Neurocomputing 133 (2014) 25–31

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Letters

Fast identification algorithms for Gaussian process model

Xia Hong [a,*], Junbin Gao [b], Xinwei Jiang [c], Chris J. Harris [d]

[a] School of Systems Engineering, University of Reading, Reading RG6 6AY, UK
[b] School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia
[c] School of Computer Science, China University of Geosciences, Wuhan 430074, China
[d] Electronics and Computer Science, University of Southampton, Southampton, UK

ARTICLE INFO

Article history:
Received 16 August 2013
Received in revised form
24 October 2013
Accepted 30 November 2013
Communicated by K. Li
Available online 11 January 2014

Keywords:
Gaussian process
Optimization
Kullback–Leibler divergence

ABSTRACT

A class of fast identification algorithms is introduced for Gaussian process (GP) models. The fundamental approach is to propose a new kernel function which leads to a covariance matrix with low rank, a property that is consequently exploited for computational efficiency for both model parameter estimation and model predictions. The objective of either maximizing the marginal likelihood or the Kullback–Leibler (K–L) divergence between the estimated output probability density function (pdf) and the true pdf has been used as respective cost functions. For each cost function, an efficient coordinate descent algorithm is proposed to estimate the kernel parameters using a one dimensional derivative free search, and noise variance using a fast gradient descent algorithm. Numerical examples are included to demonstrate the effectiveness of the new identification approaches.

1. Introduction

Gaussian process (GP) models for regression and classification have received increased attention in the past decade [1–5]. The GP provides a simple, principled probabilistic approach that is fundamentally different from other nonlinear parametric regression models such as neural networks and support vector machine (SVM) [6]. In the Gaussian process regression (GPR) modeling paradigm, a systems output is a data sample drawn from a Gaussian distribution conditional on its input. The functional mapping of the system input/output in GPR is assumed unknown, but is a random function (an infinite dimensional vector), i.e. a process with a specific covariance function of the input. Being a fully probabilistic approach, GPR allows probabilistic inference for the system output to be made by predicting the system output probability density function conditional on a given input. The GPR model has been successfully applied to a wide range of applications, e.g. latent models for dimensionality reduction [3,4] and modeling dynamical systems [5]. Note that similar to any probabilistic approaches GPR is based on some general assumptions which are appropriate for a large class of practical systems. For these systems, GPR can have additional advantage of being able to quantify the uncertainties at the sample level, over many other nonlinear functional based modeling paradigms, e.g. support vector regression.

The predictive output distribution of a GPR model is usually parameterized by a small number of parameters in the covariance function (which can be served by a typical kernel function) as well as the variance of additive noise, which can be regarded as one of the parameters. Typically, for a given data set the estimation of GPR model is in general achieved by maximum log marginal likelihood or just maximum a posteriori estimation [2]. Alternatively the GPR model estimation can be configured as a special type of probability density estimation problem concerning the conditional probability of an output variable. It is therefore a straightforward matter to construct the distance measure based objective functions between the estimated output probability density function (pdf) for a given data set and an assumed true pdf. Recently the Kullback–Leibler (K–L) divergence [7,8] was integrated with GPR model to obtain an alternative GPR parameter estimation criterion [9].

The main computational limitation of regression models based on Gaussian processes is that the associated memory requirements and computational overheads grow in the order of $O(N^3)$ due to the inversion of the covariance matrix. In order to overcome this limitation a wealth of sparse approximations to GP models have been proposed (see [2, Chapter 8; 10]). In [10], most of these have been analyzed via an effective prior placed on a set of latent variables (pseudo-input points) specifically introduced to achieve sparse GP approximation. Alternatively Williams and Seeger [11] proposed a low rank approximation to the kernel matrix and then applied to Gaussian process classification and regression. The idea was to generate a reduced-rank approximation to the kernel matrix using the Nyström method, so that by invoking the matrix

* Corresponding author.
E-mail addresses: x.hong@reading.ac.uk (X. Hong), jbgao@csu.edu.au (J. Gao), ysjxw@hotmail.com (X. Jiang), cjh@ecs.soton.ac.uk (C.J. Harris).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.11.035

inversion lemma, the matrix inversion is only applicable to a sub-matrix of a much smaller size.

In this paper we define a new kernel function which leads to the associated covariance matrix to have an exact rather than approximated low rank. By exploiting this structure property both model parameter estimation and model predictions can be carried out efficiently. The basic idea is that since a kernel function over two data points in the input space is an inner product of their respective mappings in a feature space, an exact low rank covariance matrix can be obtained by explicitly defining a finite dimensional feature space. We then propose that the finite dimensional feature vector be in the form of Gaussian radial basis functions (RBF), each of which is parameterized by a center vector in the input space. Consequently our covariance function has a larger number of parameters, including the variance of noise, RBF width in the feature (kernel) function and a set RBF center vectors, all of which can be learnt using maximum log marginal likelihood or by the recently proposed criterion of maximizing the Kullback–Leibler (K–L) divergence [9]. For each criterion, an efficient coordinate descent algorithm is proposed to estimate the kernel parameters using a derivative free one dimensional search, and the noise variance using a fast gradient descent algorithm.

This paper is organized as follows. Section 2 introduces the Gaussian process regression model and the GPR parameter estimation cost function of maximum log marginal likelihood and K–L divergence based Gaussian process regression model. Section 3 initially introduces the proposed RBF feature vectors and then analyzes the resultant computational cost reduction in parameter estimation, model prediction and marginal likelihood evaluation. Section 4 introduces the proposed fast GPR parameters estimation algorithms based on the concept of coordinate descent, which estimates the noise variance and kernel parameters in turn. Numerical experiments are utilized to illustrate the effectiveness of the proposed algorithm in Section 5, followed by our conclusions in Section 6.

2. Preliminaries

2.1. Gaussian process model

For a given data set $D_N=\{\boldsymbol{x}_n,y_n\}_{n=1}^N$, where $\boldsymbol{x}_n\in\Re^m$, let $X=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N]^{\mathrm{T}}$ denote the observed input data matrix and also input space. $Y=[y_1,\ldots,y_N]^{\mathrm{T}}$ is an observed output vector and is also the output space. Consider a nonlinear mapping $\boldsymbol{\phi}(\boldsymbol{x}):\boldsymbol{x}\in\Re^m\to F$ that may be unknown or even have infinite dimension. A kernel $k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ satisfies the property that it is an inner product in the feature space F as

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\boldsymbol{\phi}(\boldsymbol{x}_i),\boldsymbol{\phi}(\boldsymbol{x}_j)\rangle \tag{1}$$

Typical choices of kernel functions include the radial basis function $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\rho^2)$, $\rho>0$ is the width parameter.

Let $\mathcal{N}(\mu,\Sigma)$ denote the Gaussian distribution with mean μ and covariance Σ. In the classical Gaussian process regression (GPR) model, each sample y_n is generated based on

$$y=f(\boldsymbol{x})+\varepsilon \tag{2}$$

where f is drawn from a (zero-mean) Gaussian process $\boldsymbol{f}\sim\mathcal{N}(\boldsymbol{0},K_{XX})$ which is dependent only on a specific covariance/kernel function $K_{XX}=\{k(\boldsymbol{x}_i,\boldsymbol{x}_j)\}\in\Re^{N\times N}$, and $\varepsilon\sim\mathcal{N}(0,\sigma^2)$. Denote $\boldsymbol{k}_{X\boldsymbol{x}}=[k(\boldsymbol{x}_1,\boldsymbol{x}),\ldots,k(\boldsymbol{x}_N,\boldsymbol{x})]^{\mathrm{T}}\in\Re^N$.

The classical Gaussian process regression approach aims to estimate the predictive distribution $p(y|\boldsymbol{x}^*)$ for any test data $\boldsymbol{x}^*\in X$. Consider a new test observation $\boldsymbol{x}^*$. Under the Gaussian likelihood assumption, it is easy to prove [2] that the estimated predictive distribution conditioned on the given observation is

$$\hat{p}(y|\boldsymbol{x}^*,X,Y)\sim\mathcal{N}(f(\boldsymbol{x}^*),g(\boldsymbol{x}^*)) \tag{3}$$

where

$$f(\boldsymbol{x}^*)=\boldsymbol{k}_{X\boldsymbol{x}^*}^{\mathrm{T}}(K_{XX}+\sigma^2\boldsymbol{I})^{-1}Y, \tag{4}$$

$$g(\boldsymbol{x}^*)=\sigma^2+k(\boldsymbol{x}^*,\boldsymbol{x}^*)-\boldsymbol{k}_{X\boldsymbol{x}^*}^{\mathrm{T}}(K_{XX}+\sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}_{X\boldsymbol{x}^*} \tag{5}$$

with $\boldsymbol{I}$ denoting identity matrix with appropriate dimension.

Specifically let $\boldsymbol{a}=[a_1,\ldots,a_N]^{\mathrm{T}}=(K_{XX}+\sigma^2\boldsymbol{I})^{-1}Y$. The mean of (3) can be written as

$$f(\boldsymbol{x}^*)=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{k}_{X\boldsymbol{x}^*}=\sum_{i=1}^{N}a_ik(\boldsymbol{x}_i,\boldsymbol{x}^*). \tag{6}$$

This form of the prediction exhibits the fact that a GP can be represented in terms of a number of basis functions according to the representer theorem [6].

2.2. Estimation of Gaussian process model

In GPR model estimation, the variance of noise is usually regarded as a parameter and catenated with a small number of parameters in the kernel function. The mostly used criterion for the estimation of GPR model parameters is the marginal likelihood $p(Y|X)$, which is the integral of the likelihood times the prior, given as

$$p(Y|X)=\int p(Y|\boldsymbol{f},X)p(\boldsymbol{f}|X)\,d\boldsymbol{f} \tag{7}$$

and the log marginal likelihood is given by [2] as

$$J^{\mathrm{ML}}=\log\ p(Y|X)=-\frac{1}{2}Y^{\mathrm{T}}(K_{XX}+\sigma^2\boldsymbol{I})^{-1}Y -\frac{1}{2}\log\ \det(K_{XX}+\sigma^2\boldsymbol{I})-\frac{N}{2}\log(2\pi) \tag{8}$$

Alternatively, recently Hong et al. [9] proposed that the cost function of the Kullback–Leibler (K–L) divergence [7] between the true output pdf and its estimator based on both using X as prior and a GPR model can be used for GPR estimation and this is given by

$$KL=\int p(y)\log\frac{p(y)}{\hat{p}(y|X,Y)}\,dy =\int p(y)\log\ p(y)\,dy-\int\log\ \hat{p}(y|X,Y)p(y)\,dy \tag{9}$$

in which the second term $R=\int\log\ \hat{p}(y|X,Y)p(y)\,dy\approx E(\log\ \hat{p}(y|X,Y))$ needs to be maximized. From (3), we have

$$\hat{p}(y|X,Y)=\int\hat{p}(y|\boldsymbol{x},X,Y)p(\boldsymbol{x})\,d\boldsymbol{x}=E(\hat{p}(y|\boldsymbol{x},X,Y)) \approx\frac{1}{N}\sum_{j=1}^{N}\frac{1}{\sqrt{2\pi g(\boldsymbol{x}_j)}}\exp\left(-\frac{(y-f(\boldsymbol{x}_j))^2}{2g(\boldsymbol{x}_j)}\right) \tag{10}$$

where the plug-in estimator for $\int\hat{p}(y|\boldsymbol{x},X,Y)p(\boldsymbol{x})\,d\boldsymbol{x}$ with respect to the true densities $p(\boldsymbol{x})$ was used. We have

$$R\approx J^{\mathrm{KL}} =\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{N}\sum_{j=1}^{N}\frac{1}{\sqrt{2\pi g(\boldsymbol{x}_j)}}\exp\left(-\frac{(y_i-f(\boldsymbol{x}_j))^2}{2g(\boldsymbol{x}_j)}\right)\right) =\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{N}\sum_{j=1}^{N}p_{i,j}\right) \tag{11}$$

where

$$p_{i,j}=\frac{1}{\sqrt{2\pi g_j}}\exp\left(-\frac{e_{i,j}^2}{2g_j}\right). \tag{12}$$

Also, $e_{i,j}=y_i-f_j$. f_i, g_j denote $f(\boldsymbol{x}_i)$ and $g(\boldsymbol{x}_j)$ respectively. Note that in (9)–(11), since nothing is known about the true density $p(y)$ and $p(\boldsymbol{x})$, we approximate KL based on the well known principle of plug in estimator. This can be fully justified since the approximation