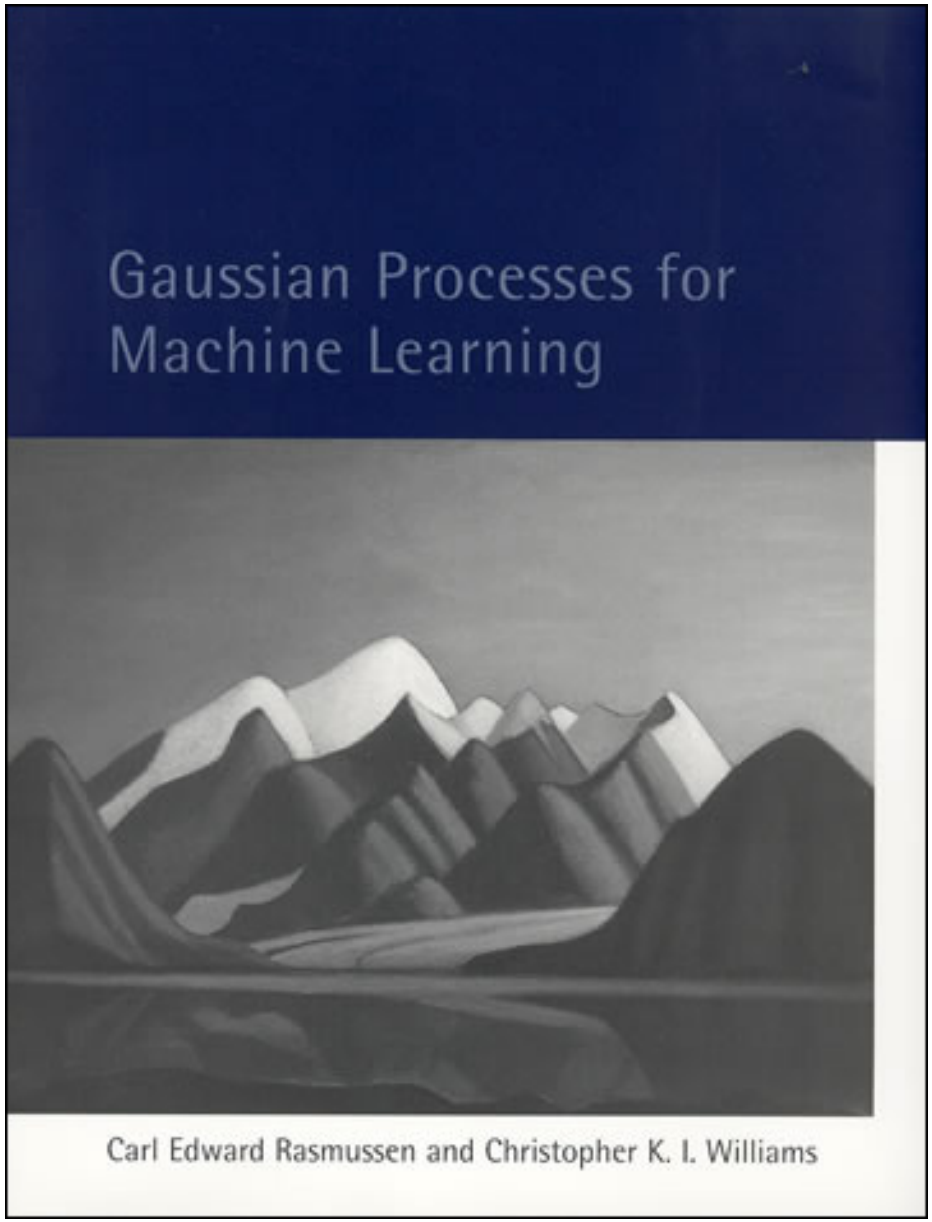
Introducción a los Procesos Gaussianos en Aprendizaje de Máquina

Mauricio A. Álvarez

Topics on Deep Probabilistic Models

Agradecimientos

Libro

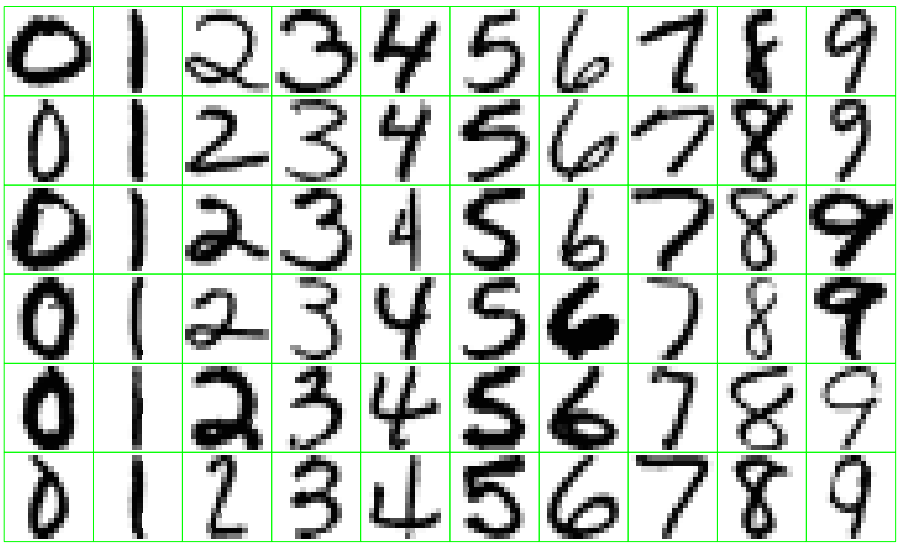

Gaussian Processes for
Machine Learning
Carl Edward Rasmussen and Christopher K. I. Williams

Contenido

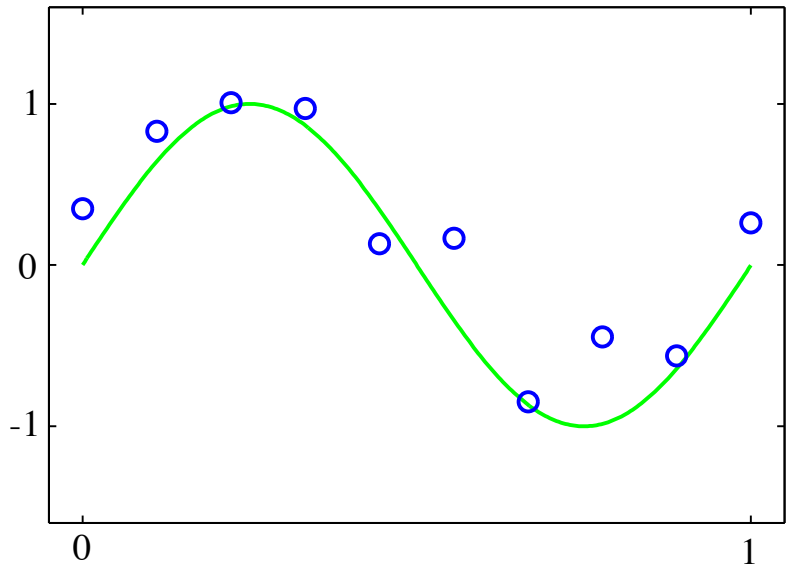
Contenido

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.



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: $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\}$.

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.

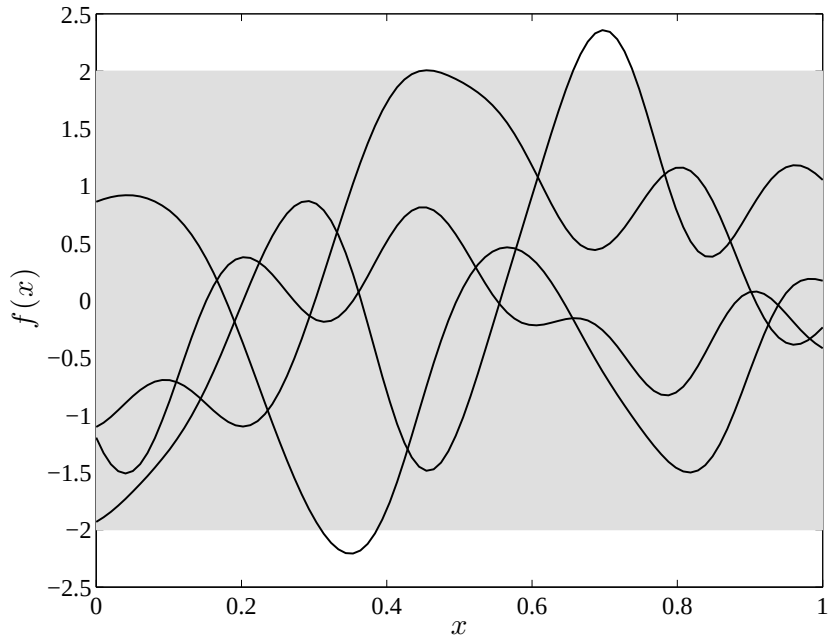
Dos alternativas

- 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.

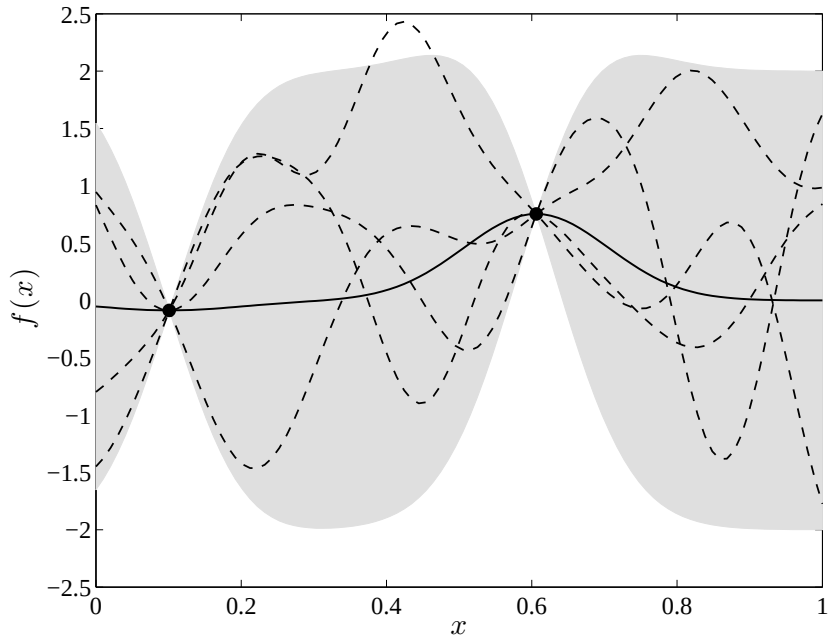
Procesos Gaussianos

- 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)



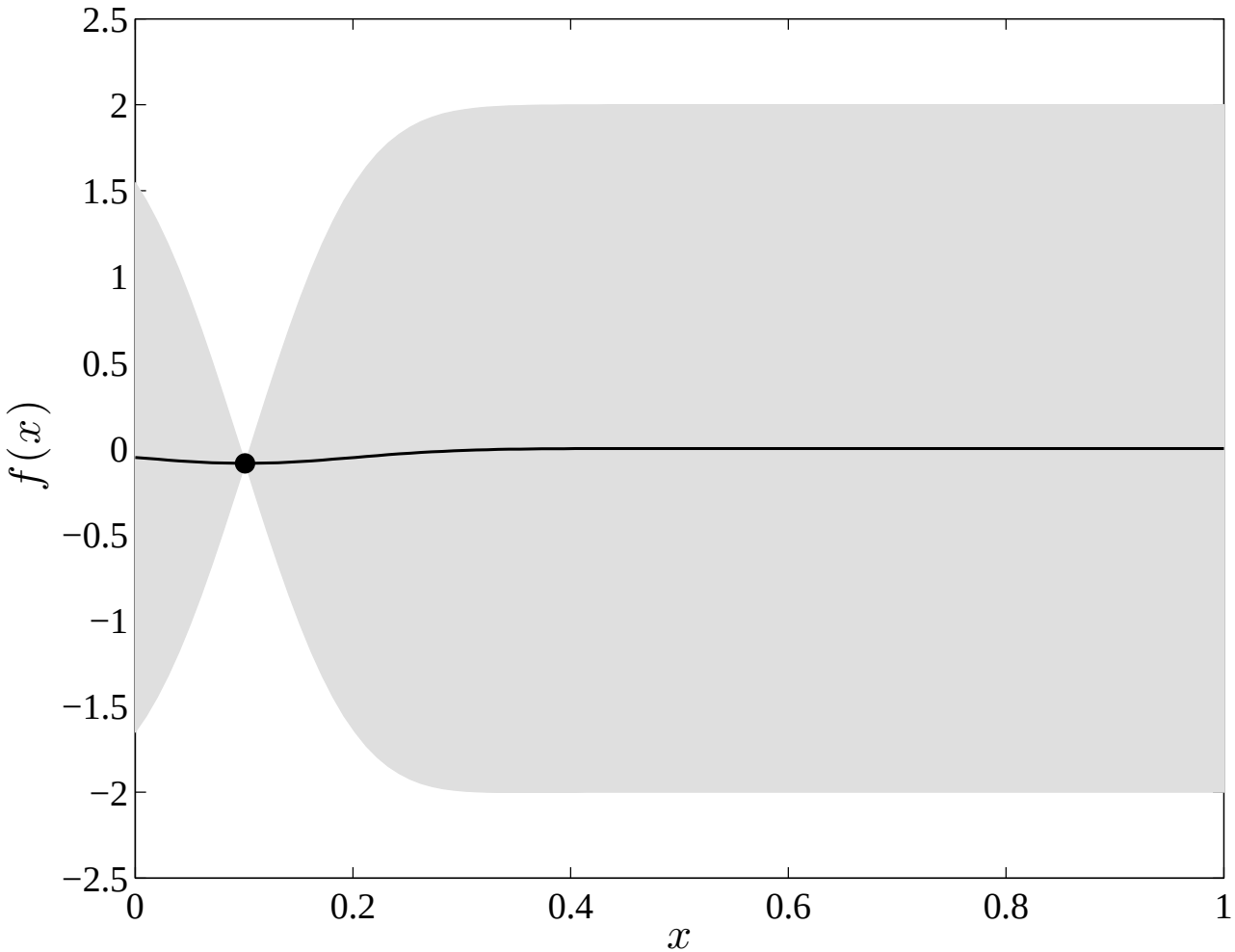
Prior



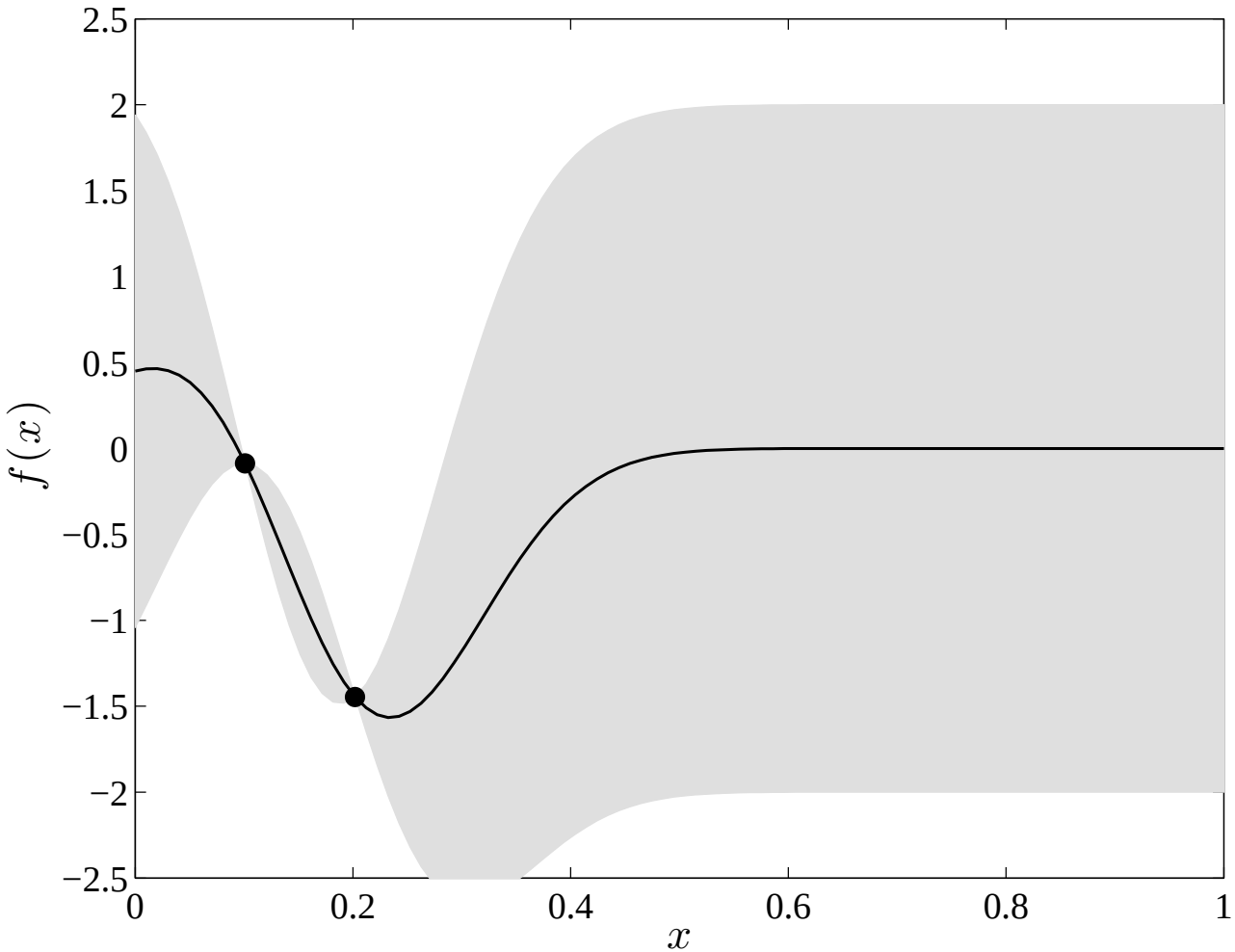
Posterior

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

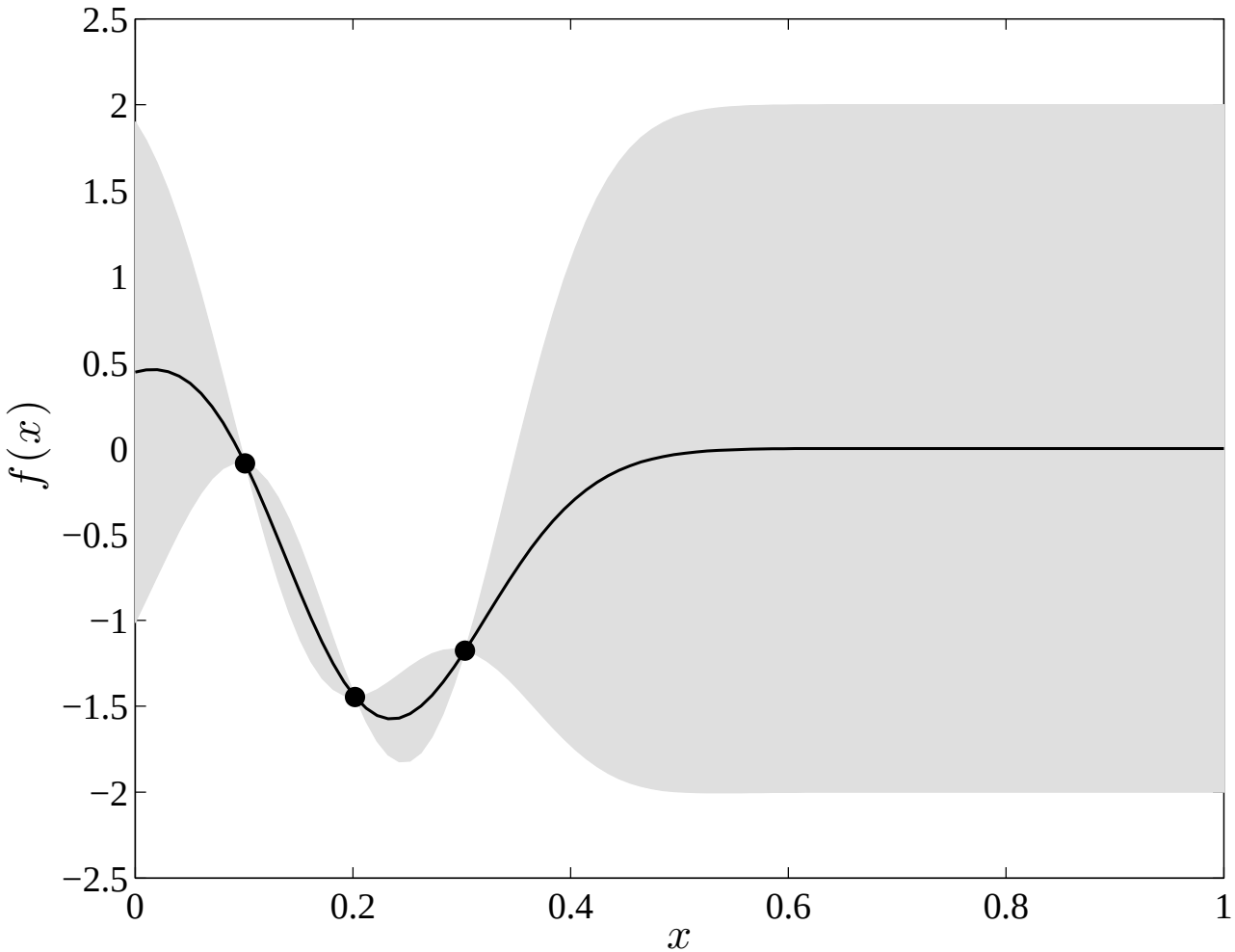
Modelamiento Bayesiano: regresión (II)



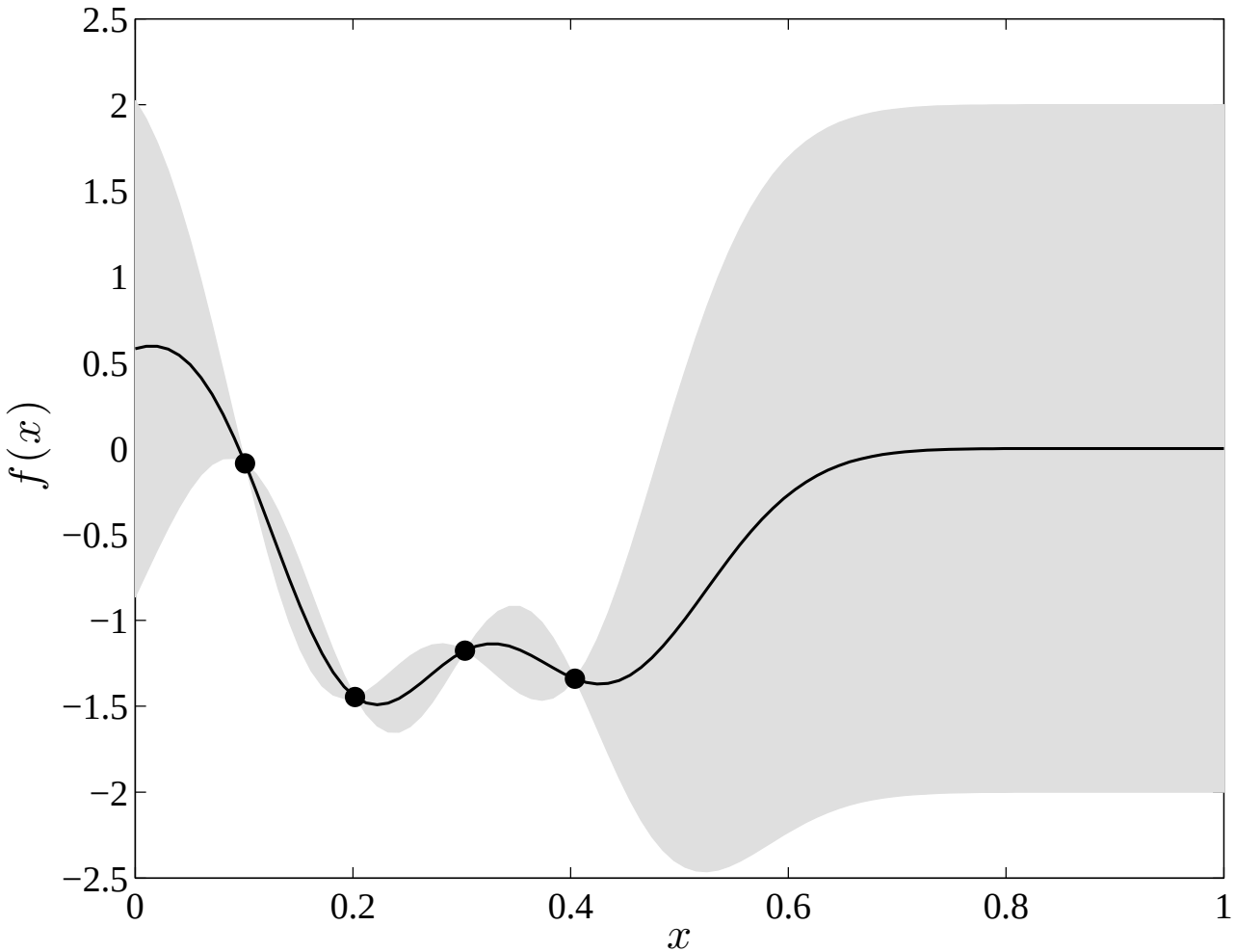
Modelamiento Bayesiano: regresión (II)



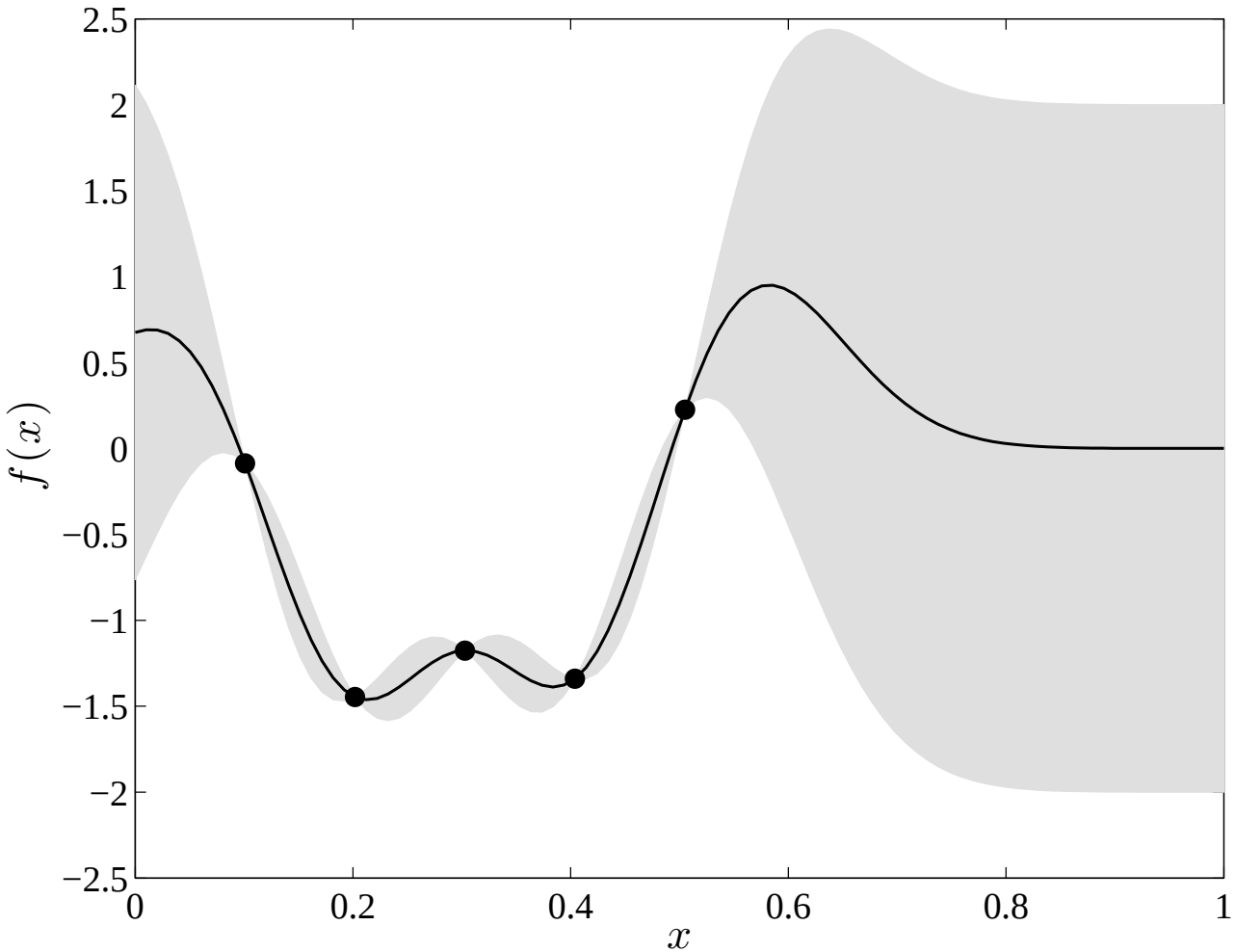
Modelamiento Bayesiano: regresión (II)



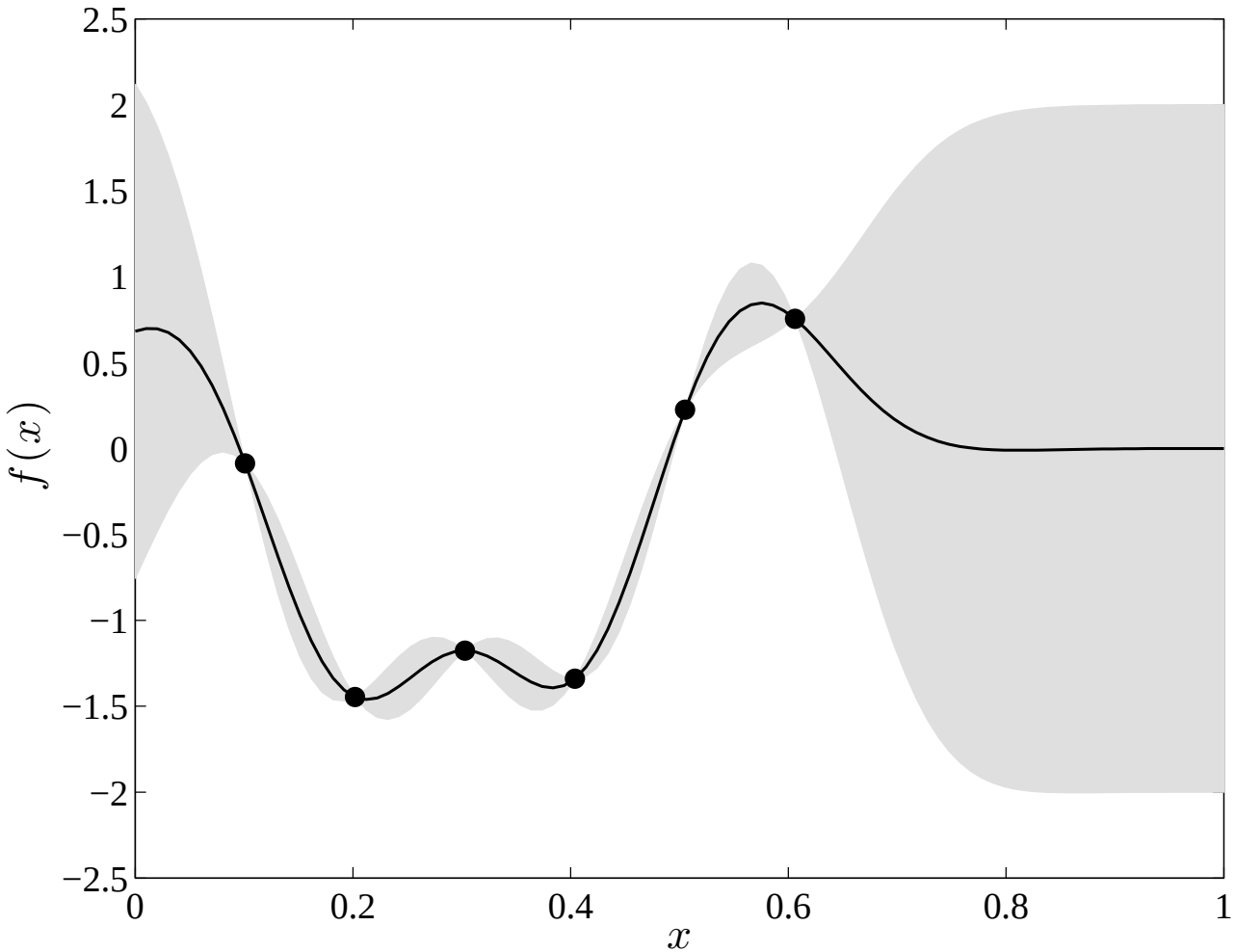
Modelamiento Bayesiano: regresión (II)



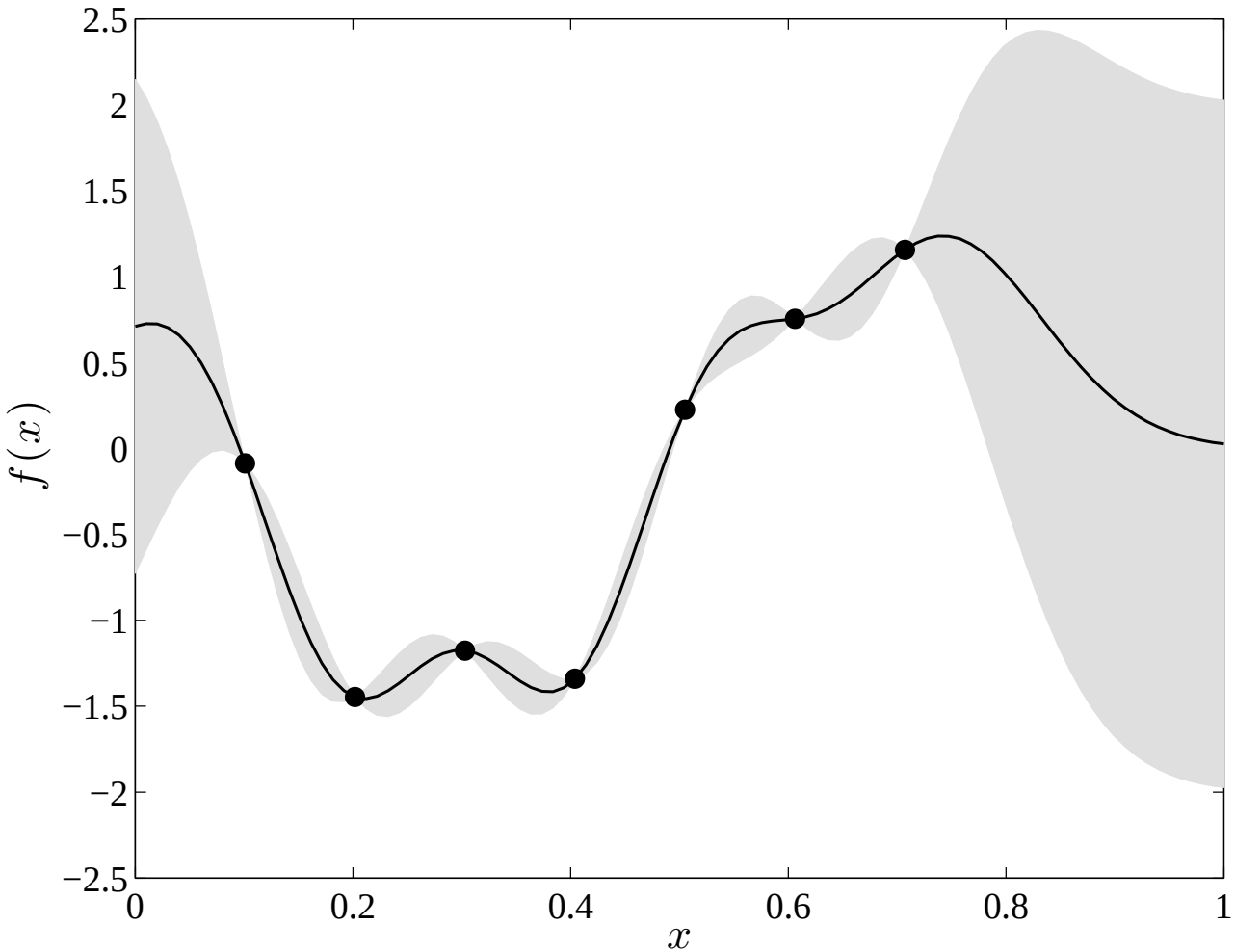
Modelamiento Bayesiano: regresión (II)



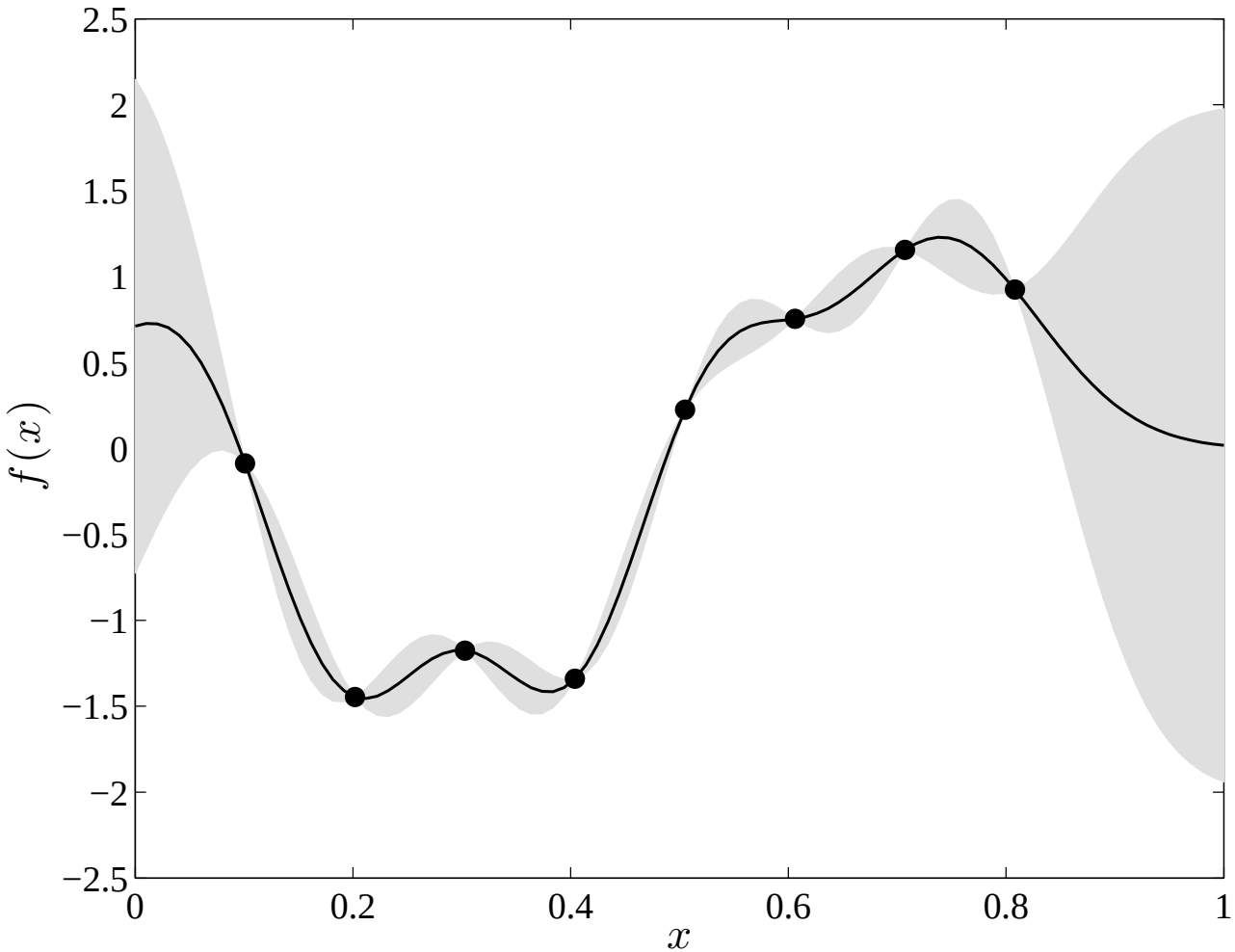
Modelamiento Bayesiano: regresión (II)



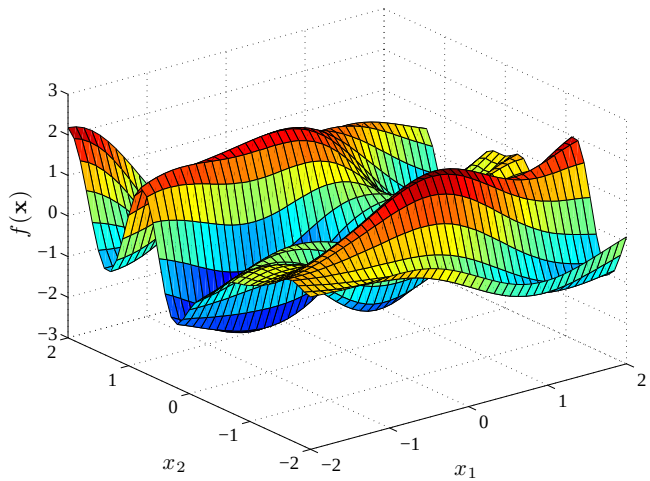
Modelamiento Bayesiano: regresión (II)



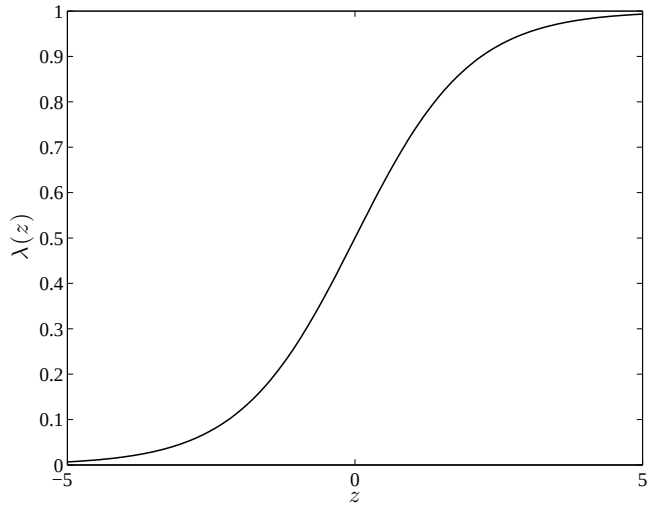
Modelamiento Bayesiano: regresión (II)



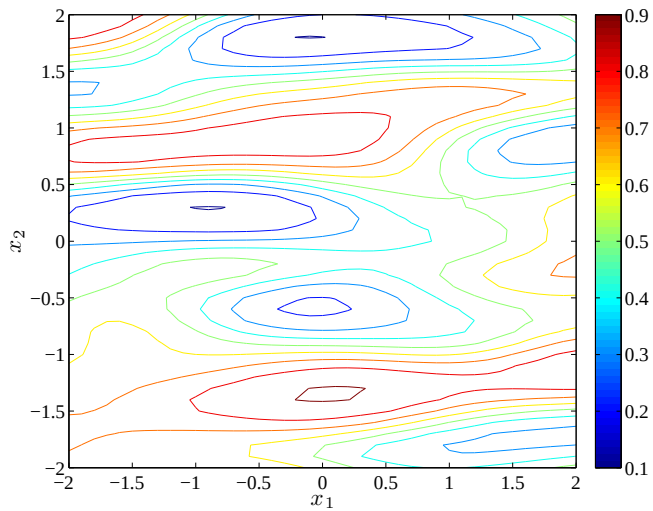
Modelamiento Bayesiano: clasificación (I)



Muestra del GP

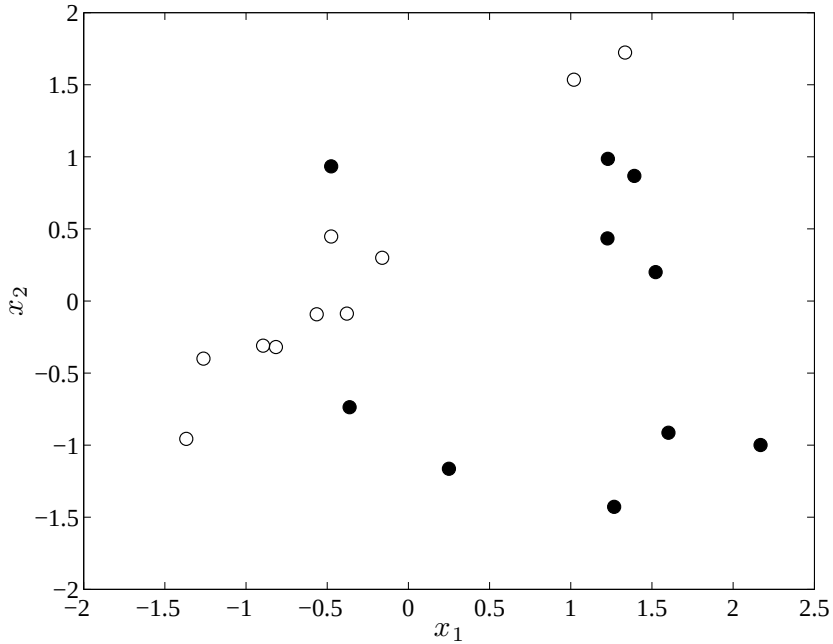


$$\lambda(z) = \frac{1}{1+\exp(-z)}$$

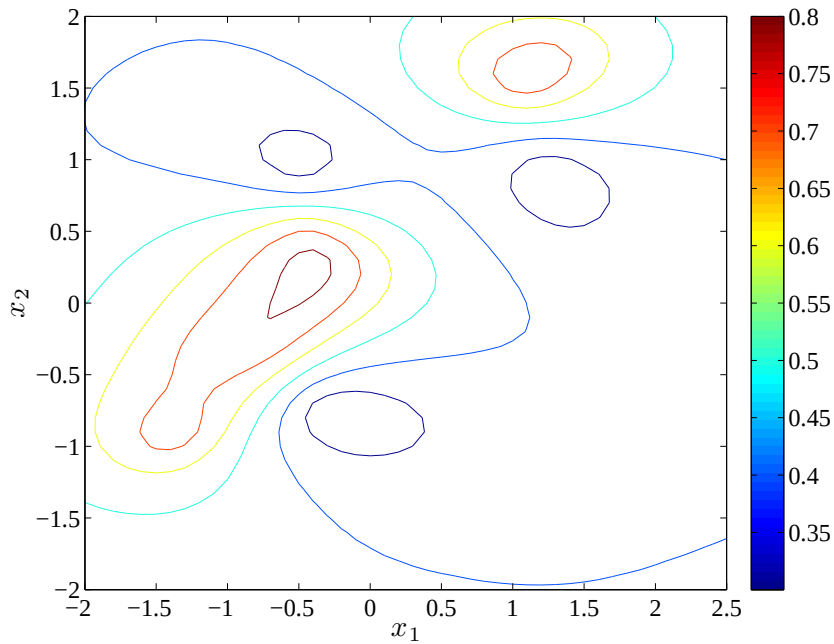


Contorno

Modelamiento Bayesiano: clasificación (II)



Datos



Contorno posterior

Contenido

Preliminares

- Aprendizaje supervisado
  - Clasificación $\rightarrow$ predicción de variables discretas.
  - Regresión $\rightarrow$ 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 $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\}$.
- $\mathbf{x}_i \in \mathbb{R}^D$, $y_i \in \mathbb{R}$.
- Se forman la matriz de diseño $\mathbf{X} \in \mathbb{R}^{D \times n}$, y el vector de salidas $\mathbf{y}$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^\top.$$

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

Modelo lineal estándar (I)

- En el modelo lineal estándar

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

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,

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

Modelo lineal estándar (II)

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

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

donde Σ_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 estándar (III)

- Para el caso del modelo lineal,

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

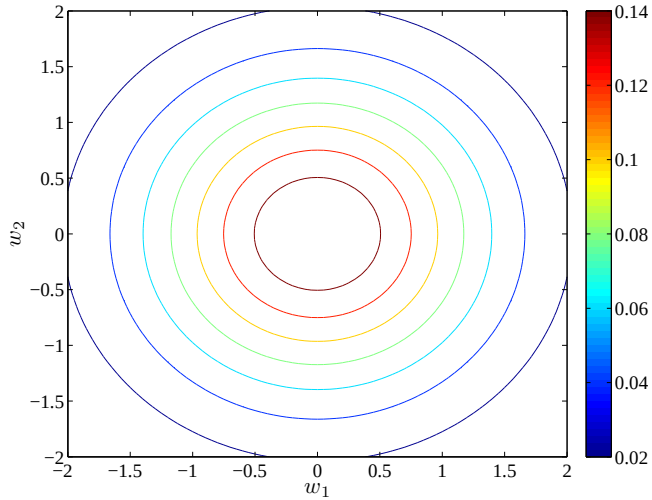
  donde $\mathbf{A} = \sigma_n^{-2}\mathbf{X}\mathbf{X}^\top + \boldsymbol{\Sigma}_p^{-1}$ es un matriz de covarianza.

- Validación $\rightarrow$ promediar sobre $\mathbf{w}$ usando la función posterior.

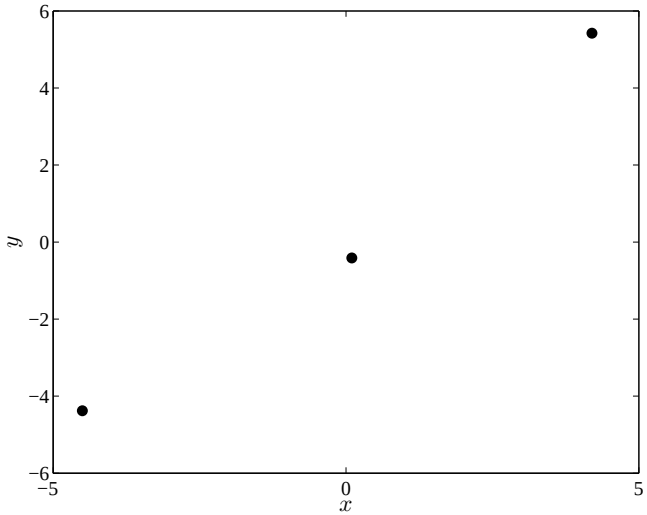
- La distribución predictiva para $f_* \equiv f(\mathbf{x}_*)$ en $\mathbf{x}_*$ está dada como

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

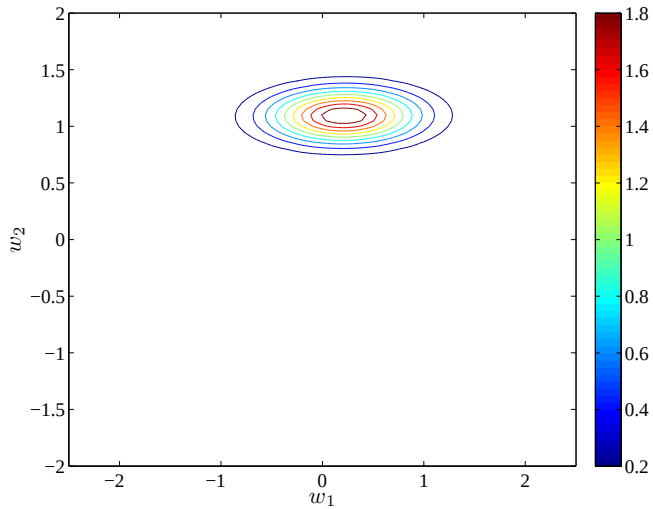
Modelo lineal estándar (IV)



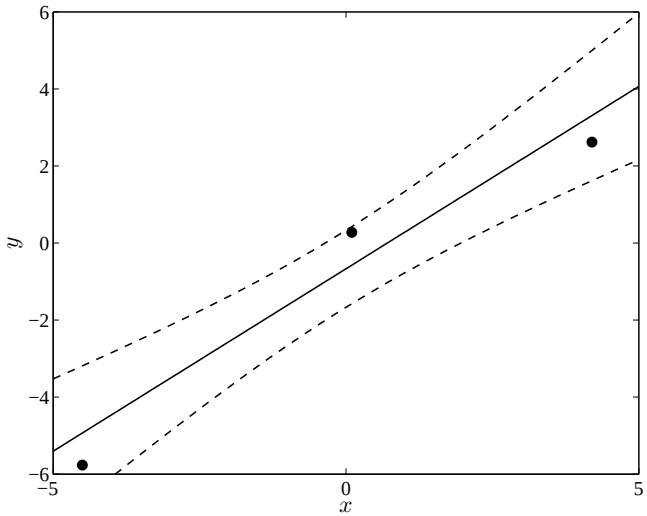
$p(\mathbf{w})$



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



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



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

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 function $\phi(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}^N$.
- $\mathbf{\Phi}(\mathbf{X}) \in \mathbb{R}^{N \times n}$.
- 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 $\mathbf{\Phi}(\mathbf{X})$.

Espacio de características (III)

- Predictiva

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

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

- Invertir **A** costoso para *N* grande.
- Se puede demostrar que

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

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

- El espacio de características siempre aparece de las formas $\phi_*^\top \mathbf{\Sigma}_p \mathbf{\Phi}$, $\phi_*^\top \mathbf{\Sigma}_p \phi_*$, y $\mathbf{\Phi}^\top \mathbf{\Sigma}_p \mathbf{\Phi}$.

Truco del kernel

- Las entradas de las matrices anteriores se pueden escribir de la forma $\phi(\mathbf{x})^\top \boldsymbol{\Sigma}_p \phi(\mathbf{x}')$.
- Se define $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \boldsymbol{\Sigma}_p \phi(\mathbf{x}')$, como un *kernel* o *función de covarianza*.
- $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \boldsymbol{\Sigma}_p \phi(\mathbf{x}') = \boldsymbol{\psi}(\mathbf{x}) \cdot \boldsymbol{\psi}(\mathbf{x})$, con $\boldsymbol{\psi}(\mathbf{x}) = \boldsymbol{\Sigma}_p^{1/2} \phi(\mathbf{x}')$.
- Algoritmo sólo depende de productos internos de vectores del espacio de entrada $\rightarrow$ 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 $(y_1, y_2) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, luego $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$, donde

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

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

Punto de vista del espacio de funciones (IV)

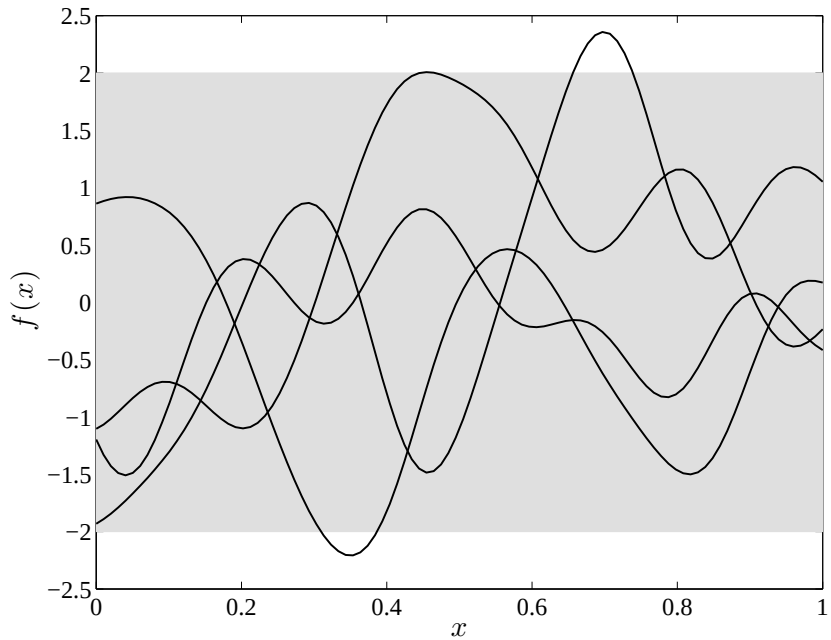
- El modelo Bayesiano lineal es un ejemplo de un proceso Gaussiano.
- Se tiene $f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w}$ con prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$.
- Luego

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

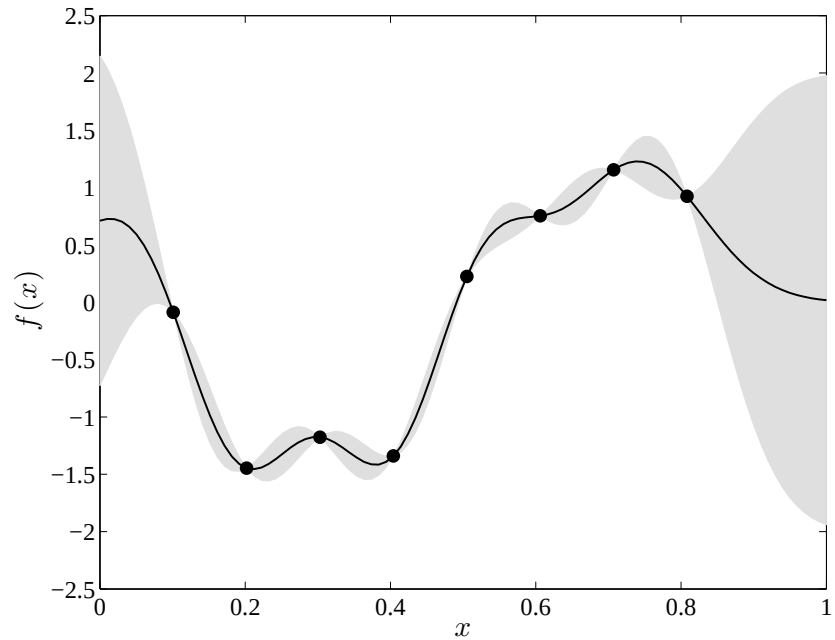
- Examinar un conjunto grande de variables no cambia la distribución de uno de sus subconjuntos.

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*.

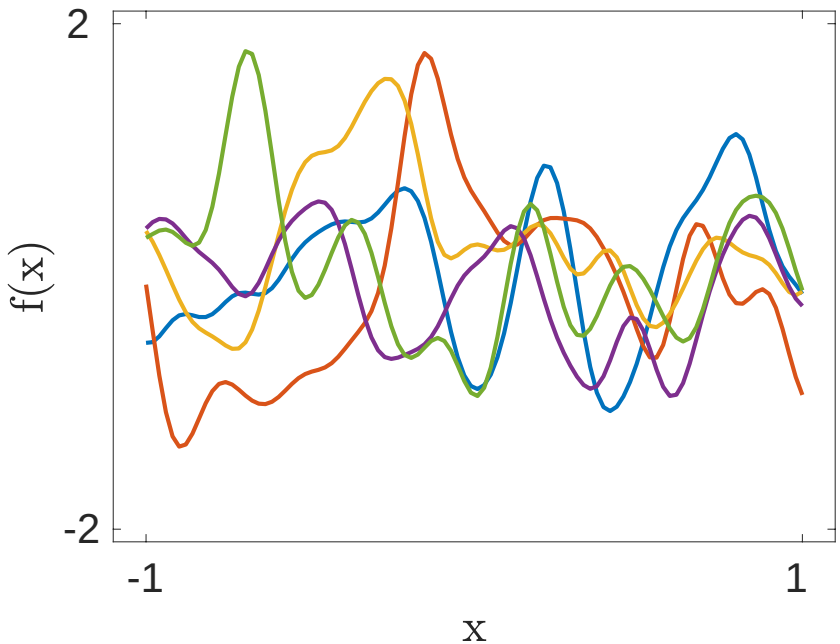
- Por definición, la función de covarianza es positiva semidefinida, lo que conduce a una matriz de covarianza que también es positiva semidefinida

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

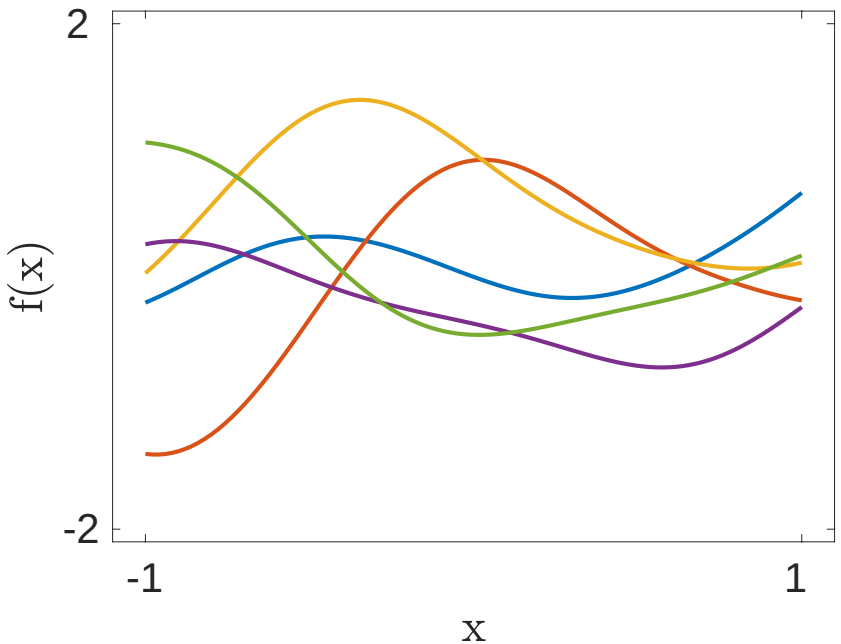
- En aplicaciones prácticas, la función de covarianza se selecciona de un conjunto de funciones disponibles.

Tipos de función de covarianza: exponencial cuadrada

$$k(\mathbf{x}, \mathbf{x}') = S_f \exp\left\{ -\frac{1}{2\ell^2} \|\mathbf{x} - \mathbf{x}'\|_2^2 \right\}.$$



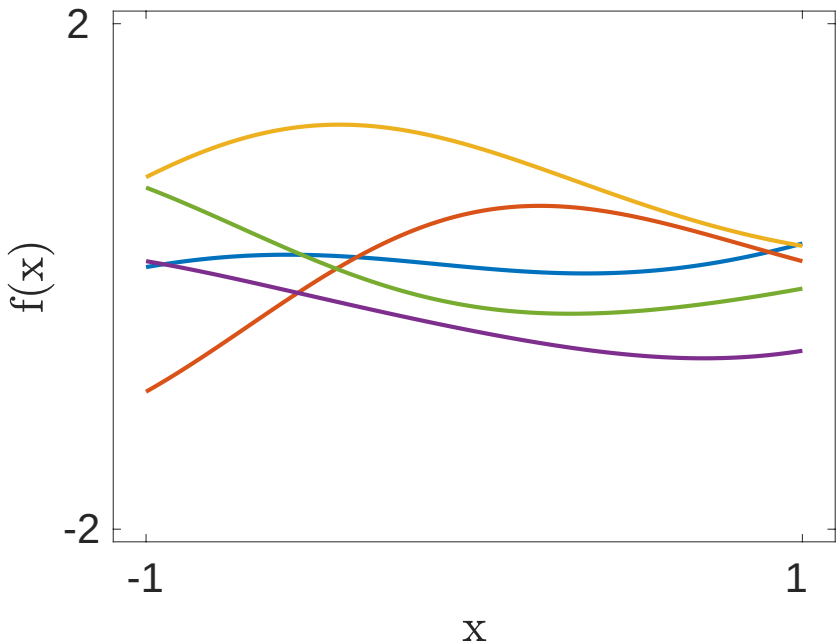
$S_f = 0.5, \ell = 0.1$



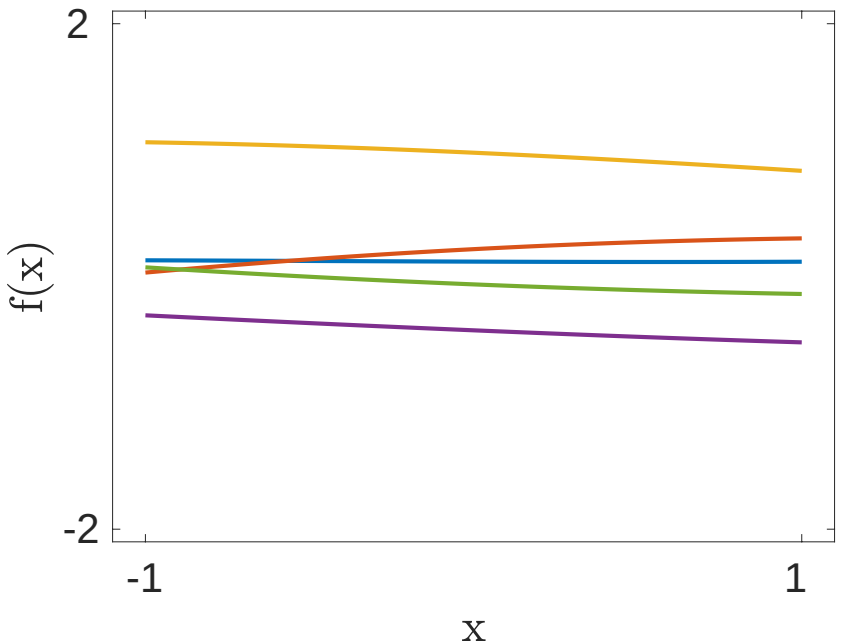
$S_f = 0.5, \ell = 0.5$

Tipos de función de covarianza: exponencial cuadrada

$$k(\mathbf{x}, \mathbf{x}') = S_f \exp\left\{ -\frac{1}{2\ell^2} \|\mathbf{x} - \mathbf{x}'\|_2^2 \right\}.$$



$S_f = 0.5, \ell = 1$

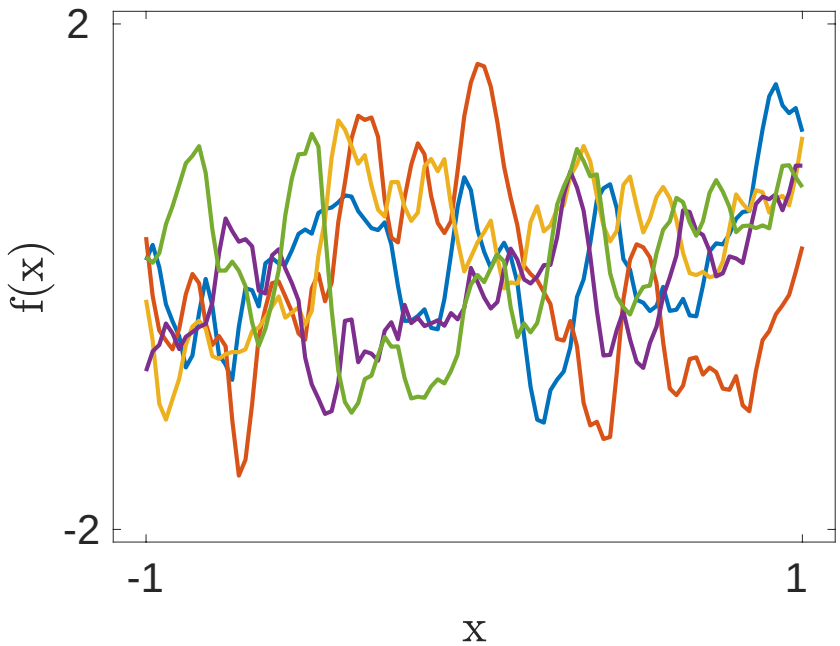


$S_f = 0.5, \ell = 5$

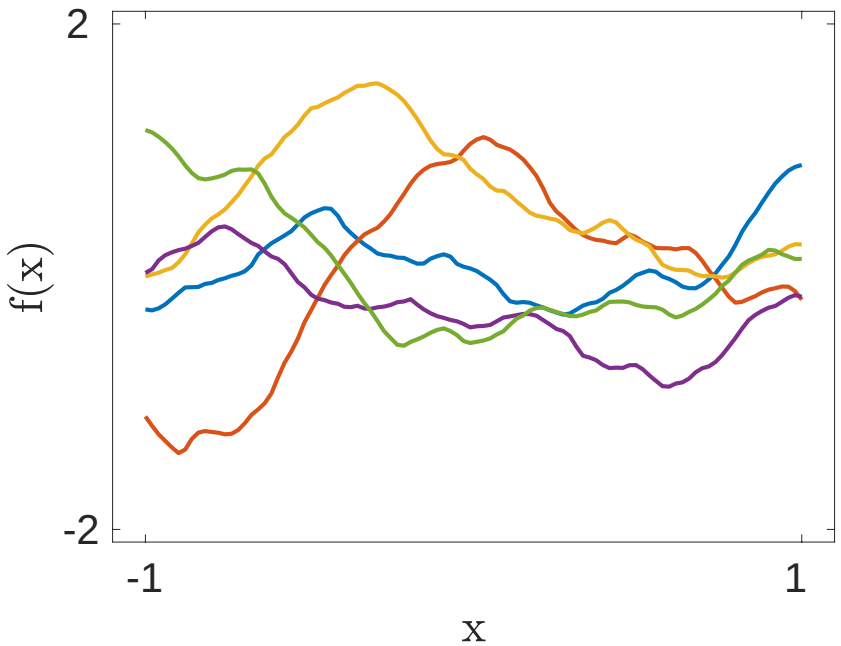
Tipos de función de covarianza: Matèrn

$$k(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu} S_f}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|\mathbf{x} - \mathbf{x}'\|_2}{\ell} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\|\mathbf{x} - \mathbf{x}'\|_2}{\ell} \right)$$

$$k(\mathbf{x}, \mathbf{x}') = S_f \left(1 + \frac{\sqrt{3}\|\mathbf{x} - \mathbf{x}'\|_2}{\ell} \right) \exp\left(- \frac{\sqrt{3}\|\mathbf{x} - \mathbf{x}'\|_2}{\ell} \right), \quad \nu = \frac{3}{2}$$



$S_f = 0.5, \ell = 0.1$



$S_f = 0.5, \ell = 0.5$

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

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

donde

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

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

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, ℓ y σ_n^2 pueden estimarse maximizando el logaritmo de la verosimilitud marginal,

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^\top \left(\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}| - \frac{n}{2} \log 2\pi.$$

Diferenciación automática (autodiff)

Automatic Differentiation in Machine Learning: a Survey

Atılım Güneş Baydin GUNES@ROBOTS.OX.AC.UK
Department of Engineering Science
University of Oxford
Oxford OX1 3PJ, United Kingdom

Barak A. Pearlmutter BARAK@PEARLMUTTER.NET
Department of Computer Science
National University of Ireland Maynooth
Maynooth, Co. Kildare, Ireland

Alexey Andreyevich Radul AXCH@MIT.EDU
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139, United States

Jeffrey Mark Siskind QOBI@PURDUE.EDU
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907, United States

Editor: Léon Bottou

Abstract

Derivatives, mostly in the form of gradients and Hessians, are ubiquitous in machine learning. Automatic differentiation (AD), also called algorithmic differentiation or simply "autodiff", is a family of techniques similar to but more general than backpropagation for efficiently and accurately evaluating derivatives of numeric functions expressed as computer programs. AD is a small but established field with applications in areas including computational fluid dynamics, atmospheric sciences, and engineering design optimization. Until very recently, the fields of machine learning and AD have largely been unaware of each other and, in some cases, have independently discovered each other's results. Despite its relevance, general-purpose AD has been missing from the machine learning toolbox, a situation slowly changing with its ongoing adoption under the names "dynamic computational graphs" and "differentiable programming". We survey the intersection of AD and machine learning, cover applications where AD has direct relevance, and address the main implementation techniques. By precisely defining the main differentiation techniques and their interrelationships, we aim to bring clarity to the usage of the terms "autodiff", "automatic differentiation", and "symbolic differentiation" as these are encountered more and more in machine learning settings.

GPs con diferenciación automática

Hay diferentes paquetes para GPs construidos sobre programas que hacen diferenciación automática como TensorFlow y PyTorch:

1. GPFlow, desarrollado sobre TensorFlow por la Universidad de Cambridge y ahora mantenido por PROWLER.io
2. GPyTorch, desarrollado sobre PyTorch por la Universidad de Cornell.

De forma alternativa, uno puedo desarrollar sus propios algoritmos de GPs y recurrir a autodiff cuando sea necesario usando una librería de Python como *autograd*.

Ejercicio regresión

Usando GPFlow (ver folder lab5)

Contenido

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).

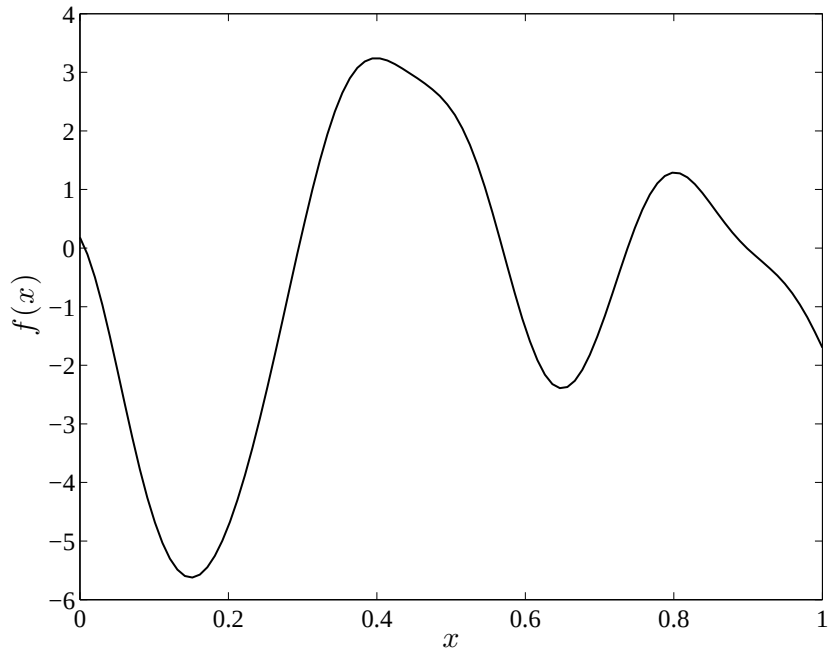
Procesos Gaussianos para clasificación binaria (I)

- Se asume que la función $f(\mathbf{x})$ sigue un proceso Gaussiano.

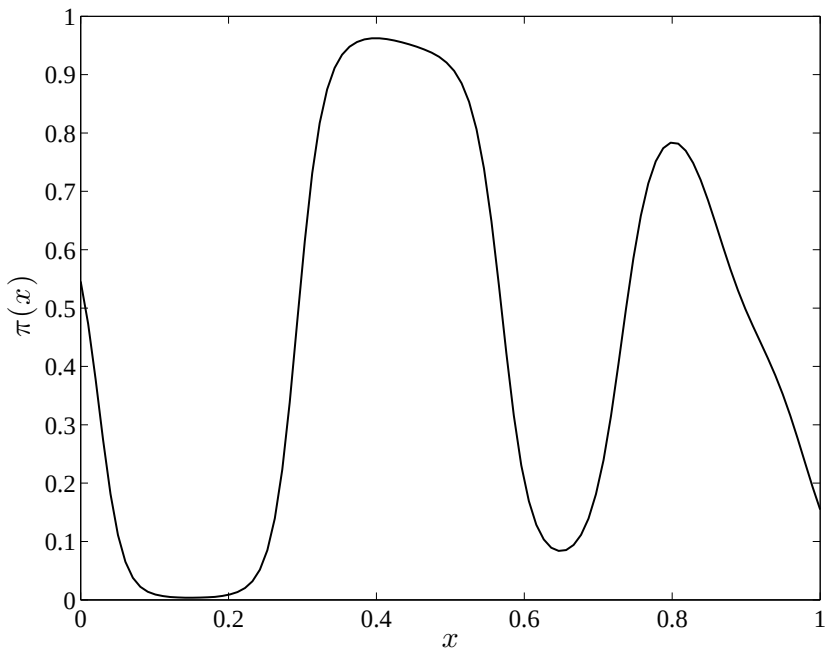
- La función $f(\mathbf{x})$ se pasa a través de la función logística $\sigma(\cdot)$

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

Procesos Gaussianos para clasificación binaria (II)



Función latente



Clase condicional

Inferencia en dos pasos

- Paso 1. Para un nuevo $\mathbf{x}_*$ se calcula para 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

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

- La integral del Paso 1 no es tratable analíticamente.

Aproximación por Laplace (I)

- El posterior $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ no es Gaussiano debido a la función de verosimilitud asociada.
- La aproximación de Laplace aproxima $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ usando una distribución normal.
- Otras aproximaciones incluyen el algoritmo de Propagación de la Esperanza (Expectation-Propagation- EP), Markov chain Monte Carlo (MCMC) e Inferencia Variacional (el tema de los próximos tres días).

Aproximación por Laplace (II)

- En la aproximación por Laplace

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

donde

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

- Para encontrar $\widehat{\mathbf{f}}$ 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 $\widehat{\mathbf{f}}$ se encuentra usando optimización por Newton.

Predicción, y estimación

- Usando la aproximación de Laplace para el posterior,

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

- 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

Contenido

Otros temas asociados a Procesos Gaussianos

- Diversidad de funciones de covarianza.
- Relaciones de procesos Gaussianos con otros modelos.
- Métodos de aproximación.
- Modelos de variable latente.
- Múltiples salidas (aprendizaje multi-tarea).