Teleinformatics Engineering Department, Federal University of Ceará

Introduction to Gaussian Processes

Filipe P. de Farias, IC
filipepfarias@fisica.ufc.br

February 19, 2019

Outline

1 Linear Regression
1.1 Curve Fitting
1.2 A probabilistic perspective

2 Bayesian Linear Regression
2.1 The D-dimensional Gaussian Distribution
2.2 Bayes' theorem for Gaussian variables

3 Gaussian Processes
3.1 Linear regression revisited

4 Appendix
4.1 Appendix A - Matrix Calculus

Linear Regression

Curve Fitting

If we have a set of points in the space that comes from observations of an experiment and we want to predict other points, this could be done with **curve fitting** .

Curve Fitting

If we have a set of points in the space that comes from observations of an experiment and we want to predict other points, this could be done with **curve fitting** .

So we could define some strategy to find our model.

Curve Fitting

If we have a set of points in the space that comes from observations of an experiment and we want to predict other points, this could be done with **curve fitting** .

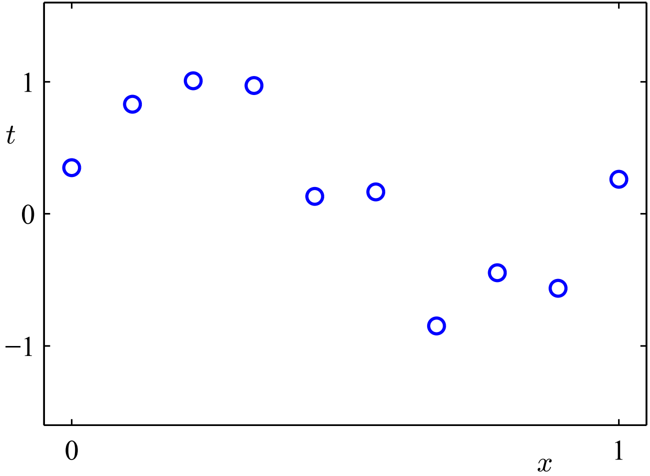
So we could define some strategy to find our model.

Strategy

1. Purpose a **model**, e.g. functions like exponential, polynomial and others.
2. Train our model with the training data set, finding the **unknown parameters**.

Curve Fitting

Let's fit the points below by polynomial curve fitting



Curve Fitting

Be the model chosen

Be the model chosen

$$y(x, \mathbf{w}) = w_0 x^0 + w_1 x^1 + w_2 x^2 + \ldots + w_{M-1} x^{M-1} = \sum_{j=1}^{M-1} w_j x^j$$

Curve Fitting

Be the model chosen

$$y(x, \mathbf{w}) = w_0 x^0 + w_1 x^1 + w_2 x^2 + \ldots + w_{M-1} x^{M-1} = \sum_{j=1}^{M-1} w_j x^j$$

For general, we could write this *weighted sum* with any other function. In other words, we can put this in terms of $\phi_n(x) = x^n$, where ϕ could be other *basis function*. For simplicity, we'll carry this notation along.

Curve Fitting

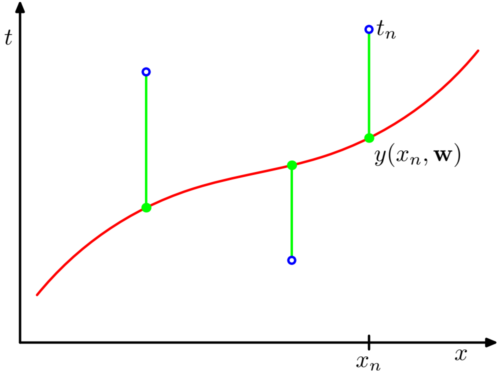
Be the model chosen

$$y(x,\mathbf{w}) = w_0x^0 + w_1x^1 + w_2x^2 + \ldots + w_{M-1}x^{M-1} = \sum_{j=1}^{M-1} w_jx^j$$

For general, we could write this *weighted sum* with any other function. In other words, we can put this in terms of $\phi_n(x) = x^n$, where ϕ could be other *basis function*. For simplicity, we'll carry this notation along.

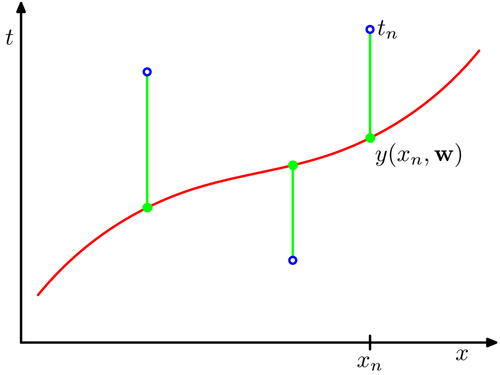
$$y(x,\mathbf{w}) = w_0\phi_0(x) + w_1\phi_1(x) + w_2\phi_2(x) + \ldots + w_{M-1}\phi_{M-1}(x) = \sum_{j=1}^{M-1} w_j\phi_j(x)$$

The chosen model will give us some curve that is needed to adjust such that we'll *minimize* his **distance** to the given points, or **targets** (t).



The chosen model will give us some curve that is needed to adjust such that we'll *minimize* his **distance** to the given points, or **targets** (t).

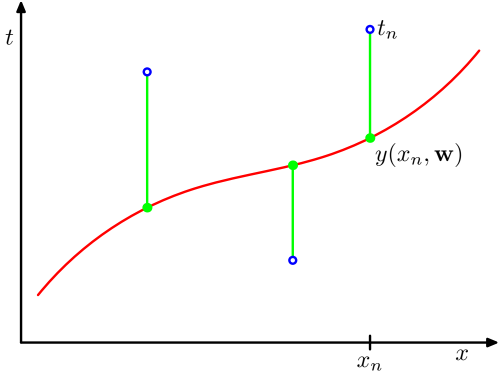
Here, let's define the sum of these distances as *cost function*, or loss function, and write as



The chosen model will give us some curve that is needed to adjust such that we'll *minimize* his **distance** to the given points, or **targets** (t).

Here, let's define the sum of these distances as *cost function*, or loss function, and write as

$$E(\mathbf{w}) \triangleq \frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2$$



Curve Fitting

Insert some *Minkowski* loss.

Curve Fitting

Remembering that

$$y_n(x_n, \mathbf{w}) = w_0\phi_0(x_n) + w_1\phi_1(x_n) + w_2\phi_2(x_n) + \ldots + w_{M-1}\phi_{M-1}(x_n)$$

We could put $y_n(x_i, \mathbf{w})$ in the matricial form and get

$$y_n = \begin{bmatrix} \phi_0(x_n) & \phi_1(x_n) & \ldots & \phi_{M-1}(x_n) \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{bmatrix}$$

and then

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \dots & \phi_{M-1}(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_{M-1}(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_{N-1}) & \phi_1(x_{N-1}) & \dots & \phi_{M-1}(x_{N-1}) \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}}_{\mathbf{w}}$$

This represents the system $\mathbf{y} = \Phi\mathbf{w}$. If

$$E(\mathbf{w}) = \frac{1}{2}\left(\mathbf{y} - \mathbf{t}\right)^T\left(\mathbf{y} - \mathbf{t}\right)$$

where $\mathbf{t} = \begin{bmatrix} t_1 & t_2 & \dots & t_n \end{bmatrix}^T$

Then we'll have

$$\begin{aligned}
E(\mathbf{w}) =& \frac{1}{2}\left(\mathbf{y}^T\mathbf{y} - \mathbf{t}^T\mathbf{y} - \mathbf{y}^T\mathbf{t} + \mathbf{t}^T\mathbf{t}\right) \\
=& \frac{1}{2}\left((\Phi\mathbf{w})^T(\Phi\mathbf{w}) - \mathbf{t}^T(\Phi\mathbf{w}) - (\Phi\mathbf{w})^T\mathbf{t} + \mathbf{t}^T\mathbf{t}\right) \\
=& \frac{1}{2}\left(\mathbf{w}^T\Phi^T\Phi\mathbf{w} - 2\mathbf{t}^T\Phi\mathbf{w} + \mathbf{t}^T\mathbf{t}\right)
\end{aligned}$$

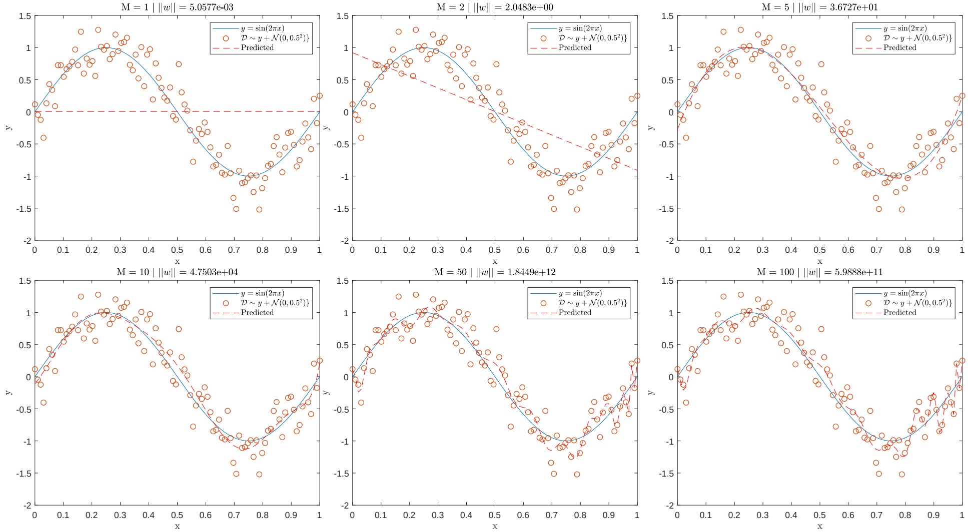
this by the fact that $\alpha = \mathbf{t}^T(\Phi\mathbf{w}) = (\Phi\mathbf{w})^T\mathbf{t}$, being α a scalar.

Curve Fitting

In sequence, we'll try to minimize it in terms of the weights ($\mathbf{w}$) by

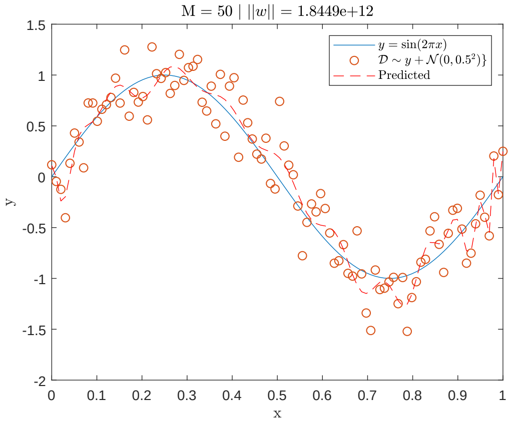
$$
\begin{aligned}
0 =& \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} \\
0 =& \frac{1}{2}\left(2\mathbf{w}^T\Phi^T\Phi - 2\mathbf{t}^T\Phi + 0\right) \\
\mathbf{w}^T =& \mathbf{t}^T\Phi\left(\Phi^T\Phi\right)^{-1} \\
\mathbf{w} =& \left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{t}
\end{aligned}
$$

Here, we've obtained $\mathbf{w}$ for the curve fitting.

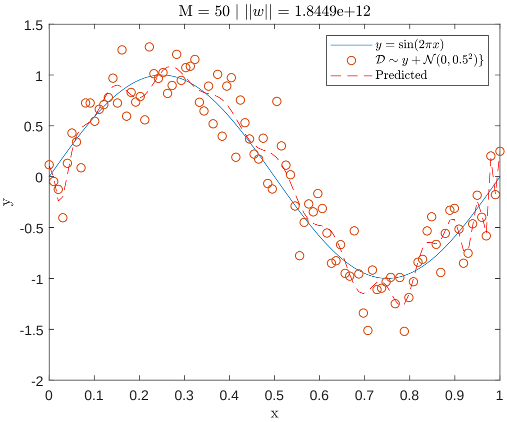

M = 1 | ||w|| = 5.0577e-03
M = 2 | ||w|| = 2.0483e+00
M = 5 | ||w|| = 3.6727e+01
M = 10 | ||w|| = 4.7503e+04
M = 50 | ||w|| = 1.8449e+12
M = 100 | ||w|| = 5.9888e+11
$y = \sin(2\pi x)$
$\mathcal{D} \sim y + \mathcal{N}(0, 0.5^2)\}$
Predicted
x
y

Curve Fitting

A visible effect of the *increase of the complexity* of the model, represented here by *M*, is the *increase of the weights*. We call it **over-fitting**.

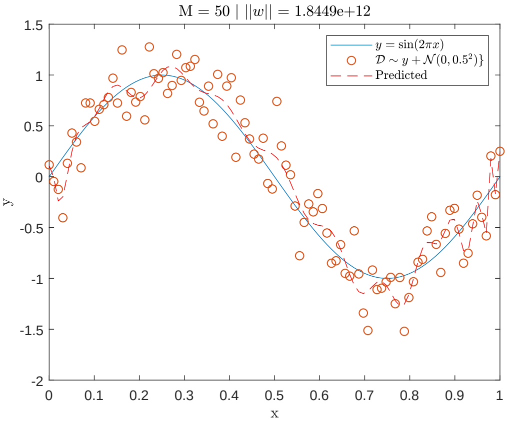


A visible effect of the *increase of the complexity* of the model, represented here by M, is the *increase of the weights*. We call it **over-fitting**.
It's reasonable to see that our model start's to differ from the y and starts to interpolate the noise.



Curve Fitting

A visible effect of the *increase of the complexity* of the model, represented here by M, is the *increase of the weights*. We call it **over-fitting**.
It's reasonable to see that our model start's to differ from the y and starts to interpolate the noise.
Last, a basic concept it's that larger weights could imply in massive system inputs, and this sometimes it's wanted to be avoided.



Curve Fitting

To control the over-fitting, we try to *regularize* the weights by adding a penalty term (λ) to error function, by this we force the coefficients to not reach high values.

Curve Fitting

To control the over-fitting, we try to *regularize* the weights by adding a penalty term (λ) to error function, by this we force the coefficients to not reach high values.

$$
\begin{aligned}
\tilde{E}(\mathbf{w}) =& \frac{1}{2}(\mathbf{y}-\mathbf{t})^T(\mathbf{y}-\mathbf{t}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \\
=& \frac{1}{2}\left(\mathbf{w}^T\Phi^T\Phi\mathbf{w} - 2\mathbf{t}^T\Phi\mathbf{w} + \mathbf{t}^T\mathbf{t} + \lambda\mathbf{w}^T\mathbf{I}\mathbf{w}\right) \\
\Rightarrow \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} =& \frac{1}{2}\left(2\mathbf{w}^T\Phi^T\Phi - 2\mathbf{t}^T\Phi + 0 + 2\lambda\mathbf{w}^T\mathbf{I}\right) \\
0 =& \mathbf{w}^T\Phi^T\Phi - \mathbf{t}^T\Phi + \lambda\mathbf{w}^T\mathbf{I} \\
\mathbf{w} =& \left(\Phi^T\Phi + \lambda\mathbf{I}\right)^{-1}\Phi^T\mathbf{t}
\end{aligned}
$$

A probabilistic perspective

A probabilistic perspective

So, we'll start to look the regression with a statistical approach. To encourage you, let's take the sentence.

A probabilistic perspective

So, we'll start to look the regression with a statistical approach. To encourage you, let's take the sentence.

Sentence

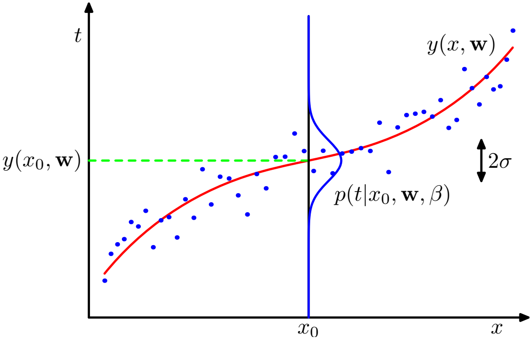
Having an ***uncertainty*** *in the measured value, we could represent it with a* ***probability distribuition****.*

A probabilistic perspective

So, we'll start to look the regression with a statistical approach. To encourage you, let's take the sentence.

Sentence

*Having an **uncertainty** in the measured value, we could represent it with a **probability distribuition**.*



A probabilistic perspective

Let's go back to the initial problem of curve fitting. Each observation of the phenomenon is described with a random variable whose *mean* is given by $y(x, \mathbf{w})$, and the *variance* by β.

A probabilistic perspective

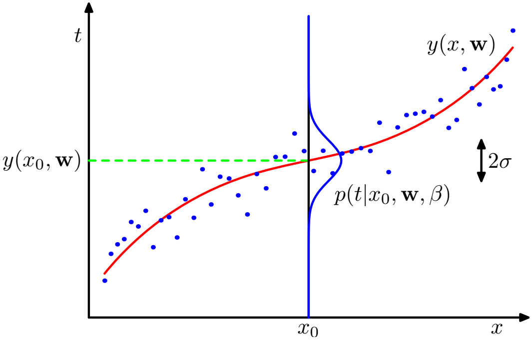
Let's go back to the initial problem of curve fitting. Each observation of the phenomenon is described with a random variable whose *mean* is given by $y(x, \mathbf{w})$, and the *variance* by β.

Then, we want to obtain the probability of the *targets*, given some parameters, in this case $\mathbf{x}$, $\mathbf{w}$ and β.

A probabilistic perspective

Let's go back to the initial problem of curve fitting. Each observation of the phenomenon is described with a random variable whose *mean* is given by $y(x, \mathbf{w})$, and the *variance* by β.

Then, we want to obtain the probability of the *targets*, given some parameters, in this case $\mathbf{x}$, $\mathbf{w}$ and β.



A probabilistic perspective

So, if we consider that our conditions are such that being the random variables independent and identically distributed, we can say that our *joint probability* is given by

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} p\left(t_n | x_n, \mathbf{w}, \beta\right)$$

A probabilistic perspective

So, if we consider that our conditions are such that being the random variables independent and identically distributed, we can say that our *joint probability* is given by

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} p\left(t_n|x_n, \mathbf{w}, \beta\right)$$

Let's assume we have a distribution such that $p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)$. Our goal is, given the *parameters*, maximize the *probability* of the *targets* given the *parameters*. An approach to do this use the fact that

$$\int_{\infty}^{-\infty} p(x)dx = 1 \text{ and } p(x) \geq 0$$

A probabilistic perspective

Seen this, we're supposing that p could assume values much smaller than one. To avoid computational singularity and for future purposes, we'll take the logarithmic probability. And then

$$\ln\left(p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)\right)$$

Reminding that

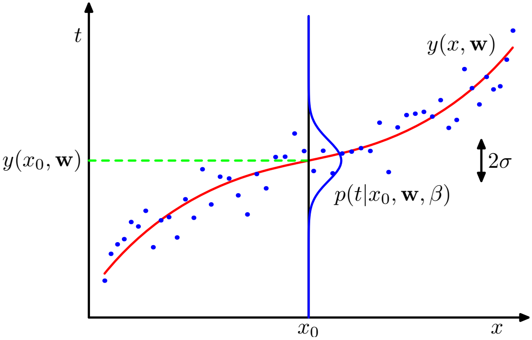
$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = \prod_{n=1}^{N} p\left(t_n|x_n,\mathbf{w},\beta\right)$$

Implies that

$$\ln\left(p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)\right) = \sum_{n=1}^{N} \ln\left(p\left(t_n|x_n,\mathbf{w},\beta\right)\right)$$

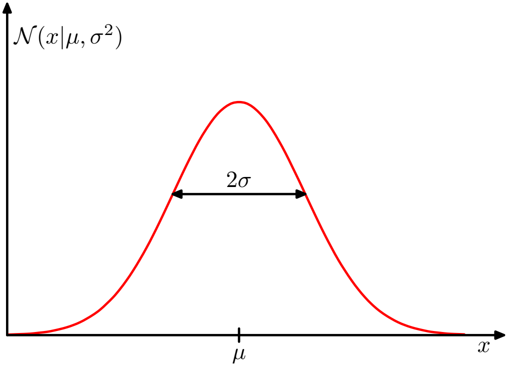
A probabilistic perspective

To proceed, we need to know what distribution p is. Let's choose the **Gaussian distribution**.



A probabilistic perspective

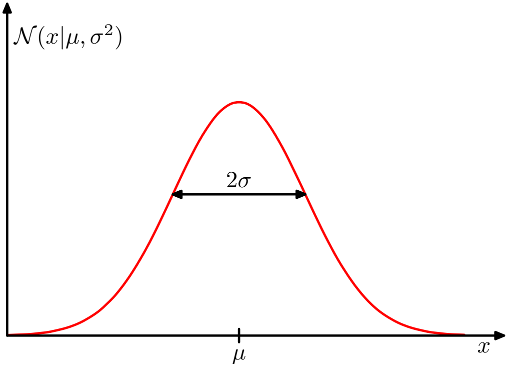
The **Gaussian distribution** comes from many different contexts, as the one that maximize the entropy among of all ones with fixed variance and from the sum of multiple random variables with finite variance.



A probabilistic perspective

The **Gaussian distribution** comes from many different contexts, as the one that maximize the entropy among of all ones with fixed variance and from the sum of multiple random variables with finite variance.

First we'll consider a geometrical approach by the quadratic distance $(x-\mu)^2$ normalized by the variance σ^2. This comprehension will help us with the D-dimensional case.



One-dimensional Gaussian distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} > 0$$

where μ is the mean and σ^2 the variance.

A probabilistic perspective

Now, back to the discussion of the maximization of

$$\ln\left(p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)\right) = \sum_{n=1}^{N} \ln\left(p\left(t_n|x_n,\mathbf{w},\beta\right)\right)$$

A probabilistic perspective

Now, back to the discussion of the maximization of

$$\ln\left(p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)\right) = \sum_{n=1}^{N} \ln\left(p\left(t_n|x_n,\mathbf{w},\beta\right)\right)$$

Reminding that

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

we can state a Gaussian distribution for each target and then

A probabilistic perspective

Now, back to the discussion of the maximization of

$$\ln\left(p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)\right) = \sum_{n=1}^{N} \ln\left(p\left(t_n|x_n,\mathbf{w},\beta\right)\right)$$

Reminding that

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

we can state a Gaussian distribution for each target and then

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

A probabilistic perspective

And then

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

We could observe that taking the derivative with respect to $\mathbf{w}$, our expression becomes closer to the *error function* presented previously, added some dependency of β

$$E(\mathbf{w}) \triangleq \frac{1}{2}\sum_{n=1}^{N}\{y_n - t_n\}^2$$

Then some behaviors could be expected, as the **over-fitting**. But there's a difference between the equations by the term β.

A probabilistic perspective

So, here our objective is to find the parameters to minimize the error and maximize the log probability. But it's reasonable to think that we'll face the same **over-fitting** problem presented before. Added to the fact that we don't apply yet a fully *bayesian* treatment of regression, let's introduce some machinery

Bayesian Linear Regression

The D-dimensional Gaussian Distribution

Remember the One-dimensional Gaussian distribution

One-dimensional Gaussian distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} > 0$$

where μ is the mean and σ^2 the variance.

The D-dimensional Gaussian Distribution

To more than one dimensions, we'll consider the points (x) distance for the mean of the distribution, as we done in the one dimensional case, by adding a term to prioritize some dimension distribution in particular. Then

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

called *Mahalanobis distance*. And it's becomes the *Euclidean distance*, when $\boldsymbol{\Sigma}$ is the indentity matrix. This means that the all the distances are equally normalized. The matrix $\boldsymbol{\Sigma}$ is the covariance matrix of the distributions, by definition.

And then

D-dimensional Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right\}$$

where $\boldsymbol{\mu}$ is the D-dimensional mean vector, $\boldsymbol{\Sigma}$ the D$\times$D-dimensional variance matrix and $|\boldsymbol{\Sigma}|$ its determinant.

Partitioned Gaussians

Given a joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ and

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}.$$

Will give us

- Conditional distribution:

$$p\left(\mathbf{x}_a|\mathbf{x}_b\right) = \mathcal{N}\left(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}\right), \ \boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)$$

- Marginal distribution:

$$p\left(\mathbf{x}_a\right) = \mathcal{N}\left(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}\right)$$

Bayes' theorem for Gaussian variables

We stated before that the Bayes' theorem could be used to **adjust** our model parameters as we obtain evidences. Let's partitionate our distribution $\mathbf{z}$ as

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$$

The strategy here is to make predictions for $\mathbf{y}$. We do this evaluating the probabilities for the whole distribution $\mathbf{z}$. And the key idea is, being $\mathbf{y}$ part of $\mathbf{z}$, we can evaluate its probabilities from $\mathbf{x}$, assuming that the partionated distributions remains Gaussian.

Bayes' theorem for Gaussian variables

In other words, we're trying to find the parginal distribution $p\left(\mathbf{y}\right)$ and the conditional distribution $p\left(\mathbf{x}|\mathbf{y}\right)$, then given

$$
\begin{aligned}
p\left(\mathbf{x}\right) =& \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \\
p\left(\mathbf{y}|\mathbf{x}\right) =& \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right)
\end{aligned}
$$

Bayes' theorem for Gaussian variables

In other words, we're trying to find the parginal distribution $p(\mathbf{y})$ and the conditional distribution $p(\mathbf{x}|\mathbf{y})$, then given

$$
\begin{aligned}
p(\mathbf{x}) =& \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \\
p(\mathbf{y}|\mathbf{x}) =& \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right)
\end{aligned}
$$

So, applying the joint distribution and the its $\ln$ after

$$
\begin{aligned}
p(\mathbf{z}) &= p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})\, p(\mathbf{x}) \\
\ln p(\mathbf{z}) &= \ln p(\mathbf{y}|\mathbf{x}) + \ln p(\mathbf{x}) \\
&= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) \\
&\quad - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^T \mathbf{L}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) + \text{const}
\end{aligned}
$$

Bayes' theorem for Gaussian variables

In other words, we're trying to find the parginal distribution $p\left(\mathbf{y}\right)$ and the conditional distribution $p\left(\mathbf{x}|\mathbf{y}\right)$, then given

$$p\left(\mathbf{x}\right) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right)$$
$$p\left(\mathbf{y}|\mathbf{x}\right) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right)$$

Bayes' theorem for Gaussian variables

In other words, we're trying to find the parginal distribution $p\left(\mathbf{y}\right)$ and the conditional distribution $p\left(\mathbf{x}|\mathbf{y}\right)$, then given

$$
\begin{aligned}
p\left(\mathbf{x}\right) =& \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \\
p\left(\mathbf{y}|\mathbf{x}\right) =& \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}\right)
\end{aligned}
$$

So, applying the joint distribution and the its $\ln$ after

$$
\begin{aligned}
p\left(\mathbf{z}\right) &= p\left(\mathbf{x},\mathbf{y}\right) = p\left(\mathbf{y}|\mathbf{x}\right) p\left(\mathbf{x}\right) \\
\ln p\left(\mathbf{z}\right) &= \ln p\left(\mathbf{y}|\mathbf{x}\right) + \ln p\left(\mathbf{x}\right) \\
&= -\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^{T}\boldsymbol{\Lambda}\left(\mathbf{x}-\boldsymbol{\mu}\right) \\
&\quad -\frac{1}{2}\left(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}\right)^{T}\mathbf{L}\left(\mathbf{y}-\mathbf{A}\mathbf{x}-\mathbf{b}\right) + \text{const}
\end{aligned}
$$

Bayes' theorem for Gaussian variables

The "const" is the term independent of $\mathbf{x}$ and $\mathbf{y}$. Then, expanding the quadratic form

$$\begin{aligned}
\ln p(\mathbf{z}) &= -\frac{1}{2}\mathbf{x}^T\left(\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A}\right)\mathbf{x} - \frac{1}{2}\mathbf{y}^T\mathbf{L}\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{L}\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{L}\mathbf{y} \\
&= -\frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^T \begin{pmatrix}\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A} & -\mathbf{A}^T\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = -\frac{1}{2}\mathbf{z}^T\mathbf{R}\mathbf{z}
\end{aligned}$$

Bayes' theorem for Gaussian variables

The "const" is the term independent of $\mathbf{x}$ and $\mathbf{y}$. Then, expanding the quadratic form

$$\ln p(\mathbf{z}) = -\frac{1}{2}\mathbf{x}^T\left(\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A}\right)\mathbf{x} - \frac{1}{2}\mathbf{y}^T\mathbf{L}\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{L}\mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{L}\mathbf{y}$$

$$= -\frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^T \begin{pmatrix}\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A} & -\mathbf{A}^T\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = -\frac{1}{2}\mathbf{z}^T\mathbf{R}\mathbf{z}$$

We'll apply the partitioned matrices inversion to obtain $\mathbf{R}^{-1}$

$$\mathbf{R}^{-1} = \begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^T\\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T\end{pmatrix}$$

Bayes' theorem for Gaussian variables

The expanded form of $\ln p\left(\mathbf{z}\right)$ give us the mean too by the linear terms, then

$$\mathbf{x}^T\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{x}^T\mathbf{A}^T\mathbf{L}\mathbf{b} + \mathbf{y}^T\mathbf{L}\mathbf{b} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^T \begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^T\mathbf{L}\mathbf{b}\\ \mathbf{L}\mathbf{b}\end{pmatrix}$$

Bayes' theorem for Gaussian variables

The expanded form of $\ln p\left(\mathbf{z}\right)$ give us the mean too by the linear terms, then

$$\mathbf{x}^T\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{x}^T\mathbf{A}^T\mathbf{L}\mathbf{b} + \mathbf{y}^T\mathbf{L}\mathbf{b} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^T\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix}$$

By inspection of the linear terms

$$\mathbb{E}[\mathbf{z}] = \mathbf{R}^{-1} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^T\mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix}$$

Bayes' theorem for Gaussian variables

The expanded form of $\ln p\left(\mathbf{z}\right)$ give us the mean too by the linear terms, then

$$\mathbf{x}^T\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{x}^T\mathbf{A}^T\mathbf{L}\mathbf{b} + \mathbf{y}^T\mathbf{L}\mathbf{b} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^T \begin{pmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^T\mathbf{L}\mathbf{b}\\ \mathbf{L}\mathbf{b}\end{pmatrix}$$

Bayes' theorem for Gaussian variables

The expanded form of $\ln p\left(\mathbf{z}\right)$ give us the mean too by the linear terms, then

$$\mathbf{x}^T \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{x}^T \mathbf{A}^T \mathbf{L}\mathbf{b} + \mathbf{y}^T \mathbf{L}\mathbf{b} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^T \mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix}$$

By inspection of the linear terms

$$\mathbb{E}[\mathbf{z}] = \mathbf{R}^{-1} \begin{pmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - \mathbf{A}^T \mathbf{L}\mathbf{b} \\ \mathbf{L}\mathbf{b} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \end{pmatrix}$$

Bayes' theorem for Gaussian variables

And then we we'll have that

$$
\begin{aligned}
\mathbb{E}[\mathbf{y}] &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \\
\operatorname{cov}[\mathbf{y}] &= \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{T}
\end{aligned}
$$

And then we we'll have that

$$\mathbb{E}[\mathbf{y}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$
$$\mathrm{cov}[\mathbf{y}] = \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T$$

$$\mathbb{E}[\mathbf{x}|\mathbf{y}] = \left(\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A}\right)^{-1}\left\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}$$
$$\mathrm{cov}[\mathbf{x}|\mathbf{y}] = \left(\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A}\right)^{-1}$$

Bayes' theorem for Gaussian variables

Marginal and Conditioned Gaussians

From the results above, we'll have

- For $\mathbf{y}$ given $\mathbf{x}$:

$$p\left(\mathbf{y}|\mathbf{x}\right) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right)$$
$$p\left(\mathbf{x}\right) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right)$$

- For $\mathbf{x}$ given $\mathbf{y}$:

$$p\left(\mathbf{x}|\mathbf{y}\right) = \mathcal{N}\left(\mathbf{y}|, \boldsymbol{\Sigma}\left\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b} + \boldsymbol{\Sigma}\boldsymbol{\mu})\right\}, \boldsymbol{\Sigma}\right)$$
$$p\left(\mathbf{y}\right) = \mathcal{N}\left(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T\right), \text{where } \boldsymbol{\Sigma} = \left(\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A}\right)^{-1}$$

Gaussian Processes

Linear regression revisited

In order to motivate the Gaussian process viewpoint, let us return to the linear regression example and re-derive the predictive distribution by working in terms of distributions over functions $y(x, \mathbf{w})$. This will provide a specific example of a Gaussian process. Consider a model defined in terms of a linear combination of M fixed basis functions given by the elements of the vector (x) so that
where x is the input vector andw is theM-dimensional weight vector.

Appendix

Definition (Matrix Multiplication)

Given $\mathbf{A}$ being $m \times n$ and $\mathbf{B}$ being $p \times q$

$$\mathbf{AB} = \left[\sum_{s=1}^{n} a_{is} b_{sj} \right], \text{with } n = p$$

$$\mathbf{BA} = \left[\sum_{k=1}^{r} b_{ik} a_{kj} \right], \text{with } m = q$$

Definition (Matrix Multiplication)

Given $\mathbf{A}$ being $m \times n$ and $\mathbf{B}$ being $p \times q$

$$[\mathbf{AB}]^T = \left[\sum_{s=1}^{m} a_{is} b_{sj}\right]^T = \left[\sum_{s=1}^{n} b_{is} a_{sj}\right] = \mathbf{B}^T \mathbf{A}^T, \text{ with } n = p$$

Proposition

Given $\mathbf{y}$ being $m \times 1$, $\mathbf{x}$ being $n \times 1$, $\mathbf{A}$ being $m \times n$ independent of $\mathbf{x}$ and

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

Then

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$$

Definition (Matrix Derivative)

Given $\mathbf{A}$ being $m \times n$ and $\mathbf{B}$ being $p \times q$

$$\mathbf{AB} = \left[\sum_{s=1}^{n} a_{is} b_{sj} \right]$$

$$\mathbf{BA} = \left[\sum_{k=1}^{r} b_{ik} a_{kj} \right]$$

Bibliography I

C.M. Bishop (2016). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer New York. ISBN: 9781493938438. URL: https://books.google.com.br/books?id=kOXDtAEACAAJ.

J.L. Crassidis and J.L. Junkins (2011). *Optimal Estimation of Dynamic Systems*. Chapman & Hall/CRC Applied Mathematics & Nonlinear Science. CRC Press. ISBN: 9781439839867. URL: https://books.google.com.br/books?id=CmjRBQAAQBAJ.

P.J. Dhrymes (2013). *Mathematics for Econometrics*. SpringerLink : Bücher. Springer New York. ISBN: 9781461481454. URL: https://books.google.com.br/books?id=HIK8BAAAQBAJ.

J. L. Doob (Sept. 1944). "The Elementary Gaussian Processes". In: *Ann. Math. Statist.* 15.3, pp. 229–282. DOI: 10.1214/aoms/1177731234. URL: https://doi.org/10.1214/aoms/1177731234.

Bibliography II

F.A. Graybill (2001). *Matrices with Applications in Statistics*. Duxbury Classic Series. Brooks/Cole. ISBN: 9780534401313. URL: https://books.google.com.br/books?id=BV3CAAAACAAJ.