Gaussian Process and Beyond

Sailun Xu
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
sailunx@andrew.cmu.edu

Kaiji Lu
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
kaijil@andrew.cmu.edu

1 Introduction

In this project, we will be mainly focusing on Gaussian Process for the regression problem. A Gaussian Process is a collection of random variables where any finite number of which have a jointly Gaussain Distribution,

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

where $x \in \mathbb{R}^p$ is an arbitrary dimensional input variable. The mean function $m(x) = \mathbb{E}[f(x)]$ and covariance kernel $k(x, x') = cov(f(x), f(x'))$ characterize the prior mean and wiggliness of function value.

Then for a concrete sample: $\{x_i\}_{i=1}^N$

$$[f(x_1), f(x_2), \cdots, f(x_n)]^T \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$$

where the $N \times N$ covariance matrix $\boldsymbol{K} : K_{ij} = k(x_i, x_j)$, and mean vector $\boldsymbol{\mu}_i = m(x_i)$

Consider the regression model

$$y_i = f(x_i) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Suppose that we observe data $X, Y = \{x_i, y_i\}_{i=1}^n$, and we are given a new test point x_{n+1}
Denote:

$$\boldsymbol{K} : K_{ij} = k(x_i, x_j), 1 \leq i, j \leq n$$
$$k : k_i = k(x_i, x_{n+1}), 1 \leq i \leq n$$
$$\mu : \mu_i = m(x_i)$$

According to [1], the joint distribution could be written as:

$$y_1, \cdots, y_n, y_{n+1} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ m(x_{n+1}) \end{bmatrix}, \begin{bmatrix} K + \sigma^2 I & k \\ k^T & k(x_{n+1}, x_{n+1}) + \sigma^2 \end{bmatrix}\right) \tag{1}$$

We recognize the predictive distribution as a conditional Gaussian Distribution:

$$y_{n+1} \sim \mathcal{N}(k^T (K + \sigma^2 I)^{-1} Y, k(x_{n+1}, x_{n+1}) + \sigma^2 - k^T (K + \sigma^2 I)^{-1} k) \tag{2}$$

2 Plan

However, there are limitations with Gaussian Process that we easily recognize:

- As pointed out by [2], since we need to take matrix inverse as in $(K + \sigma^2 I)^{-1}$, which is $O(n^3)$ complexity, making GP not scalable to larger data sets.

Preprint. Work in progress.

- As pointed out by [3], albeit the choice of kernel could dramatically affect the performance of the model, there is little attention on how to choose one. Typically people would select a fixed kernel a priori, or manually construct a kernel, both of which are rather restricted. Even when learning hyper-parameter of kernels, we are limited to a restrictive family of kernel functions. In fact, finite sample estimates will be affected by the kernel choice notwithstanding the use of an universal approximating kernel.

We will re-examine the above problems and carry out experiments on more data sets following the papers we are going to read.

3 Reading List

We will be first looking at some general Bayesian non-parametric proceedings for reference and background knowledge, e.g.: [4]

For translating between primal and spectral representation of kernel, we will be looking at: [5]

For the scalability of GP, we will be looking at: [2]

For non-parametric learning of kernels, we will be looking at: [3]

References

[1] Larry Wasserman. Nonparametric bayesian methods.

[2] Andrew Gordon Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1775–1784. JMLR.org, 2015.

[3] Arthur Gretton and Christian C. Robert, editors. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2016.

[4] Peter Orbanz. Lecture notes on bayesian nonparametrics.

[5] Andrew Gordon Wilson and Ryan Prescott Adams. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1067–III–1075. JMLR.org, 2013.