Learning from data: Gaussian processes

Christian Forssén

Department of Physics, Chalmers University of Technology, Sweden

Oct 4, 2020

1 Inference using Gaussian processes

Assume that there is a set of input vectors with independent, predictor, variables

$$\boldsymbol{X}_N \equiv \{\boldsymbol{x}^{(i)}\}_{i=1}^N$$

and a set of target values

$$\boldsymbol{t}_N \equiv \{t^{(i)}\}_{i=1}^N.$$

- Note that we will use the symbol t to denote the target, or response, variables in the context of Gaussian Processes. Here we will consider single, scalar outputs $t^{(i)}$. The extension to vector outputs $\boldsymbol{t}^{(i)}$ is straightforward.
- Furthermore, we will use the subscript N to denote a set of N vectors (or scalars): $\boldsymbol{X}_N$ ($\boldsymbol{t}_N$),
- ... while a single instance i is denoted by a superscript: $\boldsymbol{x}^{(i)}$ ($t^{(i)}$).

We will consider two different *inference problems*:

1. The prediction of a *new target* $t^{(N+1)}$ given the data $\boldsymbol{X}_N, \boldsymbol{t}_N$ and a new input $\boldsymbol{x}^{(N+1)}$.
2. The inference of a *model function* $y(\boldsymbol{x})$ from the data $\boldsymbol{X}_N, \boldsymbol{t}_N$.

The former can be expressed with the pdf

$$p\left(t^{(N+1)}|\boldsymbol{t}_N, \boldsymbol{X}_N, \boldsymbol{x}^{(N+1)}\right)$$

while the latter can be written using Bayes' formula (in these notes we will not be including information I explicitly in the conditional probabilities)

$$p\left(y(\boldsymbol{x})|\boldsymbol{t}_N, \boldsymbol{X}_N\right) = \frac{p\left(\boldsymbol{t}_N|y(\boldsymbol{x}), \boldsymbol{X}_N\right) p\left(y(\boldsymbol{x})\right)}{p\left(\boldsymbol{t}_N|\boldsymbol{X}_N\right)}$$

The inference of a function will obviously also allow to make predictions for new targets. However, we will need to consider in particular the second term in the numerator, which is the **prior** distribution on functions assumed in the model.

- This prior is implicit in parametric models with priors on the parameters.
- The idea of Gaussian process modeling is to put a prior directly on the **space of functions** without parameterizing $y(\boldsymbol{x})$.
- A Gaussian process can be thought of as a generalization of a Gaussian distribution over a finite vector space to a **function space of infinite dimension**.
- Just as a Gaussian distribution is specified by its mean and covariance matrix, a Gaussian process is specified by a **mean and covariance function**.

Gaussian process

A Gaussian process is a stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution

References:

1. Gaussian Processes for Machine Learning, Carl Edward Rasmussen and Chris Williams, the MIT Press, 2006, online version.
2. GPy: a Gaussian Process (GP) framework written in python, from the Sheffield machine learning group.

1.1 Parametric approach

Let us express $y(\boldsymbol{x})$ in terms of a model function $y(\boldsymbol{x};\boldsymbol{\theta})$ that depends on a vector of model parameters $\boldsymbol{\theta}$.

For example, using a set of basis functions $\left\{\phi^{(h)}(\boldsymbol{x})\right\}_{h=1}^{H}$ with linear weights $\boldsymbol{\theta}_H$ we have

$$y(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{h=1}^{H} \theta^{(h)}\phi^{(h)}(\boldsymbol{x})$$

Notice

The basis functions can be non-linear in $\boldsymbol{x}$ such as Gaussians (aka *radial basis functions*)

$$\phi^{(h)}(\boldsymbol{x}) = \exp\left[-\frac{\left(\boldsymbol{x}-\boldsymbol{c}^{(h)}\right)^2}{2(\sigma^{(h)})^2}\right].$$

Still, this constitutes a linear model since $y(\boldsymbol{x}, \boldsymbol{\theta})$ depends linearly on the parameters $\boldsymbol{\theta}$.

The inference of model parameters should be a well-known problem by now. We state it in terms of Bayes theorem

$$p\left(\boldsymbol{\theta}|\boldsymbol{t}_N, \boldsymbol{X}_N\right) = \frac{p\left(\boldsymbol{t}_N|\boldsymbol{\theta}, \boldsymbol{X}_N\right) p\left(\boldsymbol{\theta}\right)}{p\left(\boldsymbol{t}_N|\boldsymbol{X}_N\right)}$$

Having solved this inference problem (note that the likelihhod is Gaussian, cf linear regression) a prediction can be made through marginalization

$$p\left(t^{(N+1)}|\boldsymbol{t}_N, \boldsymbol{X}_N, \boldsymbol{x}^{(N+1)}\right) = \int d^H\boldsymbol{\theta} p\left(t^{(N+1)}|\boldsymbol{\theta}, \boldsymbol{x}^{(N+1)}\right) p\left(\boldsymbol{\theta}|\boldsymbol{t}_N, \boldsymbol{X}_N\right).$$

Here it is important to note that the final answer does not make any explicit reference to our parametric representation of the unknown function $y(\boldsymbol{x})$.

Assuming that we have a fixed set of basis functions and Gaussian prior distributions (with zero mean) on the weights $\boldsymbol{\theta}$ we will show that:

- The joint pdf of the observed data given the model $p(\boldsymbol{t}_N|\boldsymbol{X}_N)$, is a multivariate Gaussian with mean zero and with a covariance matrix that is determined by the basis functions.
- This implies that the conditional distribution $p(t^{(N+1)}|\boldsymbol{t}_N, \boldsymbol{X}_{N+1})$, is also a multivariate Gaussian whose mean depends linearly on $\boldsymbol{t}_N$.

Proof.

Sum of normally distributed random variables

If X and Y are independent random variables that are normally distributed (and therefore also jointly so), then their sum is also normally distributed. i.e., $Z = X + Y$ is normally distributed with its mean being the sum of the two means, and its variance being the sum of the two variances.

Consider the linear model and define the $N \times H$ design matrix $\boldsymbol{R}$ with elements

$$R_{nh} \equiv \phi^{(h)}\left(\boldsymbol{x}^{(n)}\right).$$

Then $\boldsymbol{y}_N = \boldsymbol{R\theta}$ is the vector of model predictions, i.e.

$$y^{(n)} = \sum_{h=1}^{H} R_{nh}\boldsymbol{\theta^{(h)}}.$$

Assume that we have a Gaussian prior for the linear model weights $\boldsymbol{\theta}$ with zero mean and a diagonal covariance matrix

$$p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}; 0, \sigma_\theta^2 \boldsymbol{I}\right).$$

Now, since y is a linear function of $\boldsymbol{\theta}$, it is also Gaussian distributed with mean zero. Its covariance matrix becomes

$$\boldsymbol{Q} = \langle \boldsymbol{y}\boldsymbol{y}^T \rangle = \langle \boldsymbol{R\theta\theta}^T \boldsymbol{R}^T \rangle = \sigma_\theta^2 \boldsymbol{R}\boldsymbol{R}^T,$$

which implies that

$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y}; 0, \sigma_\theta^2 \boldsymbol{R}\boldsymbol{R}^T\right).$$

This will be true for any set of points $\boldsymbol{X}_N$; which is the defining property of a **Gaussian process**.

- What about the target values $\boldsymbol{t}$?

Well, if $t^{(n)}$ is assumed to differ by additive Gaussian noise, i.e.,

$$t^{(n)} = y^{(n)} + \varepsilon^{(n)},$$

where $\varepsilon^{(n)} \sim \mathcal{N}\left(0, \sigma_\nu^2\right)$; then $\boldsymbol{t}$ also has a Gaussian distribution

$$p(\boldsymbol{t}) = \mathcal{N}\left(\boldsymbol{t}; 0, \boldsymbol{C}\right),$$

where the covariance matrix of this target distribution is given by

$$\boldsymbol{C} = \boldsymbol{Q} + \sigma_\nu^2 \boldsymbol{I} = \sigma_\theta^2 \boldsymbol{R}\boldsymbol{R}^T + \sigma_\nu^2 \boldsymbol{I}.$$

The covariance matrix as the central object. The covariance matrices are given by

$$Q_{nn'} = \sigma_\theta^2 \sum_h \phi^{(h)}\left(\boldsymbol{x}^{(n)}\right) \phi^{(h)}\left(\boldsymbol{x}^{(n')}\right),$$

and

$$C_{nn'} = Q_{nn'} + \delta_{nn'}\sigma_\nu^2.$$

This means that the correlation between target values $t^{(n)}$ and $t^{(n')}$ is determined by the points $\boldsymbol{x}^{(n)}$, $\boldsymbol{x}^{(n')}$ and the behaviour of the basis functions.

1.2 Non-parametric approach: Mean and covariance functions

In fact, we don't really need the basis functions and their parameters anymore. The influence of these appear only in the covariance matrix that describes the distribution of the targets, which is our key object. We can replace the parametric model altogether with a **covariance function** $C(\boldsymbol{x}, \boldsymbol{x}')$ which generates the elements of the covariance matrix

$$Q_{nn'} = C\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(n')}\right),$$

for any set of points $\left\{\boldsymbol{x}^{(n)}\right\}_{n=1}^{N}$.

Note, however, that $\boldsymbol{Q}$ must be positive-definite. This constrains the set of valid covariance functions.

Once we have defined a covariance function, the covariance matrix for the target values will be given by

$$C_{nn'} = C\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(n')}\right) + \sigma_\nu^2 \delta_{nn'}.$$

A wide range of different covariance contributions can be constructed. These standard covariance functions are typically parametrized with hyperparameters $\boldsymbol{\alpha}$ so that

$$C_{nn'} = C\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(n')}, \boldsymbol{\alpha}\right) + \delta_{nn'} \Delta\left(\boldsymbol{x}^{(n)}; \boldsymbol{\alpha}\right),$$

where Δ is usually included as a flexible noise model.

Stationary kernels. The most common types of covariance functions are stationary, or translationally invariant, which implies that

$$C\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(n')}, \boldsymbol{\alpha}\right) = D\left(\boldsymbol{x} - \boldsymbol{x}'; \boldsymbol{\alpha}\right),$$

where the function D is often referred to as a *kernel*.

A very standard kernel is the RBF (also known as Exponentiated Quadratic or Gaussian kernel) which is differentiable infinitely many times (hence, very smooth),

$$C_{\mathrm{RBF}}(\mathbf{x}, \mathbf{x}'; \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 \exp\left[-\frac{1}{2}\sum_{i=1}^{I} \frac{(x_i - x_i')^2}{r_i^2}\right]$$

where I denotes the dimensionality of the input space. The hyperparameters are: $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1, \vec{r}\}$. Sometimes, a single correlation length r is used.

2 GP models for regression

Let us return to the problem of predicting $t^{(N+1)}$ given $\boldsymbol{t}_N$. The independent variables $\boldsymbol{X}_{N+1}$ are also given, but will be omitted from the conditional pdfs below.

The joint density is

$$p\left(t^{(N+1)}, \boldsymbol{t}_N\right) = p\left(t^{(N+1)}|\boldsymbol{t}_N\right) p\left(\boldsymbol{t}_N\right) \quad \Rightarrow \quad p\left(t^{(N+1)}|\boldsymbol{t}_N\right) = \frac{p\left(t^{(N+1)}, \boldsymbol{t}_N\right)}{p\left(\boldsymbol{t}_N\right)}.$$

Since both $p\left(t^{(N+1)}, \boldsymbol{t}_N\right)$ and $p\left(\boldsymbol{t}_N\right)$ are Gaussian distributions, then the conditional distribution, obtained by the ratio, must also be a Gaussian. Let us use the notation $\boldsymbol{C}_{N+1}$ for the $(N+1) \times (N+1)$ covariance matrix for $\boldsymbol{t}_{N+1} = \left(\boldsymbol{t}_N, t^{(N+1)}\right)$. This implies that

$$p\left(t^{(N+1)}|\boldsymbol{t}_N\right) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{t}_N, t^{(N+1)}\right) \boldsymbol{C}_{N+1}^{-1} \begin{pmatrix} \boldsymbol{t}_N \\ t^{(N+1)} \end{pmatrix}\right]$$

Summary

The prediction of the (Gaussian) pdf for $t^{(N+1)}$ requires an inversion of the covariance matrix $\boldsymbol{C}_{N+1}$.

Elegant approach using linear algebra tricks. Let us split the $\boldsymbol{C}_{N+1}$ covariance matrix into four different blocks

$$\boldsymbol{C}_{N+1} = \begin{pmatrix} \boldsymbol{C}_N & \boldsymbol{k} \\ \boldsymbol{k}^T & \kappa \end{pmatrix},$$

where $\boldsymbol{C}_N$ is the $N \times N$ covariance matrix (which depends on the positions $\boldsymbol{X}_N$), $\boldsymbol{k}$ is an $N \times 1$ vector (that describes the covariance of $\boldsymbol{X}_N$ with $\boldsymbol{x}^{(N+1)}$), while κ is the single diagonal element obtained from $\boldsymbol{x}^{(N+1)}$.

We can use the partitioned inverse equations (Barnett, 1979) to rewrite $\boldsymbol{C}_{N+1}^{-1}$ in terms of $\boldsymbol{C}_N^{-1}$ and $\boldsymbol{C}_N$ as follows

$$\boldsymbol{C}_{N+1}^{-1} = \begin{pmatrix} \boldsymbol{M}_N & \boldsymbol{m} \\ \boldsymbol{m}^T & \mu \end{pmatrix},$$

where

$$\begin{aligned} \mu &= \left(\kappa - \boldsymbol{k}^T \boldsymbol{C}_N^{-1} \boldsymbol{k}\right)^{-1} \\ \boldsymbol{m} &= -\mu \boldsymbol{C}_N^{-1} \boldsymbol{k} \\ \boldsymbol{M}_N &= \boldsymbol{C}_N^{-1} + \frac{1}{\mu} \boldsymbol{m}\boldsymbol{m}^T. \end{aligned}$$

Question

What are the dimensions of the different blocks? Check that the answer.

This implies that we can make a prediction for the Gaussian pdf of $t^{(N+1)}$ (meaning that we predict its value with an associated uncertainty) for an N^3 computational cost (the inversion of an $N \times N$ matrix).

Summary

The prediction for $t^{(N+1)}$ is a Gaussian

$$p\left(t^{(N+1)}|\boldsymbol{t}_N\right) = \frac{1}{Z}\exp\left[-\frac{\left(t^{(N+1)} - \hat{t}^{(N+1)}\right)^2}{2\sigma^2_{\hat{t}_{N+1}}}\right]$$

with

$$\begin{aligned} \text{mean}: \quad & \hat{t}^{(N+1)} = \boldsymbol{k}^T \boldsymbol{C}_N^{-1} \boldsymbol{t}_N \\ \text{variance}: \quad & \sigma^2_{\hat{t}_{N+1}} = \kappa - \boldsymbol{k}^T \boldsymbol{C}_N^{-1} \boldsymbol{k}. \end{aligned}$$

In fact, since the prediction only depends on the N available data we might as well predict several new target values at once. Consider $\boldsymbol{t}_M = \{t^{(N+i)}\}_{i=1}^M$ so that

$$\boldsymbol{C}_{N+M} = \begin{pmatrix} \boldsymbol{C}_N & \boldsymbol{k} \\ \boldsymbol{k}^T & \boldsymbol{\kappa} \end{pmatrix},$$

where $\boldsymbol{k}$ is now an $N \times M$ matrix and $\boldsymbol{\kappa}$ an $M \times M$ matrix.

The prediction becomes a multivariate Gaussian

$$p\left(\boldsymbol{t}_{N+M}|\boldsymbol{t}_N\right) = \frac{1}{Z}\exp\left[-\frac{1}{2}\left(\boldsymbol{t}_M - \hat{\boldsymbol{t}}_M\right)^T \boldsymbol{\Sigma}_M^{-1}\left(\boldsymbol{t}_M - \hat{\boldsymbol{t}}_M\right)\right],$$

where the $M \times 1$ mean vector and $M \times M$ covariance matrix are

$$\begin{aligned} \hat{\boldsymbol{t}}_M &= \boldsymbol{k}^T \boldsymbol{C}_N^{-1} \boldsymbol{t}_N \\ \boldsymbol{\Sigma}_M &= \boldsymbol{\kappa} - \boldsymbol{k}^T \boldsymbol{C}_N^{-1} \boldsymbol{k}. \end{aligned}$$

Optimizing the GP model hyperparameters. Predictions can be made once we have

1. Chosen an appropriate covariance function.
2. Determined the hyperparameters.
3. Evaluated the relevant blocks in the covariance function and inverted $\boldsymbol{C}_N$.

How do we determine the hyperparameters $\boldsymbol{\alpha}$? Well, recall that

$$p\left(\boldsymbol{t}_N\right) = \frac{1}{Z}\exp\left[-\frac{1}{2}\boldsymbol{t}_N^T \boldsymbol{C}_N^{-1} \boldsymbol{t}_N\right].$$

This pdf is basically a data likelihood.

- The frequentist approach would be to find the set of hyperparameters $\boldsymbol{\alpha}^*$ that maximizes the data likelihood, i.e. that minimizes $\boldsymbol{t}_N^T \boldsymbol{C}_N^{-1} \boldsymbol{t}_N$.
- A Bayesian approach would be to assign a prior to the hyperparameters and seek a posterior pdf $p(\boldsymbol{\alpha}|\boldsymbol{t}_N)$ instead.

The former approach is absolutely dominating the literature on GP regression. The covariance function hyperparameters are first optimized and then used for regression.