Statistical Techniques in Robotics (16-831, F09) Lecture #21 (11/03/2009)

Gaussian Process - Part 2

Lecturer: Drew Bagnell *Scribe: Stephane Ross*

1 Gaussian Process

A gaussian process can be thought of as a gaussian distribution over functions (thinking of functions as infinitely long vectors containing the value of the function at every input). Formally let the input space $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$ a function from the input space to the reals, then we say f is a gaussian process if for any vector of inputs $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ such that $x_i \in \mathcal{X}$ for all i, the vector of output $f(\mathbf{x}) = [f(x_1), f(x_2), \ldots, f(x_n)]^T$ is gaussian distributed.

The gaussian process is specified by a mean function $\mu : \mathcal{X} \to \mathbb{R}$, such that $\mu(x)$ is the mean of $f(x)$ and a covariance/kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(x, x')$ is the covariance between $f(x)$ and $f(x')$. We say $f \sim GP(\mu, k)$ if for any $x_1, x_2, \ldots x_n \in \mathcal{X}$, $[f(x_1), f(x_2), \ldots, f(x_n)]^T$ is gaussian distributed with mean $[\mu(x_1), \mu(x_2), \ldots, \mu(x_n)]^T$ and $n \times n$ covariance/kernel matrix $K_{\mathbf{xx}}$:

$$K_{\mathbf{xx}} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \ldots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \ldots & k(x_2, x_n) \\ \ldots & \ldots & \ldots & \ldots \\ k(x_n, x_1) & k(x_n, x_2) & \ldots & k(x_n, x_n) \end{bmatrix}$$

The kernel function must be symmetric and positive definite. That is $k(x, x') = k(x', x)$, and the kernel matrix K induced by k for any set of input is a positive definite matrix. Example of some kernel functions are given below:

- Squared Exponential Kernel (Gaussian/RBF): $k(x, x') = \exp(\frac{-(x-x')^2}{2\gamma^2})$ where γ is the length scale of the kernel.
- Laplace Kernel: $k(x, x') = \exp(\frac{-|x-x'|}{\gamma})$.
- Indicator Kernel: $k(x, x') = I(x = x')$, where I is the indicator function.
- Linear Kernel: $k(x, x') = x^T x'$.

More complicated kernels can be constructed by adding known kernel functions together, as the sum of 2 kernel functions is also a kernel function.

2 Inference

Gaussian Processes are useful as priors over functions for doing non-linear regression. Given a set of observed input and corresponding output values $(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_n, f(x_n))$, and

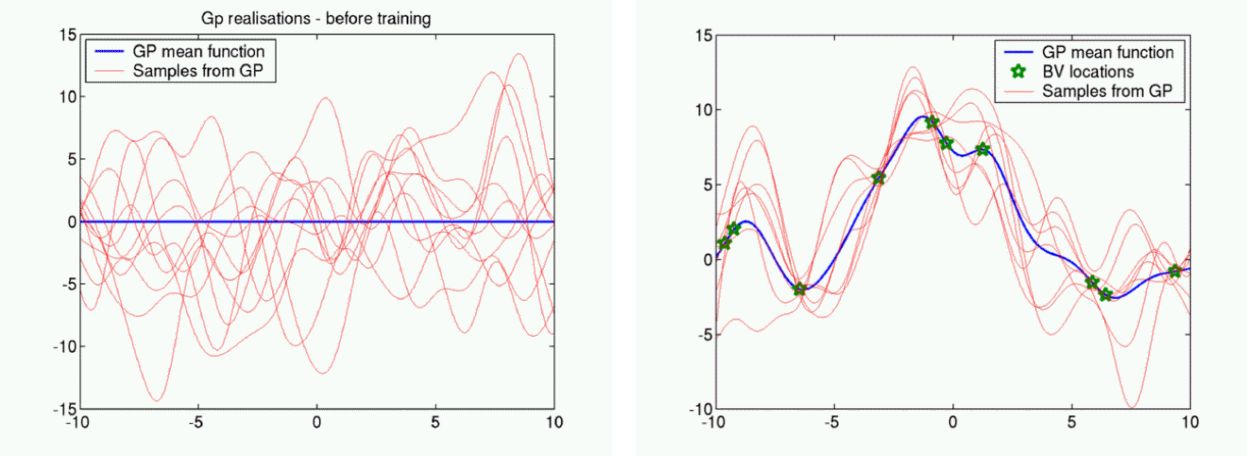

Figure 1: Samples from a zero-mean GP prior (Left) and samples from the posterior after a few observations (Right).

gaussian process prior on f, $f \sim GP(\mu, k)$, we would like to compute the posterior over the value $f(x^*)$ at any query input x^*. Figure 1 illustrates this process. Sample functions from a prior zero-mean GP are first shown on the left, and after observing a few values, the posterior mean and sample functions from the posterior are shown on the right. We can observe from this that the sample functions from the posterior passes close to the observed values but varies a lot in region where there is no observation.

2.1 Computing the Posterior

The posterior can be derived similarly to how the update equations for the Kalman filter was derived. First we will find what is the joint distribution of $[f(x^*), f(x_1), f(x_2), \ldots, f(x_n)]^T$, and then use the conditioning rule for gaussian to compute the conditional distribution of $f(x^*)|f(x_1), \ldots, f(x_n)$.

Assume for now that the prior mean function $\mu = 0$. Then the joint distribution of $[f(x^*), f(x_1), f(x_2), \ldots, f(x_n)]^T$ is gaussian:

$$\begin{bmatrix} f(x^*) \\ f(x_1) \\ \ldots \\ f(x_n) \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \ldots \\ 0 \end{bmatrix}, \begin{bmatrix} k(x^*, x^*) & k(x^*, \mathbf{x})^T \\ k(x^*, \mathbf{x}) & K_{\mathbf{xx}} \end{bmatrix} \right)$$

where

$$k(x^*, \mathbf{x}) = \begin{bmatrix} k(x^*, x_1) \\ k(x^*, x_2) \\ \ldots \\ k(x^*, x_n) \end{bmatrix}$$

Now using the conditioning rule we obtained that the posterior for $f(x^*)$ is gaussian:

$$f(x^*)|f(\mathbf{x}) \quad \sim \quad N(\quad k(x^*,\mathbf{x})^T K_{\mathbf{xx}}^{-1} f(\mathbf{x}) \quad , \quad k(x^*,x^*) + k(x^*,\mathbf{x})^T K_{\mathbf{xx}}^{-1} k(x^*,\mathbf{x}) \quad)$$

Notice that the posterior mean $\mathbb{E}(f(x^*)|f(\mathbf{x}))$ can be represented as a linear combination of the kernel function values:

$$\mathbb{E}(f(x^*)|f(\mathbf{x})) = \sum_{i=1}^{n} \alpha_i k(x^*, x_i)$$

for $\alpha = K_{\mathbf{xx}}^{-1} f(\mathbf{x})$. This means we can compute the mean without explicitly inverting K, by solving $K\alpha = f(\mathbf{x})$ instead. Similarly, it can also be represented as a linear combination of the observed function values:

$$\mathbb{E}(f(x^*)|f(\mathbf{x})) = \sum_{i=1}^{n} \beta_i f(x_i)$$

for $\beta = k(x^*,\mathbf{x})^T K^{-1}$.

2.2 Non-zero mean prior

If the prior mean function is non-zero, we can still use the previous derivation by noting that if $f \sim GP(\mu, k)$, then the function $f' = f - \mu$ is a zero-mean gaussian process $f' \sim GP(0, k)$. Hence if we have observations from the values of f, we can subtract the prior mean function values to get observations of f', do the inference on f', and finally once we obtain the posterior on $f'(x^*)$ we can simply add back the prior mean $\mu(x^*)$ to the posterior mean to obtain the posterior on f.

2.3 Noise in observed output values

If instead of having noise-free observation of f, we observe $y(x) = f(x) + \delta$, where $\delta \sim N(0, \sigma^2)$ is some zero-mean gaussian noise, then the joint distribution of $[f(x^*), y(x_1), \ldots y(x_n)]^T$ is also gaussian so that we can apply a similar derivation to compute the posterior of $f(x^*)$. More specifically if the prior mean function $\mu = 0$, we have that:

$$\begin{bmatrix} f(x^*) \\ y(x_1) \\ \ldots \\ y(x_n) \end{bmatrix} \quad \sim \quad N\left(\begin{bmatrix} 0 \\ 0 \\ \ldots \\ 0 \end{bmatrix}, \begin{bmatrix} k(x^*,x^*) + \sigma^2 & k(x^*,\mathbf{x})^T \\ k(x^*,\mathbf{x}) & K_{\mathbf{xx}} + \sigma^2 I \end{bmatrix} \right)$$

The only difference with respect to the noise-free case is that the covariance matrix of the joint now has an extra σ^2 term on its diagonal. This is because the noise is independent among different observations and also independent of f (so no covariance between noise terms, and between f and δ). So we obtain that the posterior on $f(x^*)$ is:

$$f(x^*)|y(\mathbf{x}) \quad \sim \quad N(\quad k(x^*,\mathbf{x})^T (K_{\mathbf{xx}} + \sigma^2 I)^{-1} y(\mathbf{x}) \quad , \quad k(x^*,x^*) + \sigma^2 + k(x^*,\mathbf{x})^T (K_{\mathbf{xx}} + \sigma^2 I)^{-1} k(x^*,\mathbf{x}) \quad)$$

2.4 Choosing Kernel Length Scale and Noise Variance Parameters

We can use the data to fit the kernel lenght scale (γ) and noise variance (σ^2) parameters by choosing the parameters that maximizes the log likelihood of the observed data. Assuming a gaussian kernel, then we obtain the most likely parameters γ and σ by solving:

$$\max_{\gamma,\sigma}[\log P(y(\mathbf{x})|\gamma,\sigma)] = \max_{\gamma,\sigma}[-\frac{1}{2}f(\mathbf{x})^T K_{\mathbf{xx}} f(\mathbf{x}) - \frac{1}{2}\log(\det(K_{\mathbf{xx}} + \sigma^2 I)) - \frac{1}{2}\log(2\pi)]$$

Here the determinant will be small when $K_{\mathbf{xx}}$ is almost diagonal, so this maximization will favor smoother kernel (larger γ).

Additionally σ^2 can be chosen to have a higher value to prevent overfitting, as a larger σ means each observation is less informative.

2.5 Computational Complexity

One drawback of the Gaussian Process is that it scales very badly with the number of observations N. Solving for the coefficients α defining the mean function requires $O(N^3)$ computations. Note that bayesian linear regression, which can be seen as a special case of GP with the linear kernel, has complexity of only $O(d^3)$ to find the mean weight vector, for d the dimension of the input space $\mathcal{X}$. Finally to make a prediction at any point, Gaussian Process requires $O(N\hat{d})$ (where $\hat{d}$ is the complexity of evaluating the kernel) while BLR only requires $O(d)$ computations.