

# Clase 20 - Procesos Gaussianos

## Aprendizaje de Máquinas - MA5204

Felipe Tobar

Department of Mathematical Engineering &  
Center for Mathematical Modelling  
Universidad de Chile

18 de marzo de 2021



UNIVERSIDAD  
DE CHILE

## Modelos paramétricos y no paramétricos

Entre los métodos que hasta ahora hemos visto, se encuentran el de regresión lineal y no lineal. Una característica en común que tienen estos métodos es que el proceso de entrenamiento consiste en encontrar un número fijo de parámetros, que minimicen cierta función objetivo, donde la cantidad de parámetros y la forma del modelo es parte del diseño. A este tipo de modelos se les llama *modelos paramétricos*.

En contraste, se encuentran los modelos *no paramétricos* los cuales no tienen un número fijo de parámetros, donde pueden llegar a ser en algunos casos infinito, un ejemplo clásico de esto Un ejemplo es el algoritmo de  $k$ -vecinos más cercanos (KNN), donde los puntos del conjunto de entrenamiento son usados para clasificar nuevas muestras. Otro ejemplo son las máquinas de soporte vectorial (SVM) donde al entrenar se obtienen los vectores de soporte.

En este capítulo introduciremos un método no paramétrico probabilístico de regresión no lineal, llamado procesos gaussianos ( $\mathcal{GP}$ ). Este modelo en vez de encontrar un candidato único de la función a estimar, define una distribución sobre funciones  $\mathbb{P}(f)$ , donde  $f$  es una función de un espacio de entrada  $\mathcal{X}$  a los reales,  $f : \mathcal{X} \rightarrow \mathbb{R}$ . Esto tiene la virtud de permitir cuantificar la incertidumbre puntual que existe en la predicción de nuestro modelo, la cual servirá en forma de intervalos de confianza para la distribución gaussiana.

## Modelos paramétricos y no paramétricos

Entre los métodos que hasta ahora hemos visto, se encuentran el de regresión lineal y no lineal. Una característica en común que tienen estos métodos es que el proceso de entrenamiento consiste en encontrar un número fijo de parámetros, que minimicen cierta función objetivo, donde la cantidad de parámetros y la forma del modelo es parte del diseño. A este tipo de modelos se les llama *modelos paramétricos*.

En contraste, se encuentran los modelos *no paramétricos* los cuales no tienen un número fijo de parámetros, donde pueden llegar a ser en algunos casos infinito, un ejemplo clásico de esto Un ejemplo es el algoritmo de  $k$ -vecinos más cercanos (KNN), donde los puntos del conjunto de entrenamiento son usados para clasificar nuevas muestras. Otro ejemplo son las máquinas de soporte vectorial (SVM) donde al entrenar se obtienen los vectores de soporte.

En este capítulo introduciremos un método no paramétrico probabilístico de regresión no lineal, llamado procesos gaussianos ( $\mathcal{GP}$ ). Este modelo en vez de encontrar un candidato único de la función a estimar, define una distribución sobre funciones  $\mathbb{P}(f)$ , donde  $f$  es una función de un espacio de entrada  $\mathcal{X}$  a los reales,  $f : \mathcal{X} \rightarrow \mathbb{R}$ . Esto tiene la virtud de permitir cuantificar la incertidumbre puntual que existe en la predicción de nuestro modelo, la cual servirá en forma de intervalos de confianza para la distribución gaussiana.

## Modelos paramétricos y no paramétricos

Entre los métodos que hasta ahora hemos visto, se encuentran el de regresión lineal y no lineal. Una característica en común que tienen estos métodos es que el proceso de entrenamiento consiste en encontrar un número fijo de parámetros, que minimicen cierta función objetivo, donde la cantidad de parámetros y la forma del modelo es parte del diseño. A este tipo de modelos se les llama *modelos paramétricos*.

En contraste, se encuentran los modelos *no paramétricos* los cuales no tienen un número fijo de parámetros, donde pueden llegar a ser en algunos casos infinito, un ejemplo clásico de esto Un ejemplo es el algoritmo de  $k$ -vecinos más cercanos (KNN), donde los puntos del conjunto de entrenamiento son usados para clasificar nuevas muestras. Otro ejemplo son las máquinas de soporte vectorial (SVM) donde al entrenar se obtienen los vectores de soporte.

En este capítulo introduciremos un método no paramétrico probabilístico de regresión no lineal, llamado procesos gaussianos ( $\mathcal{GP}$ ). Este modelo en vez de encontrar un candidato único de la función a estimar, define una distribución sobre funciones  $\mathbb{P}(f)$ , donde  $f$  es una función de un espacio de entrada  $\mathcal{X}$  a los reales,  $f : \mathcal{X} \rightarrow \mathbb{R}$ . Esto tiene la virtud de permitir cuantificar la incertidumbre puntual que existe en la predicción de nuestro modelo, la cual servirá en forma de intervalos de confianza para la distribución gaussiana.

# Proceso Gaussiano

## Definition (proceso gaussiano)

Un proceso gaussiano ( $\mathcal{GP}$ ) es una colección de variables aleatorias, tal que para cualquier subconjunto finito de puntos, estos tienen una distribución conjuntamente gaussiana.

Al aplicar esta definición a nuestro caso anterior,  $\mathbb{P}(f)$  será un  $\mathcal{GP}$  y para cualquier conjunto finito  $\{x_i\}_{i=1}^n \subset \mathcal{X}$ , la distribución de  $\mathbb{P}(f(\mathbf{x}))$  es Gaussiana multivariada  $f(\mathbf{x}) = (f(x_1), \dots, f(x_n))^T$ . En este caso las variables aleatorias representan el valor de la función  $f(x_i)$  en la posición  $x_i$ .

Un  $\mathcal{GP}$  queda completamente caracterizado por su función de media  $m(\cdot)$  y función de covarianza  $K(\cdot, \cdot)$ , de esta forma para cualquier conjunto finito podemos encontrar la distribución. Definimos estas funciones como

$$m(x) = \mathbb{E} \{f(x)\}$$
$$K(x, x') = \mathbb{E} \left\{ (f(x) - m(x)) (f(x') - m(x')) \right\}$$

# Proceso Gaussiano

## Definition (proceso gaussiano)

Un proceso gaussiano ( $\mathcal{GP}$ ) es una colección de variables aleatorias, tal que para cualquier subconjunto finito de puntos, estos tienen una distribución conjuntamente gaussiana.

Al aplicar esta definición a nuestro caso anterior,  $\mathbb{P}(f)$  será un  $\mathcal{GP}$  y para cualquier conjunto finito  $\{x_i\}_{i=1}^n \subset \mathcal{X}$ , la distribución de  $\mathbb{P}(f(\mathbf{x}))$  es Gaussiana multivariada  $f(\mathbf{x}) = (f(x_1), \dots, f(x_n))^T$ . En este caso las variables aleatorias representan el valor de la función  $f(x_i)$  en la posición  $x_i$ .

Un  $\mathcal{GP}$  queda completamente caracterizado por su función de media  $m(\cdot)$  y función de covarianza  $K(\cdot, \cdot)$ , de esta forma para cualquier conjunto finito podemos encontrar la distribución. Definimos estas funciones como

$$m(x) = \mathbb{E} \{f(x)\}$$
$$K(x, x') = \mathbb{E} \left\{ (f(x) - m(x)) (f(x') - m(x')) \right\}$$

# Proceso Gaussiano

## Definition (proceso gaussiano)

Un proceso gaussiano ( $\mathcal{GP}$ ) es una colección de variables aleatorias, tal que para cualquier subconjunto finito de puntos, estos tienen una distribución conjuntamente gaussiana.

Al aplicar esta definición a nuestro caso anterior,  $\mathbb{P}(f)$  será un  $\mathcal{GP}$  y para cualquier conjunto finito  $\{x_i\}_{i=1}^n \subset \mathcal{X}$ , la distribución de  $\mathbb{P}(f(\mathbf{x}))$  es Gaussiana multivariada  $f(\mathbf{x}) = (f(x_1), \dots, f(x_n))^T$ . En este caso las variables aleatorias representan el valor de la función  $f(x_i)$  en la posición  $x_i$ .

Un  $\mathcal{GP}$  queda completamente caracterizado por su función de media  $m(\cdot)$  y función de covarianza  $K(\cdot, \cdot)$ , de esta forma para cualquier conjunto finito podemos encontrar la distribución. Definimos estas funciones como

$$m(x) = \mathbb{E} \{f(x)\}$$
$$K(x, x') = \mathbb{E} \left\{ (f(x) - m(x)) (f(x') - m(x')) \right\}$$

## Proceso Gaussiano

Y de esta forma podemos escribir el proceso como:

$$f \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Donde para un conjunto finito tenemos que la marginal resulta de la forma:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

Hasta el momento hemos hablado del espacio de entrada  $\mathcal{X}$  como genérico, un caso común es definir los  $\mathcal{GP}$  sobre el tiempo ( $\mathbb{R}^+$ ), es decir que los  $x_i$  son instantes de tiempo. Es de notar que este no es el único caso, y se podría definir sobre un espacio más general, por ejemplo  $\mathbb{R}^d$ .

Otro punto a notar es que como estamos hablando de una colección (no necesariamente finita) de variables aleatorias, es necesario que se cumpla la propiedad de marginalización (o llamada consistencia). Esta propiedad se refiere a que si un  $\mathcal{GP}$  define una distribución multivariada para digamos dos variables  $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$  entonces también debe definir  $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$  donde  $\mu_1$  es la componente respectiva del vector  $\mu$  y  $\Sigma_{11}$  la submatriz correspondiente de  $\Sigma$ . En otras palabras, el tomar un subconjunto más grande de puntos no cambia la distribución de un subconjunto más pequeño. Y podemos notar que esta condición se cumple si tomamos la función de covarianza definida anteriormente.



## Proceso Gaussiano

Y de esta forma podemos escribir el proceso como:

$$f \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Donde para un conjunto finito tenemos que la marginal resulta de la forma:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

Hasta el momento hemos hablado del espacio de entrada  $\mathcal{X}$  como genérico, un caso común es definir los  $\mathcal{GP}$  sobre el tiempo ( $\mathbb{R}^+$ ), es decir que los  $x_i$  son instantes de tiempo. Es de notar que este no es el único caso, y se podría definir sobre un espacio más general, por ejemplo  $\mathbb{R}^d$ .

Otro punto a notar es que como estamos hablando de una colección (no necesariamente finita) de variables aleatorias, es necesario que se cumpla la propiedad de marginalización (o llamada consistencia). Esta propiedad se refiere a que si un  $\mathcal{GP}$  define una distribución multivariada para digamos dos variables  $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$  entonces también debe definir  $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$  donde  $\mu_1$  es la componente respectiva del vector  $\mu$  y  $\Sigma_{11}$  la submatriz correspondiente de  $\Sigma$ . En otras palabras, el tomar un subconjunto más grande de puntos no cambia la distribución de un subconjunto más pequeño. Y podemos notar que esta condición se cumple si tomamos la función de covarianza definida anteriormente.

## Proceso Gaussiano

Y de esta forma podemos escribir el proceso como:

$$f \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Donde para un conjunto finito tenemos que la marginal resulta de la forma:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

Hasta el momento hemos hablado del espacio de entrada  $\mathcal{X}$  como genérico, un caso común es definir los  $\mathcal{GP}$  sobre el tiempo ( $\mathbb{R}^+$ ), es decir que los  $x_i$  son instantes de tiempo. Es de notar que este no es el único caso, y se podría definir sobre un espacio más general, por ejemplo  $\mathbb{R}^d$ .

Otro punto a notar es que como estamos hablando de una colección (no necesariamente finita) de variables aleatorias, es necesario que se cumpla la propiedad de marginalización (o llamada consistencia). Esta propiedad se refiere a que si un  $\mathcal{GP}$  define una distribución multivariada para digamos dos variables  $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$  entonces también debe definir  $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$  donde  $\mu_1$  es la componente respectiva del vector  $\mu$  y  $\Sigma_{11}$  la submatriz correspondiente de  $\Sigma$ . En otras palabras, el tomar un subconjunto más grande de puntos no cambia la distribución de un subconjunto más pequeño. Y podemos notar que esta condición se cumple si tomamos la función de covarianza definida anteriormente.

## Proceso Gaussiano

Y de esta forma podemos escribir el proceso como:

$$f \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Donde para un conjunto finito tenemos que la marginal resulta de la forma:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

Hasta el momento hemos hablado del espacio de entrada  $\mathcal{X}$  como genérico, un caso común es definir los  $\mathcal{GP}$  sobre el tiempo ( $\mathbb{R}^+$ ), es decir que los  $x_i$  son instantes de tiempo. Es de notar que este no es el único caso, y se podría definir sobre un espacio más general, por ejemplo  $\mathbb{R}^d$ .

Otro punto a notar es que como estamos hablando de una colección (no necesariamente finita) de variables aleatorias, es necesario que se cumpla la propiedad de marginalización (o llamada consistencia). Esta propiedad se refiere a que si un  $\mathcal{GP}$  define una distribución multivariada para digamos dos variables  $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$  entonces también debe definir  $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$  donde  $\mu_1$  es la componente respectiva del vector  $\mu$  y  $\Sigma_{11}$  la submatriz correspondiente de  $\Sigma$ . En otras palabras, el tomar un subconjunto más grande de puntos no cambia la distribución de un subconjunto más pequeño. Y podemos notar que esta condición se cumple si tomamos la función de covarianza definida anteriormente.

## Proceso Gaussiano

Y de esta forma podemos escribir el proceso como:

$$f \sim \mathcal{GP}(m(\cdot), K(\cdot, \cdot))$$

Donde para un conjunto finito tenemos que la marginal resulta de la forma:

$$f(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}))$$

Hasta el momento hemos hablado del espacio de entrada  $\mathcal{X}$  como genérico, un caso común es definir los  $\mathcal{GP}$  sobre el tiempo ( $\mathbb{R}^+$ ), es decir que los  $x_i$  son instantes de tiempo. Es de notar que este no es el único caso, y se podría definir sobre un espacio más general, por ejemplo  $\mathbb{R}^d$ .

Otro punto a notar es que como estamos hablando de una colección (no necesariamente finita) de variables aleatorias, es necesario que se cumpla la propiedad de marginalización (o llamada consistencia). Esta propiedad se refiere a que si un  $\mathcal{GP}$  define una distribución multivariada para digamos dos variables  $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$  entonces también debe definir  $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$  donde  $\mu_1$  es la componente respectiva del vector  $\mu$  y  $\Sigma_{11}$  la submatriz correspondiente de  $\Sigma$ . En otras palabras, el tomar un subconjunto más grande de puntos no cambia la distribución de un subconjunto más pequeño. Y podemos notar que esta condición se cumple si tomamos la función de covarianza definida anteriormente.

# Clase 20 - Procesos Gaussianos

## Aprendizaje de Máquinas - MA5204

Felipe Tobar

Department of Mathematical Engineering &  
Center for Mathematical Modelling  
Universidad de Chile

18 de marzo de 2021



UNIVERSIDAD  
DE CHILE