Approximate Hyperparameter Marginalisation for Gaussian Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

In Bayesian methods a common problem is to choose a prior. In inference with Gaussian processes this task could involve choosing kernel function hyperparameters. In practice however, it is often unclear how to make this choice a priori. Therefore, most implementations deviate from proper Bayesian treatment by estimating the hyperparameters from the data via maximum lmarginal-ikelihood methods. In contrast, we propose to add another hierarchy of inference on top of that. In particular, we propose to place a prior distribution over the hyperparameters. Its hyperparemeters can in turn be estimated learned from the data. Since the resuting integrals of the marginalizations are non-analytic we use a Taylor expansion to yield a Gaussian process which approximates the correct process with marginalized hyperparameters. PERHAPS WRITE ABOUT INTEGRAL KERNEL AND RELATION TO RATIONAL QUADRATIC? We conduct experiments illustrating the benefits our approach on artificial as well as on real data.

1 Introduction

2 Gaussian Processes

Gaussian processes (`GPs`) constitute a powerful method for performing Bayesian inference about functions using a limited set of observations [3]. 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 multi-variate Gaussian. A vector of observations $\mathbf{y} = \{y_1, ..., y_n\}$ could be viewed as a single point sampled from a n-variate Gaussian distribution.

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 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 $f : \mathcal{X} \to \mathbb{R}$ about which we wish to perform inference and a set of input points $\mathbf{x} \subseteq \mathcal{X}$, the Gaussian process prior distribution over the function values $\mathbf{f} = f(\mathbf{x})$ is given by:

$$p(\mathbf{f}|\mathbf{x}, \boldsymbol{\theta}, I) := \mathcal{N}(\mathbf{f}; \mu_\theta(\mathbf{x}), K_\theta(\mathbf{x}, \mathbf{x})) \tag{1}$$

$$:= \frac{1}{\sqrt{\det 2\pi K_{\mathbf{f}}}} \exp\left(-\frac{1}{2}(\mathbf{f} - \mu_{\mathbf{f}})^\top K_{\mathbf{f}}^{-1}(\mathbf{f} - \mu_{\mathbf{f}})\right) \tag{2}$$

where $\boldsymbol{\theta}$ is a vector containing any parameters required by μ and K: the *hyperparameters* of the model, I. Due to the ubiquity of I we henceforth drop it from explicit representation for notational convenience. 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.

Once we have observations of the function $(\mathbf{x_s}, \mathbf{y})$ 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:

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

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

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

where the posterior mean and covariance are:

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

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

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

2.1 Marginalising out the Hyperparameters

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

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

Unfortunately, such integrals are generally non-analytic, requiring numerical approximation. Randomized Monte Carlo techniques [1] form the most popular approaches to numerical integration, although Bayesian alternatives [4, 2] also exist. These techniques estimate the integral given the value of the integrand on a set of sample points, usually via a weighted mixture

$$p(f_*|x_*, \mathbf{y}) \simeq \sum_i \rho_i p(f_*|x_*, \mathbf{y}, \boldsymbol{\theta}_i) \tag{9}$$

for some weight vector $\boldsymbol{\rho}$. Unfortunately, the computational expense of evaluating the integrand at sufficient samples to estimate high-dimensional integrals is often prohibitive. As a consequence, such approaches are rarely used for the marginalisation of GP hyperparameters.

A less computationally demanding alternative is to select only a single sample. Type II maximum likelihood, or maximum marginal likelihood, approximates as

$$p(f_*|x_*, \mathbf{y}) \simeq p(f_*|x_*, \mathbf{y}, \boldsymbol{\theta}_{\mathrm{ML}}) \tag{10}$$

where $\boldsymbol{\theta}_{\mathrm{ML}}$ is the hyperparameter vector that maximises the marginal likelihood,

$$\boldsymbol{\theta}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \;\; p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\,. \tag{11}$$

As per Figure ****, (10) is equivalent to approximating the likelihood $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ as the delta function $\delta(\theta - \theta_{\mathrm{ML}})$. This assumption is a poor representation of our true state of ignorance given only low numbers of data, for which the likelihood is typically broad and/or multi-modal. As a consequence, type II maximum likelihood can often lead to over-fitting. Nonetheless, the less onerous computational requirements of this approach have made it ubiquitous throughout machine learning.

++*+*+*+*+*+ TEXT NEEDED!!! - Insert comment about what our paper does as a gap between the two methods just discussed *+*+*+*+*+*+*+*+

3 Approximate Hyperparameter Marginalisation

In this section we introduce an approximate method for performing inference using a Gaussian Process which accommodates the inherent uncertainty one should have about the value of hyperparameters, but which does not suffer from excessive computational expense. Rather than attempt to approximate the integral in (8), we present a philosophical shift away from the standard GP prior opting instead to develop a new prior which incorporates and marginalises the uncertainty in the hyperparameters *before* fusing the prior with observations. The result is an adjusted covariance function which can be used interchangeably with a standard covariance function in GP inference, but whose parameters are now *hyper-hyperparameters* explaining the marginalised covariance function.

3.1 The Squared Exponential Covariance Function

One of the most pervasive covariance functions used in Bayesian inference is the *Squared Exponential*, or *Gaussian* kernel:

$$K(x, x') \tag{12}$$

$L = \exp(\beta)$

$$p(\beta|\nu, \Lambda) = \mathcal{N}(\beta; \nu, \Lambda) \tag{13}$$

$$= \frac{1}{\sqrt{2\pi\Lambda}} \exp\left(\frac{-(\beta-\nu)^2}{2\Lambda}\right) \tag{14}$$

$$K_\beta = \exp\left(-\ln\frac{1}{K_\beta}\right) \tag{15}$$

$$= \exp(-A) \tag{16}$$

$$A \approx \ln\frac{1}{K_\beta}\bigg|_\nu + (\beta-\nu)\frac{\partial A}{\partial\beta}\bigg|_\nu + \frac{1}{2}(\beta-\nu)^2\frac{\partial^2 A}{\partial^2\beta}\bigg|_\nu \tag{17}$$

$$K_\beta = h^2 \exp\left(-\frac{1}{2} \triangle^2 \exp(-2\beta)\right) \tag{18}$$

$$K'_\beta = K_\beta \left(\triangle^2 \exp(-2\beta)\right) \tag{19}$$

$$K''_\beta = K_\beta \left(\triangle^4 \exp(-4\beta) - \triangle^2 \exp(-2\beta)\right) \tag{20}$$

$$A = \ln\frac{1}{K_\beta} \tag{21}$$

$$\frac{\partial A}{\partial\beta} = -\frac{K'_\beta}{K_\beta} = \ \triangle^2 \exp(-2\beta) \tag{22}$$

$$\frac{\partial^2 A}{\partial^2\beta} = -\frac{K''_\beta}{K_\beta} + \frac{{K'_\beta}^2}{K_\beta^2} = \ 2\triangle^2 \exp(-2\beta) \tag{23}$$

$C = \triangle^2 \exp(-2\nu)$

$$K_\beta \approx \exp\left(-\left(\ln\frac{1}{K_\nu} + (\beta-\nu)C + (\beta-\nu)^2 C\right)\right) \tag{24}$$

$$= K_\nu \ \exp\left((\beta-\nu)C + (\beta-\nu)^2 C\right) \tag{25}$$

$$K_{\nu,\Lambda} = \int_{-\infty}^{+\infty} K_\beta \ p(\beta|\nu,\Lambda) d\beta \tag{26}$$

$$= K_\nu \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\Lambda}} \exp\left(\frac{-(\beta-\nu)^2}{2\Lambda}\right) \left((\beta-\nu)C + (\beta-\nu)^2 C\right) \tag{27}$$

$$= K_\nu \ \exp\left(\frac{\Lambda C^2}{2(1+2\Lambda C)}\right) \frac{1}{\sqrt{1+2\Lambda C}} \tag{28}$$

3.2 Proof of Positive Semi-Definiteness

++*+*+ I'll tidy this up!

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

$$K_{\nu,\Lambda} = \int K_\beta \; p(\beta) d\beta \tag{29}$$

should be a legitimate kernel if K_β is also a legitimate kernel as $p(\beta)$ just weights the contents of the integral.

$$K_\beta = K_\nu \; \exp\left((\beta-\nu)C + (\beta-\nu)^2 C\right) \tag{30}$$

completing the square:

$$K_\beta = K_\nu \; \exp\left(-\frac{1}{2}(\beta-\nu-1)^2 C\right) \exp\left(\frac{1}{2}C\right) \tag{31}$$

$$\tag{32}$$

Product of kernels is also a kernel. Therefore if all three parts of the above equation are kernel, then K_β is also a covariance function. First two parts are kernels, the last part isn't. However:

$$K_\nu = h^2 \exp\left(-\frac{1}{2} \triangle^2 \exp(-2\nu)\right) \tag{33}$$

$$= h^2 \exp\left(-\frac{C}{2}\right) \tag{34}$$

$$K_\nu \exp\left(\frac{1}{2}C\right) = h^2 \exp\left(-\frac{C}{2}\right) \exp\left(\frac{1}{2}C\right) \tag{35}$$

$$= h^2 \exp(0) \tag{36}$$

so K_β is a kernel.

++*+*+ I need to check this, as the result has changed since I did my substitution...

4 Experiments

5 Related Work

6 Conclusion

Acknowledgments

Do we have any? Aladdin / Orchid?

References

[1] M.H. Chen, Q.M. Shao, and J.G. Ibrahim. *Monte Carlo methods in Bayesian computation*. Springer Verlag, 2000.

[2] M.A. Osborne, R. Garnett, S.J. Roberts, C. Hart, S. Aigrain, N.P. Gibson, and S. Aigrain. Bayesian quadrature for ratios. *Fifteenth International Conference on Artificial Intelligence and Statistics*, 2012.

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

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