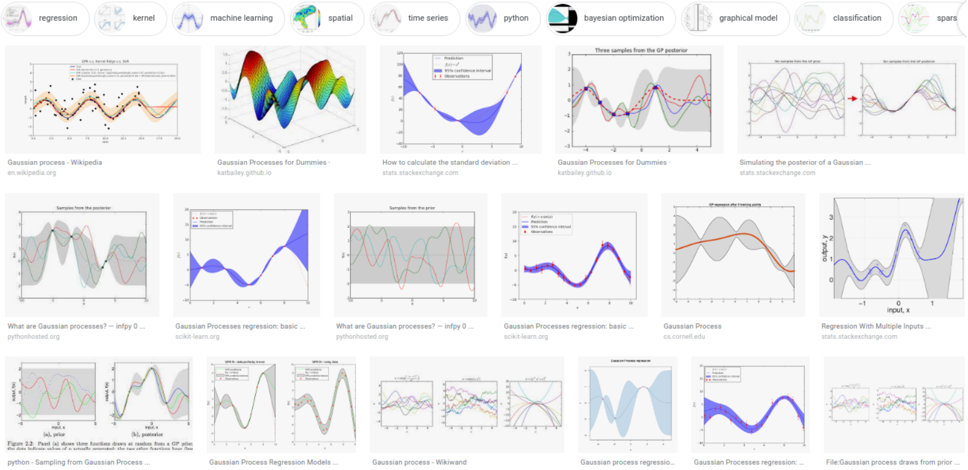
Michigan State University

February 28, 2019

Introduction

- You've all probably heard of them.
- It's a hot topic in data science, along with neural nets and machine learning.
- You might have heard that they provide a non-parametric function model.
  - Or just that they depend on hyperparameters.
- Or that they can fit an arbitrery function.
- Or seen figures like these:

Google image search for “Gaussian process”



Gaussian process - Wikipedia
en.wikipedia.org

Gaussian Processes for Dummies ·
katbailey.github.io

How to calculate the standard deviation ...
stats.stackexchange.com

Gaussian Processes for Dummies ·
katbailey.github.io

Simulating the posterior of a Gaussian ...
stats.stackexchange.com

What are Gaussian processes? — infpy 0 ...
pythonhosted.org

Gaussian Processes regression: basic ...
scikit-learn.org

What are Gaussian processes? — infpy 0 ...
pythonhosted.org

Gaussian Processes regression: basic ...
scikit-learn.org

Gaussian Process
cs.cornell.edu

Regression With Multiple Inputs ...
stats.stackexchange.com

python - Sampling from Gaussian Process ...

Gaussian Process Regression Models ...

Gaussian process - Wikiwand

Gaussian process regressio...

Gaussian Processes regression: ...

File:Gaussian process draws from prior ...

Definition

- One of the most commonly seen defenitions of a GP is this:

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

- This is like saying that a fork is something that splits at one end. An accurate, yet uninspiring statement.

- We'll return to the definition later.

The first step

- Assume we made an observation of a vector $\mathbf{d} = [y_1, y_2, \ldots, y_n]$, and that $\mathbf{d}$ is sampled from a multivariate Gaussian distribution, i.e. $\mathbf{d} \sim \mathcal{N}(\mathbf{0}, D)$.
- Assuming non-zero mean is no more general, because the mean can be absorbed into the defenition of $\mathbf{d}$.
- If we arbitrarily split $\mathbf{d}$ into subvectors $\mathbf{a}$ and $\mathbf{b}$, then we can write

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} A & C \\ C^\mathsf{T} & B \end{bmatrix}\right) \tag{1}$$

- The conditional probability of $\mathbf{b}$ given $\mathbf{a}$ is

$$p(\mathbf{b}|\mathbf{a}) = \frac{p(\mathbf{a} \cap \mathbf{b})}{p(\mathbf{a})} = \frac{p(\mathbf{d})}{\int p(\mathbf{d}) p(\mathbf{b}') \mathrm{d}\mathbf{b}'} = \boxed{\mathcal{N}\left(C^\mathsf{T} A^{-1} \mathbf{a},\, B - C^\mathsf{T} A^{-1} C\right)} \tag{2}$$

- Proof: `https://stats.stackexchange.com/q/30588/239215`

Bayesian Inference

- Recall the Bayes' theorem

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}, \quad p(\mathbf{b}|\mathbf{a}) = \frac{p(\mathbf{a}|\mathbf{b})}{p(\mathbf{a})}\, p(\mathbf{b}). \qquad (3)$$

- $p(\mathbf{b}) = \mathcal{N}\left(\mathbf{0}, B\right)$ can be viewed as the prior, and $p(\mathbf{b}|\mathbf{a}) = \mathcal{N}\left(C^\mathsf{T} A^{-1}\mathbf{a},\, B - C^\mathsf{T} A^{-1} C\right)$ as the posterior.
  * Conditioning
- GP is defined by the multivariate Gaussian distribution in Eq. 1.
- Clearly, we didn't have to have observed the $\mathbf{b}$ part of the vector to make this inference.
- In GP regression, inference is made about unobserved function values.
  - Think of a function as a (generally continuous infinite-dimensional) vector, whose elements are labeled by the coordinates on the manifold on which the function lives.
- Non-zero convariance, C, contains the additional information. Otherwise, $p(\mathbf{b}) = p(\mathbf{b}|\mathbf{a})$.

GP Regression

- Given: observed function values, y_i, at points x_i.
- Model: prior distribution of the function values, given my the mean m and covariance matrix K. This is the GP.

$$f(x) \sim \mathcal{N}\left(m(x), K(x, x')\right) \tag{4}$$

- Take $m = 0$ for brevity
- For observed values y, and unobserved values y_*, we can write

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(x,x) & K(x,x_*) \\ K(x_*,x) & K(x_*,x_*) \end{bmatrix}\right), \text{ or} \tag{5}$$

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_* \\ K_*^\mathsf{T} & K_{**} \end{bmatrix}\right). \tag{6}$$

- Then,

$$\boxed{y_*|y \sim \mathcal{N}\left(K_*^\mathsf{T} K^{-1} y,\, K_{**} - K_*^\mathsf{T} K^{-1} K_*\right)} \tag{7}$$

GP Regression

$$\boxed{y_*|y \sim \mathcal{N}\left(K_*^\mathsf{T} K^{-1} y,\, K_{**} - K_*^\mathsf{T} K^{-1} K_*\right)} \tag{8}$$

- Gives the best linear unbiased prediction.
- Direct measure of uncertainty at each point of the function.
- Function value predictions are independent.
  - Can compute point-by-point. * Lazy learning
- Also called Kriging.
  - The theoretical basis for the method was developed by the French mathematician Georges Matheron in 1960, based on the Master's thesis of Danie G. Krige, the pioneering plotter of distance-weighted average gold grades at the Witwatersrand reef complex in South Africa. [Wikipedia]

GP Regression Algorithm

$$\boxed{y_*|y \sim \mathcal{N}\left(K_*^\mathsf{T} K^{-1} y,\, K_{**} - K_*^\mathsf{T} K^{-1} K_*\right)} \tag{9}$$

$$\bar{y}_* = \mathbf{k}_*^\mathsf{T} K^{-1} \mathbf{y}, \quad \mathrm{var}(y_*) = k_{**} - \mathbf{k}_*^\mathsf{T} K^{-1} \mathbf{k}_* \tag{10}$$

- In practice, instead of inverting the K matrix, Cholesky decomposition is used: $K = LL^\mathsf{T}$.
  - K is symmetric.
  - L is triangular.

- Then $K^{-1} = (L^{-1})^\mathsf{T} L^{-1}$, so

$$\bar{y}_* = \left(L^{-1}\mathbf{k}_*\right)^\mathsf{T} \left(L^{-1}\mathbf{y}\right), \quad \mathrm{var}(y_*) = k_{**} - \left(L^{-1}\mathbf{k}_*\right)^\mathsf{T} \left(L^{-1}\mathbf{k}_*\right) \tag{11}$$

- Cholesky decomposition + back substitution generally has better numerical stability than matrix inversion + multiplication.

- My C++ implementation: https://git.io/fhN1j

What is a process?

- A stochastic process is a collection of labelled random variables.
  - Implies a certain relationship between the random variables.
  - E.g. given by mean and covariance for a GP.
- Other examples: Poisson, Markov, Wiener (Brownian motion).

- Although this is strictly a subset, I like to imagine a **random field**.
  - That's the subset useful for GP regression anyway.
- At every point on a manifold there is a distribution.
- Interesting properties arise from the properties of a particular field, relating distributions at different points.

- These things are historically called processes, because early applications dealt with random variables labelled by points in time, giving the interpretation of a stochastic process representing some system randomly changing over time. Examples include
  - growth of a bacterial population,
  - electrical current fluctuating due to thermal noise,
  - movement of gas molecules.

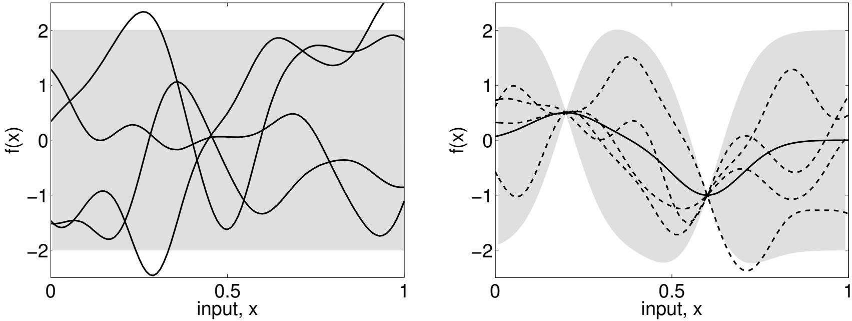
GP: Function space view

- Gaussian process is a multidimensional Gaussian distribution with each dimension corresponding a point on the domain manifold.
- $\therefore$ GP is a distribution of these vectors
- Each vector is a function on the manifold.
- $\therefore$ GP is a distribution of functions.

- These functions can be sampled.
- For a univariate normal distribution: $\mathcal{N}\left(\mu, \sigma^2\right) = \mu + \sigma\mathcal{N}\left(0, 1\right)$
- For a multivariate normal distribution: $\mathcal{N}\left(\boldsymbol{\mu}, \Sigma\right) = \boldsymbol{\mu} + B\mathcal{N}\left(\mathbf{0}, I\right)$, where $BB^{\mathsf{T}} = \Sigma$.
- To get a sample function:
  1. Generate a vector $\mathbf{v}$ of n standard normally distributed numbers;
  2. Find Cholesky decomposition of an $n \times n$, $K_{**} = BB^{\mathsf{T}}$ matrix;
  3. The function values are $\mathbf{f} = \boldsymbol{\mu} + B\mathbf{v}$.

Example of function sampling

Example from Rasmussen & Williams



(a), prior

(b), posterior

Figure 1.1: Panel (a) shows four samples drawn from the prior distribution. Panel (b) shows the situation after two datapoints have been observed. The mean prediction is shown as the solid line and four samples from the posterior are shown as dashed lines. In both plots the shaded region denotes twice the standard deviation at each input value x.

GP: Weight space view

- Model a linear function with some Gaussian noise:

$$f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{w}, \quad y = f(\mathbf{x}) + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma_n^2\right) \tag{12}$$

- This implies the likelihood, the probability density of the observations given the parameters:

$$p(\mathbf{y}|X, \mathbf{w}) = \mathcal{N}\left(X^\mathsf{T}\mathbf{w},\, \sigma_n^2 I\right) \tag{13}$$

- Putting a prior on $\mathbf{w}$ and using Bayes' theorem yields a posterior distribution:

$$\mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \Sigma_p\right) \quad \Rightarrow \quad p(\mathbf{w}|X, \mathbf{y}) = \mathcal{N}\left(\sigma_n^{-2} A^{-1} X\mathbf{y},\, A^{-1}\right), \tag{14}$$

  where $A = \sigma_n^{-2} X X^\mathsf{T} + \Sigma_p^{-1}$.
- This is like doing a linear fit using infinite-dimensional function basis. Weight is synonymous to fit parameter in this language.

Relation to Neural Networks

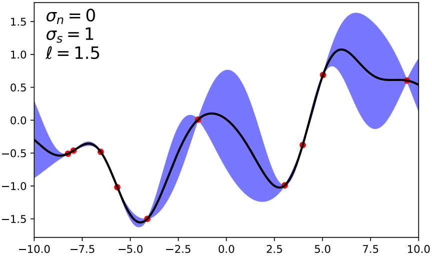
- Bayesian neural networks because Gaussian processes in the limit of the number of hidden nodes approaching infinity.
  - Neal, R.M. Bayesian Learning for Neural Networks.
    Springer Verlag, 1996. ISBN 0387947248.

Squared Exponential Kernel

Synonims: Radial Basis Function, Gaussian, Exponentiated Quadratic

$$k(x, x') = \exp\left(-\frac{|x - x'|^2}{2\ell^2}\right) \tag{15}$$

- The de-facto default kernel.
- Produces infinitely differentiable functions.
- ℓ - lengthscale. Determines the scale of function's features. Generally, cannot extrapolate farther than distance ℓ from the data.
- Stationary: depends only on $x - x'$.
- Isotropic: depends only on $|x - x'|$.

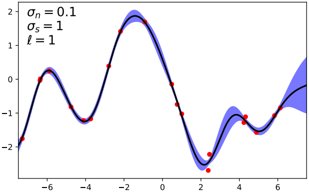


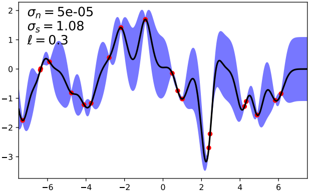
Noise and signal variances

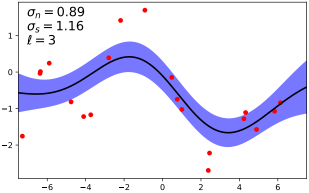
- Typically, two extra parameters are added to the kernel model.
- σ_s^2 - signal variance. This is a multiplicative factor attached to every additive term in a kernel. Determines the average distance of the function away from the mean.
- σ_n^2 - noise variance. This is added to the diagonal elements of the kernel, and provides a measure of the observation's uncertainty.
- With both terms introduced, the SE kernel looks like this

$$k(x,x') = \sigma_s^2 \exp\left(-\frac{|x-x'|^2}{2\ell^2}\right) + \sigma_n^2 \delta_{xx'} \tag{16}$$

- σ_n^2 can be different for every point, e.g. Poisson uncertainties for a histogram.



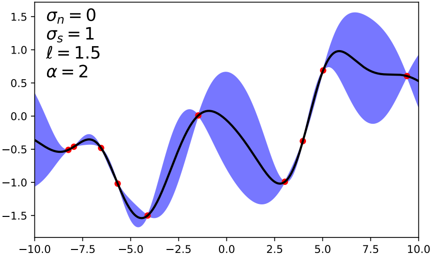




Rational Quadratic Kernel

$$k(x, x') = \left(1 + \frac{|x - x'|^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{17}$$

- Equivalent to a series of SE kernels with different length scales.
- α - determines the relative weighting of large-scale and small-scale variations.
- RQ → SE as $\alpha \to \infty$.



Periodic Kernel

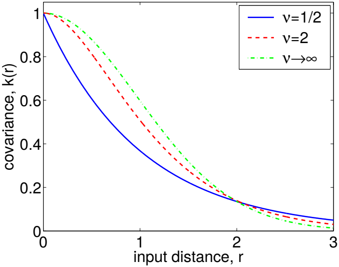
$$k(x, x') = \exp\left(-\frac{2}{\ell^2}\sin^2\left[\frac{\pi|x - x'|}{\lambda}\right]\right) \tag{18}$$

- Allows to model functions with exact periodicity.

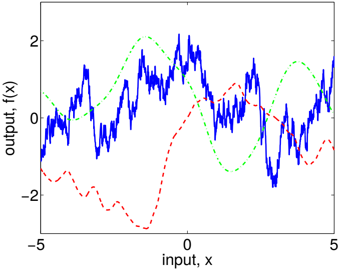
Matérn Kernel

$$k(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\, \frac{r}{\rho} \right)^{\nu} K_{\nu} \left(\sqrt{2\nu}\, \frac{r}{\rho} \right) \tag{19}$$

- $r = |x - x'|$
- Γ - gamma function
- K_ν - modified Bessel function of the second kind
- ρ and ν are non-negative parameters.
- Yields sample paths that are $\lceil \nu \rceil - 1$ times differentiable.



kernel functions



sampled functions

Figure from Rasmussen & Williams

Kernel Composition

- Allows to better capture features on different scales
- Multiplication: AND operation
- Addition: OR operation
- Example: Locally periodic kernel

$$k(x,x') = \exp\left(-\frac{2}{\ell_1^2}\sin^2\left[\frac{\pi|x-x'|}{\lambda}\right]\right)\exp\left(-\frac{|x-x'|^2}{2\ell_2^2}\right) \tag{20}$$

Optimization of Kernel Parameters

- Need to maximize $p(\theta|x, y)$.
- By the Bayes' theorem, this is the same as maximizing $p(y|x, \theta)$.
- This is the prior.

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, K\right) \tag{21}$$

- Therefore, optimal parameters are given my maximizing log prior likelihood:

$$\log p(y) = -\frac{1}{2}\mathbf{y}^{\mathsf{T}}K^{-1}\mathbf{y} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi \tag{22}$$

- Knowledge of the specific problem may dictate more appropriate kernel parameters.
  - For example, noise variance may be known.
  - Length scale may have a physical meaning.

References

- Rasmussen & Williams (the cannonical textbook)
  http://www.gaussianprocess.org/gpml/chapters/
- Mark Ebden (simpler introduction than R&W)
  https://arxiv.org/abs/1505.02965
- Katherine Bailey (simple python + numpy code)
  http://katbailey.github.io/post/gaussian-processes-for-dummies/
- CS229 Stanford Notes (Chuong Do)
  http://cs229.stanford.edu/section/cs229-gaussian_processes.pdf
- Kernel Cookbook (David Duvenaud)
  https://www.cs.toronto.edu/~duvenaud/cookbook/
- Rob Fletcher's slides (ATLAS internal)
  https://indico.cern.ch/event/744257/contributions/3077617/attachments/1688169/2715519/GP_Fletcher.pdf
- Some interesting papers:
  - arXiv:1709.05681 (Modeling smooth backgrounds)
  - arXiv:1302.4245 (Spectral mixture kernel)