
Approximate Hyperparameter Marginalisation for Gaussian Processes

Anonymous Author(s)

Affiliation

Address

email

Abstract

Jan wrote; 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 likelihood 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 hyperparameters can in turn be estimated learned from the data. Since the resulting 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 $f : \mathcal{X} \rightarrow \mathbb{R}$ such that the distribution over the possible function values on any finite set of \mathcal{X} is multi-variate Gaussian. A vector of observations $\mathbf{y} = \{y_1, \dots, y_n\}$ 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 $\mu : \mathcal{X} \rightarrow \mathbb{R}$, which describes the overall trend of the function, and a positive semidefinite covariance function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ which describes how function values are correlated as a function of their locations in the domain. Given a function $f : \mathcal{X} \rightarrow \mathbb{R}$ 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:

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

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

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.

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:

$$K(x, x') = \quad (12)$$

$$L = \exp(\beta)$$

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

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

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

$$= \exp(-A) \quad (16)$$

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

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

$$K'_\beta = K_\beta (\Delta^2 \exp(-2\beta)) \quad (19)$$

$$K''_\beta = K_\beta (\Delta^4 \exp(-4\beta) - \Delta^2 \exp(-2\beta)) \quad (20)$$

$$A = \ln \frac{1}{K_\beta} \quad (21)$$

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

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

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

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

$$= K_\nu \exp((\beta - \nu)C + (\beta - \nu)^2 C) \quad (25)$$

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

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

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

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:

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

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

$$K_{\beta} = K_{\nu} \exp((\beta - \nu)C + (\beta - \nu)^2 C) \quad (30)$$

completing the square:

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

$$(32)$$

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:

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

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

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

$$= h^2 \exp(0) \quad (36)$$

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.