Teleinformatics Engineering Department, Federal University of Ceará

Introduction to Gaussian Processes

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

October 23, 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

Probability and Random Process Theory Review

Specifying Random Experiments

Let's start with the sentence

Sentence

*A **random experiment** is specified by stating an **experimental procedure** and a **set** of one or more measurements or observations.*

The Gaussian distribution

The **Gaussian distribution** is defined as

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

Independency of two random variables

Independency 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

Independency 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

Independency 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

$$P[X \text{ in } A_1, Y \text{ in } A_2] = P[X \text{ in } A_1]P[Y \text{ in } A_2] \tag{2}$$

Independency 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

$$P[X \text{ in } A_1, Y \text{ in } A_2] = P[X \text{ in } A_1]P[Y \text{ in } A_2] \tag{2}$$

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.

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

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})\,\overbrace{p(\mathbf{w})}^{\text{the weights probability}}}{\underbrace{p(\mathcal{D})}_{\text{the data probability}}} \tag{3}$$

Let's take a look again at the Bayes Theorem

Bayes Theorem

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})\overbrace{p(\mathbf{w})}^{\text{the weights probability}}}{\underbrace{p(\mathcal{D})}_{\text{the data probability}}} \tag{3}$$

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

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w}) \overbrace{p(\mathbf{w})}^{\text{the weights probability}}}{\underbrace{p(\mathcal{D})}_{\text{the data probability}}} \tag{3}$$

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

But, how?

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

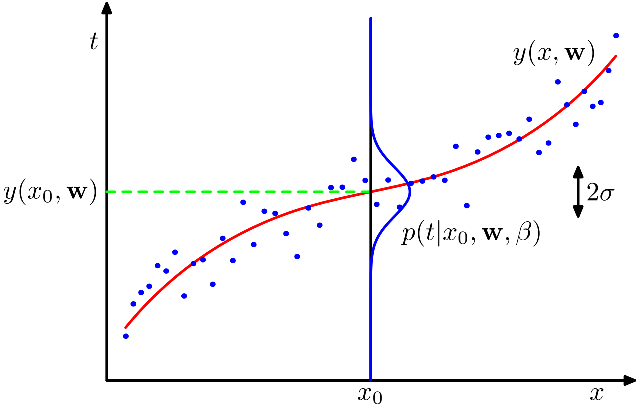

Figure: Schematic of the polynomial function $y(x, \mathbf{w})$ and the gaussian distribution p.

By the 12, we assume the relation

(5)

(6)

Bayesian Curve Fitting

By the 12, we assume the relation

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \tag{4}$$

(5)

(6)

By the 12, we assume the relation

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \tag{4}$$

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

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t_0|y(x_1, \mathbf{w}), \beta^{-1}) \cap \mathcal{N}(t_n|y(x_0, \mathbf{w}), \beta^{-1}) ... \cap \mathcal{N}(t_n|y(x_0, \mathbf{w}), \beta^{-1}) \tag{5}$$

$$\tag{6}$$

Bayesian Curve Fitting

By the 12, we assume the relation

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \tag{4}$$

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

$$\begin{aligned} p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) &= \mathcal{N}(t_0|y(x_1, \mathbf{w}), \beta^{-1}) \cap \mathcal{N}(t_n|y(x_0, \mathbf{w}), \beta^{-1}) \ldots \cap \mathcal{N}(t_n|y(x_0, \mathbf{w}), \beta^{-1}) & (5) \\ &= \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1}) & (6) \end{aligned}$$

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

$$\ln\left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = \sum_{n=1}^{N} \ln\left(\mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1})\right) \tag{7}$$

Bayesian Curve Fitting

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

$$\ln\left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = \sum_{n=1}^{N} \ln\left(\mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1})\right) \tag{7}$$

Applying the **Gaussian distribution** (see 1) will result

$$\ln\left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w} - t_n)\}^2 + \frac{N}{2}\ln(\beta) - \frac{N}{2}\ln(2\pi) \tag{8}$$

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

(9)

(11)

Bayesian Curve Fitting

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

$$\frac{\partial}{\partial \beta} \ln \left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) \right) = 0 \tag{9}$$

$$\tag{11}$$

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

$$\frac{\partial}{\partial \beta} \ln\left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)\right) = 0 \tag{9}$$

$$-\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w} - t_n)\}^2 + \frac{N}{2}\frac{1}{\beta} = 0 \tag{10}$$

$$\tag{11}$$

Bayesian Curve Fitting

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

$$\frac{\partial}{\partial \beta} \ln \left(p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) \right) = 0 \tag{9}$$

$$-\frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w} - t_n)\}^2 + \frac{N}{2} \frac{1}{\beta} = 0 \tag{10}$$

$$\frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w} - t_n)\}^2 = \frac{1}{\beta_{ML}} \tag{11}$$

Where β_{ML} is the maximum likelihood.