Gaussian Process Theory

Contents

Objectives

- Understand what a Gaussian process is
- Use a Gaussian process to define a prior probability measure on the space of functions

Readings

- Chapter 1 from C.E. Rasmussen's textbook on Gaussian processes.

Motivation: A fully Bayesian paradigm for curve fitting

Gaussian process regression is Bayesian regression on steroids. However, understanding how it works requires a change of mind. After a bit of practice it starts making sense.

Here is how it works:

- Let's say that you have to learn some function $f(\cdot)$ from some space $\mathcal{X}$ to $\mathbb{R}$ (this could either be a supervised learning problem (regression or classification) or even an unsupervised learning problem.
- You sit down and you think about $f(\cdot)$. What do you know about it? How large do you expect it be? How small do you expect it be? Is it continuous? Is it differentiable? Is it periodic? How fast does it change as you change its inputs?
- You create a probability measure on the space of functions in which $f(\cdot)$ lives which is compatible with everything you know about it. Abusing mathematical notation a lot, let's write this probability measure as $p(f(\cdot))$. Now you can sample from it. Any sample you take is compatible with your prior beliefs. You cannot tell which one is better than any other. Any of them could be the true $f(\cdot)$.
- Then, you get a little bit of data, say $\mathcal{D}$. You model the likelihood of the data, $p(\mathcal{D}|f(\cdot))$, i.e., you model how the data may have been generated if you knew $f(\cdot)$.
- Finally, you use Bayes' rule to come up with your posterior probability measure over the space of functions:

$$p(f(\cdot)|\mathcal{D}) \propto p(\mathcal{D}|f(\cdot))p(f(\cdot)),$$

which is simultaneously compatible with your prior beliefs and the data. Again, we are abusing mathematical notation here sinc you cannot really write down the probability density corresponding to a random function. But you get the point.

This is it. As Persi Diaconis' said in an 1988 paper:

> Most people, even Bayesians, think that this sounds crazy when they first hear about it.

Where do Gaussian processes come in? Well, it is just the equivalent of the multivariate Gaussian for function spaces. It defines a probability measure on the space of functions that is centered about a mean (function) and shaped by a covariance (function). In this lecture, we will show that the mean function and the covariance function is the place where you can encode your prior knowledge. We will also show how you can sample functions from this probability measure. Next time, we will show how you can condition these probability measures on observations.

Some mathematical terminology

A *stochatic process* is just a collection of random variables that are labeled by some index: $\{X_i\}$ for some i in a set I. If the set I is discrete, then we say that we have a discrete stochastic process. If the set I is continuous, then we have a continuous stochastic process. We will mostly work with continuous stochastic processes in this class.

For example, $X_t = X(t)$ is a stochastic process parameterized by time. You can also think of a stochastic process as a random function. Here is how, you sample an ω, you keep it fixed, and then $X(t,\omega)$ is just a function of time. (Remember what we learned in earlier lectures: a random variable is just a function from the event space to the real numbers).

Of course, you can have a stochastic process that is parameterized by space, for example $T(x,\omega)$, could be an unknown temperature field. This is typically called a random field. You can also have a stochastic process that is parameterized by both space and time, say $T(x,t,\omega)$. This is a unknown spatiotemporal temperature field. As a matter of fact, you can have a stochastic process parameterized by any continuous label you like. It does not have to be space or time. And also, you can have as many different labels as you like.

Gaussian processes are the simplest continuous stochastic processes you can have.

Gaussian process

A Gaussian process (GP) 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: \begin{equation} f(\cdot) \sim \mbox{GP}\left(m(\cdot), k(\cdot, \cdot) \right), \end{equation} where $m:\mathbb{R}^d\rightarrow\mathbb{R}$ is the *mean function* and $k:\mathbb{R}^d\times\mathbb{R}^d\rightarrow\mathbb{R}$ 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 $\boldsymbol{\Sigma}$.

But, what does this definition actually mean? Actually, it gets its meaning from the multivariate Gaussian distribution. Here is how:

- Let $\mathbf{x}_{1:n}=\{\mathbf{x}_1,\dots,\mathbf{x}_n\}$ be n points in $\mathbb{R}^d$.
- Let $\mathbf{f}\in\mathbb{R}^n$ 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:

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

Now that we have defined a Gaussian process (GP), let us talk about we encode our prior beliefs into a GP. We do so through the mean and covariance functions.

Interpretation of the mean function

What is the meaning of $m(\cdot)$? Well, it is quite easy to grasp. For any point $\mathbf{x}\in\mathbb{R}^d$, $m(\mathbf{x})$ should give us the value we believe is more probable for $f(\mathbf{x})$. Mathematically, $m(\mathbf{x})$ is nothing more than the expected value of the random variable $f(\mathbf{x})$. That is: \begin{equation} m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]. \end{equation}

The mean function can be any arbitrary function. Essentially, it tracks generic trends in the response as the input is varied. In practice, we try and make a suitable choice for the mean function that is easy to work with. Such choices include:

- zero, i.e.,

$$m(\mathbf{x}) = 0.$$

- a constant, i.e.,

$$m(\mathbf{x}) = c,$$

where c is a parameter.

- linear, i.e.,

$$m(\mathbf{x}) = c_0 + \sum_{i=1}^d c_i x_i,$$

where $c_i, i=0,\dots,d$ are parameters.

- using a set of m basis functions (generalized linear model), i.e.,

$$m(\mathbf{x}) = \sum_{i=1}^m c_i\phi_i(\mathbf{x}),$$

where c_i and $\phi_i(\cdot)$ are parameters and basis functions.

- generalized polynomial chaos (gPC), i.e., using a set of d polynomial basis functions upto a given degree ρ $m(\mathbf{x}) = \sum_{i=1}^d c_i\phi_i(\mathbf{x})$ where the basis functions ϕ_i are mutually orthonormal with respect to some measure μ:

$$\int \phi_i(\mathbf{x})\phi_j(\mathbf{x})d\mu(\mathbf{x}) = \delta_{ij}$$

- and many other possibilities.

Interpretation of the covariance function

What is the meaning of $k(\cdot,\cdot)$? This concept is considerably more challenging than the mean.

Let’s try to break it down:

- Let $\mathbf{x}\in\mathbb{R}^d$. Then $k(\mathbf{x},\mathbf{x})$ is the variance of the random variable $f(\mathbf{x})$, i.e.,

$$\mathbb{V}[f(\mathbf{x})] = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))^2\right].$$

In other words, we believe that there is about 95% probability that the value of the random variable $f(\mathbf{x})$ fall within the interval:

$$\left((m(\mathbf{x}) - 2\sqrt{k(\mathbf{x},\mathbf{x})}, m(\mathbf{x}) + 2\sqrt{k(\mathbf{x},\mathbf{x})}\right).$$

- Let $\mathbf{x},\mathbf{x}'\mathbb{R}^d$. Then $k(\mathbf{x},\mathbf{x}')$ tells us how the random variable $f(\mathbf{x})$ and $f(\mathbf{x}')$ are correlated. In particular, $k(\mathbf{x},\mathbf{x}')$ is equal to the covariance of the random variables $f(\mathbf{x})$ and $f(\mathbf{x}')$, i.e.,

$$k(\mathbf{x},\mathbf{x}') = \mathbb{C}[f(\mathbf{x}), f(\mathbf{x}')] = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))\left(f(\mathbf{x}') - m(\mathbf{x}')\right)\right].$$

Essentially, a covariance function (or covariance kernel) defines a nearness or similarity measure on the input space. We cannot choose any arbitrary function of two variables as a covariance kernel. How we go about choosing a covariance function is discussed in great detail here. We briefly discuss some properties of covariance functions here and then we shall move onto a discussion of what kind of prior beliefs we can encode through the covariance function.

Properties of the covariance function

Properties of the covariance function

- There is one property of the covariance function that we can note right away. Namely, that for any $\mathbf{x}\in\mathbb{R}^d$, $k(\mathbf{x},\mathbf{x}) > 0$. This is easly understood by the interpretation of $k(\mathbf{x},\mathbf{x})$ as the variance of the random variable $f(\mathbf{x})$.
- $k(\mathbf{x},\mathbf{x}')$ becomes smaller as the distance between $\mathbf{x}$ and $\mathbf{x}'$ grows.
- For any choice of points $\mathbf{X}\in\mathbb{R}^{n\times d}$, the covariance matrix: $\mathbf{K}(\mathbf{X},\mathbf{X})$ has to be positive-definite (so that the vector of outputs $\mathbf{f}$ is indeed a multivariate normal distribution).

Encoding prior beliefs in the covariance function

- **Modeling regularity**. The choice of the covariance function controls the regularity properties of the functions sampled from the probability induced by the GP. For example, if the covariance kernel chosen is the squared exponential kernel, which is infinitely differentiable, then the functions sampled from the GP will also be infinitely differentiable.
- **Modeling invariance** If the covariance kernel is invariant w.r.t. a transformation T, i.e., $k(\mathbf{x}, T\mathbf{x}')=k(T\mathbf{x},\mathbf{x}')=k(\mathbf{x},\mathbf{x}')$ then samples from the GP will be invariant w.r.t. the same transformation.
- Other possibilities include periodicity, additivity etc.

By Ilias Bilionis (ibilion[at]purdue.edu)

© Copyright 2021.