Statistical Techniques in Robotics (16-831, F10) Lecture#22 (Thursday November 4)

Guassian Processes

Lecturer: Drew Bagnell *Scribe:Joseph Lisee* [1]

1 Gaussian Processes

1.1 Introduction

There are many machine learning problems for which regression is a viable solution. Examples of linear regression methods include Bayes linear regression (BLR) and online convex programming when applied to square laws. For some problems linear approximations do not provide enough accuracy in their estimates and nonlinear methods methods like extended kalman filters are required. Gaussian processes are the state of the art in nonlinear regression methods, but unlike the previously covered methods it is a non-parametric method with infinitely many parameters.

Gaussian process regression uses a multidimensional gaussian with a dimension for each training and test point. To compute posterior probability at a test point you condition on the training data points. For example we can start with two dimensional gaussian with mean μ and variance Σ:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \tag{1}$$

We can then find the conditional distrobution $P(x_2|x_1)$ with the following equations:

$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1) \tag{2}$$

$$\sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \tag{3}$$

This entire process can be expanded to match the size of the date set and is covered in more detail in section 1.5. The next section will formally define gaussian processes and cover how the Σ matrix is computed from a data set.

1.2 Formal Definition

A gaussian process is a random stochastic process where correlation is introduced between neighboring samples (think of a stochastic process as a sequence of random variables). The same way that an instance of a random variable is a single sample, an instance of a stochastic process can be thought as vector of samples:

$$X = [x_1, x_2, x_3 \ldots] \tag{4}$$

[1] Some content adapted from previous scribes: Alberto Rodriguez, Stephane Ross

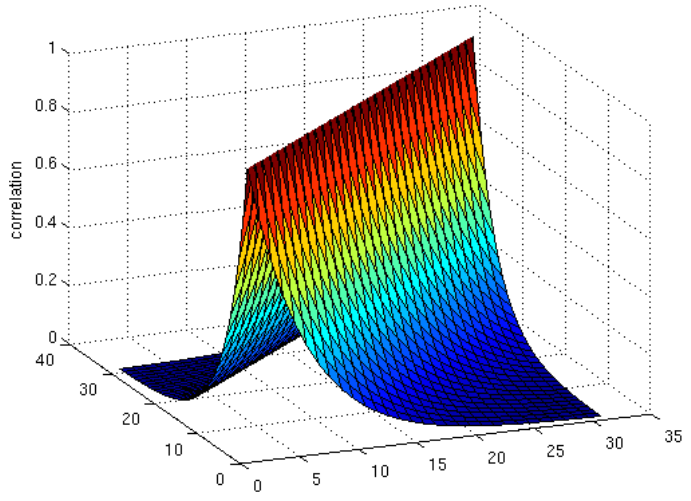

Figure 1: A plot of the laplace kernel with $\gamma = 1$ where $\mathbf{x} = (0, 0.2, 0.4, \ldots, 6)$. The correlation quickly tapers off as the distance between data points increase.

Gaussian Processes artificially introduce correlation between close samples in that vector in order to enforce some sort of smoothness on the succession of samples. The way that correlation is introduced is by constructing the joint probability distribution of the long vector of samples. Gaussian processes assume that probability distribution to be a multidimensional gaussian:

$$p(X^i) = \frac{1}{z} e^{(X-\mu)\Sigma^{-1}(X-\mu)} \tag{5}$$

The correlation between samples in the succession X^i depends on matrix Σ. In Gaussian Processes the covariance matrix is constructed as the Gram matrix of the samples with some desired kernel $\kappa(\cdot, \cdot)$ as the inner product:

$$\Sigma = \begin{pmatrix} \kappa(x_1, x_1) & \kappa(x_1, x_2) & \cdots & \kappa(x_1, x_n) \\ \kappa(x_2, x_1) & \kappa(x_2, x_2) & \cdots & \kappa(x_2, x_n) \\ \vdots & & & \vdots \\ \kappa(x_n, x_1) & \kappa(x_n, x_2) & \ldots & \kappa(x_n, x_n) \end{pmatrix} \tag{6}$$

In order to introduce correlation between neighboring samples, the kernel κ is usually designed to have small support and centered around zero (i.e. a triangular or a gaussian). The kernel usually can be described as a function of the distance $x_i - x_j$ and must be symmetric and positive definite. That is $\kappa(x, x') = \kappa(x', x)$, 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): $\kappa(x, x') = \exp(\frac{-(x-x')^2}{2\gamma^2})$ where γ is the length scale of the kernel.

- Laplace Kernel: $\kappa(x, x') = \exp(\frac{-|x-x'|}{\gamma})$ (See in figure 1 for an example).
- Indicator Kernel: $\kappa(x, x') = I(x = x')$, where I is the indicator function.
- Linear Kernel: $\kappa(x, x') = x^T x'$.

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

1.3 Gaussian Processes as a Distrobution Over Functions

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 $\kappa(x, x')$ 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_{Data}:

1.4 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 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 2 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.

1.5 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:

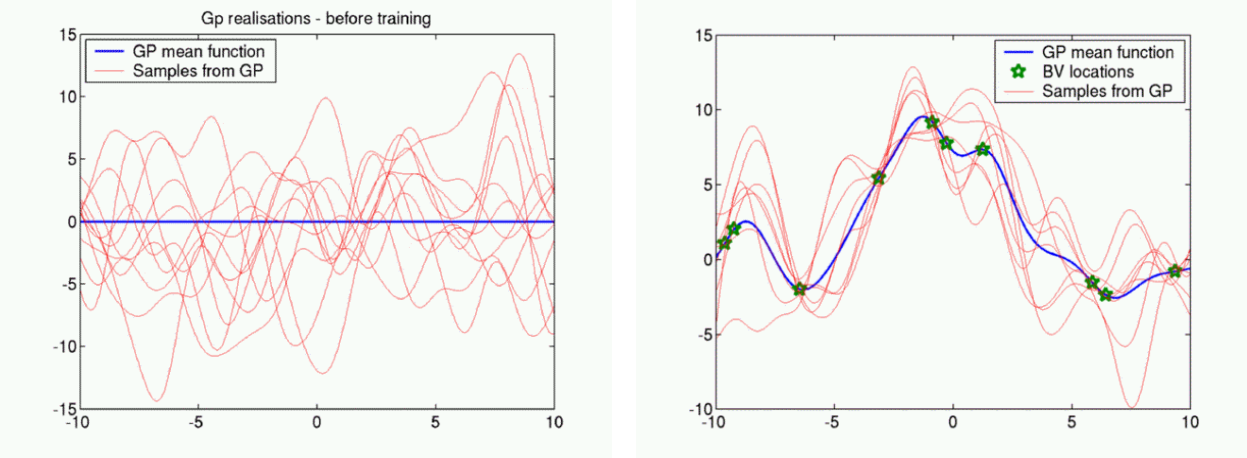

Figure 2: Samples from a zero-mean GP prior (Left) and samples from the posterior after a few observations (Right).

$$\begin{bmatrix} f(x_1) \\ \dots \\ f(x_n) \\ f(x^*) \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}, \begin{bmatrix} K_{Data} & \kappa(x^*, \mathbf{x})^T \\ \kappa(x^*, \mathbf{x}) & \kappa(x^*, x^*) \end{bmatrix} \right) \tag{7}$$

where

$$\kappa(x^*, \mathbf{x}) = \begin{bmatrix} \kappa(x^*, x_1) \\ \kappa(x^*, x_2) \\ \dots \\ \kappa(x^*, x_n) \end{bmatrix} \tag{8}$$

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

$$f(x^*)|f(\mathbf{x}) \sim N(\kappa(x^*, \mathbf{x})^T K_{Data}^{-1} f(\mathbf{x}) , \kappa(x^*, x^*) + \kappa(x^*, \mathbf{x})^T K_{Data}^{-1} \kappa(x^*, \mathbf{x})) \tag{9}$$