21: Gaussian Process Models

Taylor

University of Virginia

Introduction

We talk about Gaussian process models in this chapter. Gaussian processes describe random functions, and they can show up in statistical modeling in a few places.

If you would like to dig a little deeper, this is considered a good reference: `http://gaussianprocess.org/gpml/`. We will be using chapter 2 as an additional resource.

Definitions

Let your predictors $x_i \in \mathbb{R}^p$. We say μ follows a **Gaussian process** with mean function m and covariance function k if for any finite set of nonrandom points $x_1, \ldots, x_n$

$$\mu(x_i), \ldots, \mu(x_n) \sim \text{Normal}((m(x_1), \ldots, m(x_n)), K(x_1, \ldots, x_n)).$$

For short, we write $\mu \sim \text{GP}(m, k)$.

Definitions

Let your predictors $x_i \in \mathbb{R}^p$. We say μ follows a **Gaussian process** with mean function m and covariance function k if for any finite set of nonrandom points $x_1, \ldots, x_n$

$$\mu(x_i), \ldots, \mu(x_n) \sim \text{Normal}((m(x_1), \ldots, m(x_n)), K(x_1, \ldots, x_n)).$$

For short, we write $\mu \sim \text{GP}(m, k)$.

This means $E[\mu(x_i)] = m(x_i)$ and $\text{Cov}(\mu(x_i), \mu(x_j)) = K_{i,j} = k(x_i, x_j)$.

Confusingly, μ is also a (random) mean function, but it's the mean for the ys.

Definitions

Let's assume we're regressing univariate y_is on vector-valued x_is. Then we are interested in either

$$y_i = \mu(x_i)$$

or

$$y_i = \mu(x_i) + \epsilon_i.$$

Clearly

$$E[y_i \mid x_i] = E[\mu(x_i) \mid x_i] = m(x_i).$$

It is common to put $m(x) = 0$ (e.g. if $\mu(x) = x'\beta$ and β is given a mean zero prior).

However, you may assume β is known, or put a nonzero mean prior on it, or use a nonlinear (in x) mean function.

Definitions

If

$$\mu(x_i), \ldots, \mu(x_n) \sim \text{Normal}((m(x_1), \ldots, m(x_n)), K(x_1, \ldots, x_n)).$$

then $K(x_1, \ldots, x_n)$ is an $n \times n$ covariance matrix with its p, q element $K_{p,q} = k(x_p, x_q)$.

This k function gives you a "similarity" or "nearness" measure for any two pairs of inputs. It needs to be chosen very carefully.

A popular choice

We will often use a **squared exponential kernel**

$$k(x, x') = \tau^2 \exp\left[-\sum_{i=1}^{p} \frac{(x_j - x_j')^2}{2l_j^2}\right]$$

Each l_j determines the wiggliness in the jth direction of the predictors.

The τ^2 parameter is an overall variance for each $\mu(x)$.

Simulating from the prior

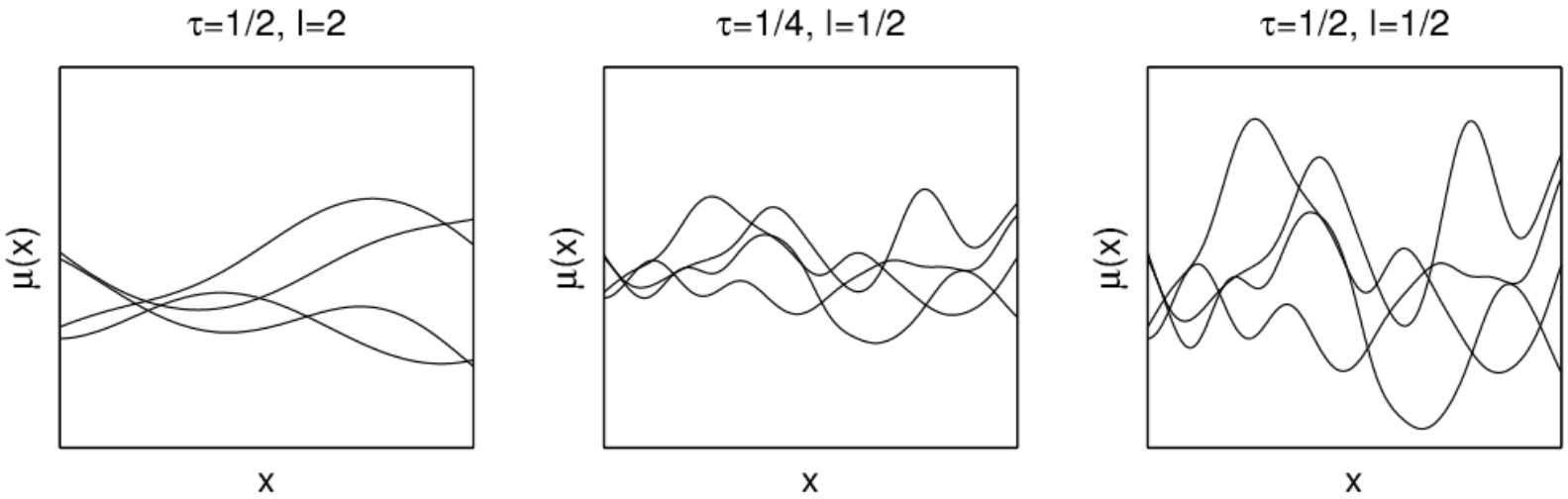

Figure 21.1 *Random draws from the Gaussian process prior with squared exponential covariance function and different values of the amplitude parameter τ and the length scale parameter l.*

There's a lot to say about many more kernels:
`https://www.cs.toronto.edu/~duvenaud/cookbook/`

Inference: conditional posterior

Let's assume the likelihood is $y_i = \mu(x_i) + \epsilon_i$ where $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, and for the prior, $m(x) = 0$.

The observed data is $\{x_i, y_i\}$, and the parameters are τ, l, σ^2. To find the conditional posterior $p(\mu(x) \mid x, y, \sigma^2, \tau, l)$, we use

$$\begin{pmatrix} y \\ \mu \end{pmatrix} \Bigg| x, \sigma^2, \tau, l \sim \text{Normal}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x,x) + \sigma^2 I & K(x,x) \\ K(x,x) & K(x,x) \end{pmatrix} \right)$$

By properties of multivariate normal random vectors $\mu \mid x, y, \tau, l, \sigma$ is normally distributed with

$$E[\mu] = K(x,x)[K(x,x) + \sigma^2 I]^{-1} y$$
$$Var[\mu] = K(x,x) - K(x,x)[K(x,x) + \sigma^2 I]^{-1} K(x,x)$$

Inference

Let's assume the likelihood is $y_i = \mu(x_i) + \epsilon_i$ where $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, and for the prior, $m(x) = 0$.

Call $\tilde{x}$ unseen data, in addition to $\{x_i, y_i\}$. Then

$$\begin{pmatrix} y \\ \tilde{\mu} \end{pmatrix} \Big| x, \tilde{x}, \sigma^2, \tau, l \sim \text{Normal}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(x,x) + \sigma^2 I & K(x,\tilde{x}) \\ K(\tilde{x},x) & K(\tilde{x},\tilde{x}) \end{pmatrix} \right)$$

By properties of multivariate normal random vectors, $\tilde{\mu} \mid x, y, \tau, l, \sigma$ is normally distributed with

$$E[\tilde{\mu}] = K(\tilde{x}, x)[K(x,x) + \sigma^2 I]^{-1} y$$
$$Var[\tilde{\mu}] = K(\tilde{x}, \tilde{x}) - K(\tilde{x}, x)[K(x,x) + \sigma^2 I]^{-1} K(x, \tilde{x})$$