Teleinformatics Engineering Department, Federal University of Ceará

Introduction to Gaussian Processes

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

October 26, 2018

Preamble

The Gaussian Processes are the widely used stochastic processes for modeling dependent data observed over time, space or even time and space. Here, we'll iniciate our study with a **Probability and Random Process Theory Review** taking some points to base our journey, going through **Linear Regression** and finally the GP.

Outline

Probability and Random Process Theory Review

Basic Concepts of Probability Theory

A key concept in the field of pattern recognition is that of **uncertainty**, that arises from both through noise on measurements, as well as through the finite size of data sets. To find this uncertainty we'll talk about a little of the **Probability Theory**. Let's begin from a simple example.

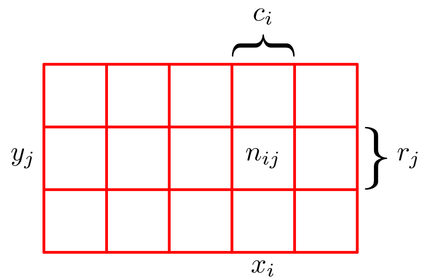


Figure: Considering in the table $X = x_i$ and $Y = y_j$

Let's choose a cell in the 6. We define the probability of choose a cell in a given column is $p(X = x_i) = c_i/N$, being N the total number of cells and $c_i = \sum_j n_{ij}$.

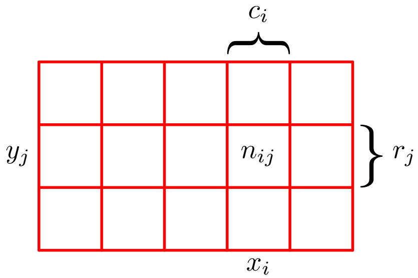


Figure: Considering in the table $X = x_i$ and $Y = y_j$

Let's choose a cell in the 6. We define the probability of choose a cell in a given column is $p(X = x_i) = c_i/N$, being N the total number of cells and $c_i = \sum_j n_{ij}$.

We could say that the probability of choose a cell in a given a row is defined as $p(Y = y_j|X = x_i) = n_{ij}/c_i$.

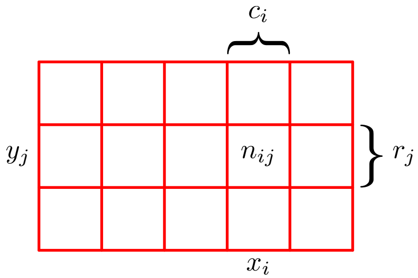


Figure: Considering in the table $X = x_i$ and $Y = y_j$

Let's choose a cell in the 6. We define the probability of choose a cell in a given column is $p(X = x_i) = c_i/N$, being N the total number of cells and $c_i = \sum_j n_{ij}$.

We could say that the probability of choose a cell in a given a row is defined as $p(Y = y_j|X = x_i) = n_{ij}/c_i$.

And so, the probability of choose a cell is defined as $p(X = x_i, Y = y_j) = n_{ij}/N$.

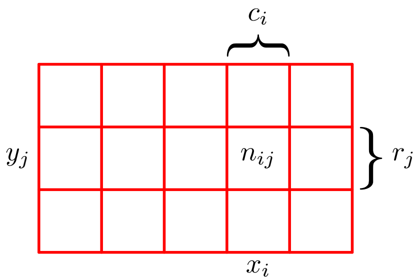


Figure: Considering in the table $X = x_i$ and $Y = y_j$

Here, we could see some properties that we call **The Rules of Probability**

The Rules of Probability

- Sum Rule:

$$p(X = x_i) = \sum_j n_{ij}/N \tag{1}$$

- Product Rule:

$$p(Y = y_j, X = x_i) = n_{ij}/N \tag{2}$$

Here, we could see some properties that we call **The Rules of Probability**

The Rules of Probability

- Sum Rule:

$$p(X = x_i) = \sum_j n_{ij}/N \Rightarrow p(X) = \sum_Y p(X, Y) \tag{1}$$

- Product Rule:

$$p(Y = y_j, X = x_i) = n_{ij}/N \tag{2}$$

Here, we could see some properties that we call **The Rules of Probability**

The Rules of Probability

- Sum Rule:

$$p(X = x_i) = \sum_{j} n_{ij}/N \Rightarrow p(X) = \sum_{Y} p(X, Y) \tag{1}$$

- Product Rule:

$$p(Y = y_j, X = x_i) = n_{ij}/N = n_{ij}/c_i \cdot c_i/N \tag{2}$$

Here, we could see some properties that we call **The Rules of Probability**

The Rules of Probability

- Sum Rule:

$$p(X = x_i) = \sum_j n_{ij}/N \Rightarrow p(X) = \sum_Y p(X, Y) \tag{1}$$

- Product Rule:

$$p(Y = y_j, X = x_i) = n_{ij}/N = n_{ij}/c_i \cdot c_i/N \Rightarrow p(X, Y) = p(Y|X)p(X) \tag{2}$$

And by the **Product Rule** we prove that

Bayes' Theorem

$$p(X, Y) = p(Y|X)p(X) = p(X|Y)p(Y) \Rightarrow p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{3}$$

So, from **The Rules of Probability**, we could show that too

Total Probability Theorem

$$p(X) = \sum_{Y} p(Y|X)p(X) \tag{4}$$

An important propriety of probability is the **Independence of events**. So, let's say that two events occurs without that one has occurred, so by the **Bayes' Theorem** we make

$$p(X|Y) = p(X) \text{ and } p(Y|X) = p(Y) \Rightarrow p(X,Y) = p(X)p(Y) \tag{5}$$

Random Variables

Simplifying, the **Random Variables** will treat the probability defined before in the *continuous domain*. So we define a random variable X as a function that assigns a real number, $X(\zeta)$, to each outcome ζ, so $X(\zeta) = x$.

The Gaussian distribution

The **Gaussian distribution** is defined as

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{6}$$

where μ is the mean and σ^2 the standard deviation.

Independence of two random variables

Sentence

X and Y are independent random variables if *any* event A_1 defined in terms of X is independent of *any* event A_2 defined in terms of Y

Independence of two random variables

Sentence

X and Y are independent random variables if *any* event A_1 defined in terms of X is independent of *any* event A_2 defined in terms of Y

The sentence above is equivalent to say mathematically that

Independence of two random variables

Sentence

X and Y are independent random variables if *any* event A_1 defined in terms of X is independent of *any* event A_2 defined in terms of Y

The sentence above is equivalent to say mathematically that

$$P[X \text{ in } A_1, Y \text{ in } A_2] = P[X \text{ in } A_1]P[Y \text{ in } A_2] \tag{7}$$

Independence of two random variables

Sentence

X and Y are independent random variables if *any* event A_1 defined in terms of X is independent of *any* event A_2 defined in terms of Y

The sentence above is equivalent to say mathematically that

$$P[X \text{ in } A_1, Y \text{ in } A_2] = P[X \text{ in } A_1]P[Y \text{ in } A_2] \tag{7}$$

that means in other words that *if X and Y are independent discrete random variables, then the* ***joint probability mass function (pmf)*** *is equal to the product of the marginal pmf's.*

Linear Regression

Curve Fitting

If we have a data set of points in the space that comes from observations of an experiment, may we want to predict other points in unknown parameters. This could be done with **curve fitting** . So the given points.

Curve Fitting

If we have a data set of points in the space that comes from observations of an experiment, may we want to predict other points in unknown parameters. This could be done with **curve fitting** . So the given points.

We're supposing that our experiment could be modeled somehow, i.e it don't be totally random. So we could define some strategy to find our model.

Curve Fitting

If we have a data set of points in the space that comes from observations of an experiment, may we want to predict other points in unknown parameters. This could be done with **curve fitting** . So the given points.

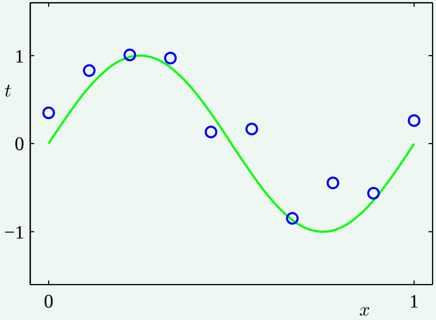
We're supposing that our experiment could be modeled somehow, i.e it don't be totally random. So we could define some strategy to find our model.

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.

Let's fit the example by polynomial curve fitting

Example



Be the model chosen

Curve Fitting

Be the model chosen

$$
\begin{aligned}
y(x, \mathbf{w}) &= w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M \\
&= \sum_{j=0}^{M} w_j x^j
\end{aligned} \tag{8}
$$

Be the model chosen

$$
\begin{aligned}
y(x, \mathbf{w}) &= w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M \\
&= \sum_{j=0}^{M} w_j x^j
\end{aligned} \tag{8}
$$

So we'll try to minimize the mean squared error between the model and the training data set. So, if the MSE is defined as

Be the model chosen

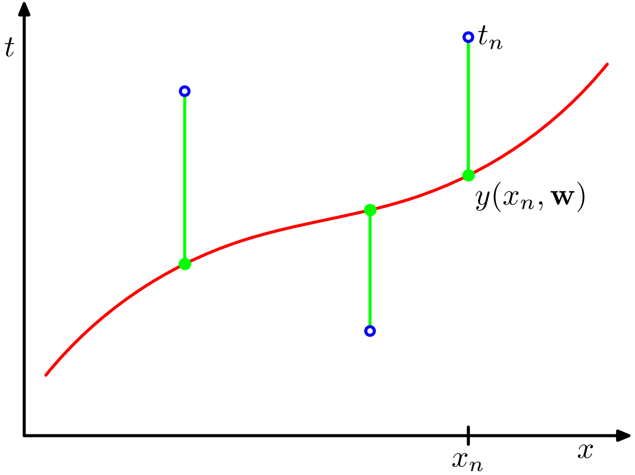
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M$$
$$= \sum_{j=0}^{M} w_j x^j \tag{8}$$

So we'll try to minimize the mean squared error between the model and the training data set. So, if the MSE is defined as

$$E(\mathbf{w}) \triangleq \frac{1}{2} \sum_{n=1}^{N} \{\hat{y}_n - y_n\}^2 \tag{9}$$

where $\hat{y}$ is our predict.

Curve Fitting



So, we evaluate the error between our model and the target **t**, that is our training data.

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y_n(x, \mathbf{w}) - t_n\}^2 \tag{10}$$

And try to minimize it by

$$\frac{\partial}{\partial \mathbf{w}} E(\mathbf{w}) = 0$$

$$\sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}} y(x, \mathbf{w}) \{y_n(x, \mathbf{w}) - t_n\} = 0 \tag{11}$$

$$\begin{bmatrix} \frac{\partial}{\partial w_1} y(x, \mathbf{w}) \left\{ y(x, \mathbf{w}) - t_n \right\} \\ \frac{\partial}{\partial w_2} y(x, \mathbf{w}) \left\{ y(x, \mathbf{w}) - t_n \right\} \\ \dots \\ \frac{\partial}{\partial w_M} y(x, \mathbf{w}) \left\{ y(x, \mathbf{w}) - t_n \right\} \end{bmatrix} = \tag{12}$$

Bayesian Curve Fitting

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

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

Sentence

If we could update the ***regression weights*** *as we acquire some new values of the experiment?*

Bayesian Curve Fitting

Let's take a look again at the Bayes Theorem

Let's take a look again at the Bayes Theorem

Bayes Theorem

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})\,\overbrace{p(\mathbf{w})}^{\text{the weights probability}}}{\underbrace{p(\mathcal{D})}_{\text{the data probability}}} \tag{13}$$

Let's take a look again at the Bayes Theorem

Bayes Theorem

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})\, \overbrace{p(\mathbf{w})}^{\text{the weights probability}}}{\underbrace{p(\mathcal{D})}_{\text{the data probability}}} \tag{13}$$

So, if **we have the probability** of the data, we'll could estimate the **future weights**.

Let's take a look again at the Bayes Theorem

Bayes Theorem

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})\,\overbrace{p(\mathbf{w})}^{\text{the weights probability}}}{\underbrace{p(\mathcal{D})}_{\text{the data probability}}} \tag{13}$$

So, if **we have the probability** of the data, we'll could estimate the **future weights**.
But, how?

Bayesian Curve Fitting

Taking some steps back, let's re-visit the **Curve Fitting**. There, the strategy was minimize the error function.

Bayesian Curve Fitting

Taking some steps back, let's re-visit the **Curve Fitting**. There, the strategy was minimize the error function.

Now we'll try to view the same problem with a *probabilistic perspective*. We're trying to make predictions for the target value $\mathbf{t}$ given some new values of x.

Bayesian Curve Fitting

Taking some steps back, let's re-visit the **Curve Fitting**. There, the strategy was minimize the error function.

Now we'll try to view the same problem with a *probabilistic perspective*. We're trying to make predictions for the target value $\mathbf{t}$ given some new values of x.

A good ideia is to express our target values $\mathbf{t}$ in terms of **gaussians distributions** with the mean equals to $y(x, \mathbf{w})$.

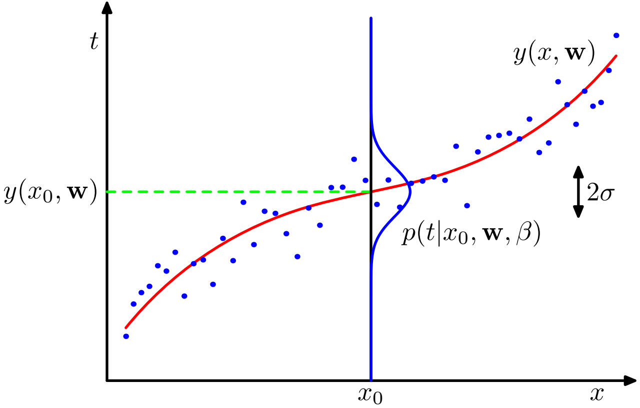


Figure: Schematic of the polynomial function $y(x, \mathbf{w})$ and the gaussian distribution p.

By the 24, we assume the relation

(15)

(16)

Bayesian Curve Fitting

By the 24, we assume the relation

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \tag{14}$$

(15)

(16)

Bayesian Curve Fitting

By the 24, we assume the relation

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \tag{14}$$

and then, assume that the training data $\{\mathbf{x}, \mathbf{t}\}$ is indenpendent and identically distributed (i.i.d.) and put on **product form**, i.e. the joint probability is

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t_0|y(x_1, \mathbf{w}), \beta^{-1}) \cap \mathcal{N}(t_n|y(x_0, \mathbf{w}), \beta^{-1}) \ldots \cap \mathcal{N}(t_n|y(x_0, \mathbf{w}), \beta^{-1})$$

(15)

(16)

Bayesian Curve Fitting

By the 24, we assume the relation

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \tag{14}$$

and then, assume that the training data $\{\mathbf{x}, \mathbf{t}\}$ is indenpendent and identically distributed (i.i.d.) and put on **product form**, i.e. the joint probability is

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t_0|y(x_1, \mathbf{w}), \beta^{-1}) \cap \mathcal{N}(t_n|y(x_0, \mathbf{w}), \beta^{-1})... \cap \mathcal{N}(t_n|y(x_0, \mathbf{w}), \beta^{-1}) \tag{15}$$

$$= \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1}) \tag{16}$$

regarding that $\beta^{-1} = \sigma^2$.

Bayesian Curve Fitting

And we'll have a function to maximize if we apply the logarithm function to p, so

Bayesian Curve Fitting

And we'll have a function to maximize if we apply the logarithm function to p, so

$$\ln\left(p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)\right) = \sum_{n=1}^{N} \ln\left(\mathcal{N}(t_n|y(x_n,\mathbf{w}),\beta^{-1})\right) \tag{17}$$

Bayesian Curve Fitting

And we'll have a function to maximize if we apply the logarithm function to p, so

$$\ln\left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = \sum_{n=1}^{N} \ln\left(\mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1})\right) \tag{17}$$

Applying the **Gaussian distribution** (see 6) will result

$$\ln\left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w} - t_n)\}^2 + \frac{N}{2}\ln(\beta) - \frac{N}{2}\ln(2\pi) \tag{18}$$

Bayesian Curve Fitting

And taking the derivatives with respect to β to minimize the error

(19)

(21)

And taking the derivatives with respect to β to minimize the error

$$\frac{\partial}{\partial \beta} \ln \left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = 0 \tag{19}$$

$$\tag{21}$$

Bayesian Curve Fitting

And taking the derivatives with respect to β to minimize the error

$$\frac{\partial}{\partial\beta} \ln\left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = 0 \tag{19}$$

$$-\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w} - t_n)\}^2 + \frac{N}{2}\frac{1}{\beta} = 0 \tag{20}$$

$$\tag{21}$$

Bayesian Curve Fitting

And taking the derivatives with respect to β to minimize the error

$$\frac{\partial}{\partial \beta} \ln \left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = 0 \tag{19}$$

$$-\frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w} - t_n)\}^2 + \frac{N}{2} \frac{1}{\beta} = 0 \tag{20}$$

$$\frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w} - t_n)\}^2 = \frac{1}{\beta_{ML}} \tag{21}$$

Where β_{ML} is the maximum likelihood.