Gaussian Process

Kangcheng Hou

Zhejiang University

May 14, 2018

Gaussian Process as function generator

Gaussian Process is a distribution over function space, it can be used to represent uncertainty. Samples from GP satisfies that any d dimensional samples $(\mathbf{x}_1, \ldots, \mathbf{x}_d)$ is from Gaussian distribution $\mathcal{N}(\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_d)^\top, \mathbf{K}(\mathbf{x}_1, \ldots, \mathbf{x}_d))$ The μ and $\mathbf{K}$ is chosen to represent our knowledge about the problem. So provided some positions to be estimated $(\mathbf{x}_1, \ldots, \mathbf{x}_d)$, GP can generate some functions and return the value estimated on those provided input points.

Inference in GP I

First we notice the properties of Gaussian distribuion,

$$p(\mathbf{f}, \mathbf{g}) = \mathcal{N}(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix})$$

$$p(\mathbf{f}|\mathbf{g}) = \mathcal{N}(\mathbf{a} + CB^{-1}(\mathbf{y} - \mathbf{b}), A - CB^{-1}C^\top)$$

In this way, if we know the mean and the covariance of joint distribution of $(\mathbf{f}, \mathbf{g})$ and the observed $\mathbf{g}$. We can then infer the distribution of $\mathbf{f}$. What if we only saw a noisy observation, $\mathbf{y} \sim \mathcal{N}(\mathbf{g}, S)$?

Inference in GP II

$$p(\mathbf{f}, \mathbf{g}, \mathbf{y}) = p(\mathbf{f}, \mathbf{g})p(\mathbf{y}|\mathbf{g})$$

is the product of two gaussian pdf, so joint distribution of $(\mathbf{f}, \mathbf{g}, \mathbf{y})$ is still Gaussian distributed. Our posterior over $\mathbf{f}$ is still Gaussian:

$$p(\mathbf{f}|\mathbf{y}) \propto \int d\mathbf{g} p(\mathbf{f}, \mathbf{g}, \mathbf{y})$$

Thus, we can compute posterior over $\mathbf{f}$ given noisy observation $\mathbf{y}$.

Inference in GP III

Assuming that the data **are really sampled from** the GP we are using, given

$$\text{function } f \sim \mathcal{GP}$$
$$\mathbf{f} \sim \mathcal{N}(\mu, K)$$
$$y_i | f_i \sim \mathcal{N}(f_i, \sigma_n^2)$$

We can do inference for test points X_*.

$$p(\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix}) = \mathcal{N}(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 \mathbb{I} & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix})$$

Using the previous formula, we can calculate the conditional distritbuion $p(\mathbf{f}_* | \mathbf{y})$.

Hyper parameters I

Now if we specify the mean and covariance function, we can then do inference of the test points given the observed value $p(\mathbf{f}_*|\mathbf{y})$. The natural question to ask is

- How to determine the mean and covariance function?
- What implication we are making if we specify the mean and covariance function?

Consider an example of Bayesian linear regression,

$$f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + b, \quad \mathbf{w} \sim \mathcal{N}(0, \sigma_w^2 \mathbb{I}), b \sim \mathcal{N}(0, \sigma_b^2)$$

$$\begin{aligned} \mathsf{cov}(f_i, f_j) &= \langle f_i, f_j \rangle - \langle f_i \rangle \langle f_j \rangle \\ &= \langle \mathbf{w}^\top \mathbf{x}_i + b, \mathbf{w}^\top \mathbf{x}_j + b \rangle \\ &= \sigma_w^2 \mathbf{x}_i^\top \mathbf{x}_j + \sigma_b^2 = k(\mathbf{x}_i, \mathbf{x}_j) \end{aligned}$$

This means the Bayesian linear regression is equivalent to the $k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_w^2 \mathbf{x}_i^\top \mathbf{x}_j + \sigma_b^2$ kernel.