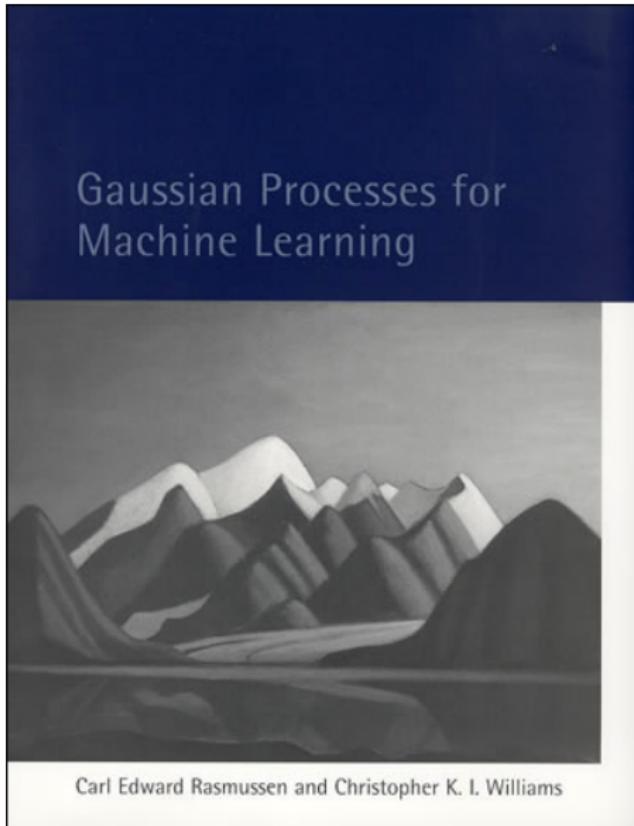


Introducción a los Procesos Gaussianos en Aprendizaje de Máquina

Mauricio A. Álvarez, PhD

Curso de entrenamiento ArcelorMittal



Contenido

Introducción

Regresión

Clasificación

Contenido

Introducción

Regresión

Clasificación

Aprendizaje supervisado

$$X = \begin{bmatrix} \quad \end{bmatrix}_{n \times p}$$

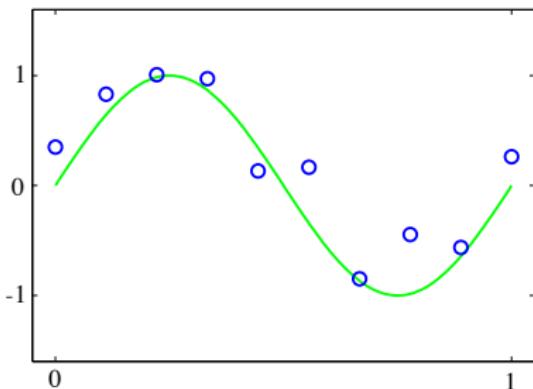
$n \gg p$ $f(x) : x \rightarrow y$ $p \gg n$ $p \rightarrow d$
 $d \ll p$

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.

$$x_1, x_2, x_3, \dots, x_n \quad x \in \mathbb{R}^p$$

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



Inducción

- Dado el conjunto de entrenamiento $\underline{\mathcal{D}}$, se desea hacer predicciones para un nuevo \underline{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

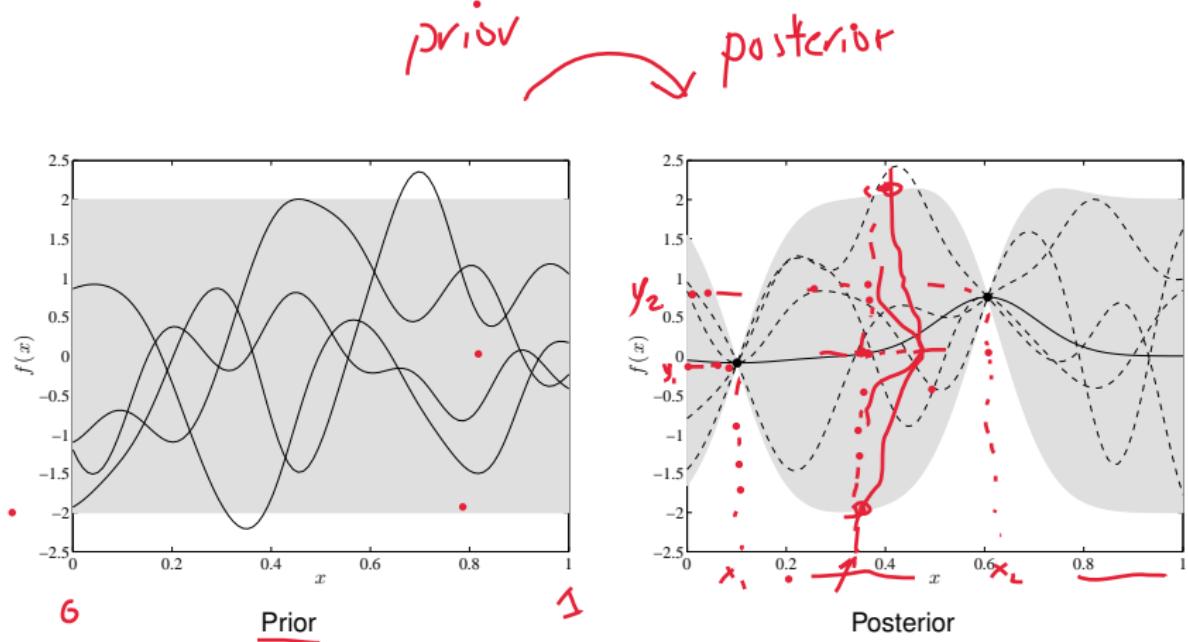
- Primer enfoque: restringir la clase de funciones que se pueden considerar.
- Problema:
 - La función objetivo podría quedar mal modelada, luego las predicciones serán pobres.
 - Aumentar la flexibilidad de la clase de funciones, con el peligro de sobre-entrenar.

- 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.

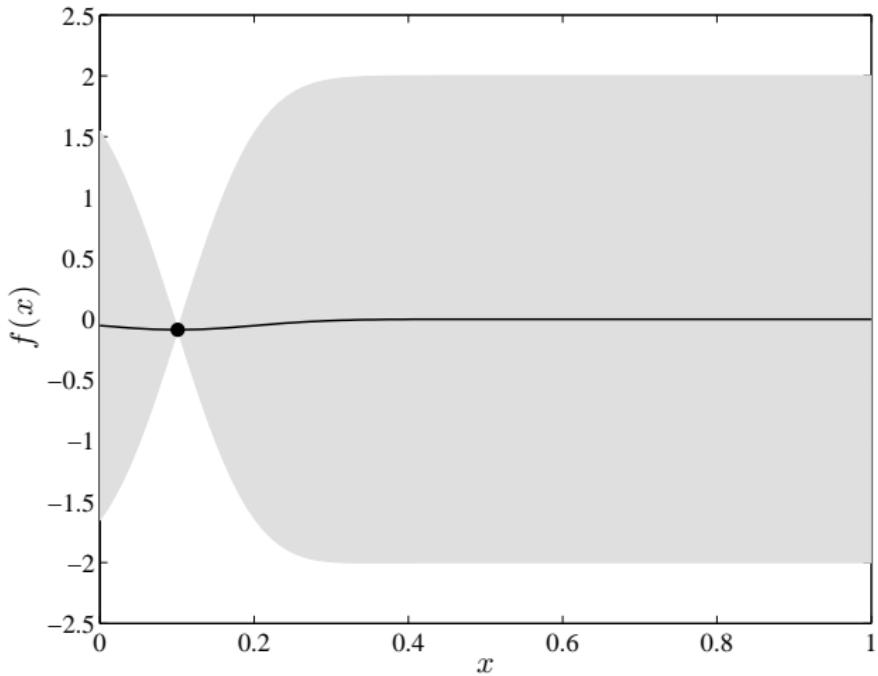
Modelamiento Bayesiano: regresión (I)



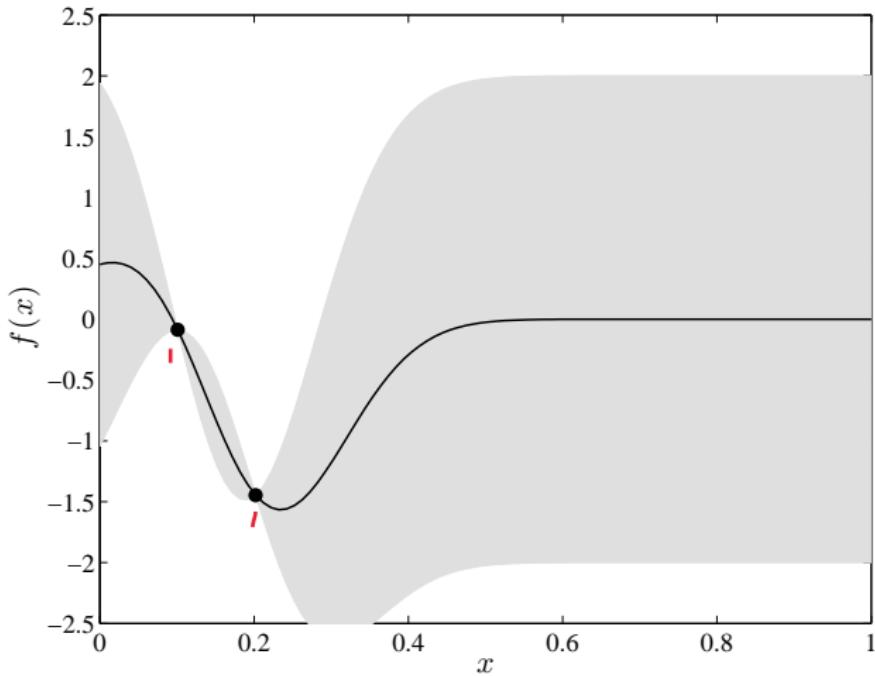
Dos observaciones $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$.

$$\left[\begin{array}{c} \cdot \\ \cdot \end{array} \right] \sim N \left(\left[\begin{array}{c} \cdot \\ \cdot \end{array} \right], \left[\begin{array}{cc} \cdot & \cdot \\ \cdot & \cdot \end{array} \right] \right)$$

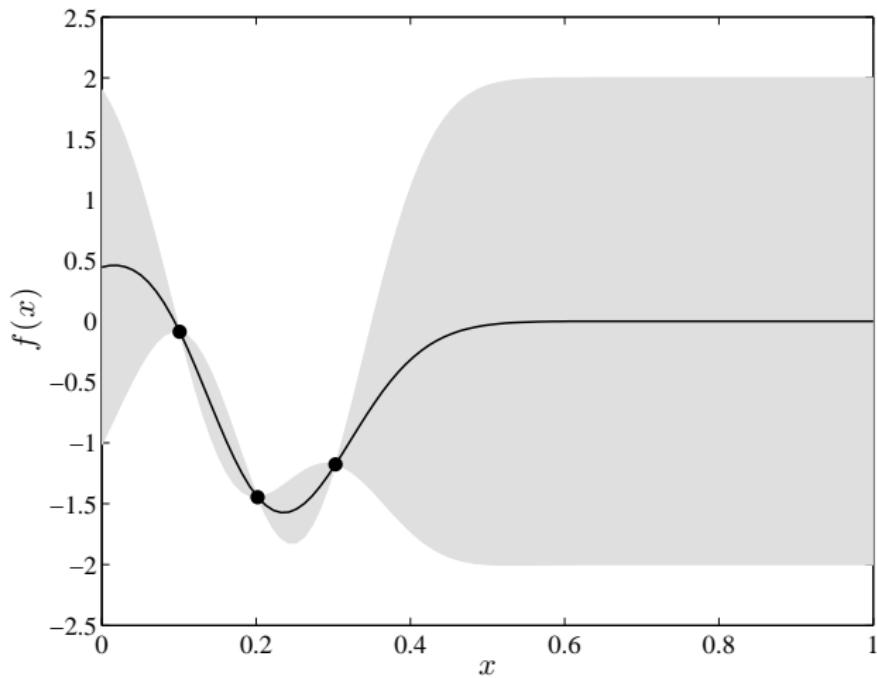
Modelamiento Bayesiano: regresión (II)



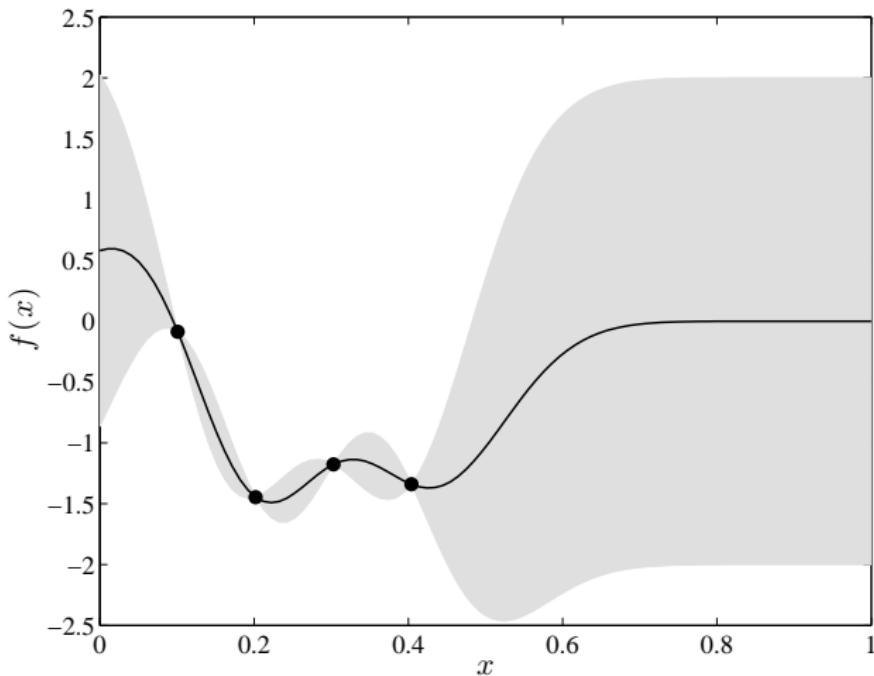
Modelamiento Bayesiano: regresión (II)



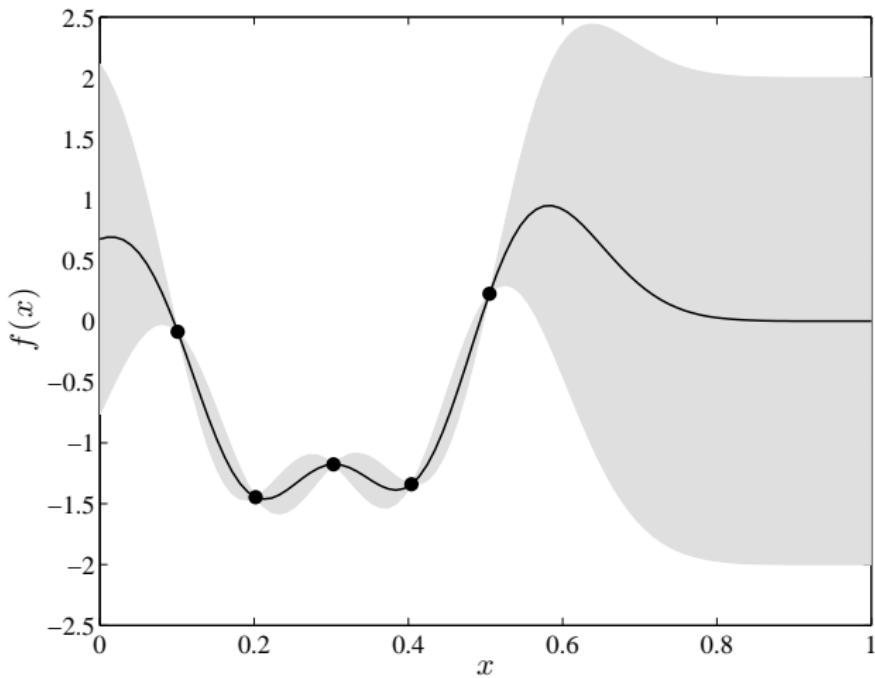
Modelamiento Bayesiano: regresión (II)



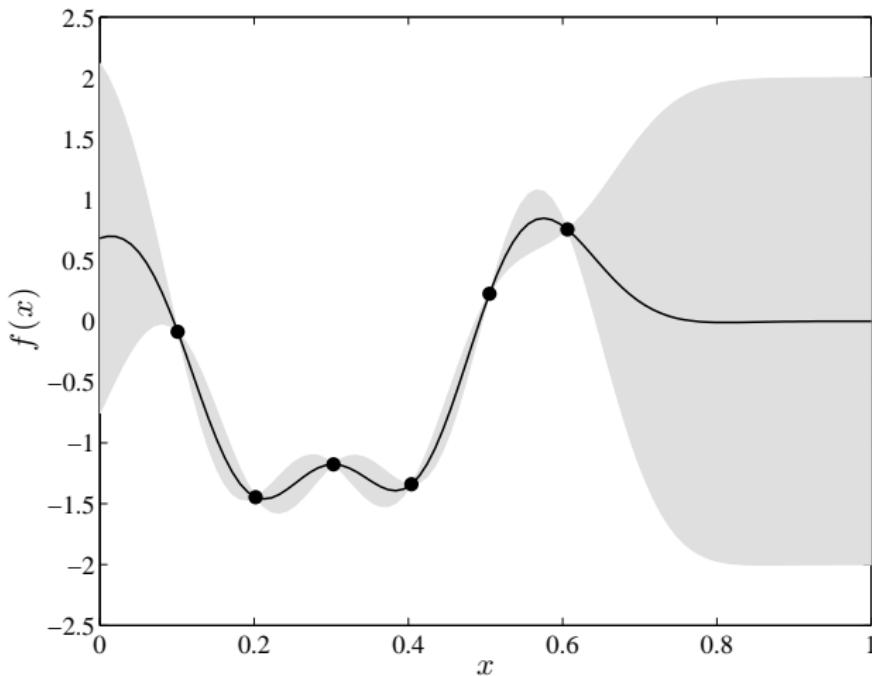
Modelamiento Bayesiano: regresión (II)



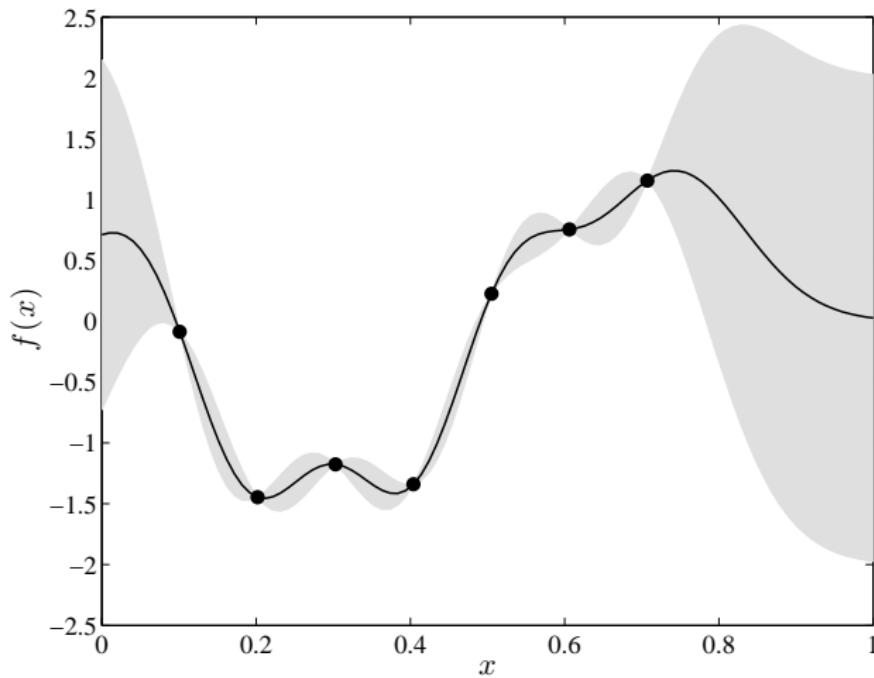
Modelamiento Bayesiano: regresión (II)



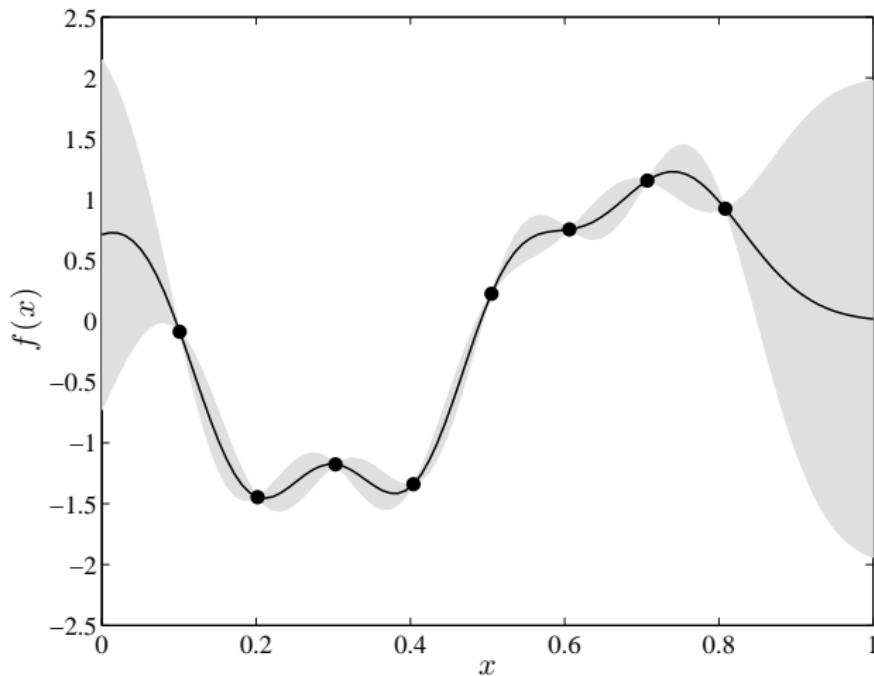
Modelamiento Bayesiano: regresión (II)



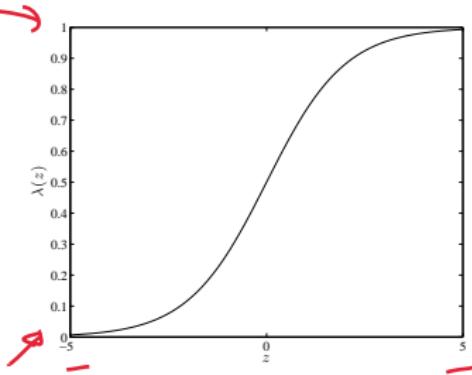
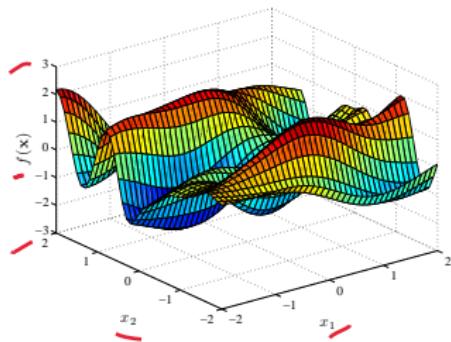
Modelamiento Bayesiano: regresión (II)



Modelamiento Bayesiano: regresión (II)

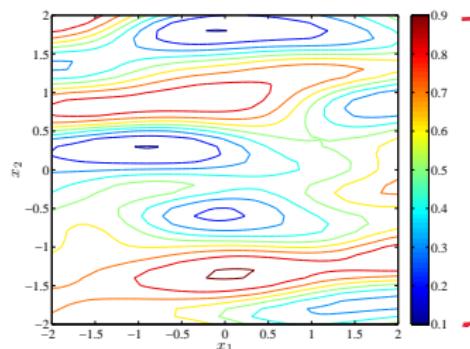


Modelamiento Bayesiano: clasificación (I)

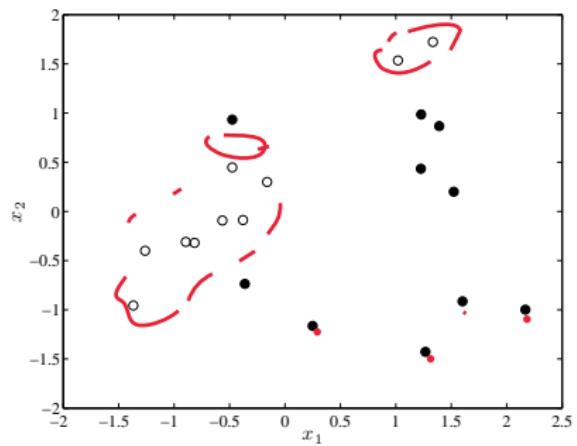


$\hat{x} = [x_1, x_2]$

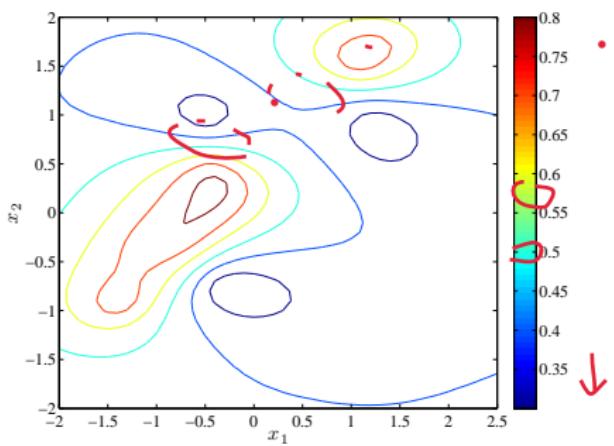
$\hat{x}_1, \dots, \hat{x}_n$



Modelamiento Bayesiano: clasificación (II)



Datos



Contorno posterior

Contenido

Introducción

Regresión

Clasificación

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 $\mathcal{D} = \{(\underline{\mathbf{x}}_i, \underline{y}_i) | i = 1, \dots, n\}$.

- $\underline{\mathbf{x}}_i \in \mathbb{R}^D$, $\underline{y}_i \in \mathbb{R}$.

$$\underline{\mathbf{X}} \in \mathbb{R}^{n \times D}.$$

- Se forman la matriz de diseño $\underline{\mathbf{X}} \in \mathbb{R}^{D \times n}$, y el vector de salidas $\underline{\mathbf{y}}$,

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

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

Modelo lineal estándar (I) $\underline{w} = [w_1 \dots w_D]^\top$

- En el modelo lineal estándar

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

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

$$\underline{y} = f(\underline{x}) + \epsilon$$

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

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

Modelo lineal est醤dar (II)

$$p(y|X, w)$$

- Se especifica una distribución prior para w, por ejemplo,

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

donde Σ_p es un matriz de covarianza.

- Teorema de Bayes,

$$\text{posterior} = \frac{\text{verosimilitud} \times \text{prior}}{\text{verosimilitud marginal}},$$

$$p(\underline{w}|y, X) = \frac{p(y|X, w)p(w)}{p(y|X)},$$

donde

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

Modelo lineal est醤dar (III)

- Para el caso del modelo lineal,

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

donde $\mathbf{A} = \underbrace{\sigma_n^{-2} \mathbf{X} \mathbf{X}^\top}_{\text{---}} + \underbrace{\Sigma_p^{-1}}_{\text{---}}$ 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

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

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

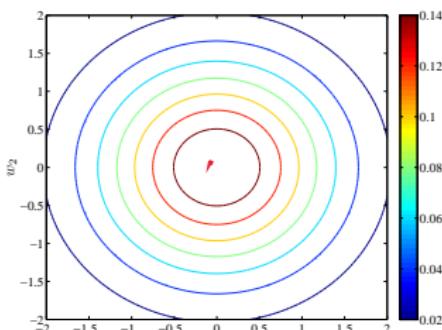
Modelo lineal est醤dar (IV)

$$f(x) = \underline{w_1} + \underline{w_2}x \quad \bar{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

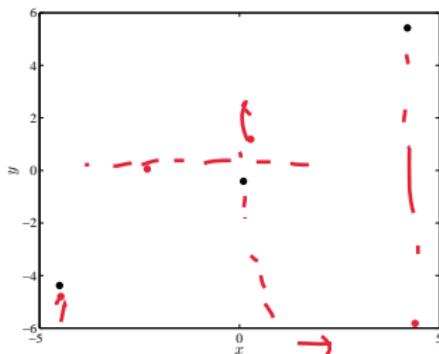
$$f(x) = \underline{w}^T \bar{x}$$

$$\underline{\zeta_p} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

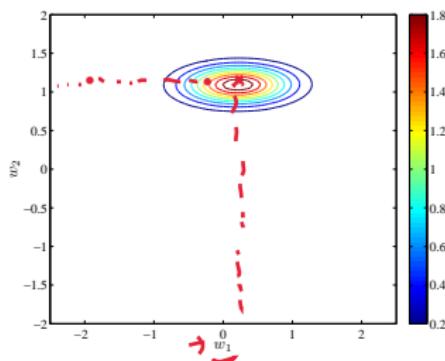
$$\underline{w} = \begin{bmatrix} \underline{w}_1 & \underline{w}_2 \end{bmatrix}^T$$



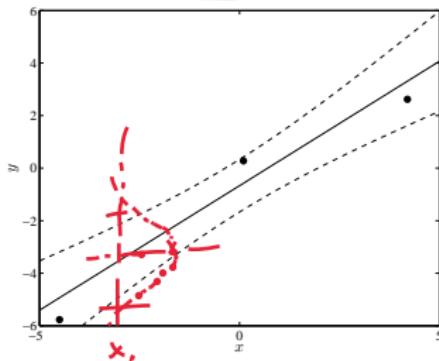
$$p(\underline{w})$$



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



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



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

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)

$$\underline{\underline{x}} \in \mathbb{R}^D \quad \underline{\underline{X}} \in \mathbb{R}^{D \times n}$$

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

Espacio de características (III)

O(N³)

N > D

- Predictiva

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

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

$$p(w) = \mathcal{N}(w | 0, \Sigma_p)$$

- Invertir $\underline{\mathbf{A}}$ costoso para \underline{N} grande.

- Se puede demostrar que

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

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

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

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*.
 - $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \Sigma_p \phi(\mathbf{x}') = \psi(\mathbf{x})^\top \psi(\mathbf{x}')$, con $\psi(\mathbf{x}) = (\Sigma_p^{1/2} \phi(\mathbf{x}'))^\top$.
 - 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.

$$f = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \xrightarrow{\quad} []$$

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

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

- 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 $f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$.

$$0 \quad y \quad x$$

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

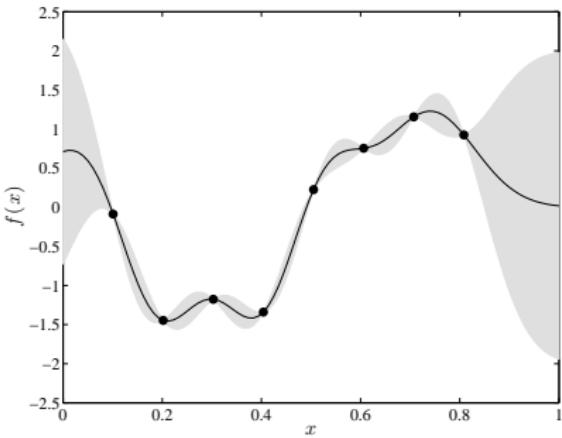
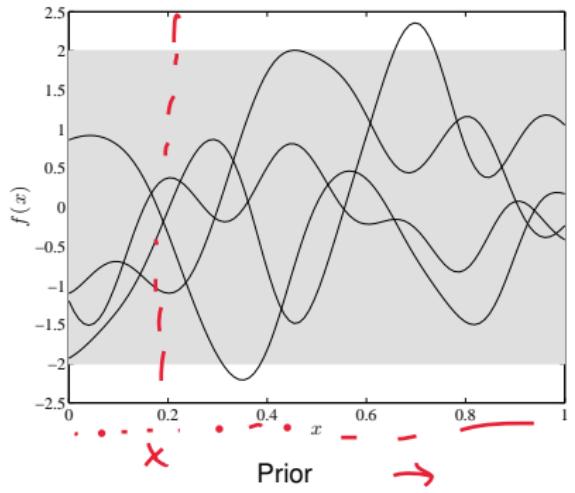
$$\mathbb{E}[f(\mathbf{x})] = \phi(\mathbf{x})^\top \mathbb{E}[\mathbf{w}] = \mathbf{0},$$

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

Ejemplo

$$K(\mathbf{x}, \mathbf{x}') = \begin{bmatrix} k(x_1, x'_1) & k(x_1, x'_2) & k(x_1, x'_3) & \dots & k(x_1, x'_m) \\ k(x_2, x'_1) & k(x_2, x'_2) & k(x_2, x'_3) & \dots & k(x_2, x'_m) \\ k(x_3, x'_1) & k(x_3, x'_2) & k(x_3, x'_3) & \dots & k(x_3, x'_m) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ k(x_m, x'_1) & k(x_m, x'_2) & k(x_m, x'_3) & \dots & k(x_m, x'_m) \end{bmatrix}$$

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



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

$$K(\mathbf{x}, \mathbf{x}) \in \mathbb{R}^{n \times n}$$

RBF
EQ

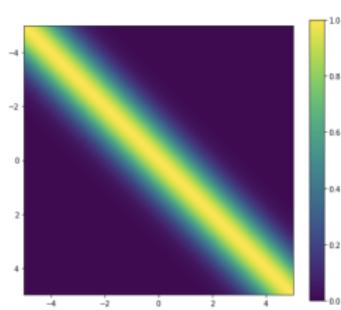
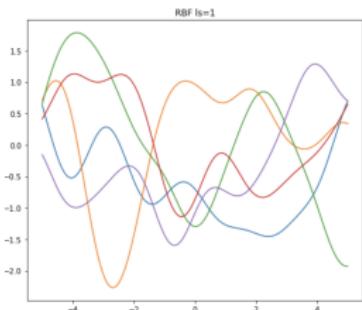
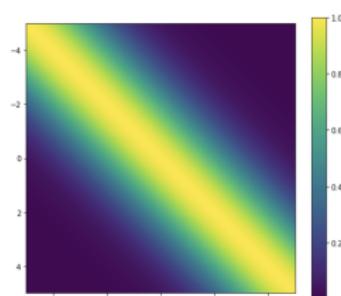
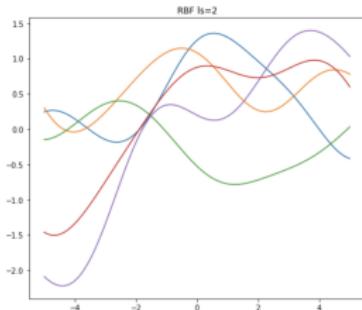
Función de covarianza

- 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

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

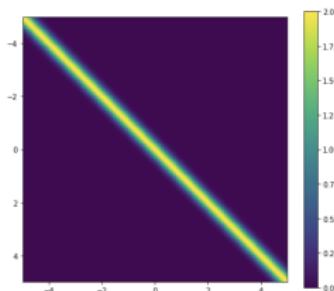
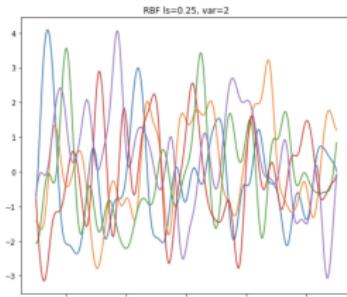
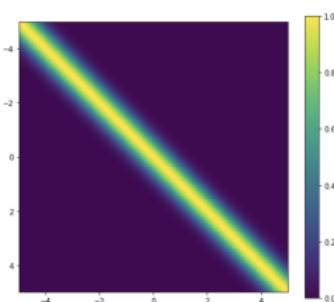
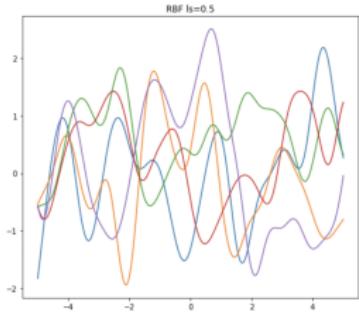
- En aplicaciones prácticas, la función de covarianza se selecciona de un conjunto de funciones disponibles

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



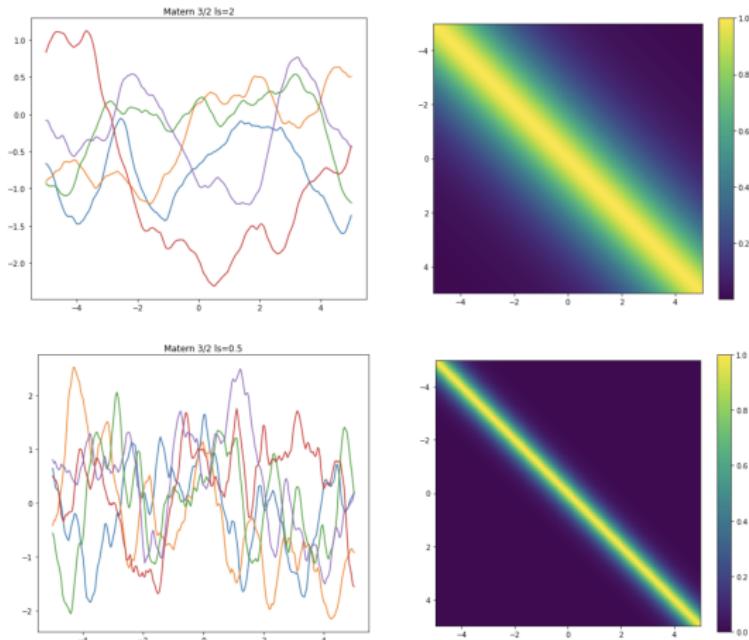
$$k(\mathbf{x}, \mathbf{x}') = s_f \exp \left(-\frac{r^2}{2\ell^2} \right), \quad r = |\mathbf{x} - \mathbf{x}'|.$$

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 , ℓ y σ_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

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

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

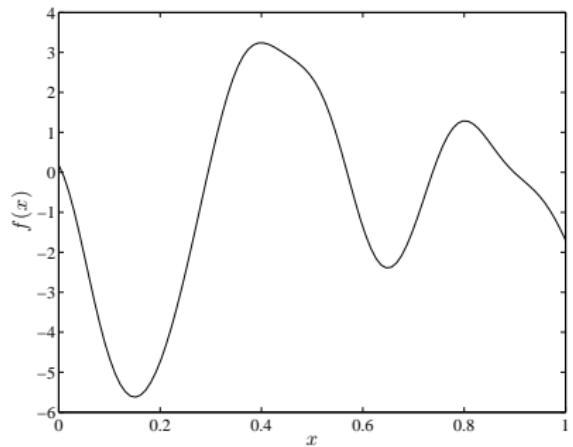
- En clasificación, el posterior no tiene una forma analítica simple.
- Algoritmo IRLS (iteratively reweighted least squares).

Procesos Gaussianos para clasificación binaria (I)

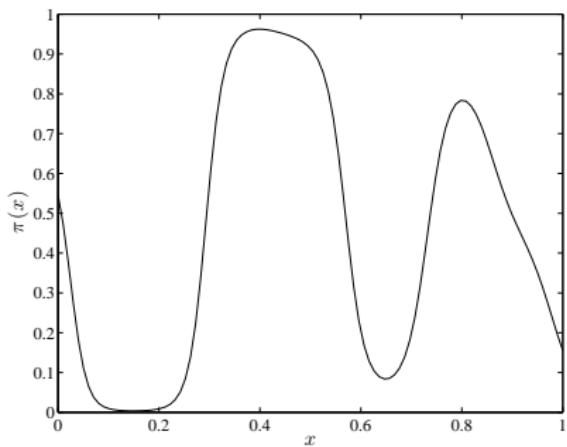
- Se asume que la función $f(\mathbf{x})$ sigue un proceso Gaussiano.
- La función $f(\mathbf{x})$ se pasa a través de la función logística $\sigma(\cdot)$

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

Procesos Gaussianos para clasificación binaria (II)



Función latente



Clase condicional

Inferencia en dos pasos

- ❑ Paso 1. Para un nuevo \mathbf{x}_* , 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.

Predicción, y estimación

- Usando la aproximación de Laplace para el posterior,

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

- 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