

GP Marginal Likelihood and Hyperparameters

Carl Edward Rasmussen

October 13th, 2016

Key concepts

- We give an interpretation of the marginal likelihood in terms of
 - a data fit
 - a complexity penalty
- covariance functions can be parameterized using hyperparameters
- hyperparameters can be fit by optimizing the marginal likelihood
 - this is a form of model selection
- Occam's razor is automatic and avoids overfitting

The Gaussian process marginal likelihood

Log marginal likelihood has a closed form

$$\log p(\mathbf{y}|\mathbf{x}, \mathcal{M}_i) = -\frac{1}{2}\mathbf{y}^\top [\mathbf{K} + \sigma_n^2 \mathbf{I}]^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K} + \sigma_n^2 \mathbf{I}| - \frac{n}{2} \log(2\pi)$$

and is the combination of a **data fit** term and **complexity penalty**. Occam's Razor is automatic.

Hyperparameters: properties of covariance functions

The covariance function which we have seen before

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^2),$$

encodes that $f(\mathbf{x})$ and $f(\mathbf{x}')$ have large covariance if \mathbf{x} is **close to** \mathbf{x}' , but it doesn't really quantify what is means by **close to**?

We can parameterize the covariance function using **hyperparameters** such as ℓ , in

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{(\mathbf{x} - \mathbf{x}')^2}{2\ell^2}\right).$$

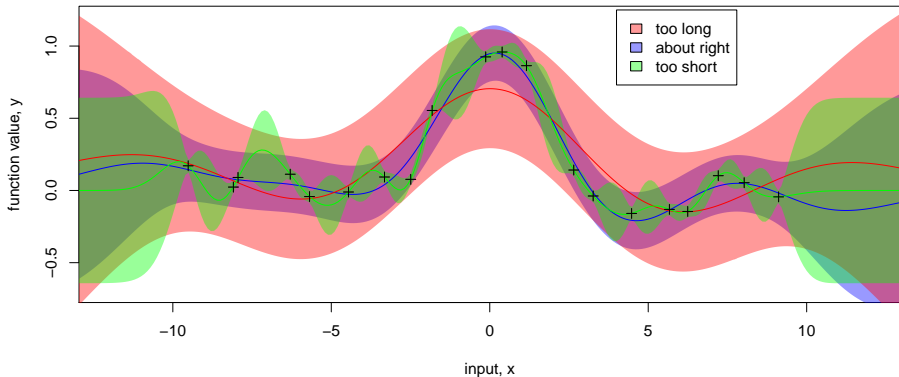
Learning in Gaussian process models involves finding

- the form of the covariance function, and
- any unknown (hyper-) parameters θ .

Example: Fitting the length scale parameter

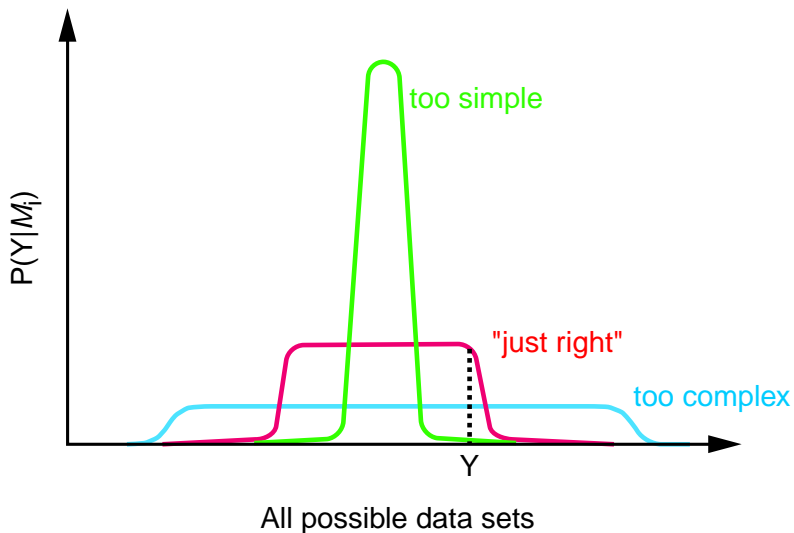
Parameterized covariance function: $k(x, x') = \nu^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right) + \sigma_{\text{noise}}^2 \delta_{xx'}$.

Characteristic Lengthscales



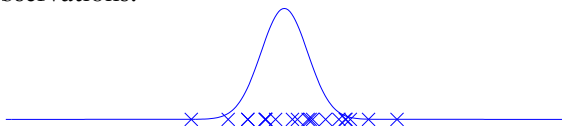
The mean posterior predictive function is plotted for 3 different length scales (the blue curve corresponds to optimizing the marginal likelihood). **Notice, that an almost exact fit to the data can be achieved by reducing the length scale – but the marginal likelihood does not favour this!**

How can Bayes rule help find the right model complexity? Marginal likelihoods and Occam's Razor



An illustrative analogous example

Imagine the simple task of fitting the variance, σ^2 , of a zero-mean Gaussian to a set of n scalar observations.



The log likelihood is $\log p(\mathbf{y}|\mu, \sigma^2) = -\frac{1}{2}\mathbf{y}^\top \mathbf{I} \mathbf{y} / \sigma^2 - \frac{1}{2} \log |\mathbf{I} \sigma^2| - \frac{n}{2} \log(2\pi)$