Efficient Probabilistic Machine Learning Using Newton’s Identities

Kishan Sharma, supervised by Henry Moss

Motivation for using Gaussian processes

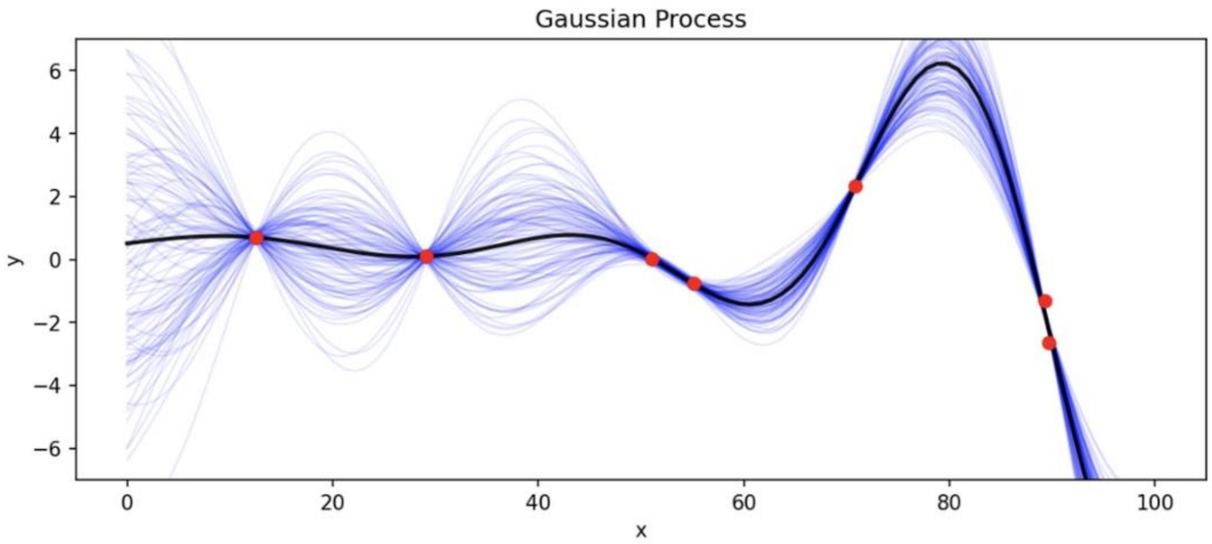


Image source: https://medium.com/geekculture/what-is-gaussian-process-intuitive-explaination-fcee3c78c587

Definition of Gaussian processes

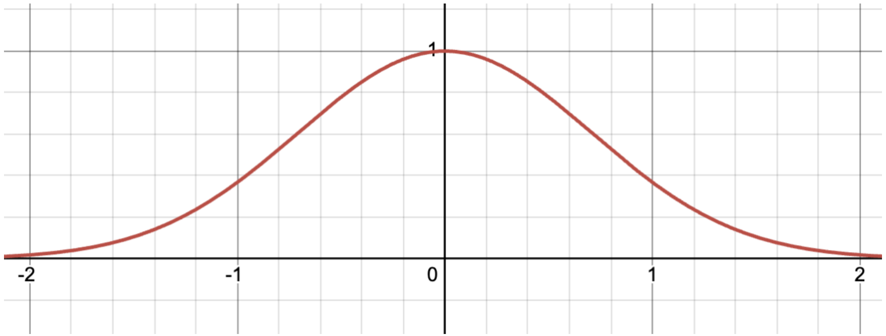
A Gaussian process f(**x**) is a collection of random variables, any finite number of which are Gaussian.

$$
\begin{aligned}
m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\
k(\mathbf{x},\mathbf{x}') &= \mathrm{Cov}\left[f(\mathbf{x}), f(\mathbf{x}')\right] \\
&= \mathbb{E}\left[(f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})])\left(f(\mathbf{x}') - \mathbb{E}[f(\mathbf{x}')]\right)\right]
\end{aligned}
$$

Kernel functions

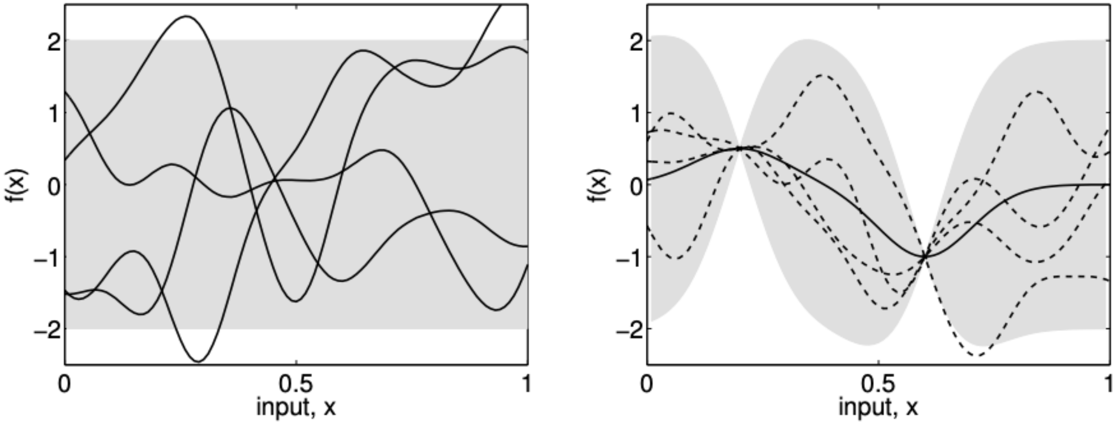
- Radial basis function

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{(\mathbf{x} - \mathbf{x}')^2}{2l^2}\right)$$



- Other examples such as polynomial, periodic

Example of Gaussian processes



(a), prior (b), posterior

Image source: C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X. c 2006 Massachusetts Institute of Technology. www.GaussianProcess.org/gpml

Likelihood function

Describes the noise in the observations

$$y = f(\mathbf{x}) + \varepsilon$$

$$\varepsilon \sim N\left(0, \sigma_n^2\right)$$

Additive Gaussian processes

Assign each dimension a base kernel $k_i(x_i, x_i')$

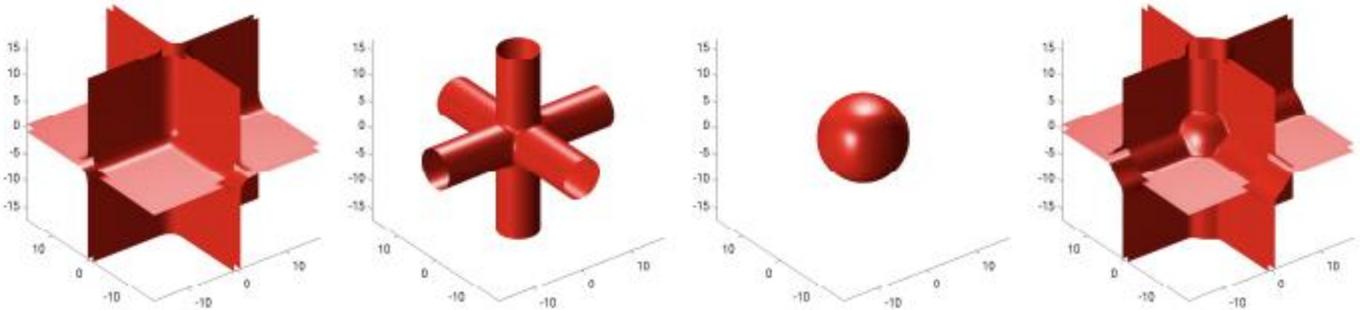
$$k_i(x_i, x_i') = \exp\left(-\frac{(x_i - x_i')^2}{2l_i^2}\right)$$

Additive Gaussian processes

Assign each dimension a base kernel $k_i(x_i, x_i') = \exp\left(-\frac{(x_i - x_i')^2}{2l_i^2}\right)$

$$
\begin{aligned}
k_{add_1}(\mathbf{x}, \mathbf{x}') &= \sigma_1^2 \sum_{i=1}^{D} k_i(x_i, x_i') \\
k_{add_2}(\mathbf{x}, \mathbf{x}') &= \sigma_2^2 \sum_{i=1}^{D} \sum_{j=i+1}^{D} k_i(x_i, x_i') k_j(x_j, x_j') \\
k_{add_n}(\mathbf{x}, \mathbf{x}') &= \sigma_n^2 \sum_{1 \le i_1 < i_2 < \ldots < i_n \le D} \left[\prod_{d=1}^{n} k_{i_d}(x_{i_d}, x_{i_d}') \right]
\end{aligned}
$$

Additive Gaussian processes



1st order interactions $k_1 + k_2 + k_3$ | 2nd order interactions $k_1k_2 + k_2k_3 + k_1k_3$ | 3rd order interactions $k_1k_2k_3$ (Squared-exp kernel) | All interactions (Additive kernel)

$$k_{add_n}(\mathbf{x}, \mathbf{x}') = \sigma_n^2 \sum_{1 \le i_1 < i_2 < \ldots < i_n \le D} \left[\prod_{d=1}^{n} k_{i_d}(x_{i_d}, x'_{i_d}) \right]$$

Newton's Identities

Let $p_n(x_1, \ldots, x_D) = \sum_{i=1}^{D} x_i^n$ be n-th power sum.

Let $e_n(x_1, \ldots, x_D) = \sum_{1 \le i_1 \le \ldots \le i_n \le D} \left[\prod_{d=1}^{n} x_{i_d} \right]$ be n-th elementary symmetric polynomial.

Then $e_n(x_1, \ldots, x_D) = \frac{1}{n} \sum_{i=1}^{n} (-1)^{i-1} e_{n-i}(x_1, \ldots, x_D) p_i(x_1, \ldots, x_D)$

Model fitting

- Overfitting
- Choice of train-test split
- Multiple local minima

Attempt #2: Holder Automatic Additivity Determination (HAAD) Kernel

$$k(x, x') = \left[k_1(x_1, x_1')^p + \ldots + k_n(x_n, x_n')\right]^{1/p} \qquad p \in [0, 1]$$

Attempt 3#: ExtendoLengthscalo (EL) Kernel

$$k_{el_n}(\mathbf{x}, \mathbf{x}') = \sigma_n^2 \sum_{1 \leq i_1 \leq i_2 < \ldots < i_n \leq D} \left[\prod_{l=1}^{n} k_{i_l}\left(x_{i_l}, x'_{i_l}\right)^{1/n} \right]$$

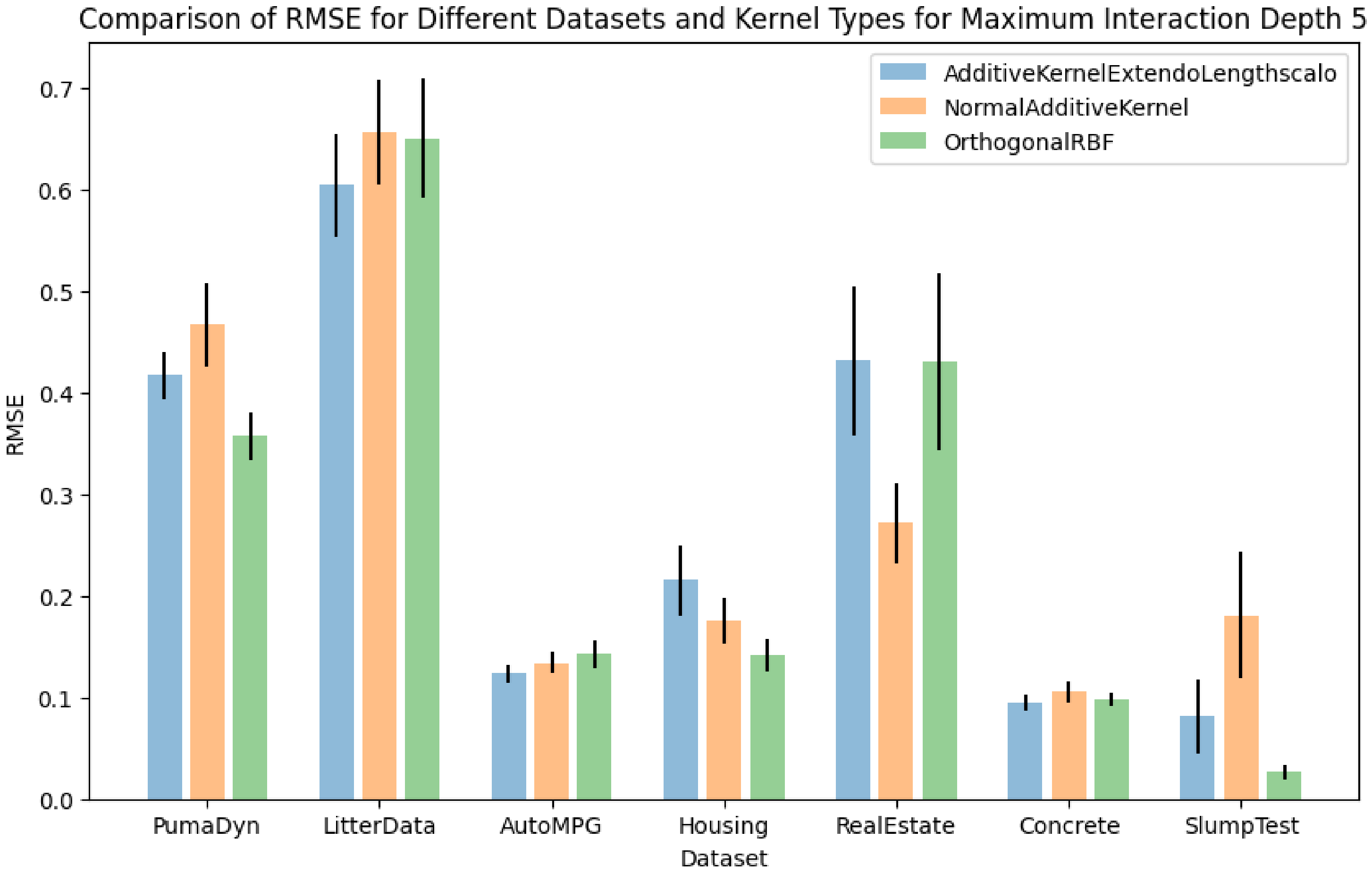
Orthogonal RBF

Consider decomposing the Gaussian process as

$$f(\mathbf{x}) = f_1(x_1) + f_2(x_2) + \cdots + f_{12}(x_1, x_2) + \cdots + f_{12\ldots D}(x_1, x_2, \cdots x_D).$$

Introduce constraint: $$\int f_i(x_i) p_i(x_i) dx_i = 0$$

$p_i(x_i)$ is the density for the corresponding input feature


Comparison of RMSE for Different Datasets and Kernel Types for Maximum Interaction Depth 5
AdditiveKernelExtendoLengthscalo
NormalAdditiveKernel
OrthogonalRBF
RMSE
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
PumaDyn
LitterData
AutoMPG
Housing
RealEstate
Concrete
SlumpTest
Dataset

Questions

Thank you for listening!