Introduction to Gaussian Processes

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

Abstract

A wide variety of methods exists to deal with supervised learning, as restrict a class of linear functions of the inputs, as linear regression, or give a prior probability to every possible function, giving high probability to the functions we consider more likely. The second approach is a way to Gaussian process itself. We will make the pathway through a intuitive construction of this framework.

1 Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2 Linear Regression

Starting with a simple regression problem. Be the data set $\mathcal{D} = \{x_i, y_i | i = 0, \dots, N-1\}$, where we observe a real-valued input variable x and a measured real-valued variable y. Then, we'll use synthetically generated data for comparison against any learned *model*. And N will be the number of observations of the value y. Our objective is make predictions of the new value $\hat{y}$ for some new input $\hat{x}$.

For this example, we'll use a simple approach based on curve fitting by the polynomial model, i.e., being the function

$$f(x, \mathbf{w}) = \sum_{j=0}^{M-1} w_j x^j \tag{2.1}$$

where M is the order of the polynomial and $\mathbf{w} = [w_0, \dots, w_M]$ its coefficients. It's important to note that the f isn't linear in x but in $\mathbf{w}$. These functions which are linear on the unknown parameters are called *linear models*. Section 1.1 - Bishop (pg 4).

We can extend the class of models considering linear combinations of nonlinear functions of the input variables, i.e.

$$f(x, \mathbf{w}) = \sum_{j=0}^{M-1} \phi_j(x) w_j \tag{2.2}$$

where $\phi_j(x)$ are known as *basis functions*, and then the total number of parameters for this model will be M. We can evaluate the same operation of (2.2) in the matrix form by

$$f(x, \mathbf{w}) = \boldsymbol{\phi}(x)^\top \mathbf{w} \tag{2.3}$$

where $\boldsymbol{\phi}(x) = [\phi_0(x), \ldots, \phi_{M-1}(x)]^\top$. In the example of the curve fitting, the polynomial regression implies that $\phi_j(x) = x^j$. It's important to note that these linear models are needed to define its basis functions before the training data set is observed. Section 1.4 - Bishop (pg 33).

The values of $\mathbf{w}$ are obtained by minimizing the *error function*, a measure of the distance between the training data set and f, given values of $\mathbf{w}$. By the way, the chosen error function will be

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{f(x_i, \mathbf{w}) - y_i\}^2 \tag{2.4}$$

This indicate that if E is zero, f passes exactly through each training data point. Observe that E do not assume negative values because of its quadratic form, then we can find $\mathbf{w}$ by finding the minimum value of E, denoted $\mathbf{w}^*$, by

$$\frac{\partial E}{\partial \mathbf{w}} = 0 \tag{2.5}$$

We can rewrite (2.4) in the matrix form, considering $\mathbf{f} = f(\mathbf{x}, \mathbf{w})$, where $\mathbf{x} = [x_0, \ldots, x_{N-1}]^\top$, i.e. f evaluated for all input variables (See Appendix A.1). Then we have

$$\mathbf{f} = \Phi \mathbf{w} \tag{2.6}$$

where Φ is the *design matrix* such that $\boldsymbol{\phi}(x)$ is evaluated for all $\mathbf{x}$. Proceeding with the minimization, we obtain the optimal $\mathbf{w}$, or $\mathbf{w}^*$ by

$$\mathbf{w}^* = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y} \tag{2.7}$$

which are the parameters that best fit the model to the data.

Insert over-fitting.

As we increase the number of parameters, our model becomes more flexible and then our error function approximates of zero for the training data. But when compared to the test data, the error increases. This is known as *over-fitting*.

2.1 Regularized Linear Regression

An approach to minimize the over-fitting problem is to control the flexibility of the model. This can be done by controlling the norm of $\mathbf{w}^*$ as the number of parameters increases. By (2.4) we can add the penalty term $||\mathbf{w}||^2$ scaled by the factor $\lambda/2$, then

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{f(x_i, \mathbf{w}) - y_i\}^2 + \frac{\lambda}{2} ||\mathbf{w}||^2 \tag{2.8}$$

whats means that our error increases as the norm of $\mathbf{w}$ grows. This will lead us to the matrix form

$$\mathbf{w}^* = (\Phi^\top \Phi + \lambda \mathbf{I})^{-1} \Phi^\top \mathbf{y} \tag{2.9}$$

This allow us to increase the number of parameters trying to control the over-fitting. More, add parameters will allows us to capture different aspects of the data set, what we will see later.

3 Bayesian Linear Regression

3.1 A Bayesian view of Linear Regression

Until now, we see the curve fitting problem in terms of error minimization. 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.

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{3.1}$$

We can use the Bayes' theorem (3.1) to convert a *prior* probability into a *posterior* probability at the light of some evidence. We can make inferences 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 a probability distribution over the parameters, 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

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w} \tag{3.2}$$

3.2 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 y. This uncertainty can be represented as a probability distribution p, in this particular case a Gaussian distribution, with a mean equal to the model $f(x, \mathbf{w})$. Thus we have

$$p(y|x, \mathbf{w}, \beta) = \mathcal{N}\left(y|f(x, \mathbf{w}), \beta^{-1}\right) \tag{3.3}$$

Where β is the variance of the distribution. Note that a large β will give is more uncertainty about the measured value y, then we can call it of *precision parameter*, i.e. how much certain we are about y. As we done in linear regression, we are trying to obtain the parameters for the model. In other words, given a value y, 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

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{i=0}^{N-1} \mathcal{N}\left(y_i|f(x_i, \mathbf{w}), \beta^{-1}\right) \tag{3.4}$$

When viewed as function of $f(x_i, \mathbf{w})$, 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, 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[†]

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{i=1}^{N}\left\{f\left(x_i, \mathbf{w}\right) - y_i\right\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) \tag{3.5}$$

The maximization of (3.5) taking the derivative with respect to $\mathbf{w}$ will lead us back to the same of the minimization of (2.4), 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 (3.5), what gives

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{i=0}^{N-1}\left\{f\left(x_i, \mathbf{w}_{\mathrm{ML}}\right) - y_i\right\}^2 \tag{3.6}$$

Now we have a probabilistic view of the regression and then we can make predictions for new values of x. This became possible because we construct probabilistic model and we can make these predictions with a *predictive distribution*, i.e. a probability distribution over y, what gives us not just simply a point.

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

$$p\left(y|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}\right) = \mathcal{N}\left(y|f\left(x, \mathbf{w}_{\mathrm{ML}}\right), \beta_{\mathrm{ML}}^{-1}\right) \tag{3.7}$$

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

$$p(\mathbf{w}|\alpha) = \mathcal{N}\left(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\right) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2}\exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\} \tag{3.8}$$

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

There are many example of choices for basis functions, as

$$\phi_j(x) = \exp\left\{-\frac{\left(x - \mu_j\right)^2}{2s^2}\right\} \tag{3.9}$$

known as *squared exponential*, where μ_j controls the location of the basis function in the *input space*, and s the spatial scale. It's usually referred as 'Gaussian' basis function because of its similarity with the Gaussian distribution function, although there is no probabilistic interpretation here.

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

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Appendix

A Derivations

A.1 Matrix Form

Be the linear model $f(x, \mathbf{w}) = \mathbf{w}^\top \boldsymbol{\phi}(x)$. Suppose $\Phi = [\boldsymbol{\phi}(x_1), \ldots, \boldsymbol{\phi}(x_N)]^\top$, then Φ will be of the form

$$\Phi = \begin{bmatrix} \phi_0(x_0) & \cdots & \phi_{M-1}(x_0) \\ \vdots & \ddots & \vdots \\ \phi_0(x_{N-1}) & \cdots & \phi_{M-1}(x_{N-1}) \end{bmatrix} \tag{A.1}$$

called *design matrix*. Then the model turns to $\mathbf{f} = \Phi\mathbf{w}$. This will lead us to the matrix form for the quadratic error function

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}(\mathbf{f} - \mathbf{y})^\top(\mathbf{f} - \mathbf{y}) \\ &= \frac{1}{2}(\Phi\mathbf{w} - \mathbf{y})^\top(\Phi\mathbf{w} - \mathbf{y}) \\ &= \frac{1}{2}(\mathbf{w}^\top\Phi^\top\Phi\mathbf{w} - \mathbf{y}^\top\Phi\mathbf{w} - \mathbf{w}^\top\Phi^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}) \end{aligned}$$

Observe that even in the matrix form, the error function remains scalar, which implies that $\mathbf{y}^\top\Phi\mathbf{w} = \mathbf{w}^\top\Phi^\top\mathbf{y}$ by the transpose of the product rule. Then

$$E(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^\top\Phi^\top\Phi\mathbf{w} - 2\mathbf{y}^\top\Phi\mathbf{w} + \mathbf{y}^\top\mathbf{y})$$

Then we proceed by the minimization by $\frac{\partial E}{\partial \mathbf{w}} = 0$

$$\begin{aligned} 0 &= \frac{1}{2}(2\mathbf{w}^\top\Phi^\top\Phi - 2\mathbf{y}^\top\Phi)^\dagger \\ \mathbf{w}^{*\top} &= \mathbf{y}^\top\Phi(\Phi^\top\Phi)^{-1} \\ \mathbf{w}^* &= (\Phi^\top\Phi)^{-1}\Phi^\top\mathbf{y} \end{aligned} \tag{A.2}$$

For the regularized linear regression, we do $\frac{\lambda}{2}||\mathbf{w}||^2 = \mathbf{w}^\top\mathbf{w}$, then

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}(\mathbf{f} - \mathbf{y})^\top(\mathbf{f} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} \\ &= \frac{1}{2}(\mathbf{w}^\top\Phi^\top\Phi\mathbf{w} - \mathbf{y}^\top\Phi\mathbf{w} - \mathbf{w}^\top\Phi^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} \end{aligned}$$

And with the minimization we do $\frac{\partial E}{\partial \mathbf{w}} = 0$, then

[†] Using two facts. First, if $\alpha = \mathbf{x}^\mathrm{T}\mathbf{A}\mathbf{x}$, then $\frac{\partial \alpha}{\partial \mathbf{x}} = 2\mathbf{x}^\top\mathbf{A}$, being α scalar. Second, if $\alpha = \mathbf{y}^\mathrm{T}\mathbf{A}\mathbf{x}$, then $\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{y}^\mathrm{T}\mathbf{A}$. For both, $\mathbf{A}$ is independent of $\mathbf{x}$ and $\mathbf{y}$ [Gra83].

$$
\begin{aligned}
0 &= \mathbf{w}^\top \Phi^\top \Phi - \mathbf{y}^\top \Phi + \lambda \mathbf{w}^\top \\
\mathbf{w}^{*\top} &= \mathbf{y}^\top \Phi (\Phi^\top \Phi + \lambda \mathbf{I})^{-1} \\
\mathbf{w}^* &= (\Phi^\top \Phi + \lambda \mathbf{I})^{-1} \Phi^\top \mathbf{y}
\end{aligned} \tag{A.3}
$$

where $\mathbf{I}$ is the identity matrix.

B Mauris euismod

References

[Gra83] F.A. Graybill. Matrices with applications in statistics. Wadsworth statistics - probability series. Wadsworth International Group, 1983.