International Journal of Approximate Reasoning 78 (2016) 153–171

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A prior near-ignorance Gaussian process model for nonparametric regression

Francesca Mangili

Istituto Dalle Molle di studi sull'Intelligenza Artificiale (IDSIA), Scuola universitaria professionale della Svizzera italiana (SUPSI), Università della Svizzera italiana (USI), Switzerland

ARTICLE INFO

Article history:
Received 5 December 2015
Received in revised form 30 May 2016
Accepted 10 July 2016
Available online 15 July 2016

Keywords:
Gaussian process
Prior near-ignorance
Nonparametric regression
Hypothesis testing
Bayesian nonparametrics

ABSTRACT

This paper proposes a prior near-ignorance model for regression based on a set of Gaussian Processes (GP). GPs are natural prior distributions for Bayesian regression. They offer a great modeling flexibility and have found widespread application in many regression problems. However, a GP requires the prior elicitation of its mean function, which represents our prior belief about the shape of the regression function, and of the covariance between any two function values.

In the absence of prior information, it may be difficult to fully specify these infinite dimensional parameters. In this work, by modeling the prior mean of the GP as a linear combination of a set of basis functions and assuming as prior for the combination coefficients a set of conjugate distributions obtained as limits of truncate exponential priors, we have been able to model prior ignorance about the mean of the GP. The resulting model satisfies translation invariance, learning and, under some constraints, convergence, which are desirable properties for a prior near-ignorance model. Moreover, it is shown in this paper how this model can be extended to allow for a weaker specification of the GP covariance between function values, by letting each basis function to vary in a set of functions.

Application to hypothesis testing has shown how the use of this model induces the capability of automatically detecting when a reliable decision cannot be made based on the available data.

1. Introduction

A Gaussian Process (GP) extends multivariate Gaussian distributions to infinite dimensionality, thus defining a distribution over functions. Therefore, it is a natural prior distribution in Bayesian analysis for learning an unknown real-valued function $f(x)$ from a set of noisy data. GPs have found widespread use in different application domains such as classification, regression etc. [1–6]. The reason of such success can be attributed to the great modelling flexibility of GPs, which are often used in situations where little is known about $f(x)$.

The probabilistic formulation of GPs and the simple closed form expression of their posterior inferences, makes them a good starting point to develop prior near-ignorance models for nonparametric regression. There are many techniques other than GPs available for nonparametric regression, e.g., splines, relevance vector machines, kernel smoothers, etc., some of which share strong analogies with Gaussian Processes [4, Sec. 6]. Their relative strengths and weaknesses w.r.t. GPs are

E-mail address: francesca@idsia.ch.

http://dx.doi.org/10.1016/j.ijar.2016.07.005
0888-613X/

discussed in [4, Sec. 7]. As they are all precise methods, we can expect them to suffer from the same weaknesses outlined below for the precise GPs.

A GP is completely specified by its mean function (encoding our prior belief about the shape of the regression function) and its kernel $k(x, x')$, used to define the covariance between any two function values: $Cov(f(x), f(x')) = k(x, x')$. A multitude of possible families exists for the covariance function, including squared exponential, polynomial, periodic, etc. (see [4]), among which the squared exponential family is by far the most popular. In the absence of prior knowledge, it can be difficult to make well grounded choices about the mean function and the kernel. A solution, proposed, among others, in [4, Ch. 2.7] to allow for a weaker specification of the prior mean function, is to use a linear combination $\mathbf{h}(x)\boldsymbol{w}$ of a set of fixed basis functions $\mathbf{h}(x) = [h_1(x), \ldots, h_p(x)]$ whose coefficients $\boldsymbol{w} = [w_1, \ldots, w_p]^T$ are assumed to have an improper uniform prior distribution. Such prior belongs to the family of the so-called *non-informative* priors, which are commonly used in objective Bayesian analysis based on the fact that they satisfy some desirable invariance property, like, for instance, translation invariance. However, the improper uniform prior is just one among the priors presented in [7], all of which verify translation invariance and conjugacy with the likelihood of the GP regression model. Choosing a different prior in this family would lead to different posterior inferences. Therefore, the choice of the improper uniform prior should not be considered fully uninformative.

A way to remove this arbitrariness in the choice of the prior is to use a set of prior distributions, rather than a single distribution, and to update each of them by Bayes rule, producing a set of posterior distributions. This approach proceeds after Bayesian sensitivity analysis or Bayesian robustness [8], but with a different viewpoint, as it does not assume the existence of a correct, although unknown, prior distribution. Instead, following the theory of imprecise probabilities or coherent lower (and upper) previsions [9,10], only upper and lower bounds for the posterior inferences of interest (expressed as expectations) are retained as valid representation of our state of knowledge. In lack of prior knowledge, to reflect this state of prior ignorance, the set of priors $\mathcal{M}$ should be as large as possible to be *vacuous* for the inferences of interest, i.e., it should provide upper and lower bounds that encompass all admissible values of such inferences. On the other side, it has to be small enough to guarantee learning from a finite number of observations. As prior ignorance and learning from data are usually conflicting properties [9,11,12], prior ignorance is actually required only for a limited number of basic inferences, thus modeling a state of *near*-ignorance.

In this work, we show that a regression model verifying prior near-ignorance and learning can be obtained by assuming for $\boldsymbol{w}$ the set of priors $\mathcal{M}$ presented in [7], which includes finitely additive probabilities obtained as limits of truncated exponential functions. We call this model an Imprecise GP (IGP). This set of priors can be interpreted as the set of all GPs with fixed kernel $k(x, x')$ and mean function free to vary in the set of all possible linear combinations of the set of basis functions $\boldsymbol{h}(x)$. This model improves with respect to a precise GP prior, as it models prior ignorance about the mean of the Gaussian process, i.e., about the value of the regression function. Moreover, it verifies translation invariance and, under some assumptions, convergence, which are desirable properties for a prior near-ignorance model, as discussed in [7]. Notice, however, that this model still requires to specify the covariance between any two function values, for which a good amount of prior knowledge is necessary. To address this issue, some preliminary work aimed to weaken the prior specification of the covariance is also presented. It builds on the idea of letting the basis function free to vary in a set of admissible functions, starting from the simple case of an IGP with single basis function free to vary in a set of functions obtained as linear combination of the basis functions in $\mathbf{h}(x)$.

To demonstrate the properties of the proposed approach, the IGP model has been applied to statistical hypothesis testing, focusing on a test for the difference between two regression functions given two samples of noisy observations. A nonparametric Bayesian test for the equality of regression functions based on GPs is described in [13]. In that work it is assumed that the covariates of the two samples cover the same range of values, and the comparison between the regression functions is limited to that range of values, because, having no data outside of it, nothing can be stated about the difference or equality of the two functions in other ranges of values. Instead, using the IGP it is possible to perform the equality test without worrying about the training covariate values, as the imprecise approach is able to identify those instances where the decision is prior dependent and thus it automatically detects when a reliable decision cannot be made.

2. Gaussian process

Consider the regression model

$$y = f(x) + v, \tag{1}$$

where $x \in \mathcal{X} \subseteq \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ and $v \sim \mathcal{N}(0, \sigma_v^2)$ is a white Gaussian noise with variance σ_v^2, and assume that we observe the data (x_i, y_i) for $i = 1, \ldots, n$. Our goal is to employ these observations to make inferences about the unknown function $f(x)$. Following the Bayesian estimation approach, we place a prior distribution on $f(x)$, and employ the observations to compute its posterior distribution; finally we use this posterior to make inferences about $f(x)$. Since $f(x)$ is a function, the Gaussian process is a natural prior distribution for it [3,4]. Formally,