Teleinformatics Engineering Department, Federal University of Ceará

Introduction to Gaussian Processes

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

October 24, 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 point to base our journey, going through **Linear Regression** and finally the **Gaussian Processes**.

The material here presented isn't sufficient to guide you over basic probability, so it's recommended to have some knowledge, once we'll just take a simple review.

Outline

1. Probability and Random Process Theory Review
   - 1.1 Basic Concepts of Probability Theory
   - 1.2 Random Variables
   - 1.3 The Gaussian distribution
   - 1.4 Independence of two random variables
2. Linear Regression
   - 2.1 Basic Concepts of Curve Fitting
   - 2.2 Bayesian Curve Fitting

Probability and Random Process Theory Review

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.

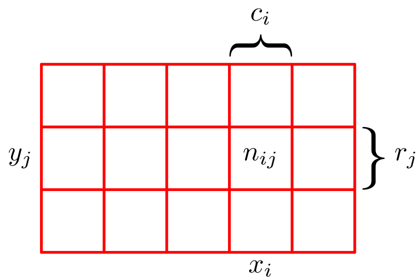

Figure: Considering in the table $X = x_i$ and $Y = y_j$

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 $c_i = \sum_j n_{ij}$.

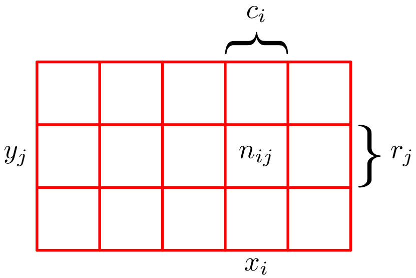

Figure: Considering in the table $X = x_i$ and $Y = y_j$

Basic Concepts of Probability Theory

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 $p(Y = y_j|X = x_i) = n_{ij}/c_i$.

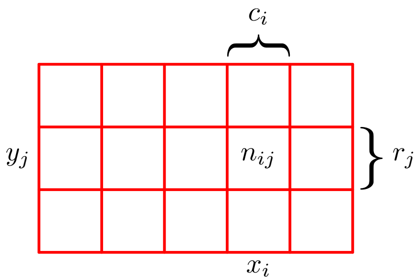

Figure: Considering in the table $X = x_i$ and $Y = y_j$

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

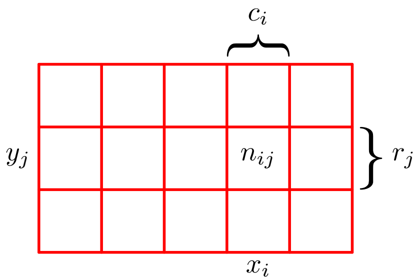

Figure: Considering in the table $X = x_i$ and $Y = y_j$

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

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

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

Bayesian Curve Fitting

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

Bayesian Curve Fitting

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

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w}) \overbrace{p(\mathbf{w})}^{\text{the weights probability}}}{\underbrace{p(\mathcal{D})}_{\text{the data probability}}} \tag{8}$$

Let's take a look again at the Bayes Theorem

Bayes Theorem

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})\overbrace{p(\mathbf{w})}^{\text{the weights probability}}}{\underbrace{p(\mathcal{D})}_{\text{the data probability}}} \tag{8}$$

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

Bayesian Curve Fitting

Let's take a look again at the Bayes Theorem

Bayes Theorem

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})\,\overbrace{p(\mathbf{w})}^{\text{the weights probability}}}{\underbrace{p(\mathcal{D})}_{\text{the data probability}}} \tag{8}$$

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 $y(x, \mathbf{w})$.

Bayesian Curve Fitting

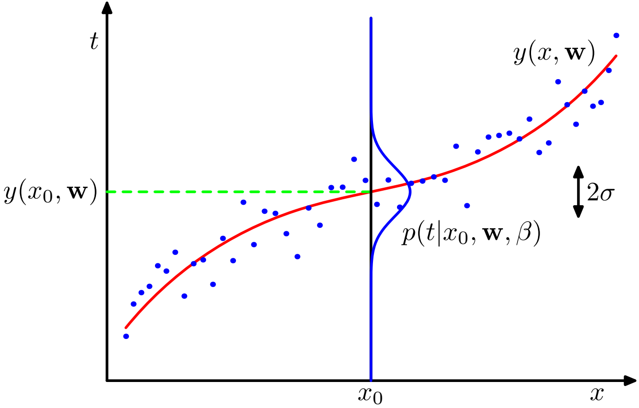

Figure: Schematic of the polynomial function $y(x, \mathbf{w})$ and the gaussian distribution p.

By the 18, we assume the relation

(10)

(11)

Bayesian Curve Fitting

By the 18, we assume the relation

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \tag{9}$$

(10)

(11)

Bayesian Curve Fitting

By the 18, we assume the relation

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \tag{9}$$

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

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

(10)

(11)

Bayesian Curve Fitting

By the 18, we assume the relation

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \tag{9}$$

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

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t_0|y(x_1, \mathbf{w}), \beta^{-1}) \cap \mathcal{N}(t_n|y(x_0, \mathbf{w}), \beta^{-1}) \ldots \cap \mathcal{N}(t_n|y(x_0, \mathbf{w}), \beta^{-1}) \tag{10}$$

$$= \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1}) \tag{11}$$

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

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

$$\ln\left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = \sum_{n=1}^{N} \ln\left(\mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1})\right) \tag{12}$$

Bayesian Curve Fitting

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

$$\ln\left(p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)\right) = \sum_{n=1}^{N} \ln\left(\mathcal{N}(t_n|y(x_n,\mathbf{w}),\beta^{-1})\right) \tag{12}$$

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

$$\ln\left(p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)\right) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}-t_n)\}^2 + \frac{N}{2}\ln(\beta) - \frac{N}{2}\ln(2\pi) \tag{13}$$

Bayesian Curve Fitting

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

(14)

(16)

Bayesian Curve Fitting

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

$$\frac{\partial}{\partial \beta} \ln\left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = 0 \tag{14}$$

$$\tag{16}$$

Bayesian Curve Fitting

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

$$\frac{\partial}{\partial \beta} \ln \left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = 0 \tag{14}$$

$$-\frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w} - t_n)\}^2 + \frac{N}{2} \frac{1}{\beta} = 0 \tag{15}$$

$$\tag{16}$$

Bayesian Curve Fitting

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

$$\frac{\partial}{\partial \beta} \ln \left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = 0 \tag{14}$$

$$-\frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w} - t_n)\}^2 + \frac{N}{2} \frac{1}{\beta} = 0 \tag{15}$$

$$\frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w} - t_n)\}^2 = \frac{1}{\beta_{ML}} \tag{16}$$

Where β_{ML} is the maximum likelihood.