Introduction to Gaussian Processes - RAW

Filipe P. Farias
Teleinformatics Engineering Department
Federal University of Ceará
filipepfarias@fisica.ufc.br

Abstract

The Gaussian processes have proven to be a powerfull framework for robust estimation and a flexible model for non-linear *regression*, case which will be the main object of this work, with some implementations of real situations.

1 Introduction

First, we'll overview some initial concepts that will set the background for the GP framework. Let's suppose a data set $\mathcal{D} = \{(x_i, t_i)\}_{i=0}^{N-1}$ which we denote the input as x, the output (or target) as t and N as the number of observations. The first step in our workflow is define a *training* set, i.e. some data that is given to make our first assumptions of the *model*. The model y can be defined as a guess of the law that rules the phenomenon of which our data was observed. This law can be, by example a senoid function as represented in Figure 1. Given this training data we wish to make *predictions* for new inputs x_* that we have not observed in the training set.

We'll assume an parametric approach, then the model is said to contain *parameters* $\mathbf{w}$, that will be adjusted during the *training phase*, when those are modified aiming to reduce the mismatch with the training set. In general we define a *loss function* $L(y, t)$ which increases itself as the mismatch becomes larger, in other words the *error* of the model. Then our work will be to reduce this error such that the smallest one will be the which defines when our model has learned the parameters of the law of the phenomenon. This turns possible to make our predictions where the data was not observed.

Unfortunately, in this trying of obtain the model by the smallest error, we may lose the capability of generalize it, i.e. our model could learned well for the training set only. So, if new data arrive or a new realization of the phenomenon occurs, that smallest error may increase for the same model. With this we define that our model isn't flexible. Then we can increase this flexibility by accepting some *uncertainty* above it. More, sometimes a good first assumption can make the difference to the estimation, and one may want to put its beliefs in the model even before to observe the data, i.e. make a *prior* assumption. These both strategies of uncertainty and prior assumptions are well defined by the Bayesian inference and could help if we assume a *probabilistic model*.

The Bayesian inference can handle with the classical approaches of search the model which has the smallest error, but our objective is achieve one step ahead. We can not only obtain one model, but a *distribution* of possible models. And with this, all the probabilistic meaning of distribution is carried with it, that is we can obtain both the model which *minimize* the error or the statistics of the distribution of models. A more explaining view of what this really means will be given in the next sections.

Furthermore, the concept of *infer* is similar to what we have done since beginning. We maded a guess of the law which rules the phenomenon, i.e. a prior assumption. Then we turns our model more plausible by reducing its error, or more *likely*. After, we obtained a result of these assumptions, a *posterior* assumption. These steps are similar in concept when dealing with Bayesian inference, except that, as we will deal with probability distributions, then some rules must be established for the method to be concise.

Finally, we'll deal with a specific class of models in which we assume not a distribution of parameters but functions in general. By example, in a space with infinite possible functions, we'll evaluate how much possible which one are to be generated the data by its statistics, what is similar to what was done for the parameters. And in this part we make the fully use of the Gaussian process.

2 Linear Regression

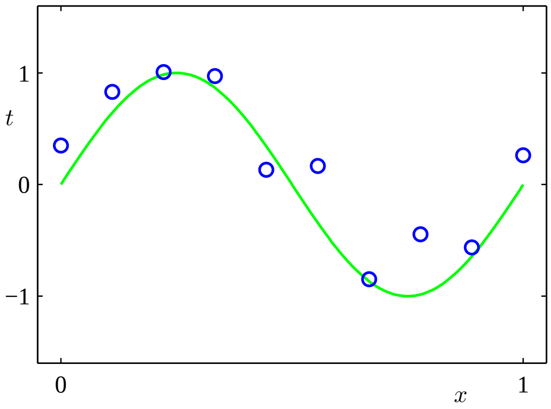

Figure 1: Training data set with $n = 10$ points in blue. The green curve shows the function $sin(2\pi x)$ used to generate the data. Our goal is to predict the value of t for some new value of x, without knowledge of the green curve [Bishop, 2006].

The following scenario is given. The data set of observations is given by $\mathcal{D} = \{(x_i, t_i)\}_{i=0}^{N-1}$, hence we make a guess about the law that rules the phenomenon behind the data. In general, we define that model as a mathematical function y whose parameters $\mathbf{w}$ will be adjusted in trying of learn the phenomenon law. By example, we may choose a polynomial as a model and try to adjust its coefficients, that will be the parameters for the model in this case.

$$y\left(x, \mathbf{w}\right) = w_0 x^0 + \cdots + w_{M-1} x^{M-1} \tag{2.1}$$

This approach is not linear in the inputs, i.e. the model output y it isn't a linear transformation of the inputs. But we can say that is *linear on the parameters*. To visualize this, considering the parameters array $\mathbf{w} = \left(w_0, \cdots, w_{M-1}\right)^\top$, we put in the matrix form as

$$y\left(x, \mathbf{w}\right) = \mathbf{w}^\top \boldsymbol{\phi}(x). \tag{2.2}$$

Here we have defined the array $\boldsymbol{\phi}(x) = \left(x^0, \cdots, x^{M-1}\right)^\top$ in order to keep the linearity of the model. This approach becomes more clear if we take $\boldsymbol{\phi}$ for all the space where we have the observations, by this we mean evaluate $\boldsymbol{\phi}$ for all the $\{x_i\}_{i=0}^{N-1}$ inputs, and we'll obtain

$$\begin{pmatrix} x_0^0 & \cdots & x_0^{M-1} \\ x_1^0 & \cdots & x_1^{M-1} \\ \vdots & \ddots & \vdots \\ x_{N-1}^0 & \cdots & x_{N-1}^{M-1} \end{pmatrix}. \tag{2.3}$$

This matrix is called the *design matrix* Φ. Hence for the vector with all the outputs, we would obtain of $\mathbf{y} = \mathbf{w}^\top \Phi$. These tools make possible gain some insight that help us in the follow steps. Backing to the data, let's take the targets in the vector form, this is $\mathbf{t} = \left(t_0, \cdots, t^{N-1}\right)^\top$, so we can consider that $\mathbf{t}$ is a *vector in space* [Bishop, 2006]. If we take the columns of Φ, denoting each one by $\boldsymbol{\varphi}_j$, what defines the vector $\mathbf{y}$ by its linear combination. If we take a polynomial of M^{th} order, being M smaller than the number of points N, than we say that $\boldsymbol{\varphi}$ will span a linear subspace $\mathcal{S}$ of dimensionality M. With this, our problem turns to actually train the parameters, what can be done defining some metric that say to us when the model is missing, as a *loss function*.

Maybe an intuitive way to say to the model if he's going far away from the target is measure this error, i.e. how much the model is *distant* from the target. We may generalize the distance itself by the *Minkowsky distance* L_p between $\mathbf{u}$ and $\mathbf{v}$, defined as

$$L_p\left(\mathbf{u}, \mathbf{v}\right) = \left(\sum_{i=1}^{n} |u_i - v_i|^p\right)^{1/p} \tag{2.4}$$

where we have the Euclidean distance for $p = 2$ and the also known Manhattan distance for $p = 1$. The question here is define what distance use, by example the Euclidean distance is the most common due to its analytical tractability and practical importance. With this choice the error between the model and the targets will be define as the distance between $\mathbf{y}$ and $\mathbf{t}$, and where the orthogonal projection of $\mathbf{t}$ onto the subspace $\mathcal{S}$ is the minimal distance. The *learning* step of the parameters $\mathbf{w}$ lies on choose those whose make that distance minimal. Hence, being $\langle \cdot, \cdot \rangle$ the inner product, we have

$$L_2(\mathbf{t} - \mathbf{y}) = \langle \mathbf{t} - \mathbf{y}, \mathbf{t} - \mathbf{y} \rangle^{1/2}. \tag{2.5}$$

In general we denote L as a *loss function* and add an $1/2$ term multiplying the inner product, what will be shown next. The minimal distance makes itself clear the process of minimization for looking the best choice of parameters. In this case, the minimization is taken by the derivative of L_2 with $\partial L_2 / \partial \mathbf{w} = 0$, then

$$\mathbf{w}^* = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top \mathbf{t} \tag{2.6}$$

for $\mathbf{y} = \mathbf{w}^\top \Phi$. Making use of the optimization notation, we denoted $\mathbf{w}^*$ as the best choice of the parameters. Note that, even working with a polynomial model, we do not manipulate it in fact. This approach enable us to generalize for a several types of *basis functions*, i.e. functions that will define our design matrix.

The linear regression is one of the simplest methods for estimation and bring us an idea to find the law of the phenomenon, but one of its problems is its flexibility and non-robustness. By the example of the polynomial model, if we construct the polynomial for the learned parameters, we note that its coefficients will have high values. This is because when learning the targets, if we consider the order of the polynomial the same as the number of data, the model starts to interpolate the noise, i.e. the aspects captured by the model may not be precise with the law of the phenomenon and with this, the derivatives of the polynomial may grow, as shown in the Figure ??. Some questions may arise as how know exactly this law, what in fact we may never know. The next step lead us to a probabilistic point of view with some more tools to deal with the problem.

3 Bayesian Linear Regression

The approach is to give a *prior* probability to every possible function, where higher probabilities are given to functions that we consider to be *more likely*, for example because they are smoother than other functions. [Rasmussen and Williams, 2005]. This appears to have a serious problem, in that surely there are an infinite set of possible functions to compute in

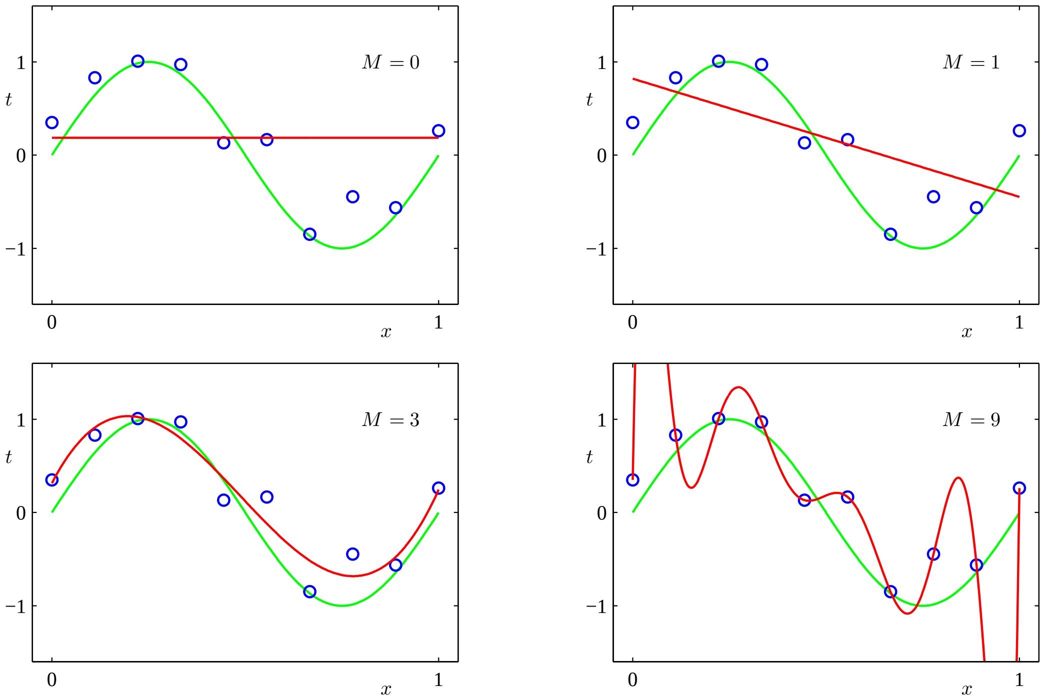

Figure 2: Plots of polynomials for the model in (**??**) having various orders M, shown as red curves [Bishop, 2006].

a finite time. This is where the Gaussian *process* comes as a possible approach. A Gaussian process is a generalization of the Gaussian probability *distribution*. Whereas a probability distribution describes random variables which are scalars or vectors (for multivariate distributions), a random process governs the properties of functions [Getoor, 2009].

We can think of a function as a very long vector, each entry in the vector specifying the function value $f(x)$ at a particular input x. Indeed, the question of how we deal computationally with these infinite dimensional objects has the resolution: if you ask only for the properties of the function at a finite number of points, then *infer* the properties for the predictions in the Gaussian process will give you the same answer if you ignore the infinitely many other points, as if you would have taken them all into account. One of the main attractions of the Gaussian process framework is precisely that it unites a sophisticated and consistent view with computational tractability.

4 Bayesian Linear Regression

4.1 A Bayesian view of Linear Regression

Until now, we see the curve fitting problem in terms of the minimization of the error function. Then we will see the same by a probabilistic perspective gaining some insights into error minimization and regularization, leading us to a full Bayesian treatment.

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{4.1}$$

We can use the Bayes' theorem (4.1) to convert a *prior* probability into a *posterior* probability at the light of some evidence. We can make assumptions about quantities such as the parameters $\mathbf{w}$ in the form of a prior distribution $p(\mathbf{w})$. The observation of the data $\mathcal{D}$ and what its implies in the parameters is expressed as a conditional probability $p(\mathcal{D}|\mathbf{w})$.

Then we can evaluate the uncertainty about $\mathbf{w}$ after observed the data $\mathcal{D}$ as a posterior probability $p(\mathbf{w}|\mathcal{D})$.

The quantity $p(\mathcal{D}|\mathbf{w})$ expresses how probable the observed data $\mathcal{D}$ is for different settings of $\mathbf{w}$. Then, not being necessarily a probability distribution, but a function over the parameters [DeGroot and Schervish, 2012], its integral with respect to $\mathbf{w}$ could not be equal one, then to normalize the equation with respect to the left-side there's a term $p(\mathcal{D})$. This distribution is called *likelihood function*.

Integrating the both sides with respect to $\mathbf{w}$, we obtain the denominator, then considering that integrating a probability distribution over itself is equal to one, we have

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w} \tag{4.2}$$

4.2 Bayesian curve fitting

Let's consider the same data set $\mathcal{D}$ presented before, but now we have some uncertainty over the value of the measured value t. This uncertainty can be represented as a probability distribution function p, in this particular case a Gaussian distribution, with a mean equal to the model $y(x, \mathbf{w})$. Thus we have

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right) \tag{4.3}$$

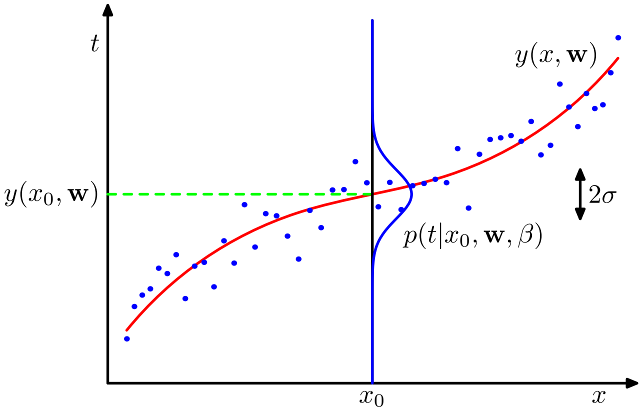

Figure 3: Schematic illustration of a Gaussian conditional distribution for t given x given by (4.3), in which the mean is given by the polynomial function y(x, w), and the precision is given by the parameter β, which is related to the variance $\beta^{-1} = \sigma^2$ [Bishop, 2006].

Where β is the variance of the distribution. Note that a large β will give us more imprecision about the measured value t, then we can call it of *precision parameter*, i.e. how much certain we are about t. As we done in linear regression, we are trying to obtain the parameters for the model. In other words, given a value t, we trying to obtain the *mean* and the *variance* which maximize the probability of the measured value. Assuming the data set being independent and identically distributed, the joint probability of the whole data set will be

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{i=0}^{N-1} \mathcal{N}\left(t_i|y(x_i, \mathbf{w}), \beta^{-1}\right) \tag{4.4}$$

When viewed as function of $y(x_i, \mathbf{w})$, the model, and β^{-1}, this is the likelihood function for the Gaussian. The parameters of the distribution can be determined by maximizing the likelihood function. It is convenient to maximize the log of the likelihood function, or minimize the negative log whats is equivalent, this implies that the maximization of the log of the function is equivalent to the maximization of the function itself, because the logarithm is a monotonically increasing of its argument. This helps the mathematical analysis and helps numerically because the small probabilities can easily underflow the numerical precision of the computer. Then[†]

$$\ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = -\frac{\beta}{2}\sum_{i=1}^{N}\left\{y\left(x_i,\mathbf{w}\right)-t_i\right\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) \tag{4.5}$$

The maximization of (4.5) taking the derivative with respect to $\mathbf{w}$ will lead us back to the same of the minimization of (**??**), the error function of the linear regression. Here, just by notation, we will call the resulting parameters of the maximization of $\mathbf{w}_{\mathrm{ML}}$, what it is called *maximum likelihood.*

We can determine the precision parameter using the maximum likelihood by taking the derivative with respect to β of (4.5), what gives

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{i=0}^{N-1}\left\{y\left(x_i,\mathbf{w}_{\mathrm{ML}}\right)-t_i\right\}^2 \tag{4.6}$$

Now we have a probabilistic view of the regression and then we can make predictions for new values of x, given that our model is capable of learn the parameters. And not just one collection of them, but a distribution probability.

In other words, after find the maximum likelihood parameters $\mathbf{w}_{\mathrm{ML}}$ and β_{ML}, we have the parameters distribution by

$$p\left(t|x,\mathbf{w}_{\mathrm{ML}},\beta_{\mathrm{ML}}\right) = \mathcal{N}\left(t|y\left(x,\mathbf{w}_{\mathrm{ML}}\right),\beta_{\mathrm{ML}}^{-1}\right) \tag{4.7}$$

Aiming to apply a "more Bayesian" approach, we not have yet a prior distribution to make the inference using the Bayes' rule. We can now introduce here the probability distribution over the parameters $p(\mathbf{w})$ as presented in the Section 3.1. The choice is arbitrary, but for this particular case we will consider

$$p(\mathbf{w}|\alpha) = \mathcal{N}\left(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I}\right) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2}\exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\top}\mathbf{w}\right\} \tag{4.8}$$

where α is the variance, or precision parameter, of the distribution and $M+1$ is the number of parameters of the model, i.e. the length of $\mathbf{w}$. We call *hyperparameters* the variables such α who control the model parameters distribution. And now we have by the Bayes' theorem considering that our *posterior* distribution is proportional to the product between the *likelihood function* and the assumed *prior*, as seen in (4.2) then, assuming the observation of the whole data set

$$p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \propto p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}|\alpha) \tag{4.9}$$

As done before, we maximize the posterior probability, i.e. find the most probable value given the data by the term $p(\mathbf{w}|\mathbf{x},\mathbf{t})$ aside of the distribution parameters. This will result a particular choice of $\mathbf{w}$. We call this approach of *maximum posterior*, or MAP. Then taking the negative logarithm

$$\ln p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \propto \ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) + \ln p(\mathbf{w}|\alpha) \tag{4.10}$$

Then we substitute the probability distributions founded before. Note that the first term in the right side is the error function founded in (4.5). Then, the terms which the minimization depends of $\mathbf{w}$ are

[†] Consider the Gaussian distribution as $\mathcal{N}\left(x|\mu,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$

$$\frac{\beta}{2}\sum_{i=1}^{N}\left\{y\left(x_i,\mathbf{w}\right)-t_i\right\}^2+\frac{\alpha}{2}\mathbf{w}^{\top}\mathbf{w} \tag{4.11}$$

And we can note the similarity with the regularized linear regression in (**??**) aside of the term λ, what can be founded by $\lambda = \alpha/\beta$. It's important to note that even called maximum posterior, here it was presented a minimization in terms of the negative logarithm, but this equals to the maximization of the positive logarithm. The signal was chosen just for similarity with the error function.

References

[Bishop, 2006] Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

[DeGroot and Schervish, 2012] DeGroot, M. and Schervish, M. (2012). Probability and Statistics. Addison-Wesley.

[Getoor, 2009] Getoor, R. (2009). J. l. doob: Foundations of stochastic processes and probabilistic potential theory. Ann. Probab., 37(5):1647–1663.

[Hennig, 2013] Hennig, P. (2013). Gaussian processes. Machine Learning Summer School 2013.

[Rasmussen and Williams, 2005] Rasmussen, C. E. and Williams, C. K. I. (2005). Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.