Lomonosov Moscow State University
Faculty of Computer Science

Review of matherials on

Gaussian Processes for Machine Learning

Pavel Izmailov

Moscow, 2016

1 Theory

In this section an introduction to Gaussian process theory is provided.

1.1 Gaussian Process

Consider the following definition

Definition 1. *A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

A Gaussian process is completely specified by it's mean function and covariance function. These functions are defined as follows

Definition 2. *Let $f(x)$ be a real-valued Gaussian process. Then the functions*

$$m(x) = \mathbb{E}[f(x)],$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))],$$

are the mean function and the covariance function of the process f respectively.

We will write the Gaussian process as $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$.

1.2 GP-regression

Consider the following task. We have a dataset $\{(x_i, f_i)|i = 1, \ldots, n\}$, generated from a Gaussian process $f \sim \mathcal{GP}(m(x), k(x, x'))$, let $x \in \mathbb{R}^d$. We will denote the matrix comprised of points $x_1, \ldots, x_n$ by $X \in \mathbb{R}^{n \times d}$ and the vector of corresponding values $f_1, ..., f_n$ by $f \in \mathbb{R}^n$. We want to predict the values $f_* \in \mathbb{R}^m$ of this random process at a set of other m points $X_* \in \mathbb{R}^{m \times d}$. The joint distribution of f and f_* is given by

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right),$$

where $K(X, X) \in \mathbb{R}^{n \times n}$, $K(X, X_*) = K(X^*, X)^T \in \mathbb{R}^{n \times m}$, $K(X^*, X^*) \in \mathbb{R}^{m \times m}$ are the matrices comprised of pairwise values of the covariance function k for the given sets.

The conditional distributin

$$f_*|X_*, X, f \sim \mathcal{N}(\hat{m}, \hat{K}),$$

where

$$\mathbb{E}[f_*|f] = \hat{m} = K(X_*, X)K(X, X)^{-1}f,$$

$$\text{cov}(f_*|f) = \hat{K} = K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*).$$

Thus, predicting the values of the Gaussian process at a new data point requires solving a linear system with a matrix of size $n \times n$ and thus scales as $O(n^3)$.

1.2.1 Noisy case

Consider the following model. We now have a dataset $\{(x_i, y_i)|i = 1, \dots n\}$, where $y_i = f(x_i) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma_n)$. This means that we only have access to the noisy observations and not the true values of the process at data points. With the notation and logics similar to the one we used it the previous section we can find the conditional distribution for the values f_* of the process at new points X_* in this case:

$$f_*|y \sim \mathcal{N}(\hat{m}, \hat{K}),$$

$$\mathbb{E}[f_*|y] = \hat{m} = K(X_*, X)^{-1}(K(X, X) + \sigma_n^2 I)^{-1} y,$$
$$\operatorname{cov}(f_*|y) = \hat{K} = K(X_*, X_*) - K(X_*, X)(K(X, X) + \sigma_n^2 I)^{-1} K(X, X_*).$$

1.3 GP-classification

To be wrritten.

1.4 Kernel functions

To be wrritten.

1.5 Hyper-parameter estimation

Bayesian paradigm provides a way of estimating the kernel hyper-parameters of the GP-model through th maximizization of the marginal likelihood of the model. Marginal likelihood is given by

$$p(y|X) = \int p(y|f, X)p(f|X)df,$$

which is the likelihood, marginalized over the hidden values f of the underlying process.

1.6 Theoretical perspectives

To be wrritten.

2 Review of existing methods

It follows from the discussion above, that full Gaussian process regression scales as $O(n^3)$ and thus cannot be applied to big datasets. In this section we will review several approximate methods, that make Gaussian processes practical.

2.1 Methods, based on inducing inputs

Most of the existing methods are based on introducing a set of m function points that are called inducing inputs. Using these inputs one can make approximate predictions of the values of the hidden process at test points with a complexity of $O(nm^3)$ instead of $O(n^3)$.

Consider the following situation. We have a dataset of n examples x_i with corresponding values y_i. We will denote the matrix of pairwise values of the covariance function by K_{nn}. Now we introduce a set of m inducing points. We will denote the corresponding covariance matrix by K_{mm} and the matrices of covariances between the inducing points and training points by K_{nm} and K_{mn}.

The methods mostly differ in the way they use to find the inducing points. The most basic idea is to choose this points at random or by some criterion from the training data, or apply some clustering procedure and use the cluster centers as the inducing points. More advanced methods use variational inference (or it's stochastic variant) to find these points. In this section a description of several methods is be provided.

2.1.1 Variational learning of inducing points

The method discussed here was introduced in [1]. To be written.

2.1.2 Stochastic variational inference

The method discussed here was proposed in [2].

Literature

[1] Titsias M. K. (2009). Variational Learning of Inducing Variables in Sparse Gaussian Processes. In: *International Conference on Artificial Intelligence and Statistics*, pp. 567–574.

[2] Hensman J., Fusi N., Lawrence D. (2013). Gaussian Processes for Big Data. In: *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*.