Posterior Gaussian Process

Carl Edward Rasmussen

October 17th, 2022

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!
The joint distribution

$$p(f, \mathbf{y}) = p(f)\, p(\mathbf{y}|f) = p(\mathbf{y})p(f|\mathbf{y})$$
$$\implies \mathcal{N}(f|\mathbf{m},\ k)\, \mathcal{N}(\mathbf{y}|f) = Z_{|\mathbf{y}}\mathcal{N}(f|\mathbf{m}_{|\mathbf{y}}, k_{|\mathbf{y}}).$$

Gaussian process prior with zero mean and covariance function k

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

Gaussian likelihood, with noise variance σ_{noise}^2
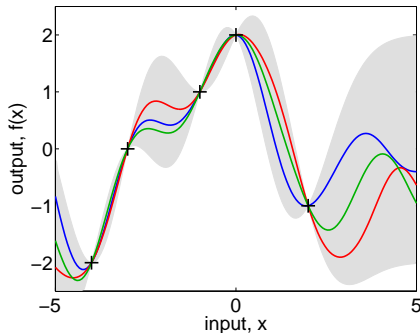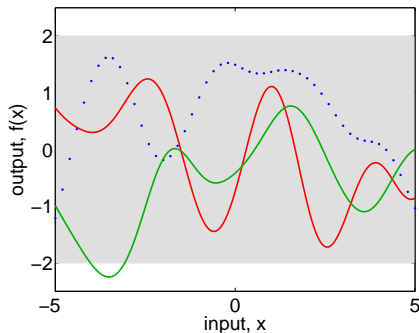
$$p(\mathbf{y}|f, \mathcal{M}_i) \sim \mathcal{N}(f,\ \sigma_{\text{noise}}^2 I),$$

leads to a Gaussian process posterior

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

Prior and Posterior

Predictive distribution:

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

Some interpretation

Recall our main result:

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

The mean is linear in two ways:

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

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

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

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