Scalable Gaussian Processes: Project Proposal

Flavio Argentieri, Ritwik Vashistha

Gaussian Processes Natural Cubic Time

We are interested in modeling the relationship between a response variable Y, and a finite set of explanatory variables collected in a vector X of dimension d. We consider Y as a random variable and X as a random vector, and we assume that there exists a relationship between them, let's say

$$Y = f(X)$$

Equivalently, we can write $Y = f(X^T)$. We are interested in making inference about f, so we consider it as a random function - i.e. a stochastic process - and put a Gaussian Process prior on it.

$$f(s) \sim GP\left(m(s), k\left(s, s'\right)\right) \tag{1}$$

Gaussian process are very flexible and allow for estimation of many different curves. For the sake of simplicity, we fix the mean function and the kernel as follows

$$\begin{aligned} m(s) &= 0 \\ k\left(s, s'\right) &= \lambda^2 \exp\left\{-\frac{1}{2l^2}\left\|s - s'\right\|^2\right\} \end{aligned}$$

The conditional distribution of y follows naturally

$$Y \mid \{f(s) = \tilde{f}(s), X = x\} \sim \delta_{\tilde{f}(x)}$$

Next, we add noise to our model, so as to allow for noisy measurements of Y. We write

$$Y = f(X) + \varepsilon$$

Where $f(X)$ is the regression function of Y on X conditional on f, and ε is a mean zero normally distributed vector. More explicitly

$$\begin{aligned} &\mathbb{E}\left(Y \mid \{f(s) = \tilde{f}(s), X = x\}\right) = f(x) \\ &\mathbb{E}\left(\varepsilon \mid \{f(s) = \tilde{f}(s), X = x\}\right) = 0 \\ \implies &\mathbb{E}(\varepsilon) = 0 \end{aligned}$$

where the last property holds because of the law of total expectations. By assuming that, conditionally on f and X, the error term is normal with mean zero, we get

$$Y \mid \{f(s) = \tilde{f}(s), X = x\} \sim \mathcal{N}(\tilde{f}(x), \sigma_\varepsilon^2)$$

where σ_ε^2 is the variance of ε.

Next, we assume to have data on Y, and X. In particular, we assume to have an independent and identically distributed training set $\{(y_i, x_i)\,,\, i = 1, \ldots, N\}$. For all $i \in \{1, \ldots, N\}$, we assume

$$Y_i = f\left(X_i^\top\right) + \varepsilon_i$$

and denote

$$\mathbf{Y} := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{X} := \begin{pmatrix} X_1^\top \\ \vdots \\ X_N^\top \end{pmatrix} \quad \mathbf{y} := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{x} := \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix} \quad \mathbf{f} := \begin{pmatrix} f\left(x_1^\top\right) \\ \vdots \\ f\left(x_N^\top\right) \end{pmatrix}$$

We assume to have a single data point in the test set. For such data point we only have the value of $X^* = x^*$, and we are interested in making inference about $f^* := f\left(x^*\right)$. Our assumptions imply the following likelihood

$$\mathbf{Y} \mid \{f(s) = \tilde{f}(s), \mathbf{X} = \mathbf{x}\} \sim \mathcal{N}\left(\tilde{\mathbf{f}}, \sigma_\varepsilon^2 I\right) \tag{2}$$

where

$$\tilde{\mathbf{f}} := \begin{pmatrix} \tilde{f}\left(x_1^\top\right) \\ \vdots \\ \tilde{f}\left(x_N^\top\right) \end{pmatrix}$$

We will work in a probability space where we have conditioned on the values of $\mathbf{X}$. Thus, the quantity we are interested in is

$$p\left(f^* \mid \mathbf{y}, \mathbf{x}, x^*\right) = \int p\left(f^*, \tilde{\mathbf{f}} \mid \mathbf{y}, \mathbf{x}, x^*\right) d\tilde{\mathbf{f}} \tag{3}$$

We are ready to compute our posterior distribution

$$p\left(\tilde{\mathbf{f}}, f^* \mid \mathbf{y}, \mathbf{x}, x^*\right) \propto p\left(\mathbf{y} \mid \tilde{\mathbf{f}}, \mathbf{x}, x^*\right) \cdot p\left(\tilde{\mathbf{f}}, f^* \mid \mathbf{x}, x^*\right) \tag{4}$$

We note that, under our assumptions, $p\left(\mathbf{y} \mid \tilde{\mathbf{f}}, \mathbf{x}, x^*\right) = p\left(\mathbf{y} \mid \tilde{\mathbf{f}}, \mathbf{x}\right) = p\left(\mathbf{y} \mid \tilde{f}, \mathbf{x}\right)$, which is simply the likelihood in (2). Next, we note that a Gaussian Process Prior implies

$$\tilde{\mathbf{f}} \mid \mathbf{x} \sim \mathcal{N}\left(0, K_{\tilde{\mathbf{f}}\tilde{\mathbf{f}}}\right)$$

where the $K_{\tilde{\mathbf{f}}\tilde{\mathbf{f}}}$ is the covariance matrix implied by our Kernel for the vector $\tilde{\mathbf{f}}$. Similarly,

$$f^* \mid x^\star \sim \mathcal{N}\left(0, K_{**}\right)$$

where $K_{**} = \lambda$. We stick to the general notation not to make it dependant on the kernel. The assumptions about our data, then imply

$$\begin{aligned} \tilde{\mathbf{f}} \mid \mathbf{x}, x^* &\sim \mathcal{N}\left(0, K_{\tilde{\mathbf{f}}\tilde{\mathbf{f}}}\right) \\ f^* \mid \mathbf{x}, x^\star &\sim \mathcal{N}\left(0, K_{**}\right) \end{aligned}$$

And our Gaussian Process Prior implies

$$\tilde{\mathbf{f}}, f^* \mid \mathbf{x}, x^* \sim N\left(0, \begin{pmatrix} K_{\tilde{\mathbf{f}}\tilde{\mathbf{f}}} & K_{*\tilde{\mathbf{f}}} \\ K_{\tilde{\mathbf{f}}*} & K_{**} \end{pmatrix}\right) \tag{5}$$

where $K_{\tilde{\mathbf{f}}*}$ is the vector of covariances implied by the kernel. By combining (2) and (5), and plugging them back into (4) and then into (3), the nice properties of normal distributions imply

$$f^* \mid \mathbf{y}, \mathbf{x}, x^* \sim N\left(K_{\tilde{\mathbf{f}}*}\left(K_{\tilde{\mathbf{f}}\tilde{\mathbf{f}}} + \sigma_\varepsilon^2 I\right)^{-1} \mathbf{y}, K_{**} - K_{\tilde{\mathbf{f}}*}\left(K_{\tilde{\mathbf{f}}\tilde{\mathbf{f}}} + \sigma_\varepsilon^2 I\right)^{-1} K_{*\tilde{\mathbf{f}}}\right) \tag{6}$$

The posterior distribution for f^* derived in (6) is exact and comes from a nonparametric, very flexible model. Thus, it is highly attractive. However, actually computing such posterior requires the inversion of a matrix of dimension n, which involves a cubic cost $O(n^3)$. Our project will primarily deal with ways to make the computation of (6) feasible, so as to allow for the use of Gaussian Process priors in problems with a very large sample size.

Literature on Scalable Gaussian Processes

The literature has identified three main approaches that allow to reduce the computational burden of estimating Gaussian Processes (GPs):

- Sparse GPs, as presented by Quiñonero Candela and Rasmussen (2005). Such approach requires to select a set $\mathbf{u}$ of inducing variables of size m and assume conditional independence of $\tilde{\mathbf{f}}$ and f^* given $\mathbf{u}$. This allows to reduce the computation time to $O(N \cdot m^2)$, where $m \ll N$.

- Spatio-temporal GPs, as presented by Särkkä, Solin, and Hartikainen (2013). This approach is applicable when the vector $\mathbf{x}$ includes time, and it requires to rewrite the model as a state space model whose solution is the posterior distribution. Filtering is then used to make inference, so that the computational time is $O(N \cdot d^3)$, where d is the dimensionality of the state-space.

- Variational GPs, as presented by Hensman, Fusi, and Lawrence (2013). This approach allows to exploit the advantages of variational inference and can lead to linear time computations.

Our main focus will be a combination of these approaches, as presented by Hamelijnck et al. (2021). The key insight will be the use of state-space models combined with variational inference, which leads to a logartihmic computation time.

References

Hamelijnck, O., W. J. Wilkinson, Niki A. Loppi, Arno Solin, and Theodoros Damoulas (2021). "Spatio-temporal variational Gaussian processes." In: *Advances in Neural Information Processing Systems* 34, pp. 23621–23633.

Hensman, J., N. Fusi, and N.D. Lawrence (2013). "Gaussian processes for big data." In: *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 282–290.

Quiñonero Candela, J. and C. E. Rasmussen (2005). "A unifying view of sparse approximate Gaussian process regression." In: *Journal of Machine Learning Research* 6, pp. 1939–1959.

Särkkä, S., A. Solin, and J. Hartikainen (2013). "Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing." In: *IEEE Signal Processing Magazine* 30.4, pp. 51–61.