GP Marginal Likelihood and Hyperparameters

Carl Edward Rasmussen

October 18th, 2022

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

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

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

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

encodes that $f(x)$ and $f(x')$ have large covariance if x is close to 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

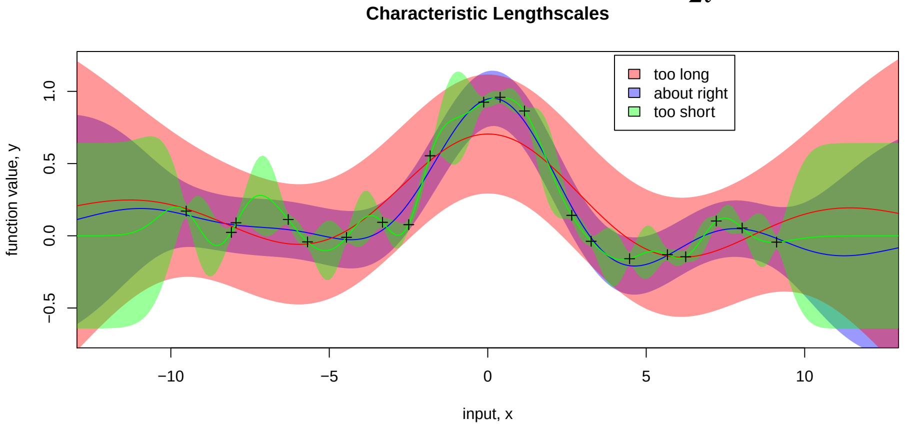
$$k(x, x') = \exp\big(-\frac{(x - x')^2}{2\ell^2}\big).$$

Learning in Gaussian process models involves finding

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

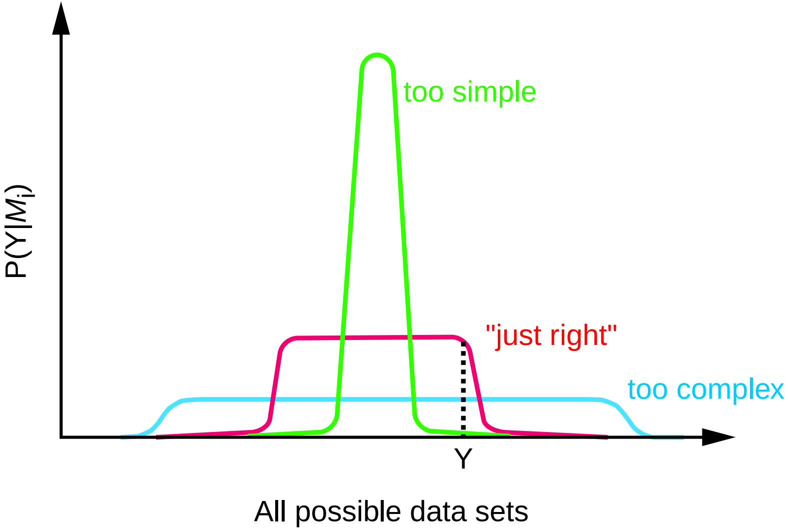
Example: Fitting the length scale parameter

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



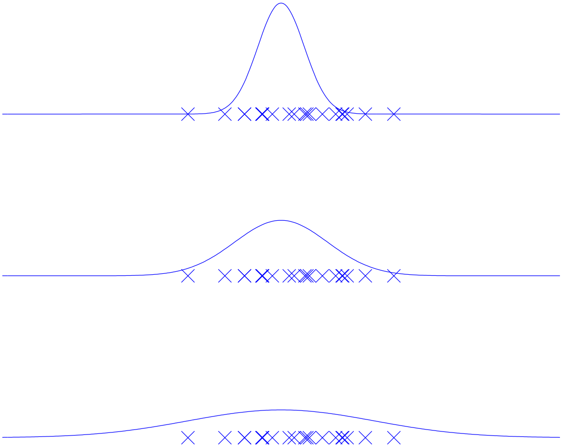
The mean posterior predictive function is plotted for 3 different length scales (the blue curve corresponds to optimizing the marginal likelihood). Notice, that an almost exact fit to the data can be achieved by reducing the length scale – but the marginal likelihood does not favour this!

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