Approximate Hyperparameter Marginalisation for Gaussian Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Enter Abstract here

1 Introduction

2 Gaussian Processes

Gaussian processes (GPs) constitute a powerful method for performing Bayesian inference about functions using a limited set of observations [1]. A GP is defined as a distribution over the functions $f : \mathcal{X} \to \mathbb{R}$ such that the distribution over the possible function values on any finite set of $\mathcal{X}$ is multivariate Gaussian. Given some arbitrary size n dataset, the observations $\mathbf{y} = \{y_1, ..., y_n\}$ could be viewed as a single point sampled from a n-variate Gaussian distribution and can be partnered with a GP.

A GP is completely defined by its first and second moments: a mean function $\mu : \mathcal{X} \to \mathbb{R}$ which describes the overall trend of the function, and a symmetric positive semidefinite covariance function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which describes how function values are correlated as a function of their locations in the domain. Given a function $f : \mathcal{X} \to \mathbb{R}$ which we would like to perform inference about, and a set of input points $\mathbf{x} \subset \mathcal{X}$, the Gaussian process prior distribution over the function values $\mathbf{f} = f(\mathbf{x})$ is given by:

$$
\begin{aligned}
p(\mathbf{f}|\mathbf{x}, \boldsymbol{\theta}, I) &:= \mathcal{N}\left(\mathbf{f}; \mu_\theta(\mathbf{x}), K_\theta(\mathbf{x}, \mathbf{x})\right) && (1) \\
&:= \frac{1}{\sqrt{\det 2\pi K_{\mathbf{f}}}} \exp\left(-\frac{1}{2}(\mathbf{f} - \mu_{\mathbf{f}})^\top K_{\mathbf{f}}^{-1}(\mathbf{f} - \mu_{\mathbf{f}})\right) && (2)
\end{aligned}
$$

where $\boldsymbol{\theta}$ is a vector containing any parameters required by μ and K, and which form the *hyperparameters* of the model, I. There exist a wide variety of mean and covariance functions which can be chosen in order to reflect any prior knowledge available about the function of interest. Note also that we need not place a GP directly on the function, for example a function known to be strictly positive might benefit from a GP over its logarithm.

Once we have observations of the function $(\mathbf{x_s}, \mathbf{y_s})$ we can make predictions about the function value f_* at input x_*. As exact measurements of the function are often not available we assume a noise model, such that:

$$p(y|f, x, \sigma_n^2, I) := \mathcal{N}(y; f, \sigma_n^2) \tag{3}$$

which represents i.i.d Gaussian observation noise with variance σ_n^2. As should be expected, the predictive distribution over f_* is Gaussian:

$$p(f_*|x_*, \mathbf{y_s}, \boldsymbol{\theta}, I) := \mathcal{N}(f_*; m_\theta(f_*|x_*, \mathbf{y_s}), C_\theta(f_*|x_*, \mathbf{y_s})) \tag{4}$$

where the posterior mean and covariance are:

$$m_\theta(f_*|x_*, \mathbf{y_s}) := \mu_\theta(x_*) + K_\theta(x_*, \mathbf{x_s})V^{-1}(\mathbf{y_s} - \mu_\theta(x_*)) \tag{5}$$

$$C_\theta(f_*|x_*, \mathbf{y_s}) := K_\theta(x_*, x_*) - K_\theta(x_*, \mathbf{x_s})V^{-1}K_\theta(\mathbf{x_s}, x_*) \tag{6}$$

$$\text{where} \quad V := K_\theta(\mathbf{x_s}, \mathbf{x_s}) + \sigma_n^2\mathbf{I} \tag{7}$$

2.1 Dealing with Hyperparameters

The previous equations assume that the hyperparameters $\boldsymbol{\theta}$ are known; in fact we can rarely be certain about $\boldsymbol{\theta}$ *a priori* which we can represent by proposing a suitably uninformative prior distribution $p(\boldsymbol{\theta}|I)$. Given such a hyper-prior, the hyperparameters should be marginalised to calculate the predictive distribution over f_*:

$$p(f_*|x_*, \mathbf{y_s}, I) = \frac{\int p(f_*|x_*, \mathbf{y_s}, \boldsymbol{\theta}, I)p(\mathbf{y_s}|\mathbf{x_s}, \boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)d\boldsymbol{\theta}}{\int p(\mathbf{y_s}|\mathbf{x_s}, \boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)d\boldsymbol{\theta}} \tag{8}$$

Unfortunately such integrals are generally non-analytic, but can be well approximated by use of Bayesian Monte Carlo [2] techniques. This involves evaluating our predictions for a range of hyperparameter samples $\{\boldsymbol{\theta_i} : i \in S\}$, with a different mean $m_{\theta_i}(f_*|x_*, \mathbf{y_s})$ and covariance $C_{\theta_i}(f_*|x_*, \mathbf{y_s})$ for each, which are then combined in a weighted mixture:

$$p(f_*|x_*, \mathbf{y_s}, I) \simeq \sum_{i\in S} \rho_i N(f_*; m_{\theta_i}(f_*|x_*, \mathbf{y_s}), C_{\theta_i}(f_*|x_*, \mathbf{y_s})) \tag{9}$$

with weights $\boldsymbol{\rho}$ as detailed in [3]. This approach will give us a close approximation to full marginalisation, but in order to take meaningful number of samples suffers from high computational costs. A far less demanding approach is to choose a single $\boldsymbol{\theta}$ which maximises the marginal likelihood

3 Approximate Hyperparameter Marginalisation

3.1 Proof of Positive Semi-Definiteness

A sum of kernels is itself a kernel, which by definition fulfils the necessary condition of positive semi-definiteness. Therefore:

$$K = \int k(\beta)p(\beta)d\beta \tag{10}$$

4 Experiments

5 Related Work

6 Conclusion

Acknowledgments

Do we have any? Aladdin / Orchid?

References

[1] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press Cambridge, MA, USA, 2006.

[2] C.E. Rasmussen and Z. Ghahramani. Bayesian monte carlo. *Advances in neural information processing systems*, 15:489–496, 2003.

[3] M.A. Osborne, SJ Roberts, A. Rogers, SD Ramchurn, and N.R. Jennings. Towards real-time information processing of sensor network data using computationally efficient multi-output gaussian processes. In *Proceedings of the 7th international conference on Information processing in sensor networks*, pages 109–120. IEEE Computer Society, 2008.