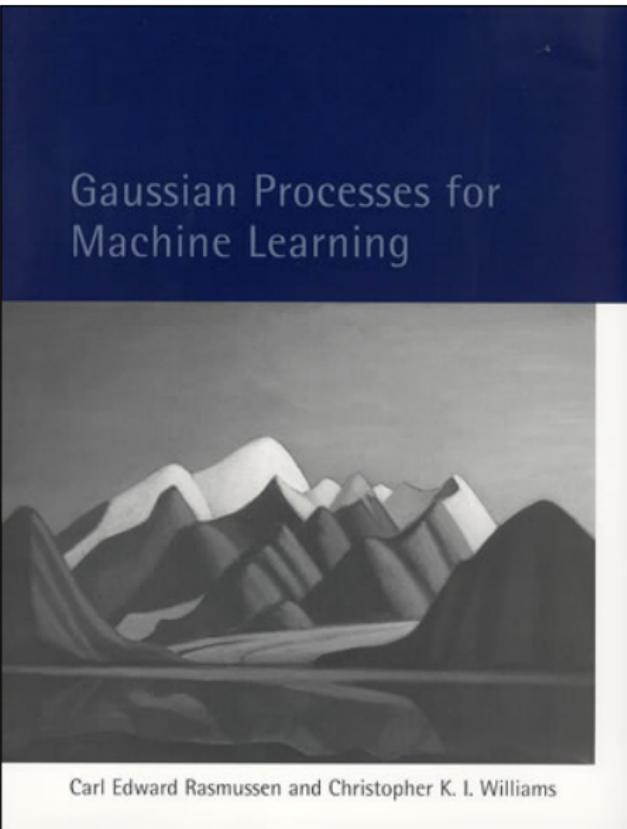


# Introducción a los Procesos Gaussianos en Aprendizaje de Máquina

Mauricio A. Álvarez, PhD

Curso de entrenamiento ArcelorMittal



# Contenido

Introducción

Regresión

Clasificación

Maximum a posterior

Aproximación de Laplace

Modelo lineal

usando procesos Gaussianos

# Contenido

Introducción

Regresión

Clasificación

Maximum a posterior

Aproximación de Laplace

Modelo lineal

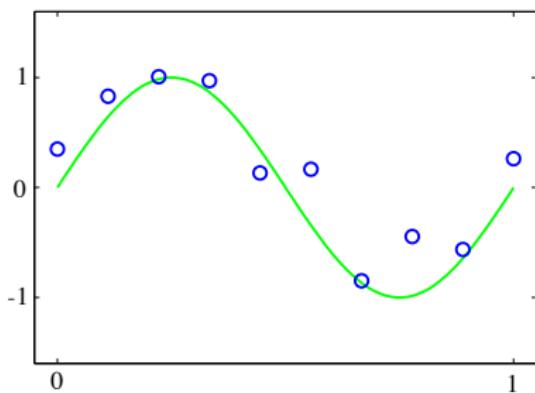
usando procesos Gaussianos

# Aprendizaje supervisado

Aprender el mapeo de un conjunto de variables de entrada a una o más variables de salida, a partir de un conjunto finito de datos.

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

Clasificación



Regresión

# Notación

- En general, la entrada se denota como  $\mathbf{x}$ , y la salida u objetivo como  $y$ .
- La variable objetivo  $y$  puede ser continua (regresión), o discreta (clasificación).
- Base de datos de  $n$  observaciones:  $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \dots, n\}$ .
- Se busca diseñar un modelo  $f(\mathbf{x}) : \mathbf{x} \rightarrow y$ .

# Inducción

- Dado el conjunto de entrenamiento ( $\mathcal{D}$ ), se desea hacer predicciones para un nuevo  $\mathbf{x}_*$ .
- Inducción: pasar de un conjunto de datos  $\mathcal{D}$  a una función  $f$ .
- Generalización: buen desempeño sobre datos nuevos.
- Presunciones acerca de  $f$ .

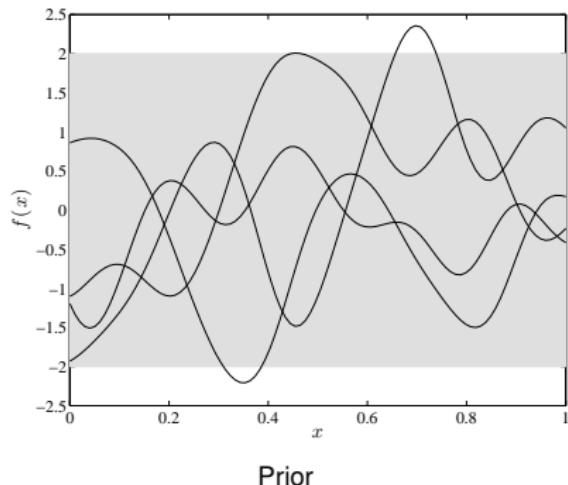
## Dos alternativas

- Primer enfoque: restringir la clase de funciones que se pueden considerar.
- Problema:
  - La función objetivo podría quedar mal modelada, luego las predicciones serán pobres.
  - Aumentar la flexibilidad de la clase de funciones, con el peligro de sobre-entrenar.
- Segundo enfoque: darle a cada función posible una probabilidad prior.
- Problema: cómo asignarle una probabilidad a un conjunto infinito de posibles funciones.

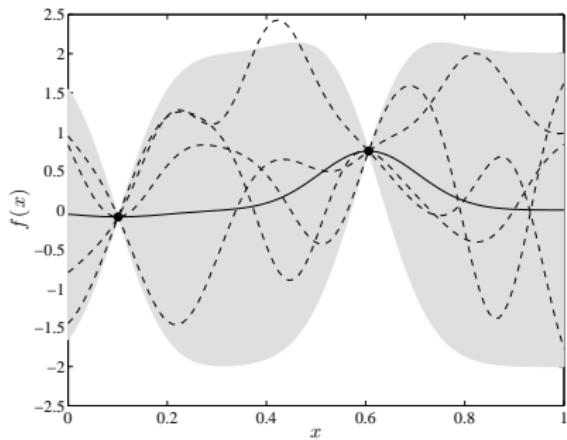
# Procesos Gaussianos

- Un proceso Gaussiano (GP) es un proceso estocástico, generalización de la distribución de probabilidad Gaussiana.
- Un proceso Gaussiano le asigna una probabilidad a una función  $f$ .
- Tratabilidad computacional: sólo es necesario conocer las propiedades de la función en un conjunto finito de puntos.

# Modelamiento Bayesiano: regresión (I)



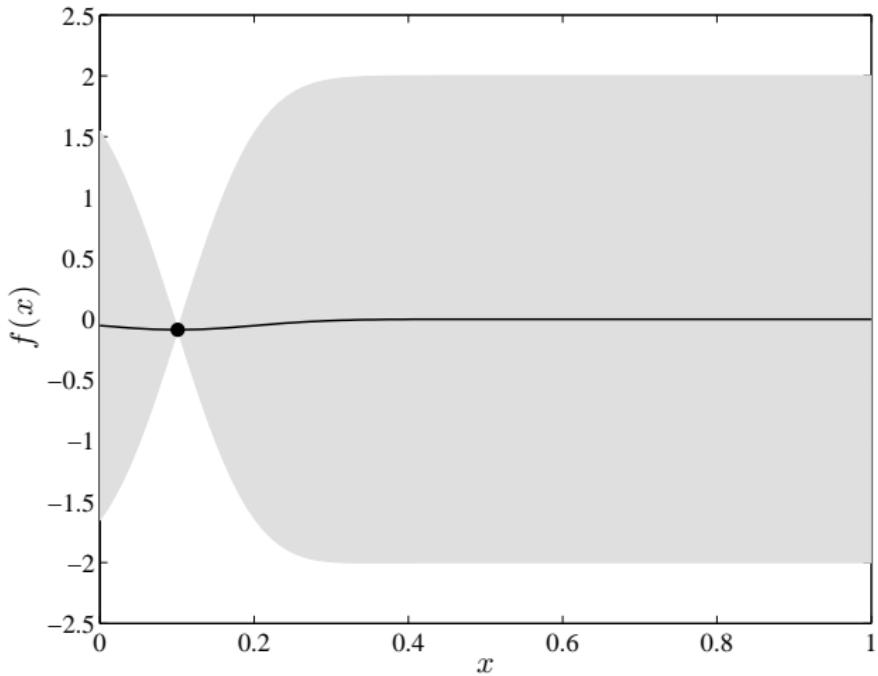
Prior



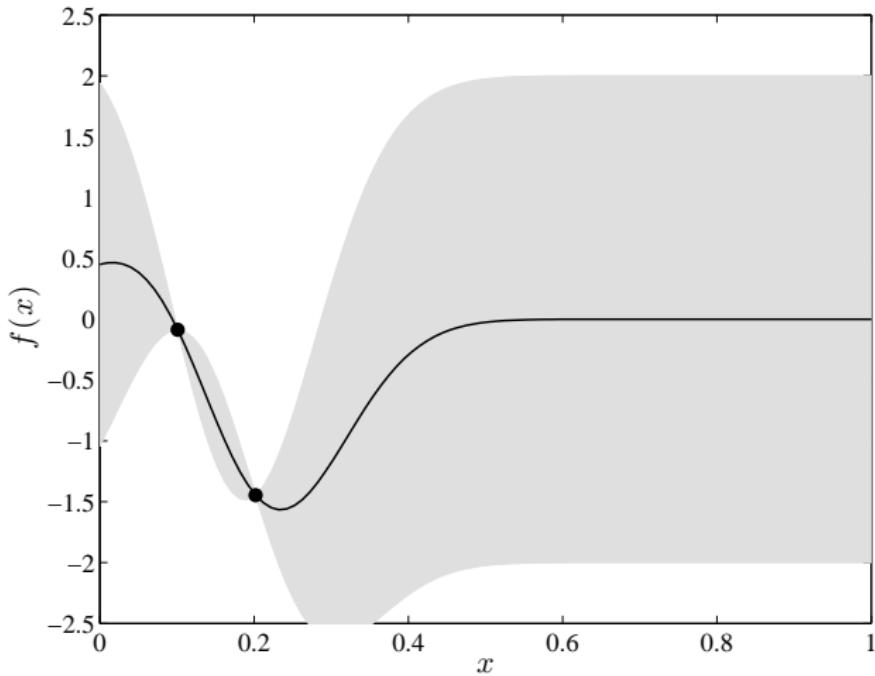
Posterior

Dos observaciones  $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\}$ .

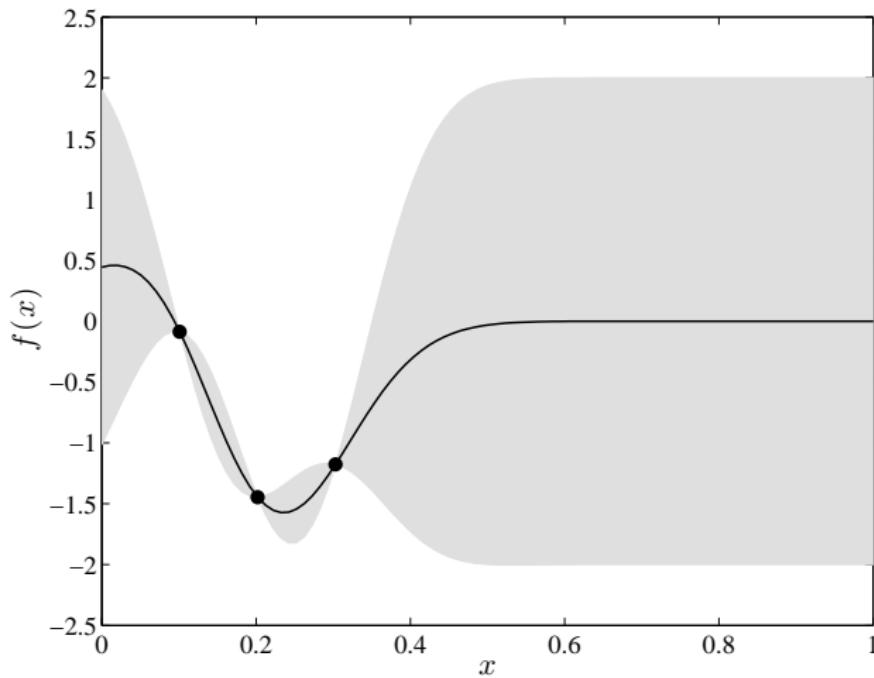
# Modelamiento Bayesiano: regresión (II)



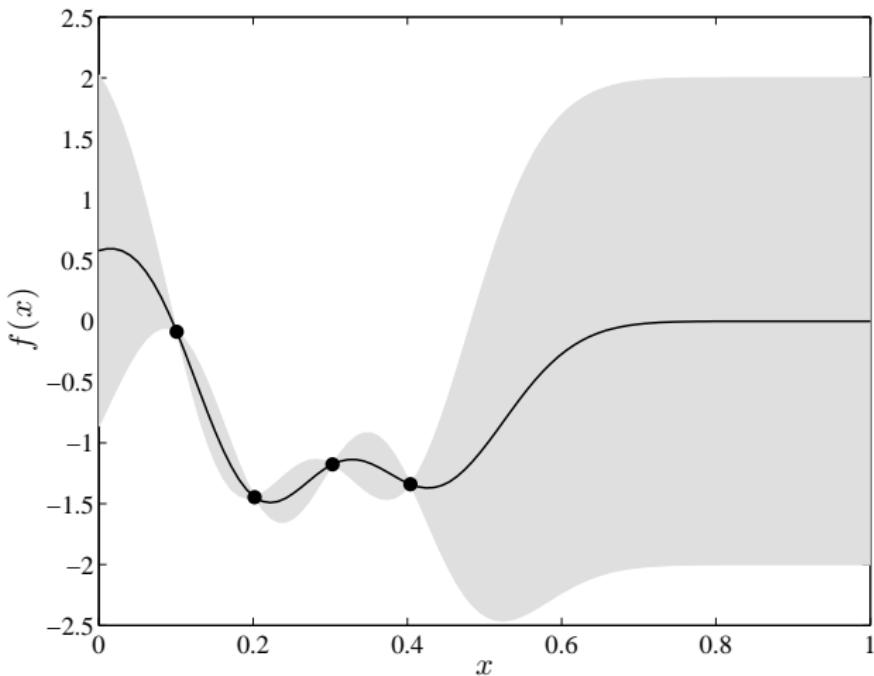
# Modelamiento Bayesiano: regresión (II)



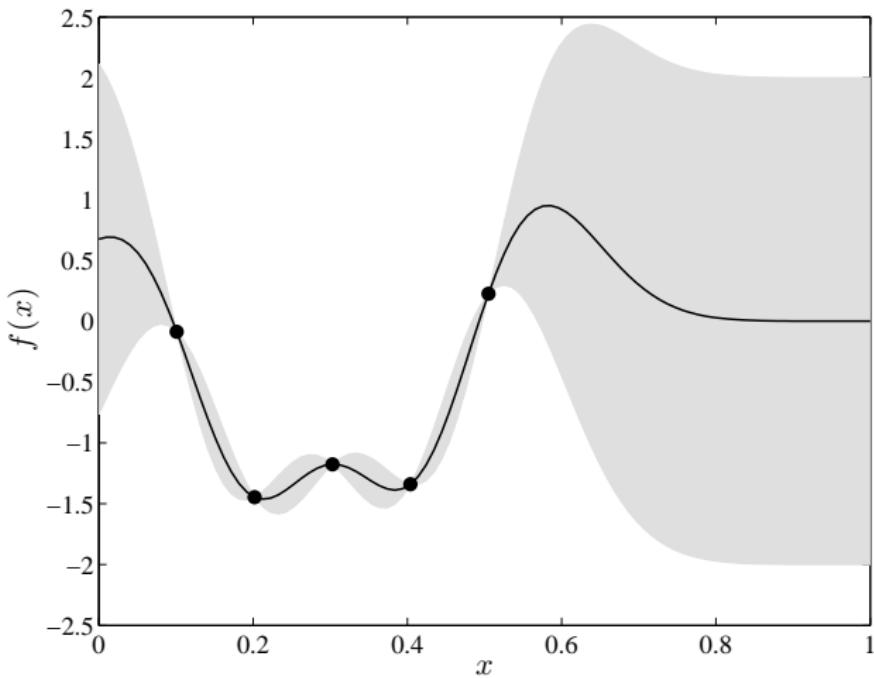
# Modelamiento Bayesiano: regresión (II)



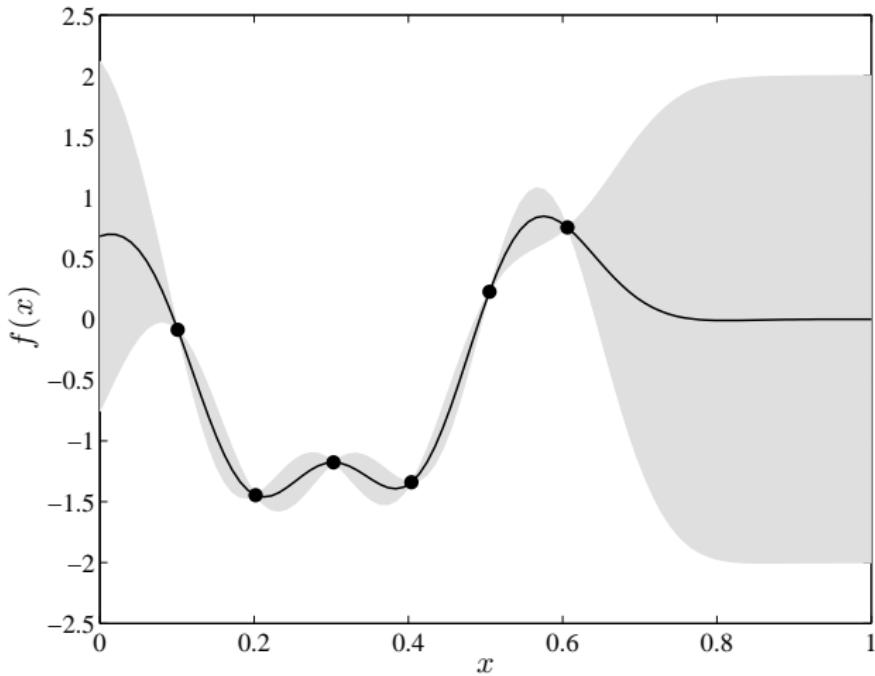
# Modelamiento Bayesiano: regresión (II)



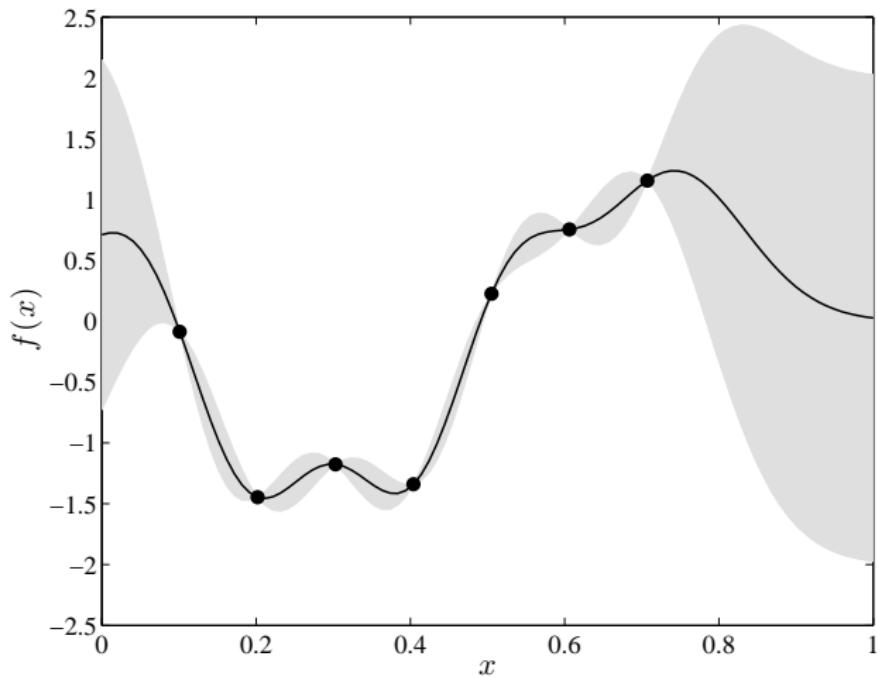
# Modelamiento Bayesiano: regresión (II)



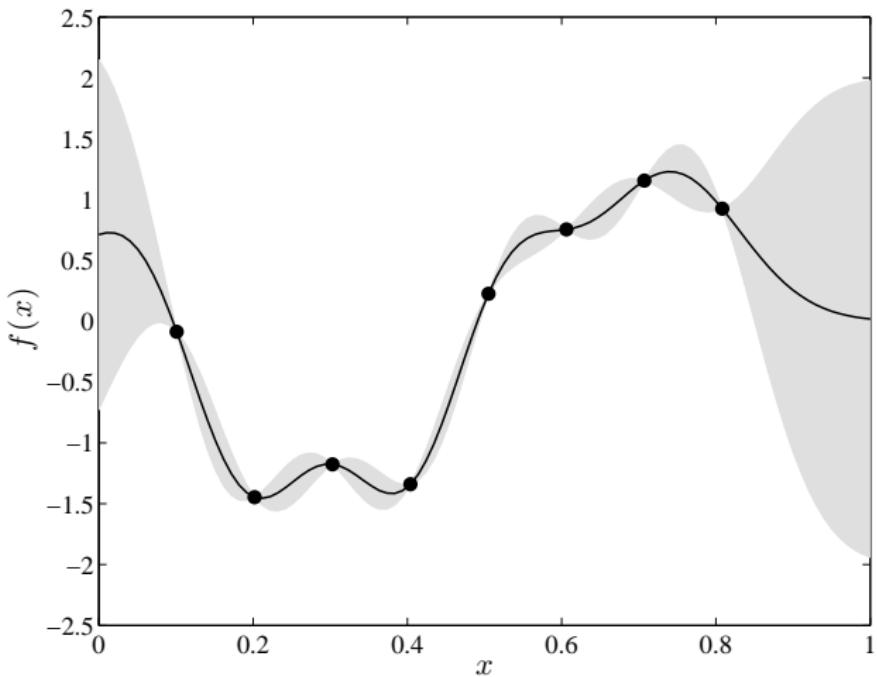
# Modelamiento Bayesiano: regresión (II)



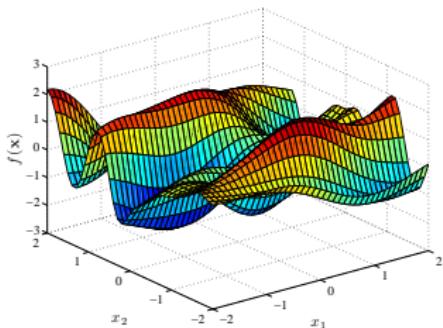
# Modelamiento Bayesiano: regresión (II)



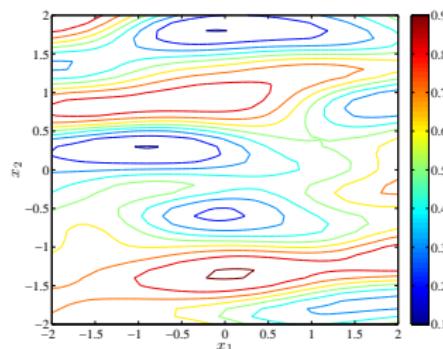
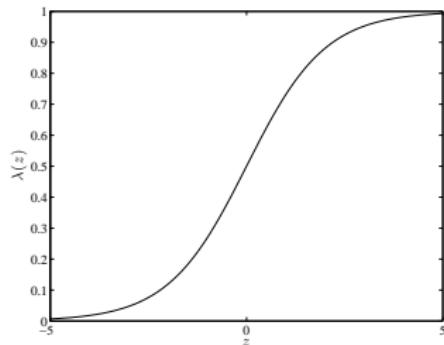
# Modelamiento Bayesiano: regresión (II)



# Modelamiento Bayesiano: clasificación (I)

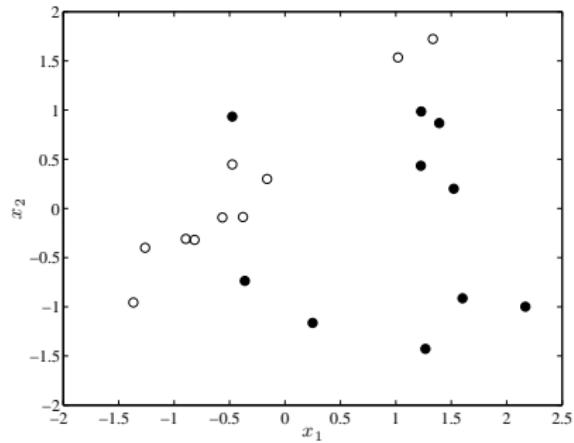


Muestra del GP

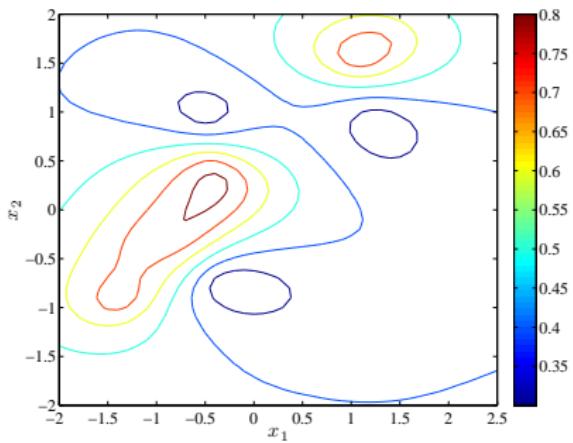


Contorno

# Modelamiento Bayesiano: clasificación (II)



Datos



Contorno posterior

# Contenido

Introducción

Regresión

Clasificación

Maximum a posterior

Aproximación de Laplace

Modelo lineal

usando procesos Gaussianos

# Preliminares

- Aprendizaje supervisado
  - Clasificación → predicción de variables discretas.
  - Regresión → predicción de variables continuas.
- Ejemplos regresión
  - Predecir precio de una mercancía, con base en la tasa de interés, la demanda, la oferta, entre otros.
  - Predecir tamaño de área de un incendio forestal, con base en datos metereológicos.
- Dos formas de estudiarlo,
  - Punto de vista del espacio de pesos (*weight-space view*).
  - Punto de vista del espacio de funciones (*function-space view*).

## Punto de vista del espacio de pesos (I)

- El modelo lineal de regresión ha sido estudiado extensivamente.
- Se discute a continuación el tratamiento Bayesiano del modelo lineal.
- El modelo se expande proyectando el espacio de entrada a un espacio de mayor dimensionalidad.
- El nuevo espacio se conoce como el espacio de características (*feature space*).

## Punto de vista del espacio de pesos (II)

- Conjunto de entrenamiento  $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \dots, n\}$ .
- $\mathbf{x}_i \in \mathbb{R}^D$ ,  $y_i \in \mathbb{R}$ .
- Se forman la matriz de diseño  $\mathbf{X} \in \mathbb{R}^{D \times n}$ , y el vector de salidas  $\mathbf{y}$ ,

$$\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_n], \quad \mathbf{y} = [y_1 \quad y_2 \quad \cdots \quad y_n]^\top.$$

- Luego  $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ .

# Modelo lineal estándar (I)

- En el modelo lineal estándar

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}, \quad y = f(\mathbf{x}) + \epsilon,$$

donde  $\mathbf{w}$  es un vector de parámetros,  $y$  es la observación para  $\mathbf{x}$ , y  $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ .

- Las observaciones en el conjunto de entrenamiento se asumen **iid**.
- Verosimilitud,  $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$ : probabilidad de las observaciones dados los parámetros,

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) &= \prod_{i=1}^n p(y_i|\mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma_n^2}\right] \\ &= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2} |\mathbf{y} - \mathbf{X}^\top \mathbf{w}|^2\right) = \mathcal{N}(\mathbf{y}|\mathbf{X}^\top \mathbf{w}, \sigma_n^2 \mathbf{I}). \end{aligned}$$

## Modelo lineal est醕ar (II)

- Se especifica una distribución prior para  $\mathbf{w}$ , por ejemplo,

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p),$$

donde  $\Sigma_p$  es un matriz de covarianza.

- Teorema de Bayes,

$$\text{posterior} = \frac{\text{verosimilitud} \times \text{prior}}{\text{verosimilitud marginal}}, \quad p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})},$$

donde

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w}.$$

## Modelo lineal est醕ar (III)

- Para el caso del modelo lineal,

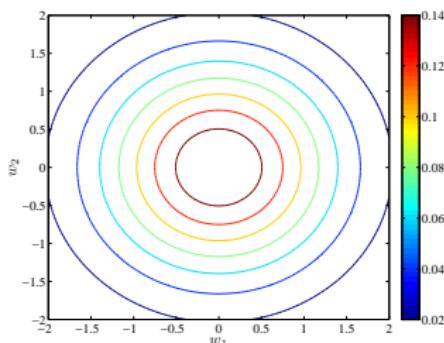
$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}(\mathbf{w} \Big| \underbrace{\frac{1}{\sigma_n^2} \mathbf{A}^{-1} \mathbf{X} \mathbf{y}}_{\hat{\mathbf{w}}}, \mathbf{A}^{-1}).$$

donde  $\mathbf{A} = \sigma_n^{-2} \mathbf{X} \mathbf{X}^\top + \boldsymbol{\Sigma}_p^{-1}$  es una matriz de covarianza.

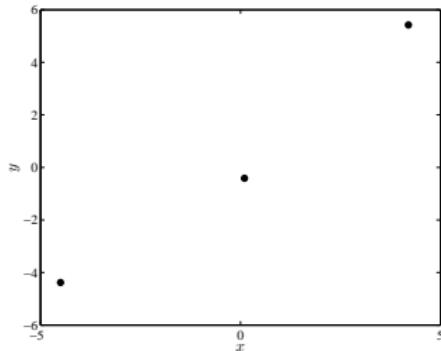
- Validación → promediar sobre  $\mathbf{w}$  usando la función posterior.
- La distribución predictiva para  $f_* \equiv f(\mathbf{x}_*)$  en  $\mathbf{x}_*$  está dada como

$$\begin{aligned} p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) &= \int p(f_* | \mathbf{x}_*, \mathbf{w}) p(\mathbf{w} | \mathbf{y}, \mathbf{X}) d\mathbf{w}, \\ &= \mathcal{N}\left(f_* \Big| \underbrace{\frac{1}{\sigma_n^2} \mathbf{x}_*^\top \mathbf{A}^{-1} \mathbf{X} \mathbf{y}}_{\hat{f}_*}, \mathbf{x}_*^\top \mathbf{A}^{-1} \mathbf{x}_*\right). \end{aligned}$$

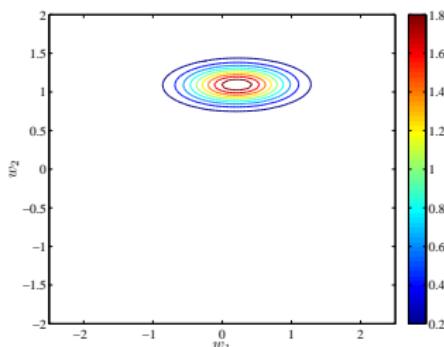
# Modelo lineal est醤dar (IV)



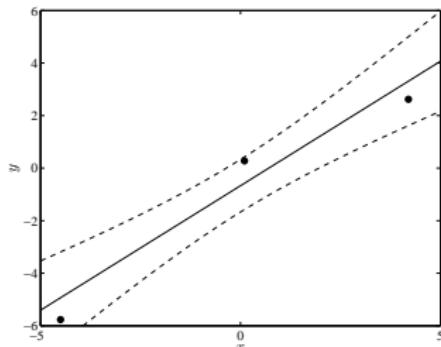
$p(\mathbf{w})$



$\mathcal{D} = (\mathbf{X}, \mathbf{y})$



$p(\mathbf{w} | \mathbf{y}, \mathbf{X})$



$p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y})$

# Espacio de características (I)

- El modelo lineal Bayesiano sufre de una expresividad limitada.
- Funciones base para proyectar las entradas a un espacio de mayor dimensionalidad.
- Aplicar el modelo lineal en ese espacio.

## Espacio de características (II)

- Se introduce la función  $\phi(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}^N$ .
- $\Phi(\mathbf{X}) \in \mathbb{R}^{N \times n}$ .
- El modelo es igual a  $f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w}$ , con  $\mathbf{w} \in \mathbb{R}^N$ .
- Ecuaciones del modelo lineal permanecen, cambiando  $\mathbf{X}$  por  $\Phi(\mathbf{X})$ .

## Espacio de características (III)

- Predictiva

$$p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N} \left( f_* \middle| \frac{1}{\sigma_n^2} \phi(\mathbf{x}_*)^\top \mathbf{A}^{-1} \Phi \mathbf{y}, \phi(\mathbf{x}_*)^\top \mathbf{A}^{-1} \phi(\mathbf{x}_*) \right),$$

donde  $\Phi = \Phi(\mathbf{X})$ , y  $\mathbf{A} = \sigma_n^{-2} \Phi \Phi^\top + \Sigma_p^{-1}$ .

- Invertir  $\mathbf{A}$  costoso para  $N$  grande.
- Se puede demostrar que

$$p(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N} \left( f_* \middle| \phi_*^\top \Sigma_p \Phi (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}, \phi_*^\top \Sigma_p \phi_* - \phi_*^\top \Sigma_p \Phi (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \Phi^\top \Sigma_p \phi_* \right),$$

donde  $\phi(\mathbf{x}_*) = \phi_*$ , y  $\mathbf{K} = \Phi^\top \Sigma_p \Phi$ .

- El espacio de características siempre aparece de las formas  $\phi_*^\top \Sigma_p \Phi$ ,  $\phi_*^\top \Sigma_p \phi_*$ , y  $\Phi^\top \Sigma_p \Phi$ .

# Truco del kernel

- Las entradas de las matrices anteriores se pueden escribir de la forma  $\phi(\mathbf{x})^\top \Sigma_p \phi(\mathbf{x}')$ .
- Se define  $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \Sigma_p \phi(\mathbf{x}')$ , como un *kernel* o *función de covarianza*.
- $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \Sigma_p \phi(\mathbf{x}') = \psi(\mathbf{x}) \cdot \psi(\mathbf{x}')$ , con  $\psi(\mathbf{x}) = \Sigma_p^{1/2} \phi(\mathbf{x}')$ .
- Algoritmo sólo depende de productos internos de vectores del espacio de entrada → se reemplazan las ocurrencias de esos productos internos por  $k(\mathbf{x}, \mathbf{x}')$ .

# Punto de vista del espacio de funciones (I)

- La idea es realizar inferencia directamente sobre el espacio de funciones.
- Se usa un proceso Gaussiano para describir una distribución sobre funciones.
- **Definición.** Un proceso Gaussiano es una colección de variables aleatorias, tal que un conjunto finito de ellas sigue una distribución Gaussiana conjunta.

## Punto de vista del espacio de funciones (II)

- Se especifica por una función media y una función de covarianza.
- La función media,  $m(\mathbf{x})$ , y la función de covarianza,  $k(\mathbf{x}, \mathbf{x}')$ , del proceso real  $f(\mathbf{x})$  se definen como

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \\ k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))].$$

- El proceso Gaussiano se denota como

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')).$$

- Sin pérdida de generalidad en lo que sigue se asume que  $f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ .

## Punto de vista del espacio de funciones (III)

- El proceso Gaussiano es *consistente*.
- Esto significa que si  $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$ , luego  $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ , donde

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

- Examinar un conjunto grande de variables no cambia la distribución de uno de sus subconjuntos.

## Punto de vista del espacio de funciones (IV)

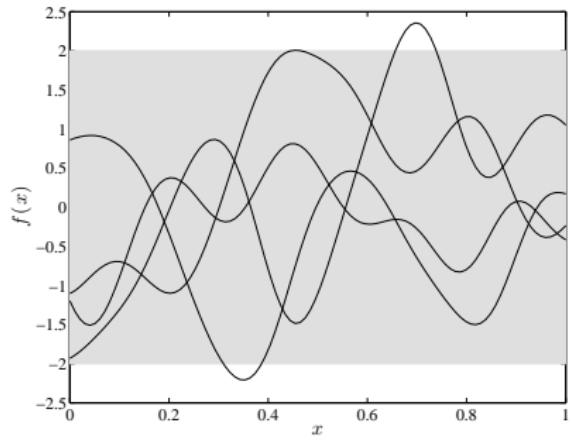
- El modelo Bayesiano lineal es un ejemplo de un proceso Gaussiano.
- Se tiene  $f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w}$  con prior  $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$ .
- Luego

$$\mathbb{E}[f(\mathbf{x})] = \phi(\mathbf{x})^\top \mathbb{E}[\mathbf{w}] = 0,$$

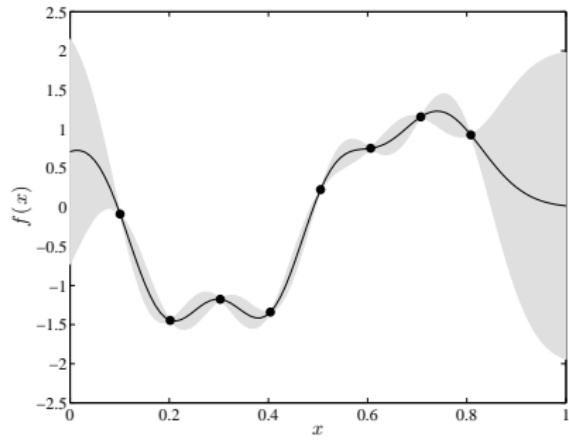
$$\mathbb{E}[f(\mathbf{x})f(\mathbf{x}')^\top] = \phi(\mathbf{x})^\top \mathbb{E}[\mathbf{w}\mathbf{w}^\top] \phi(\mathbf{x}') = \phi(\mathbf{x})^\top \Sigma_p \phi(\mathbf{x}).$$

# Ejemplo

La especificación de una función de covarianza implica una distribución sobre funciones.



Prior



Posterior

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}(\mathbf{X}, \mathbf{X})), \text{ usando } k(\mathbf{x}, \mathbf{x}') = s_f \exp\left(-\frac{|\mathbf{x}-\mathbf{x}'|^2}{2\ell^2}\right).$$

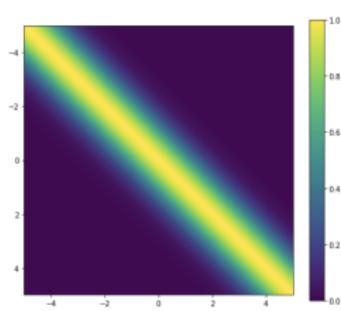
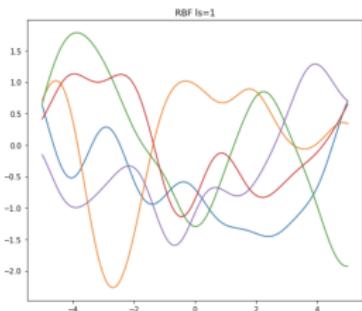
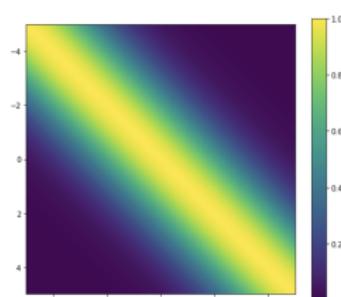
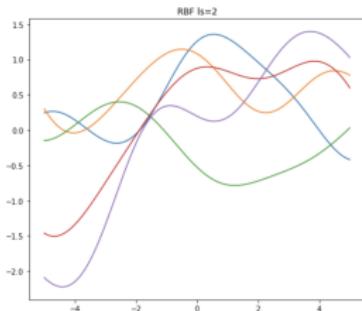
# Función de covarianza

- La función de covarianza  $k(\mathbf{x}, \mathbf{x}')$  se conoce en muchos contextos como la *función kernel*.
- Por definición, la función de covarianza es positiva semidefinida, lo que conduce a una matriz de covarianza que también es positiva semidefinida

$$\mathbf{v}^\top \mathbf{K}(\mathbf{X}, \mathbf{X}) \mathbf{v} > 0, \quad \forall \mathbf{v} \in \mathbb{R}^n.$$

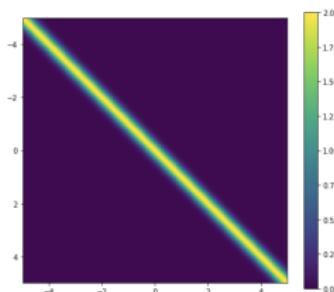
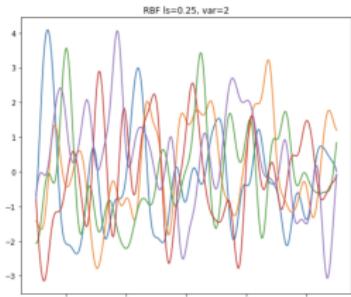
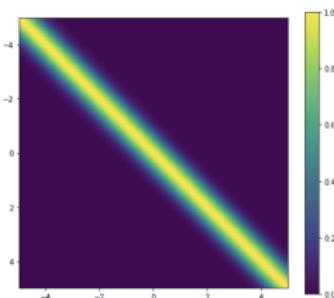
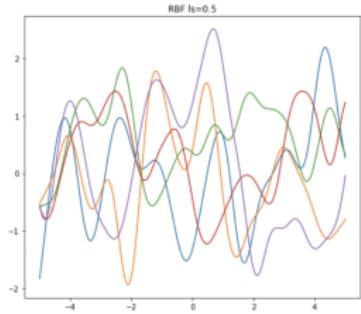
- En aplicaciones prácticas, la función de covarianza se selecciona de un conjunto de funciones disponibles

# Tipos de función de covarianza: exponencial cuadrada



$$k(\mathbf{x}, \mathbf{x}') = s_f \exp\left(-\frac{r^2}{2\ell^2}\right), \quad r = |\mathbf{x} - \mathbf{x}'|.$$

# Tipos de función de covarianza: exponencial cuadrada



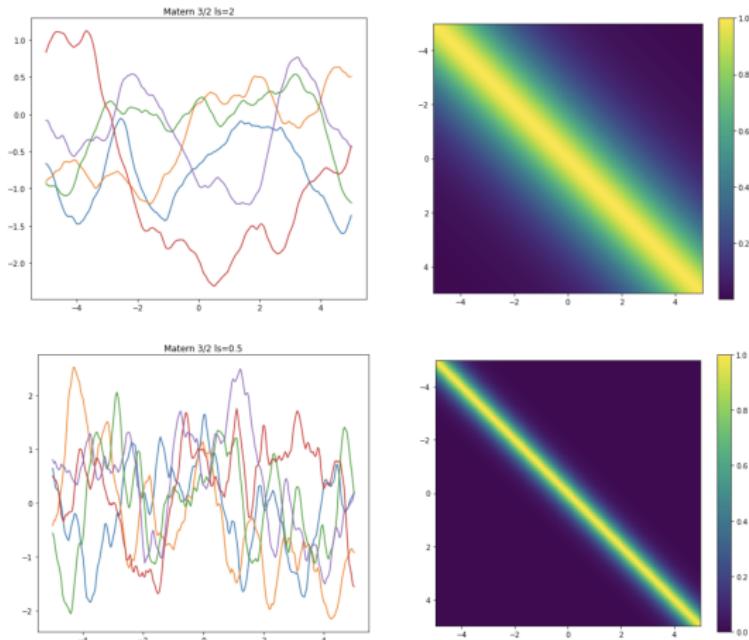
$$k(\mathbf{x}, \mathbf{x}') = s_f \exp \left( -\frac{r^2}{2\ell^2} \right), \quad r = |\mathbf{x} - \mathbf{x}'|.$$

## Tipos de función de covarianza: Matérn

$$k(r) = s_f \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu}r}{\ell} \right)^\nu K_\nu \left( \frac{\sqrt{2\nu}r}{\ell} \right)$$
$$k(r) = s_f \left( 1 + \frac{\sqrt{3}r}{\ell} \right) \exp \left( -\frac{\sqrt{3}r}{\ell} \right), \quad \nu = \frac{3}{2},$$

donde  $r = |\mathbf{x} - \mathbf{x}'|$  y  $K_\nu(\cdot)$  es la función modificada de Bessel.

# Tipos de función de covarianza: Matérn



$$k(r) = s_f \left( 1 + \frac{\sqrt{3}r}{\ell} \right) \exp \left( -\frac{\sqrt{3}r}{\ell} \right), \quad r = |\mathbf{x} - \mathbf{x}'|$$

## Construcción de nuevos kernels

Dados dos kernels válidos  $k_1(\mathbf{x}, \mathbf{x}')$  y  $k_2(\mathbf{x}, \mathbf{x}')$ , los siguientes kernels nuevos también son válidos

$$k(\mathbf{x}, \mathbf{x}') = c k_1(\mathbf{x}, \mathbf{x}')$$

$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}) k_1(\mathbf{x}, \mathbf{x}') f(\mathbf{x}')$$

$$k(\mathbf{x}, \mathbf{x}') = q(k_1(\mathbf{x}, \mathbf{x}'))$$

$$k(\mathbf{x}, \mathbf{x}') = \exp(k_1(\mathbf{x}, \mathbf{x}'))$$

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$$

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') k_2(\mathbf{x}, \mathbf{x}')$$

$$k(\mathbf{x}, \mathbf{x}') = k_3(\phi(\mathbf{x}), \phi(\mathbf{x}'))$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{A} \mathbf{x}'$$

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}_b)$$

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a) k_b(\mathbf{x}_b, \mathbf{x}_b),$$

donde  $c > 0$  es una constante,  $f(\cdot)$  es cualquier función,  $q(\cdot)$  es un polinomio con coeficientes no negativos,  $\phi(\cdot)$  es una función de  $D$  a  $N$ ,  $k_3(\cdot, \cdot)$  es un kernel válido en  $\mathbb{R}^N$ ,  $\mathbf{A}$  es una matriz simétrica positiva semidefinida,  $\mathbf{x}_a$  y  $\mathbf{x}_b$  son variables  $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$ , y  $k_a(\cdot, \cdot)$  y  $k_b(\cdot, \cdot)$  son kernels válidos sobre sus espacios respectivos.

# Predicción (I)

- Usando  $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ , predecir  $f_* = f_*(\mathbf{x}_*)$  para valores de entrada  $\mathbf{x}_*$ .
- Se asume que  $y = f(\mathbf{x}) + \epsilon$ , con  $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ .
- La función de covarianza para  $y$  está dada entonces como

$$\text{cov}(y_p, y_q) = k(\mathbf{x}_p, \mathbf{x}_q) + \sigma_n^2 \delta_{pq}, \quad \text{cov}(\mathbf{y}) = \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}.$$

## Predicción (II)

- La distribución conjunta de los valores observados  $\mathbf{y}$ , y de la función en las entradas de test,  $\mathbf{f}_*$ , está dada por

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N} \left( \mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I} & \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \\ \mathbf{K}(\mathbf{X}_*, \mathbf{X}) & \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right)$$

- Se puede demostrar que la ecuación de predicción para regresión con procesos Gaussianos está dada como

$$\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*))$$

donde

$$\bar{\mathbf{f}}_* = \mathbf{K}(\mathbf{X}_*, \mathbf{X}) [\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}]^{-1} \mathbf{y}$$

$$\text{cov}(\mathbf{f}_*) = \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*, \mathbf{X}) [\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}]^{-1} \mathbf{K}(\mathbf{X}_*, \mathbf{X}).$$

# Verosimilitud Marginal

- La verosimilitud marginal,  $p(\mathbf{y}|\mathbf{X})$ , está dada como

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X}) p(\mathbf{f}|\mathbf{X}) d\mathbf{f} = \mathcal{N}(\mathbf{0}, \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}).$$

- Los parámetros  $s_f$ ,  $\ell$  y  $\sigma_n^2$  pueden estimarse maximizando el logaritmo de la verosimilitud marginal,

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}) &= -\frac{1}{2} \mathbf{y}^\top (\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}| \\ &\quad - \frac{n}{2} \log 2\pi.\end{aligned}$$

# Contenido

Introducción

Regresión

Clasificación

Maximum a posterior

Aproximación de Laplace

Modelo lineal

usando procesos Gaussianos

# Contents

Introducción

Regresión

Clasificación

Maximum a posterior

Aproximación de Laplace

Modelo lineal

usando procesos Gaussianos

## Maximum a posterior

- En la técnica de máximo a posterior (MAP) se busca encontrar el parámetro  $\theta$  o el vector de parámetros  $\theta$  que maximice

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta).$$

- En la práctica, se maximiza  $\ln p(\theta|\mathcal{D})$ , que equivale a maximizar

$$\ln p(\mathcal{D}|\theta) + \ln p(\theta).$$

- Obsérvese que la estimación es puntual.
- El estimador MAP de  $\theta$  se denota como  $\theta_{\text{MAP}}$ .

# Contents

Introducción

Regresión

Clasificación

Maximum a posterior

Aproximación de Laplace

Modelo lineal

usando procesos Gaussianos

## Preliminares

- Un aproximación simple, pero bastante empleada, se conoce como la aproximación de Laplace.
- La aproximación de Laplace busca encontrar una aproximación Gaussiana a la densidad de probabilidad definida sobre un conjunto de variables continuas.
- Consideremos el caso de una variable escalar continua  $z$ , y supongamos que la distribución  $p(z)$  está definida como

$$p(z) = \frac{1}{Z} f(z),$$

donde  $Z$  es el coeficiente de normalización.

- Se asume que el valor de  $Z$  es desconocido.

## Serie de Taylor

- En el método de Laplace el objetivo es encontrar una aproximación Gaussiana  $q(z)$  a  $p(z)$ , centrada en un modo de  $p(z)$ .
- En otras palabras, se desea encontrar un punto  $z_0$  tal que  $p'(z_0) = 0$  o de forma equivalente,

$$\left. \frac{df(z)}{dz} \right|_{z=z_0} = 0$$

- En la práctica se trabaja con  $\ln f(z)$ .
- Miremos la expansión en serie de Taylor para  $\ln f(z)$

$$\ln f(z_0) + \left. \frac{d \ln f(z)}{dz} \right|_{z=z_0} (z - z_0) + \frac{1}{2} \left. \frac{d^2 \ln f(z)}{dz^2} \right|_{z=z_0} (z - z_0)^2 + \mathcal{O}.$$

## Gaussiana aproximada

- Asumiendo que los términos de orden mayor son despreciables, y que  $z_0$  es un modo de  $f(z)$ , la expansión en serie de Taylor se simplifica a

$$\ln f(z) \approx \ln f(z_0) - \frac{1}{2} A(z - z_0)^2,$$

donde se ha definido

$$A = -\left. \frac{d^2}{dz^2} \ln f(z) \right|_{z=z_0}$$

- Tomando exponentiales, se tiene

$$f(z) \approx f(z_0) \exp \left\{ -\frac{A}{2}(z - z_0)^2 \right\},$$

que luce como una Gaussiana.

# Gaussiana normalizada

La Gaussiana normalizada se obtiene como

$$q(z) = \left(\frac{A}{2\pi}\right)^2 \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}$$

# Proceso para encontrar la aproximación de Laplace

1. Encontrar un máximo local  $z_0$  de la fdp sin normalizar  $p(z) \propto f(z)$ . En la práctica se maximiza  $\ln f(z)$ .
2. Calcular la varianza

$$\sigma^2 = \frac{1}{A} = -\frac{1}{\left. \frac{d^2}{dz^2} \ln f(z) \right|_{z=z_0}}.$$

3. Aproximar  $p(z)$  como  $p(z) = \mathcal{N}(z|z_0, \sigma^2)$ .

## Ejemplo

La distribución Chi está dada como

$$p(z) = \frac{z^{k-1} \exp(-z^2/2)}{Z}, \quad z > 0$$

donde  $k > 0$  y  $Z$  está dada como

$$Z = 2^{\frac{k}{2}-1} \Gamma\left(\frac{k}{2}\right).$$

Usar una distribución Gaussiana para aproximarla.

## Caso multidimensional (I)

- El método de Laplace se puede extender para aproximar la distribución

$$p(\mathbf{z}) = \frac{f(\mathbf{z})}{Z},$$

definida sobre un espacio M-dimensional para  $\mathbf{z}$ .

- En un punto estacionario  $\mathbf{z}_0$ , el gradiente  $\nabla f(\mathbf{z})$  desaparece.
- Expandiendo alrededor de este punto estacionario se tiene

$$\ln f(\mathbf{z}) \approx \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^\top \mathbf{A}(\mathbf{z} - \mathbf{z}_0)$$

donde la matriz Hessiana  $\mathbf{A}$  está definida como

$$\mathbf{A} = -\nabla \nabla \ln f(\mathbf{z}) \Big|_{\mathbf{z}=\mathbf{z}_0},$$

y  $\nabla$  es el operador gradiente.

## Caso multidimensional (II)

- Tomando exponencial en ambos lados se tiene

$$f(\mathbf{z}) \approx f(\mathbf{z}_0) \exp \left\{ -\frac{1}{2} (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) \right\}.$$

- El posterior se puede aproximar como

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}} \exp \left\{ -\frac{1}{2} (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) \right\} = \mathcal{N}(\mathbf{z} | \mathbf{z}_0, \mathbf{A}^{-1})$$

- La distribución Gaussiana que se encuentra tendrá un buen comportamiento si la matriz  $\mathbf{A}$  es positiva definida, lo que implica que el punto estacionario  $\mathbf{z}_0$  sea un máximo local (no un mínimo, no un punto de inflexión).

## Observaciones (I)

- Con el objetivo de aplicar la aproximación de Laplace se necesita primero encontrar un modo  $\mathbf{z}_0$ , y luego evaluar la Hessiana en ese punto.
- En la práctica este modo se puede encontrar usando algoritmos de optimización numérica.
- El método de Laplace tiene sentido en tanto la fdp que se quiera aproximar sea unimodal.
- Si ese no es el caso, la aproximación de Laplace será diferente dependiendo del modo que se encuentre.

## Observaciones (II)

- Nótese que no es necesario conocer la constante de normalización  $Z$  de la distribución verdadera, para poder aplicar el método de Laplace.
- Como está basado en una distribución Gaussiana sólo es aplicable a variables reales.
- La aproximación de Laplace sólo se basa en un aspecto local de la función de distribución.

# Aproximación de Laplace en inferencia Bayesiana

En inferencia Bayesiana, el posterior  $p(\theta|\mathcal{D})$ , del que sólo se conoce

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta) = f(\theta),$$

se approxima empleando una Gaussiana

$$p(\theta|\mathcal{D}) = \mathcal{N}(\theta|\theta_{\text{MAP}}, \mathbf{A}^{-1}),$$

donde

$$\mathbf{A} = -\nabla\nabla \ln f(\theta) \Big|_{\theta=\theta_{\text{MAP}}}.$$

## Aproximación de la evidencia (I)

- Así como se puede encontrar una aproximación para  $p(\mathbf{z})$ , también se puede encontrar un aproximación para la constante de normalización  $Z$ .
- Se tiene

$$\begin{aligned}Z &= \int f(\mathbf{z}) d\mathbf{z} \\&\approx f(\mathbf{z}_0) \int \exp \left\{ -\frac{1}{2} (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) \right\} d\mathbf{z} \\&\approx f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}\end{aligned}$$

## Aproximación de la evidencia (II)

- La expresión anterior se puede emplear para calcular una aproximación de la evidencia.
- Del teorema de Bayes, la evidencia está definida como

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta.$$

- Haciendo  $f(\theta) = p(\mathcal{D}|\theta)p(\theta)$ , y  $Z = p(\mathcal{D})$ , y aplicando el resultado anterior, se tiene

$$\ln p(\mathcal{D}) = \ln p(\mathcal{D}|\theta_{\text{MAP}}) + \ln p(\theta_{\text{MAP}}) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{A}|,$$

donde  $\theta_{\text{MAP}}$  es el valor de  $\theta$  en el modo de la distribución posterior, y  $\mathbf{A}$  es la matriz Hessiana de las segundas derivadas del log-posterior negativo

$$\mathbf{A} = -\nabla\nabla \ln p(\mathcal{D}|\theta_{\text{MAP}})p(\theta_{\text{MAP}}) = -\nabla\nabla \ln p(\theta_{\text{MAP}}|\mathcal{D}).$$

# Contents

Introducción

Regresión

Clasificación

Maximum a posterior

Aproximación de Laplace

**Modelo lineal**

usando procesos Gaussianos

# Modelo lineal para clasificación (I)

- Problema biclase. Las clases se codifican como  $y = +1$ , y  $y = -1$ .
- La probabilidad de  $y = +1$ , se representa con un modelo lineal generalizado

$$p(y = +1 | \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{x}^\top \mathbf{w}),$$

donde  $\sigma(z) = 1 / (1 + \exp(-z))$ , es la función logística sigmoidal.

- La probabilidad de  $y = -1$ , es igual a  $1 - p(y = +1 | \mathbf{x}, \mathbf{w})$ .
- Como  $\sigma(-z) = 1 - \sigma(z)$ , ambas probabilidades se pueden escribir

$$p(y_i | \mathbf{x}_i, \mathbf{w}) = \sigma(y_i f_i),$$

donde  $f_i = \mathbf{x}_i^\top \mathbf{w}$ .

## Modelo lineal para clasificación (II)

- El logaritmo de la distribución posterior sin normalizar está dado como

$$\log p(\mathbf{w}|\mathbf{X}, \mathbf{y}) \propto -\frac{1}{2}\mathbf{w}^\top \Sigma_p^{-1} \mathbf{w} + \sum_{i=1}^n \log \sigma(y_i f_i).$$

- En clasificación, el posterior no tiene una forma analítica simple.
- Algoritmo IRLS (iteratively reweighted least squares).

# Contents

Introducción

Regresión

Clasificación

Maximum a posterior

Aproximación de Laplace

Modelo lineal

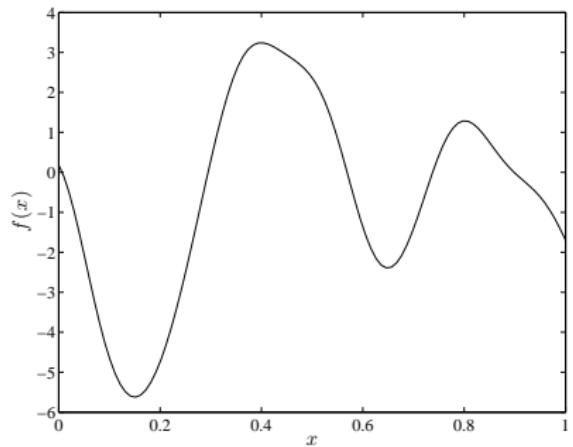
usando procesos Gaussianos

# Procesos Gaussianos para clasificación binaria (I)

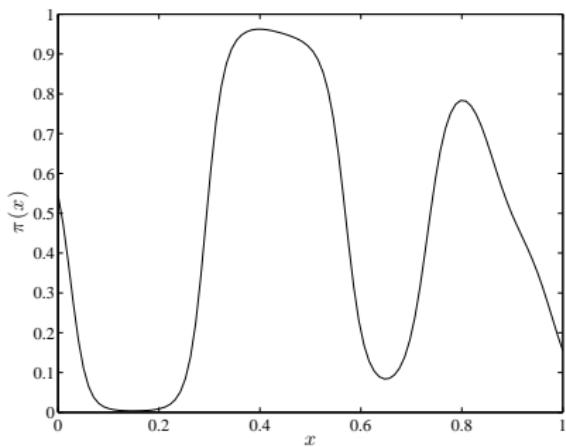
- Se asume que la función  $f(\mathbf{x})$  sigue un proceso Gaussiano.
- La función  $f(\mathbf{x})$  se pasa a través de la función logística  $\sigma(\cdot)$

$$\pi(\mathbf{x}) \equiv p(y = +1 | \mathbf{x}) = \sigma(f(\mathbf{x})).$$

# Procesos Gaussianos para clasificación binaria (II)



Función latente



Clase condicional

## Inferencia en dos pasos

- Dado un conjunto de datos  $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ , el posterior de  $\mathbf{f}$  se calcula como

$$p(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})},$$

donde

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f}$$

## Inferencia en dos pasos

- ❑ Paso 1. Para un nuevo  $\mathbf{x}_*$ , primero se calcula la distribución sobre la variable latente  $f_*$ .

$$p(f_* | \mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int p(f_* | \mathbf{X}, \mathbf{x}_*, \mathbf{f}) p(\mathbf{f} | \mathbf{X}, \mathbf{y}) d\mathbf{f}.$$

- ❑ Paso 2. Predicción probabilística

$$\hat{\pi}_* \equiv p(y_* = +1 | \mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int \sigma(f_*) p(f_* | \mathbf{X}, \mathbf{y}, \mathbf{x}_*) df_*.$$

- ❑ El posterior  $p(\mathbf{f} | \mathbf{X}, \mathbf{y})$  en el paso 1 no es Gaussiano debido a la función de verosimilitud asociada. Y luego la integral del Paso 1 no es tratable analíticamente.

# Aproximación por Laplace (I)

- La aproximación de Laplace aproxima  $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$  usando una distribución normal.
- Otras aproximaciones incluyen Bayes variacional, algoritmo de Propagación de la Esperanza (Expectation-Propagation- EP), y Markov chain Monte Carlo (MCMC).

## Aproximación por Laplace (II)

- En la aproximación por Laplace

$$p(\mathbf{f}|\mathbf{X}, \mathbf{y}) \approx q(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \mathcal{N}(\hat{\mathbf{f}}, \mathbf{A}^{-1}),$$

donde

$$\hat{\mathbf{f}} = \arg \max_{\mathbf{f}} \log p(\mathbf{f}|\mathbf{X}, \mathbf{y})$$

$$\mathbf{A} = -\nabla \nabla \log p(\mathbf{f}|\mathbf{X}, \mathbf{y}) \Big|_{\hat{\mathbf{f}}}$$

- Para encontrar  $\hat{\mathbf{f}}$  se maximiza la siguiente función

$$\begin{aligned}\psi(\mathbf{f}) &\equiv \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}|\mathbf{X}) \\ &= \log p(\mathbf{y}|\mathbf{f}) - \frac{1}{2} \mathbf{f}^\top \mathbf{K}^{-1} \mathbf{f} - \frac{1}{2} \log |\mathbf{K}| - \frac{n}{2} \log 2\pi.\end{aligned}$$

## Aproximación por Laplace (III)

- Diferenciando  $\psi(\mathbf{f})$  con respecto a  $\mathbf{f}$  se tiene

$$\nabla \psi(\mathbf{f}) = \nabla \log p(\mathbf{y}|\mathbf{f}) - \mathbf{K}^{-1}\mathbf{f},$$

$$\nabla \nabla \psi(\mathbf{f}) = \nabla \nabla \log p(\mathbf{y}|\mathbf{f}) - \mathbf{K}^{-1} = -\mathbf{W} - \mathbf{K}^{-1},$$

donde  $\mathbf{W} = -\nabla \nabla \log p(\mathbf{y}|\mathbf{f})$  es diagonal porque  $y_i$  sólo depende de  $f_i$ .

- Si  $p(y_i = +1|f_i, \mathbf{x}_i) = \sigma(y_i f_i)$ , luego

$$\frac{\partial}{\partial f_i} \log p(\mathbf{y}|\mathbf{f}) = t_i - \pi_i,$$

$$\frac{\partial^2}{\partial f_i^2} \log p(\mathbf{y}|\mathbf{f}) = -\pi_i(1 - \pi_i),$$

donde  $t_i = (y_i + 1)/2$ , y  $\pi_i = p(y_i = +1|f_i)$ .

- El valor de  $\hat{\mathbf{f}}$  se encuentra usando optimización por Newton.

# Posterior

- El posterior tiene entonces la forma

$$q(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \mathcal{N}(\hat{\mathbf{f}}, (\mathbf{K}^{-1} + \mathbf{W})^{-1})$$

# Predicción, y estimación

- Usando la aproximación de Laplace para el posterior,

$$\mathbb{E}_q[f_* | \mathbf{X}, \mathbf{y}, \mathbf{x}_*] = \mathbf{k}(\mathbf{x}_*)^\top \mathbf{K}^{-1} \hat{\mathbf{f}}$$

$$\text{var}_q[f_* | \mathbf{X}, \mathbf{y}, \mathbf{x}_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}(\mathbf{x}_*)^\top (\mathbf{K} + \mathbf{W}^{-1})^{-1} \mathbf{k}(\mathbf{x}_*).$$

- Usando estas cantidades, la predicción se aproxima como

$$\hat{\pi}_* \approx \int \sigma(f_*) q(f_* | \mathbf{X}, \mathbf{y}, \mathbf{x}_*) df_* \approx \sigma(\kappa(f_* | \mathbf{y}) \bar{f}_*),$$

donde

$$\kappa(f_* | \mathbf{y}) = (1 + \pi \text{var}_q[f_* | \mathbf{X}, \mathbf{y}, \mathbf{x}_*]/8)^{-1}.$$

- La estimación de los parámetros se realiza optimizando el logaritmo de la verosimilitud marginal

# Logaritmo de la verosimilitud marginal

- El logaritmo de la verosimilitud marginal está dado como

$$\log q(\mathbf{y}|\mathbf{X}, \theta) = -\frac{1}{2}\widehat{\mathbf{f}}^\top \mathbf{K}^{-1} \widehat{\mathbf{f}} + \log p(\mathbf{y}|\widehat{\mathbf{f}}) - \frac{1}{2} \log |\mathbf{K}| |\mathbf{K}^{-1} + \mathbf{W}|,$$

donde  $\theta$  son los hiper-parámetros del kernel.