Gaussian Process Models

Sanelma Heinonen, Emanuel Nussli, Niál Perry

ETH StatsLab: April 2023

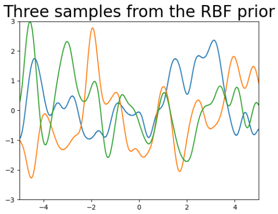
1 Gaussian Process Models - Overview

We begin by defining Gaussian processes from probability theory: A **Gaussian process** is a stochastic process – infinite collection of random variables, often indexed by time or space – such that any finite subset of them has a multivariate normal distribution. Gaussian processes are characterized by a mean function $\mu : \mathcal{X} \to \mathbb{R}$ and a covariance (kernel) function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which together describe the multivariate normal distribution of a finite set $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subseteq \mathcal{X}$.

Gaussian process models are a Bayesian, supervised machine learning method that can be used for regression. As typical with regression, the task is as following: given data $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \in \mathbb{R}^d$ and noisy labels $y_i \in \mathbb{R}$ of the form $y_i = f(\mathbf{x}_i) + \epsilon_i$ where ϵ_i is independent Gaussian noise $\epsilon_i \backsim \mathcal{N}(0, \sigma^2)$, estimate the function f.

Gaussian process models assume a Gaussian process as the prior for f: For any $A = \{\mathbf{x}_1, ..., \mathbf{x}_n\} \subseteq X$, $\mathbf{f}_A \backsim \mathcal{N}(0, \mathbf{K}_A)$, where $\mathbf{f}_A = [f(\mathbf{x}_1), ..., f(\mathbf{x}_n)]^\top$. $\mathbf{K}_A \in \mathbb{R}^{n \times n}$ is called the covariance (or gram or kernel) matrix of $\mathbf{f}_A$ and its elements are defined as $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the covariance (kernel) function of the a priori Gaussian process.

We can show using properties of the Gaussian distribution that the posterior distribution of $f|X, y$ is also Gaussian. This leads to a convenient hypothesis testing of the predictions. For each prediction, we are able to obtain a confidence interval.



2 Kernels & fitting kernels

Kernels are functions $k(\mathbf{x}_1, \mathbf{x}_2) = \mathrm{Cov}[f(\mathbf{x}_1), f(\mathbf{x}_2)]$ satisfying two conditions:

- $k(\mathbf{x}_1, \mathbf{x}_2) = k(\mathbf{x}_2, \mathbf{x}_1) \;\; \forall \mathbf{x}_1, \mathbf{x}_2 \;\in \mathcal{X}$ (symmetry)
- $\mathbf{K}_A$ is positive semi-definite for any $A \subseteq \mathcal{X}$

These conditions guarantee that $\mathbf{K}_A$ is a valid covariance matrix. The kernel determines the shape of the functions produced by the Gaussian process model. Intuitively, if the kernel deems two points $\mathbf{x}_i, \mathbf{x}_j$ to be 'similar', then we expect the output of f to be similar at those points too. When fitting a Gaussian Process, one thus *selects* a kernel based on one's beliefs about the shape of f. E.g.,

- Radial basis function (aka Gaussian or squared exponential): $k(\mathbf{x}_1, \mathbf{x}_2; h) = \exp\left(-\frac{||\mathbf{x}_1 - \mathbf{x}_2||_2^2}{2h^2}\right)$. Models smooth, ∞-diff'ble f (above plot)
- Laplace kernel (aka exponential): $k(\mathbf{x}_1, \mathbf{x}_2; h) = \exp\left(-\frac{||\mathbf{x}_1 - \mathbf{x}_2||_1}{h}\right)$. Models non-smooth but continuous f

One can also create new kernels from existing kernels k_1, k_2. A linear combination $\alpha k_1 + \beta k_2$ with $\alpha, \beta \geq 0$ constitutes a valid kernel, as do certain non-linear functions of k_1, k_2.

3 Applications and drawbacks

Gaussian process models are flexible, relatively easy to implement, and fully probabilistic (which enables uncertainty estimates). This makes them powerful tools. A common use case for Gaussian process models is geographic data (such as air pollution, weather, or soil quality).

One of the biggest drawbacks to Gaussian process models is computational demands. Gaussian process models require $\mathcal{O}(n^3)$ running time in the number of training data points. One way to overcome this is by using *sparse approximation*. The idea is to introduce $m \ll n$ inducing variables: latent variables $\mathbf{u} = [u_1, \ldots, u_m]^\top$ which are values of the Gaussian process corresponding to a set of input locations $X_\mathbf{u}$, which we call *inducing inputs*. Then we assume that $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top$ and $\mathbf{f}^* = [f(\mathbf{x}_1^*), \ldots, f(\mathbf{x}_b^*)]^\top$ (a vector of b arbitrary test points in $\mathcal{X}$) are *conditionally independent* given $\mathbf{u}$. This gives rise to methods for approximating the joint prior $p(\mathbf{f}, \mathbf{f}^*)$, and thus reduces time complexity to $\mathcal{O}(nm^2)$. See [7] for more details.

References

[1] A. Krause. Probabilistic artificial intelligence. ETH Zurich, September 2022. Lecture Notes.

[2] K. Bailey. Gaussian processes for dummies. *Towards Data Science*, 2016.

[3] M. Krasser. Sparse gaussian processes. *GitHub*, 2020.

[4] M. Kuss. *Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning*. PhD thesis, Technische Universität Darmstadt, 2006.

[5] Y. Natsume. Gaussian process models. *Towards Data Science*, 2021.

[6] O. Stegle and K.Borgwardt. An introduction to gaussian processes. Universität Tübingen. Lecture Notes.

[7] J. Quiñonero-Candela and C.E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.