Gaussian Processes for Time Series

Graham Gibson

July 22, 2017

Abstract

There is abstract text that you should replace with your own.

Contents

1   Gaussian Process Background

Gaussian processes can be viewed as infinite dimensional Gaussian distributions.

2   Example

Suppose we observe the following data points

$$\{(1,1)\}$$

and we wish to estimate a regression function $f(x)$ that fits the data. Although a line of the form $y = mx + b$ seems reasonable, it certainly isn't the only possible regression function to pass through this point. If we allow ourselves to consider any non-linear functional form for $f(x)$ we quickly see that there are an infinite number of functions passing through these points. For instance take the function

$$f(x) = \begin{cases} 1, if \ x = 1 \\ sin(x), otherwise \end{cases}$$

We can define that last case using any function out of the infinite number of functions, so therefore the total number of functions passing through

$$D = \{(1,1)\}$$

is infinite. Gaussian processes are a way to formalize our belief that certain functions are more likely to have generated D than others. Formally, a gaussian process is defined.

Definition: A Gaussian Process is a stochastic process that assumes for all finite collection of points $\vec{x} = x_1, x_2, ..., x_n$ we have that

$$p(f(x_1), f(x_2), ..., f(x_n)) \sim N(\mu(\vec{x}), K(x_i, x_j))$$

for some mean function μ and some covariance kernel function K.

For example if we take our kernel function to be the commonly chosen gaussian kernel we let

$$k(x_i, x_j) = \sigma^2 exp(\frac{-(x_i - x_j)^2}{2l^2})$$

where l is called the length-scale parameter. For now let's set $l = \sqrt{\frac{1}{2}}$ for simplicity,since the denominator in the gaussian will be 1. We will investigate both parameters σ^2, l later on. For now let us take our example from above and translate it into GP notation.

$$p(f(1)) \sim N(0, \sigma^2)$$

where we have chose $\mu(x) = 0$ for our mean function. Since we can represent a function as just a finite set of $(x, f(x))$ pairs this really is a prior over functions defined on $x = \{1\}$.

We can write the above as

$$p(f) \sim N(0, \sigma^2)$$

Now we have a prior over all possible functions! Now, by simply sampling from this normal distribution we can generate functions. What does that really mean? Well think about what samples from the normal will look like, they will be real numbers $\{y_1, y_2, ...y_n\}$. This is now a series of functions! defined by $f(x_1) \rightarrow y_i$

So samples are functions defined over the domain of our x points in our observed data. So what?

Now we can make inferences by examining the form of the posterior, given our observed data. If we go back to the definition 1 we see that a GP is a distribution over any arbitrary set of points. Consider the new data set

$$D = \{(1, 1), (2, ?)\}$$

We now want to make an inference on the unknown value of $f(2)$ given our prior distribution on f. By the GP definition we can write this as

$$p(f(2), f(1)) \sim N([0, 0], , \begin{bmatrix} \sigma^2, \sigma^2 e \\ \sigma^2 e, \sigma^2 \end{bmatrix})$$

where the covariance matrix is generated by simply applying the kernel function to all x values in the input. We now have a joint distribution over our observed data points and our unobserved data points. We can appeal to the properties

of a multivariate gaussian: the conditional distribution given $\{f(1)\}$ is again gaussian with the following parameters

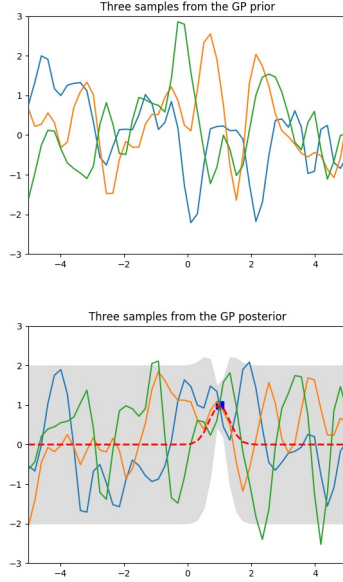$$p(f(2)|f(1),1) \sim N(\tilde{\mu}, \tilde{\sigma}^2)$$

where

$$\tilde{\mu} = K(2,1)(K(1,1) + \sigma^2 I)^{-1} f(2)$$

and

$$\tilde{\sigma}^2 = K(2,2) - K(2,1)(K(1,1) - \sigma^2 I)^{-1} K(1,2)$$

We can now sample from the posterior to find a distribution over all possible values $f(2)$, which in turns defines all functions that pass-though $(1,1)$ and are defined on $(2, f(2))$.

See below for an example on the dataset above.

Three samples from the GP prior

Three samples from the GP posterior

References

[1] Gelman, Andrew; Carlin, John B.; Stern, Hal S.; Dunson, David B.; Vehtari, Aki; Rubin, Donald B.. Bayesian Data Analysis, Third Edition (Chapman Hall/CRC Texts in Statistical Science) (Page 504). CRC Press.

[2] Murphy, Kevin P.. Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series) (Page 518). The MIT Press.