Posterior Gaussian Process

Carl Edward Rasmussen

October 13th, 2016

Key concepts

- we are not interested in random functions
- we want to *condition* on the training data
- when both prior and likelihood are Gaussian, then
    - posterior is a Gaussian process
    - predictive distributions are Gaussian
- pictorial representation of prior and posterior
- interpretation of predictive equations

Gaussian Process Inference

Recall Bayesian inference in a parametric model.
The posterior is proportional to the prior times the likelihood.
The predictive distribution is the predictions marginalized over the parameters.
How does this work in a Gaussian Process model?
Answer: in our non-parametric model, the "parameters" are the function itself!

Non-parametric Gaussian process models

In our non-parametric model, the "parameters" are the function itself!
Gaussian likelihood, with noise variance σ^2_{noise}

$$p(\mathbf{y}|\mathbf{x}, f, \mathcal{M}_i) \sim \mathcal{N}(\mathbf{f},\ \sigma^2_{\text{noise}} I),$$

Gaussian process prior with zero mean and covariance function k

$$p(f|\mathcal{M}_i) \sim \mathcal{GP}(m \equiv 0,\ k),$$

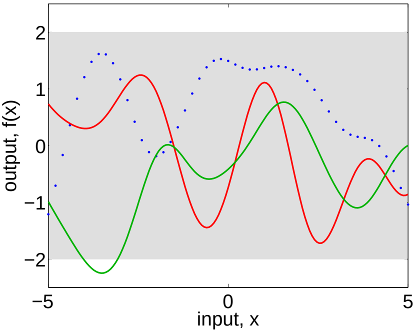
Leads to a Gaussian process posterior

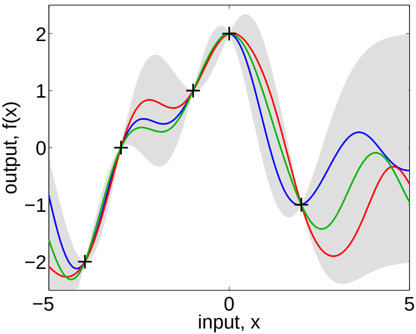
$$p(f|\mathbf{x}, \mathbf{y}, \mathcal{M}_i) \sim \mathcal{GP}(m_{\text{post}},\ k_{\text{post}}),$$
$$\text{where} \begin{cases} m_{\text{post}}(x) = k(x, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1}\mathbf{y}, \\ k_{\text{post}}(x, x') = k(x, x') - k(x, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1} k(\mathbf{x}, x'), \end{cases}$$

And a Gaussian predictive distribution:

$$p(y_*|x_*, \mathbf{x}, \mathbf{y}, \mathcal{M}_i) \sim \mathcal{N}\big(\mathbf{k}(x_*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1}\mathbf{y},$$
$$k(x_*, x_*) + \sigma^2_{\text{noise}} - \mathbf{k}(x_*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1}\mathbf{k}(x_*, \mathbf{x})\big).$$

Prior and Posterior





Predictive distribution:

$$p(y_*|x_*, \mathbf{x}, \mathbf{y}) \; \sim \; \mathcal{N}\Big(\mathbf{k}(x_*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1} \mathbf{y},$$
$$k(x_*, x_*) + \sigma^2_{\text{noise}} - \mathbf{k}(x_*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1} \mathbf{k}(x_*, \mathbf{x})\Big)$$

Some interpretation

Recall our main result:

$$f_* | x_*, \mathbf{x}, \mathbf{y} \sim \mathcal{N}\big(K(x_*, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1}\mathbf{y},$$
$$K(x_*, x_*) - K(x_*, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1} K(\mathbf{x}, x_*)\big).$$

The mean is linear in two ways:

$$\mu(x_*) = k(x_*, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1}\mathbf{y} = \sum_{n=1}^{N} \beta_n y_n = \sum_{n=1}^{N} \alpha_n k(x_*, x_n).$$

The last form is most commonly encountered in the kernel literature.
The variance is the difference between two terms:

$$V(x_*) = k(x_*, x_*) - \mathbf{k}(x_*, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1}\mathbf{k}(\mathbf{x}, x_*),$$

the first term is the *prior variance*, from which we subtract a (positive) term, telling how much the data $\mathbf{x}$ has explained.
Note, that the variance is independent of the observed outputs $\mathbf{y}$.