Gaussian Processes for Machine Learning Chapter 4 Summary*

CSCI-GA 3033 Bayesian Machine Learning
New York University
Yilun Kuang

Oct 26, 2022

4.0 Introduction

"At the heart of every Gaussian process model - controlling all the modeling power - is a covariance kernel" (Andrew Wilson's Thesis, p. 39). A covariance kernel $k(\mathbf{x}, \mathbf{x}')$ encodes inductive biases into our model. Indeed, if a neural network / brain $f \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ with SSL representations, then we would expect the covariance kernel to encodes notions of similarity in the representations.

4.1 Preliminaries

A stationary covariance function is a function of $\tau := \mathbf{x} - \mathbf{x}'$ and an isotropic covariance function is a function of $|\mathbf{x} - \mathbf{x}'|$. The kernel arises in the theory of integral operators: $(T_k f)(\mathbf{x}) = \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') d\mu(\mathbf{x}')$.

4.1.1 Mean Square Continuity and Differentiability

For a sequence of points $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$ which converges to a fixed point $\mathbf{x}_* \in \mathbb{R}^D$, a stochastic process $f(\mathbf{x})$ is said to be continuous in mean squared at $\mathbf{x}_*$ if $\lim_{k \to \infty} \mathbb{E}[|f(\mathbf{x}_k) - f(\mathbf{x}_*)|^2] = 0$.

4.2 Examples of Covariance Functions

4.2.1 Stationary Covariance Functions

By Bochner's Theorem, the spectral density / power spectrum and the stationary kernel k are Fourier duals

$$k(\tau) = \int S(s) e^{2\pi i s^\top \tau} ds S(s) = \int k(\tau) e^{-2\pi i s^\top \tau} d\tau$$

So the spectral density determines the inductive biases we're encoding in our models. Some example isotropic kernels include the Squared Exponential (SE) kernel (also called the RBF kernel), Matérn class of covariance functions, OU process, Rational Quadratic Kernel etc.

4.2.2 Dot Product Covariance Functions

Kernels in the form of $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x} \cdot \mathbf{x}'$ are dot product covariance functions.

4.2.3 Other Non-Stationary Covariance Functions

NNGP kernel for an infinitely-wide neural network is given by

$$\mathbb{E}_{\mathbf{w}}[f(\mathbf{x}) f(\mathbf{x}')] = \sigma_b^2 + N_H \sigma_v^2 \mathbb{E}_{\mathbf{u}}[h(\mathbf{x}; \mathbf{u}) h(\mathbf{x}'; \mathbf{u})]$$

*Combined with Andrew Wilson's Thesis Chapter 2.4

4.3 Eigenfunction Analysis of Kernels

A function $\phi(\cdot)$ is called an eigenfunction of the kernel k with eigenvalue λ if $\int k(\mathbf{x}, \mathbf{x}')\phi(\mathbf{x})d\mu(\mathbf{x}) = \lambda\phi(\mathbf{x}')$.

By the Mercer's Theorem, we have the following decomposition of kernels into components in infinite basis: $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x})\phi_i^*(\mathbf{x}')$

The eigenfunctions can be approximated using common methods in numerical analysis regarding the eigenvalue problems.

4.4 Kernels for Non-Vectorial Inputs

Kernels can also be defined over structured object. String kernels and Fisher kernels are examples.