On Gaussian Processes for Regression

Jeffrey Alido
Department of Electrical and Computer Engineering
Boston University
jalido@bu.edu

Shashank Manjunath
Department of Electrical and Computer Engineering
Boston University
manjuns@bu.edu

Abstract

Gaussian processes emerged in machine learning as a powerful tool for regression and classification that provides interpretability through kernel choice and uncertainty quantification. By leveraging properties of multivariate normal distributions and Bayes's rule, we may infer a probability distribution over possible functions when fitting a dataset. This Bayesian framework allows flexibility through choosing a covariance function as a prior belief about the dataset, which can provide further insight into the trends of the training data. We implement a multi-dimensional Gaussian process regressor and evaluate its performance on the Boston Housing dataset, which is comparable to those in the top 25 of the Kaggle competition. Furthermore, we perform optimization on the hyperparameters through maximum likelihood estimation, to remove the need for manual tuning of the hyperparameters.

1 Gaussian Random Variables

A random variable is a function that maps from an event space to a measurable space. The event space represents a set of all possible outcomes that the random variable may take, and the measurable space is a probability measure between 0 and 1 (inclusive). We say that a random variable X is normally distributed if the event space has a Gaussian probability distribution, fully characterized by two parameters: a mean μ and variance σ^2:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

For a one-dimensional Gaussian random variable, we refer to its distribution as a univariate Gaussian distribution. A set of Gaussian random variables may be characterized jointly as a multivariate Gaussian distribution, with joint probability distribution fully characterized by a mean vector and a covariance matrix:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

where $\boldsymbol{\mu}$ is the mean vector, and Σ is the covariance matrix whose entries describe the covariance between each pair of random variables.

Preprint. Under review.

2 Gaussian Process

A random process is essentially a collection of random variables jointly characterized as a set or vector of random variables with a multivariate joint probability distribution. A Gaussian process $f(\boldsymbol{x})$ is defined as a a random process where each set of random variable in the random process is has a multivariate Gaussian distribution. $f(\boldsymbol{x})$ is fully characterized by a mean function $m(\boldsymbol{x})$ and covariance function, $K(\boldsymbol{x}, \boldsymbol{x'})$:

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

Typically, the mean function is zero. The kernel for the covariance is chosen based on some prior belief about the dataset; more on kernels is discussed in 3.1.

3 Regression

Suppose we observe training data $\boldsymbol{t}$ and choose kernel κ. Then the mean and covariance functions are given by

$$m(\boldsymbol{x}) = C_{\boldsymbol{xt}}^{\top} C_{\boldsymbol{t}}^{-1} \boldsymbol{t}$$
$$K(\boldsymbol{x}, \boldsymbol{x'}) = C_{\boldsymbol{xx'}} - C_{\boldsymbol{xt}}^{\top} C_{\boldsymbol{t}}^{-1} C_{\boldsymbol{xt}}$$

where $C_{\boldsymbol{xt}} = \kappa(\boldsymbol{x}, \boldsymbol{t})$, $C_{\boldsymbol{t}} = \kappa(\boldsymbol{t}, \boldsymbol{t})$, and $C_{\boldsymbol{xx'}} = \kappa(\boldsymbol{x}, \boldsymbol{x'})$. Those interested in the derivation of the results are encouraged to consult section 6.4.2 of [1].

3.1 Kernels

Covariance Functions or kernels, denoted $\kappa(\boldsymbol{x}, \boldsymbol{x}')$, form the core of a Gaussian process. Kernels allow projection of input data into an alternate feature space, allowing easier separability of data in this new feature space. Gaussian processes leverage kernels to featurize input data. In particular, if we have a function $\Phi(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^n$, we can write the kernel defined by this function as:

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle = \Phi(\boldsymbol{x})^{\top} \Phi(\boldsymbol{x}')$$

In order to illustrate how kernels fit into Gaussian Processes, we will

finish derivation of kernel trick in GP

We illustrate specific kernels that we used below. These kernels can additionally be combined through addition and multiplication of the kernels together, though we do not explore this as part of this project.

3.1.1 Radial Basis Function (RBF) Kernel

The Radial Basis Function (RBF) Kernel, also known as the Squared Exponential Kernel, is given by:

$$\kappa_{RBF}(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{2\ell^2}\right)$$

This kernel is parametrized by two parameters, the lengthscale ℓ and the variance σ^2. The lengthscale determines the width of the kernel, and the variance scales the kernel[2]. We provide an images of the RBF Kernel with various lengthscales and variances in Figure 1

3.1.2 Rational Quadratic Kernel

The Rational Quadratic Kernel is another standard kernel is similar to the RBF kernel. It can be constructed from summing RBF kernels with varying lengthscales. The kernel is given by:

$$\kappa_{RQ}(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \left(1 + \frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{2\alpha\ell^2}\right)^{-\alpha}$$

This kernel is parametrized by three parameters, the lengthscale ℓ, the variance σ^2, and the lengthscale weighting parameter α[2]. We provide images of the Rational Quadratic Kernel with various α values in Figure 2

3.1.3 Periodic Kernel

The periodic kernel allows us to model periodic functions. The kernel is given by:

$$\kappa_P(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp(-\frac{2\sin^2(\pi\|\boldsymbol{x} - \boldsymbol{x}'\|)}{p\ell^2})$$

This kernel is parametrized by two parameters, p which describes the period of the function, and ℓ which is the lengthscale[2]. We provide images of the Periodic Kernel with various p values in Figure 3

4 A Simple Demo

fit some GPs on sine waves or something

5 Boston Housing Dataset

The Boston Housing Dataset, originally published in 1978 contains 506 data points, each containing 13 features and 1 label for regression[3]. The dataset provides the median value of houses in Boston suburbs. This dataset is particularly suitable for Gaussian processes, as the dataset is quite small. Use of Gaussian processes allows us to accurately quantify our uncertainty for each The label is the median value of owner-occupied homes in $1000s, and all other features are used for model fitting. The features included in the dataset are given in the table in Table 1.

Table 1: Table of Boston Housing Dataset feature names and features

Feature Name	**Feature Description**
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	Proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	Nitric oxides concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centres
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per $10,000
PTRATIO	Pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of Black people by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

For the application of Gaussian Processes, we use the regression task, i.e. fitting to the MEDV feature. Prior to fitting on the data, we normalize the data per feature. Specifically, for each feature in the dataset, we perform the following operation:

$$X_{\text{feat}} = \frac{X_{\text{feat}} - \mu(X_{\text{feat}})}{\sigma(X_{\text{feat}})}$$

where $\mu(X)$ is the mean value of that feature in the training set, and $\sigma(X)$ is the standard deviation of that feature in the training set. We additionally normalize the MEDV feature, and convert it back to non-normalized units before calculating our RMSE.

6 Results

7 Figures

Figure 1: Graph of a square exponential kernel for various lengthscales and variances

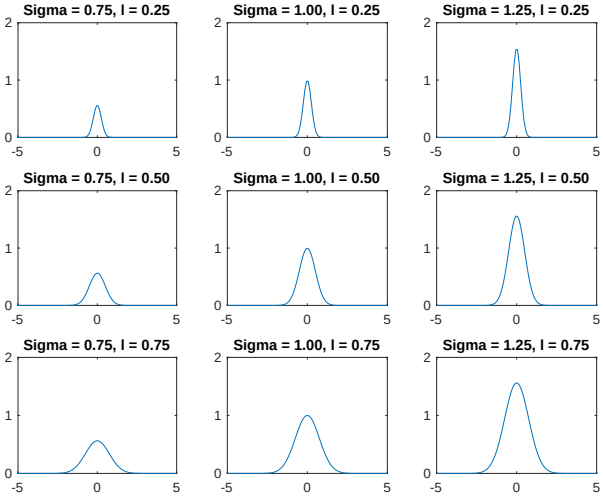


Figure 2: Graph of a rational quadratic kernel for $\sigma = 1$, $\ell = 1$, and various α values

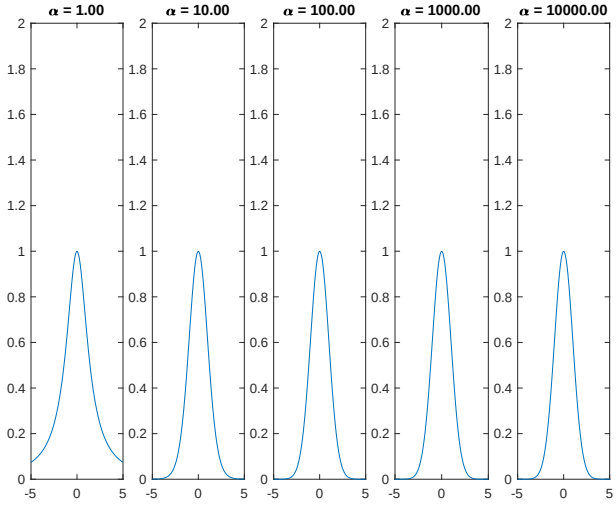
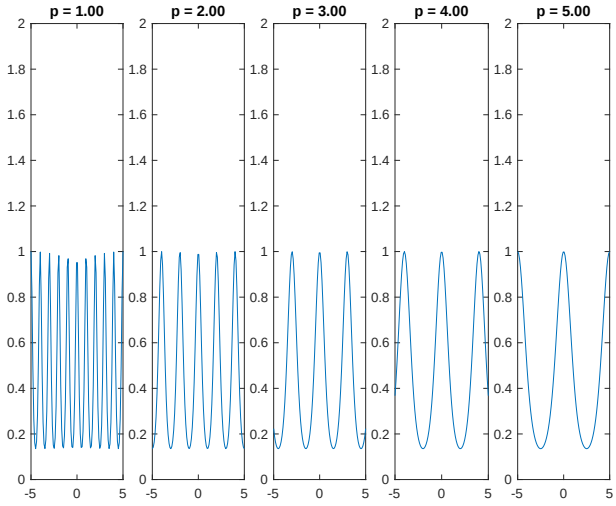


Figure 3: Graph of a rational quadratic kernel for $\sigma = 1$, $\ell = 1$, and various p values



8 References

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] David Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD Thesis, Computational and Biological Learning Laboratory, University of Cambridge, 2014.

[3] David Harrison and Daniel L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.