Chapter 1

Discussion

1.1 Why Assume Zero Mean?

In literature, as well as in practice, it is common to construct GP priors with a zero mean function. This might seem strange, since it is presumably a good place to put prior information, or if we are comparing models, to express since marginalizing over an unknown mean function can be equivalently expressed as a different GP with zero-mean, and another term added to the kernel.

A known mean function can be moved into the covariance function Specifically, if we wish to model an unknown function $f(\mathbf{x})$ with known mean $m(\mathbf{x})$, (with unknown magnitude $c \sim \mathcal{N}(0, \sigma_c^2)$), we can equivalently express this model using another GP with zero mean:

$$f \sim \mathcal{GP}(cm(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \quad c \sim \mathcal{N}\Big(0, \sigma_c^2\Big) \iff f \sim \mathcal{GP}\Big(\mathbf{0}, c^2 m(\mathbf{x})m(\mathbf{x}') + k(\mathbf{x}, \mathbf{x}')\Big) \tag{1.1}$$

By moving the mean function into the covariance function, we get the same model, but we can integrate over the magnitude of the mean function at no additional cost. This is one advantage of moving as much structure as possible into the covariance function.

An unknown mean function can be moved into the covariance function If we wish to express our ignorance about the mean function, one way would be by putting a

GP prior on it.

$$m \sim \mathcal{GP}(\mathbf{0}, k_1(\mathbf{x}, \mathbf{x}')), \quad f \sim \mathcal{GP}(m(\mathbf{x}), k_2(\mathbf{x}, \mathbf{x}')) \iff f \sim \mathcal{GP}(\mathbf{0}, k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')) \tag{1.2}$$

1.2 Why Not Learn the Mean Function Instead?

One might ask: besides integrating over the magnitude, what is the advantage of moving the mean function into the covariance function? After all, mean functions are certainly more interpretable than a posterior distribution over functions.

Instead of searching over a large class of covariance functions, which seems strange and unnatural, we might consider simply searching over a large class of structured mean functions, assuming a simple i.i.d. noise model. This is the approach taken by practically every other regression technique: neural networks, decision trees, boosting, etc. . If we could integrate over a wide class of possible mean functions, we would have a very powerful learning an inference method. The problem faced by all of these methods is the well-known problem *overfitting*. If we are forced to choose just a single function with which to make predictions, we must carefully control the flexibility of the model we learn, generally prefering "simple" functions, or to choose a function from a restricted set.

If, on the other hand, we are allowed to keep in mind many possible explanations of the data, *there is no need to penalize complexity*. [cite Occam's razor paper?] The power of putting structure into the covariance function is that doing so allows us to implictly integrate over many functions, maintaining a posterior distribution over infinitely many functions, instead of choosing just one. In fact, each of the functions being considered can be infinitely complex, without causing any form of overfitting. For example, each of the samples shown in figure **??** varies randomly over the whole real line, never repeating, each one requiring an infinite amount of information to describe. Choosing the one function which best fits the data will almost certainly cause overfitting. However, if we integrate over many such functions, we will end up with a posterior putting mass on only those functions which are compatible with the data. In other words, the parts of the function that we can determine from the data will be predicted with certainty, but the parts which are still uncertain will give rise to a wide range of predictions.

To repeat: *there is no need to assume that the function being modeled is simple, or to prefer simple explanations* in order to avoid overfitting, if we integrate over many possible explanations rather than choosing just the one.

In Chapter **??**, we will compare a system which estimates a parametric covariance function against one which estimates a parametric mean function.

1.3 Signal versus Noise

In most derivations of Gaussian processes, the model is given as

$$y = f(x) + \epsilon, \quad \text{where} \quad \epsilon \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2_{\text{noise}}\right) \tag{1.3}$$

However, ϵ can equivalently be thought of as another Gaussian process, and so this model can be written as $y(x) \sim \mathcal{GP}\left(0, k + \delta\right)$. The lack of a hard distinction between the noise model and the signal model raises the larger question: Which part of a model represents signal, and which represents noise?

We believe that it depends on what you want to do with the model - there is no hard distinction between signal and noise in general.

For example: often, we don't care about the short-term variations in a function, and only in the long-term trend. However, in many other cases, we with to de-trend our data to see more clearly how much a particular part of the signal deviated from normal.

1.4 Why does everyone use squared-exp?

In some instances, using a 'default' kernel yields acceptable performance. Many frequentist methods assume a characteristic kernel, such as the squared-exp. This choice is motivated by the fact that, in the limit of infinite data, and a shrinking lengthscale, the estimate of the function will converge asymptotically to the truth. [citation needed]