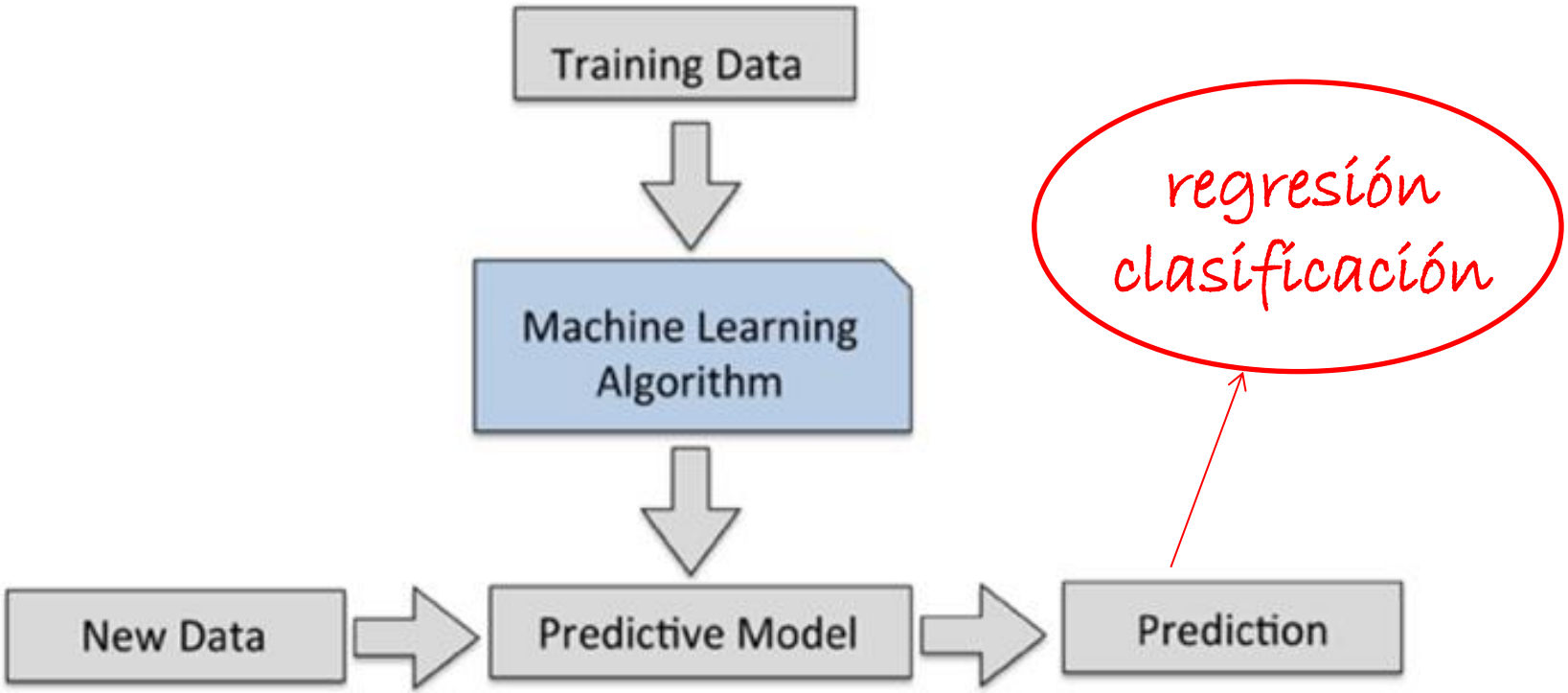
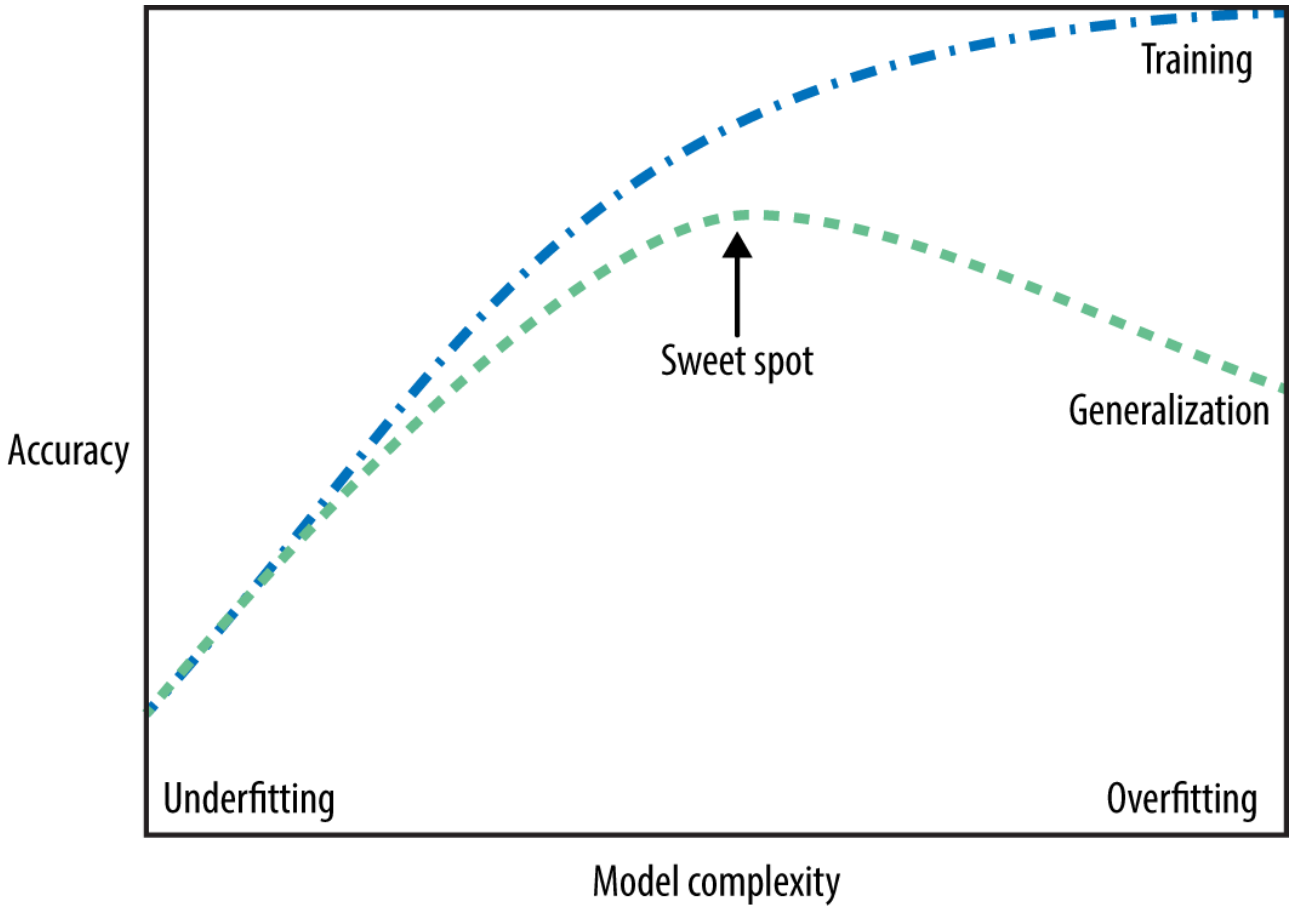
Minería de Datos: Regresión

Luis Avila
loavila@unsl.edu.ar

Aprendizaje supervisado

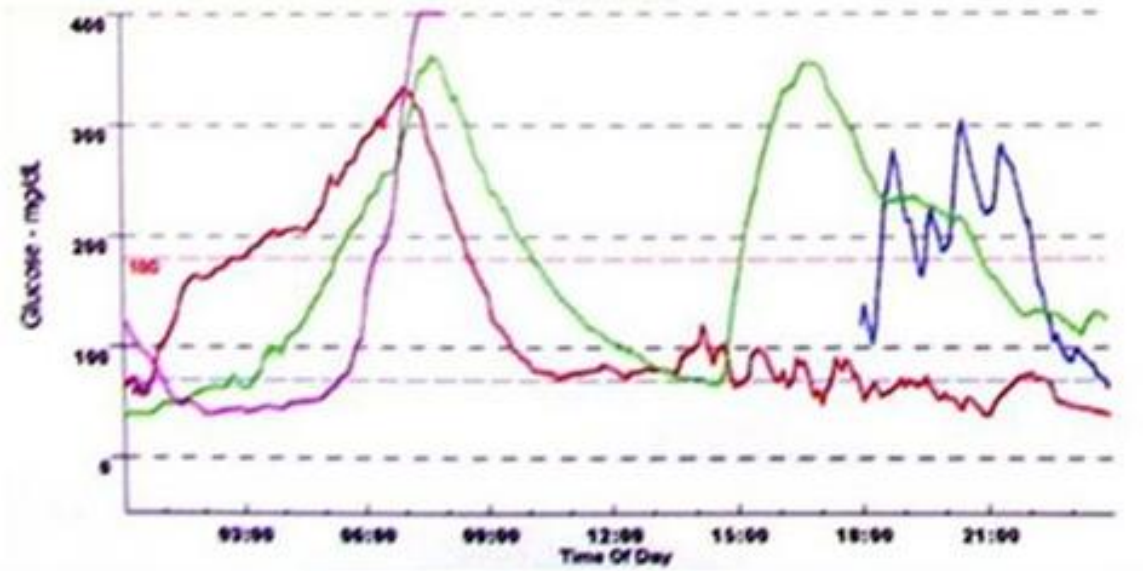


Complejidad del modelo



Incertidumbre

Variabilidad

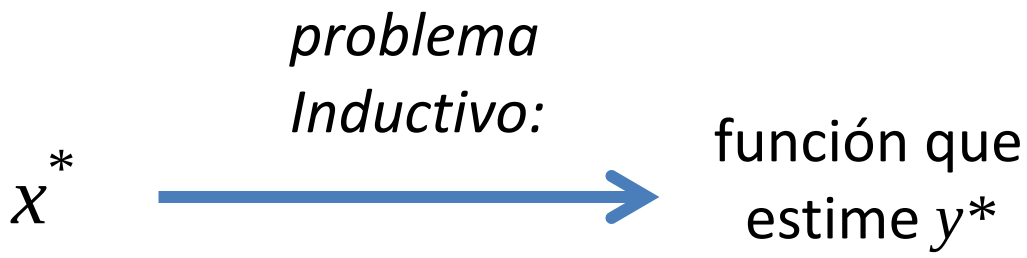


Patient 2:
HbA1c = 7.3 %
Mean glucose: 149 mg/dl
SD: 94 mg/dl

Modelos de regresión

Surge la idea de que cada variable tiene cierta dependencia, intrínsecamente relacionada a las demás variables.

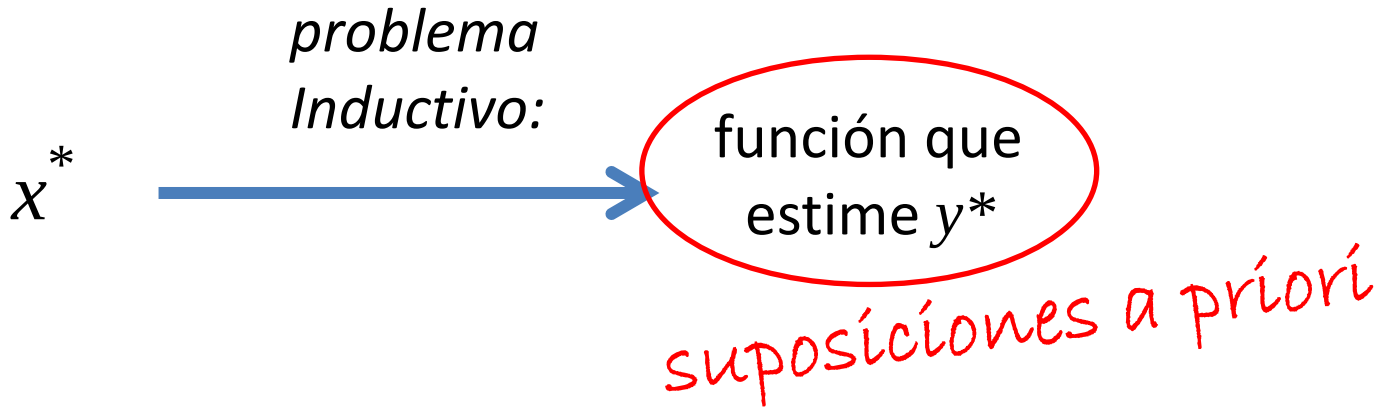
A partir de N observaciones $D = \{ x_i, y_i \mid i = 1, 2, ..., N \}$



Modelos de regresión

Surge la idea de que cada variable tiene cierta dependencia, intrínsecamente relacionada a las demás variables.

A partir de N observaciones $D = \{ x_i, y_i \mid i = 1, 2, \ldots, N \}$



Modelos de regresión

$$y_i = r\ x_{i1},\ldots,x_{in}\ + \varepsilon_i$$

especifica al modelo *especifica a la observación*

- r es la parte determinista y estructural, que permite explicar el comportamiento de los datos.
- ε_i representa la parte impredecible aleatoria y se denomina término de error.

Función de regresión

Para estimar $y(x)$, buscaremos otra función tal que:

$$\min_r E[(y_i - \hat{y}(x_{i1}, \ldots, x_{in}))^2]$$

Medidas de error

Error cuadrático medio

$$MSE = \frac{1}{n}\sum_{x \in D} \hat{y}\ x\ - y\ x\ ^2$$

Raíz cuadrada del MSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{x \in D} \hat{y}\ x\ - y\ x\ ^2}$$

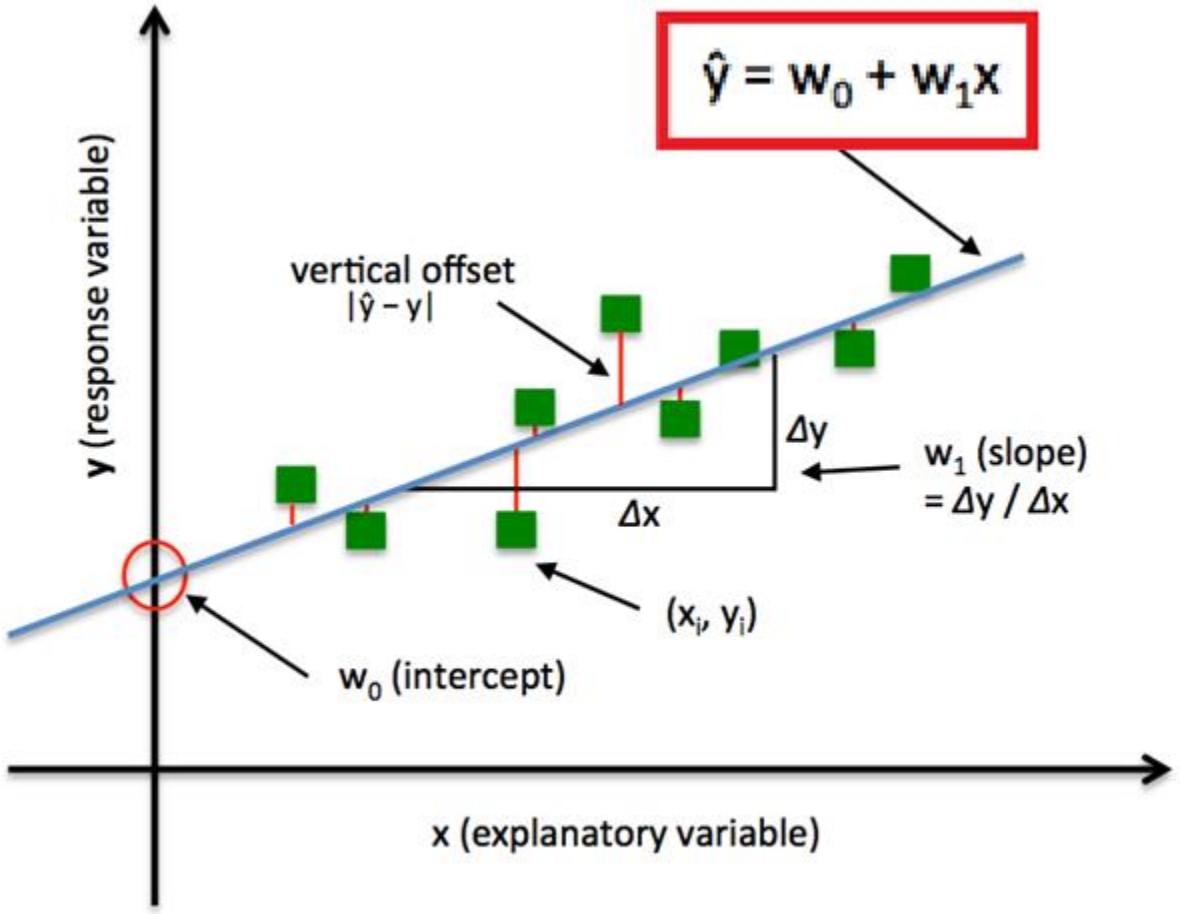
Error absoluto medio

$$MAE = \frac{1}{n}\sum_{x \in D} \left| \hat{y}\ x\ - y\ x\ \right|$$

Error cuadrático relativo

$$MSE = \frac{1}{n}\sum_{x \in D} \frac{\hat{y}(x) - y(x)\ ^2}{\hat{y}(x) - \bar{y}\ ^2}, \quad \text{donde} \quad \bar{y} = \frac{1}{n}\sum_{x \in D} y(x)$$

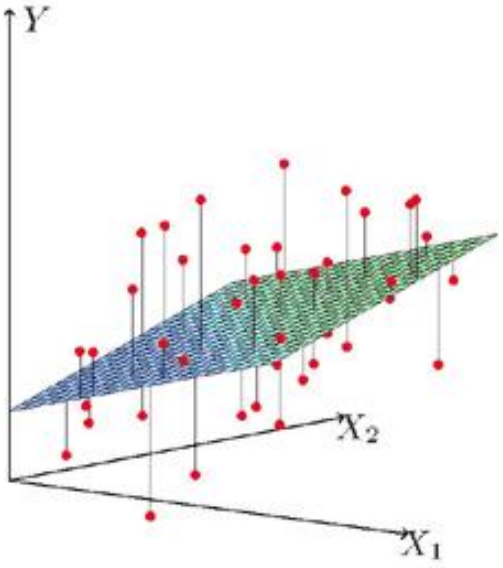
Regresión lineal simple

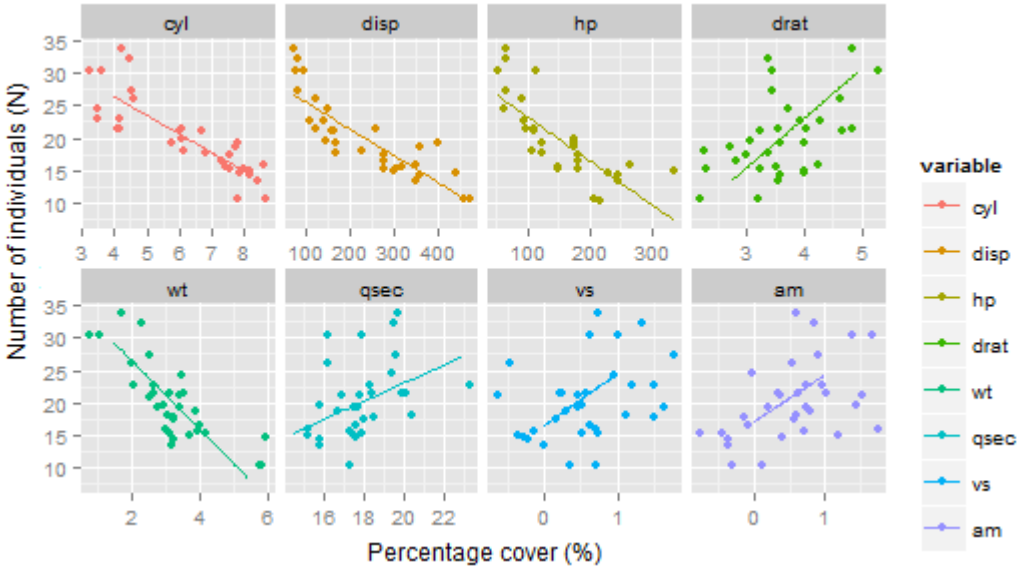


Regresión lineal multivariable

Podemos generalizar el modelo de regresión lineal a múltiples variables explicativas

$$y = w_0 x_0 + w_1 x_1 + \ldots + w_m x_m = \sum_{i=0}^{n} w_i x_i = w^T x$$





Método de mínimos cuadrados

Queremos ajustar un set $\{(x_1,y_1),\ldots,(x_N,y_N)\}$
al modelo lineal $y = w_0 + w_1 x$

$$SE = \sum_{i=1}^{N} \left[y_i - \left(w_0 + w_1 x_i \right) \right]^2$$

La objetivo es hallar los valores w_0 y w_1 que minimizan el error

$$\frac{\partial E}{\partial w_0} = -2 \sum_{i=1}^{N} \left[y_i - \left(w_0 + w_1 x_i \right) \right] x_i = 0$$

$$\frac{\partial E}{\partial w_1} = -2 \sum_{i=1}^{N} \left[y_i - \left(w_0 + w_1 x_i \right) \right] = 0$$

$$\begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix}$$

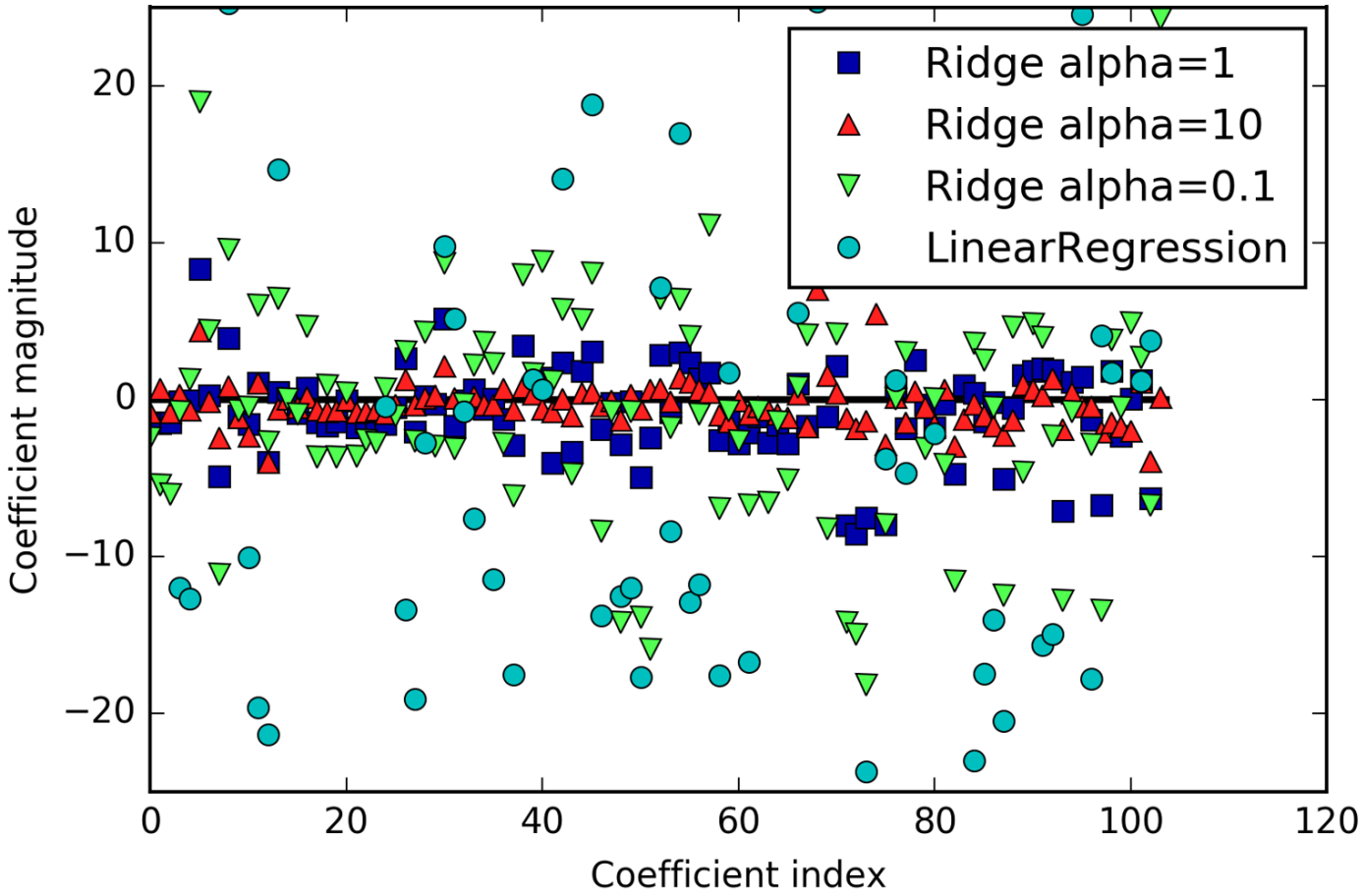
Regresión Ridge

Introducimos una restricción adicional L_2-norm

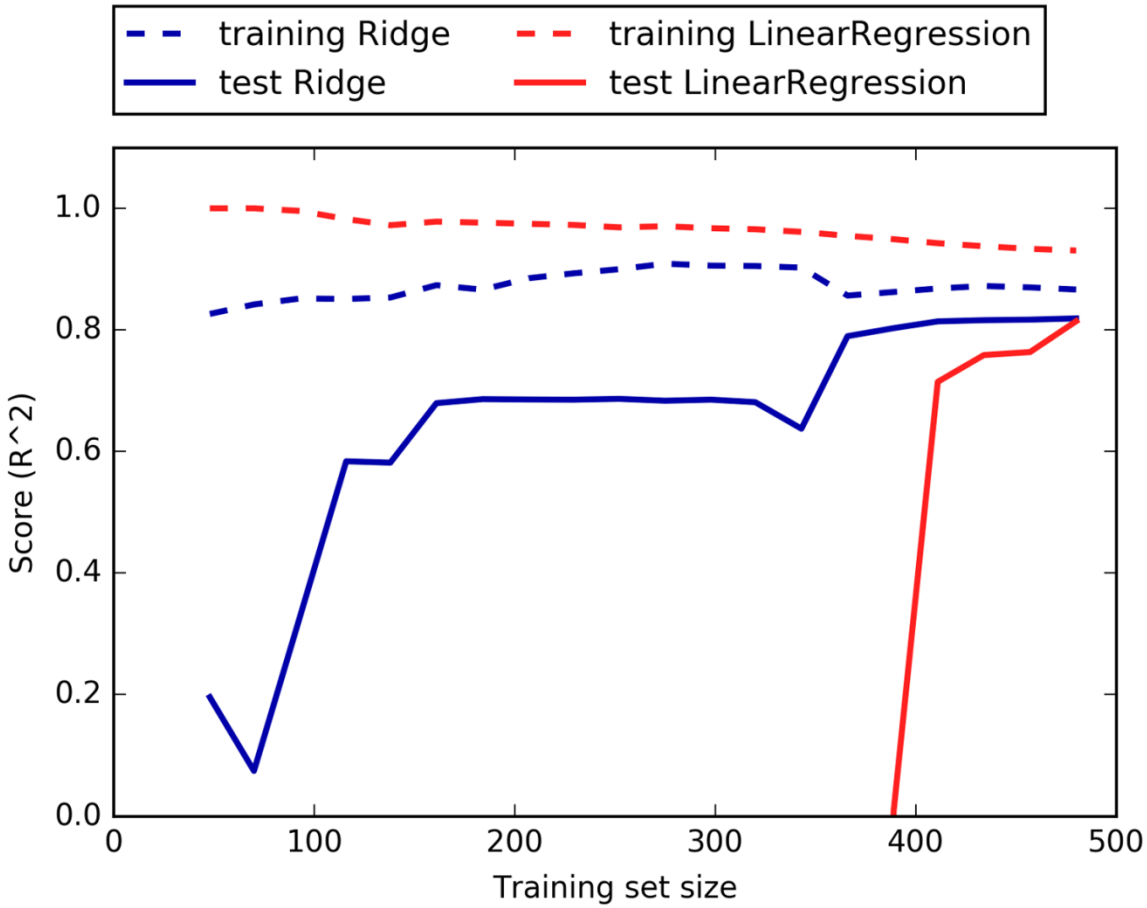
$$\hat{\beta}^{\mathbf{ridge}} = \mathbf{arg}\min_{\beta} \sum_{\mathbf{i=1}} (\mathbf{y_i} - (\beta_\mathbf{0} + \beta^\mathbf{T}\mathbf{x_i}))^\mathbf{2} + \lambda \|\beta\|_\mathbf{2}^\mathbf{2}$$

Penalidad por complejidad

Regresión Ridge



Regresión Ridge



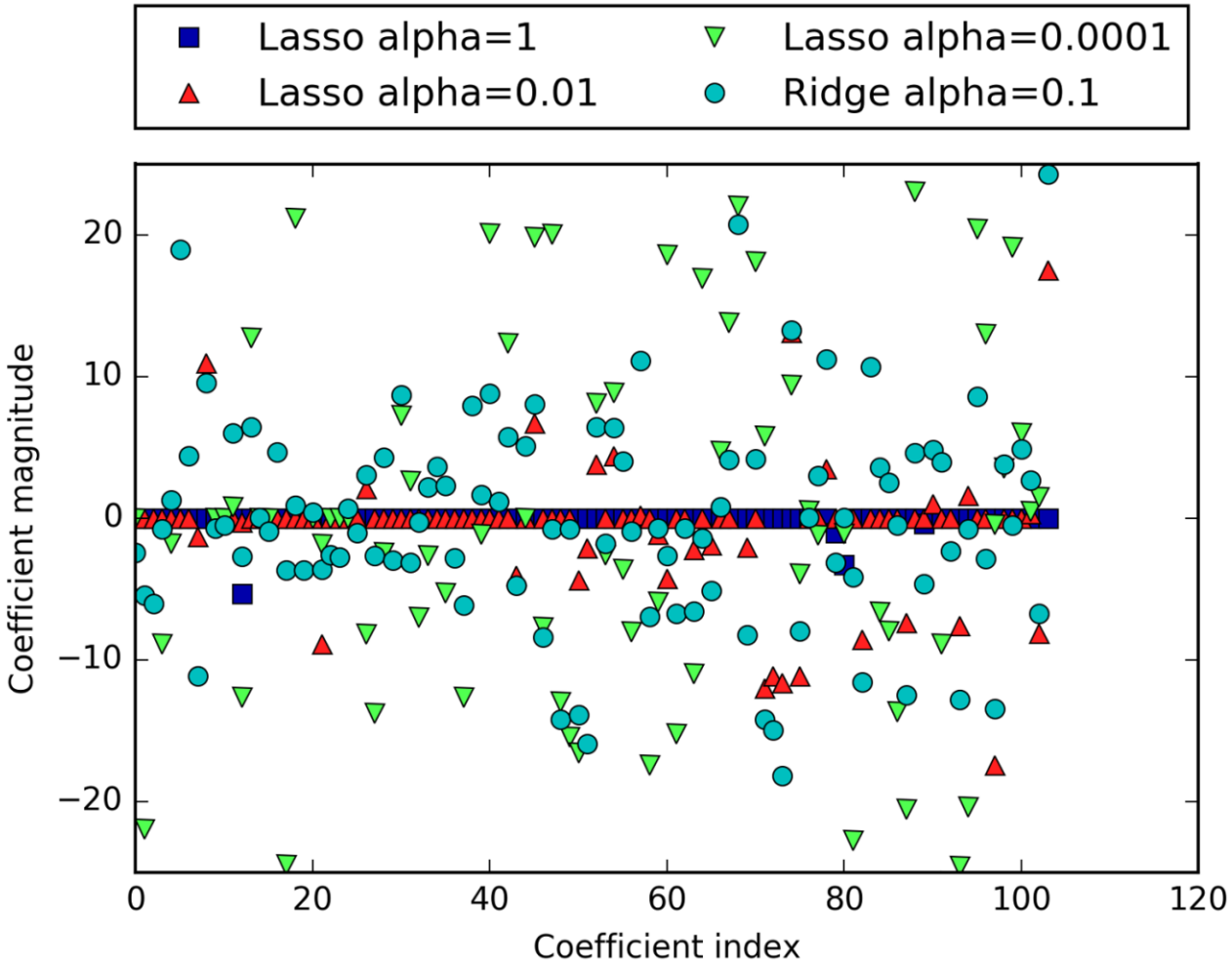
Regresión Lasso

Introducimos una restricción adicional L_1-norm

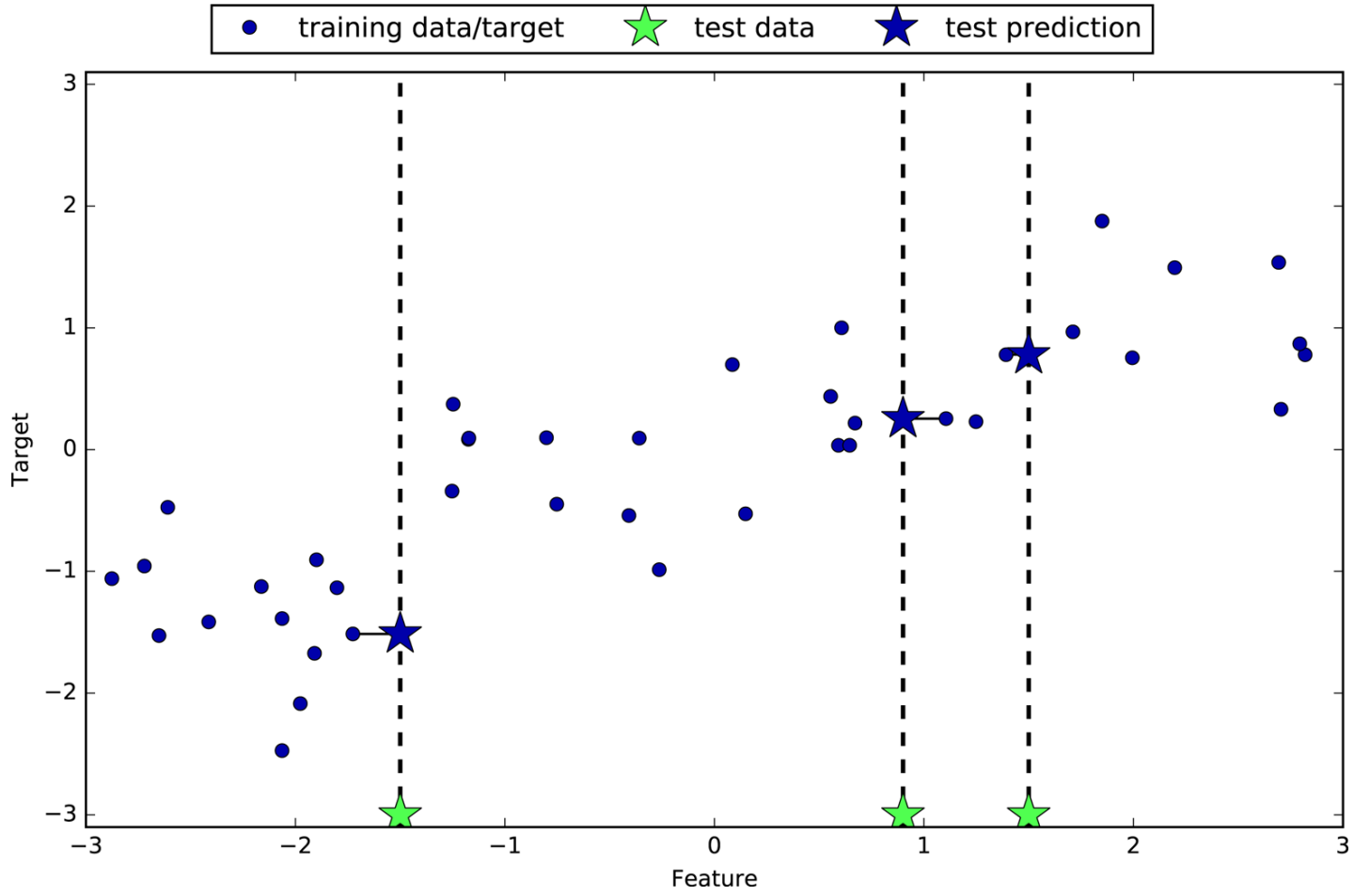
$$\hat{\beta}^{\mathbf{lasso}} = \mathbf{arg}\min_{\beta} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - (\beta_\mathbf{0} + \beta^\mathbf{T}\mathbf{x_i}))^\mathbf{2} + \lambda \|\beta\|_\mathbf{1}$$

Penalidad por complejidad

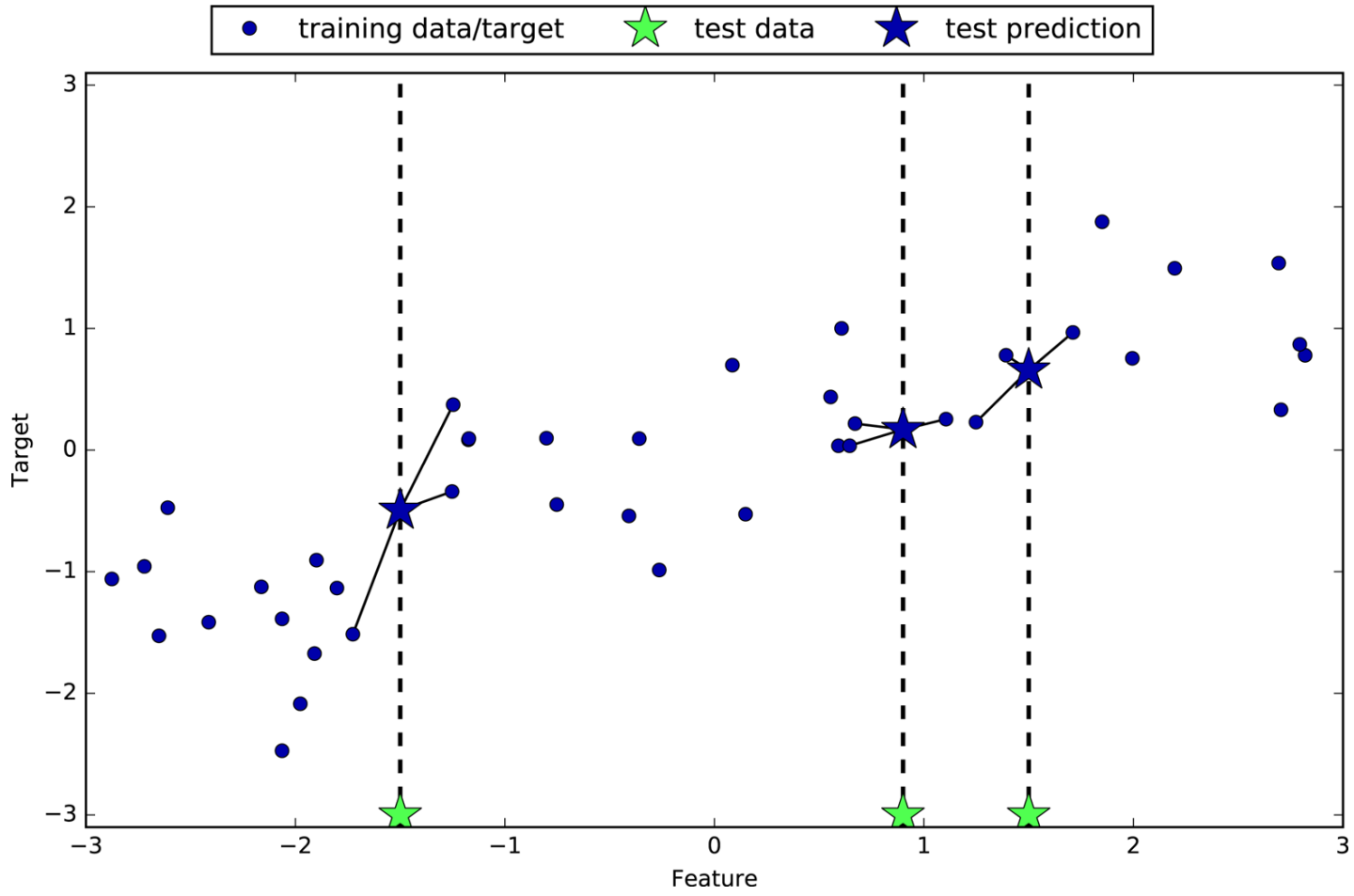
Regresión Lasso



Regresión *k*-neighbors



Regresión *k*-neighbors



Modelización paramétrica

Los modelos que dependen de un número finito de parámetros se denominan paramétricos

componente sistémica

componente aleatoria

$$y_i = w_0 + w_1 x_{i1} + \ldots + w_n x_{in} + \varepsilon_i \quad \text{con} \quad \varepsilon_i \sim N\ 0, \sigma^2$$

donde $w_0, w_1, \ldots, w_n, \sigma^2$ son los parámetros del modelo

Modelización no-paramétrica

$$y_i = r\ x_i\ + \varepsilon_i$$

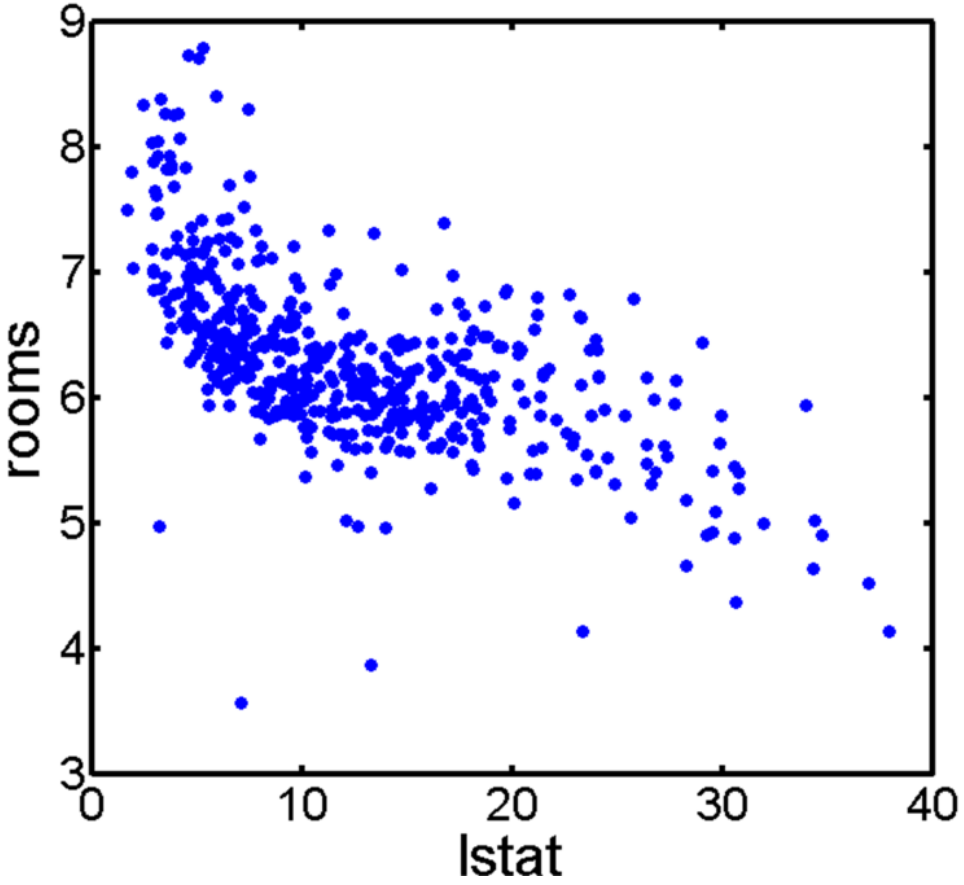
continua y suave
no se especifica su forma funcional

variable aleatoria
$E(\varepsilon_i)=0$
$V(\varepsilon_i)=\sigma^2$

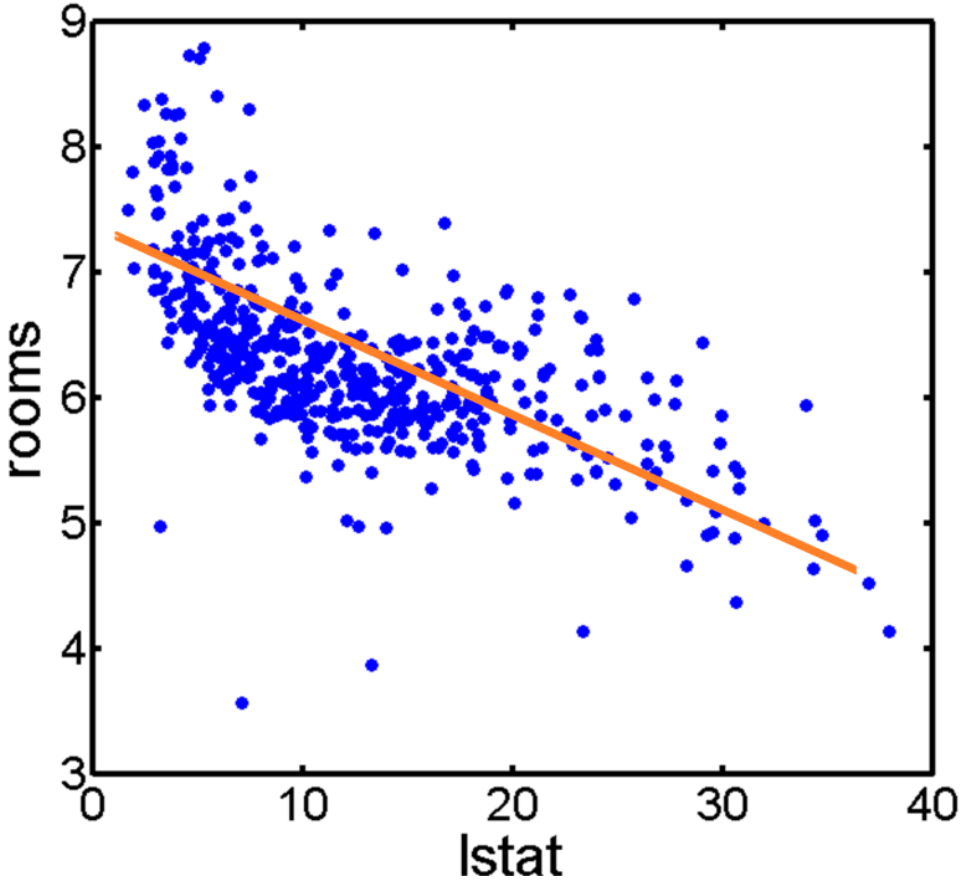
Estimador de la función de regresión $\hat{r}\ x$

Estimador de la varianza del error $\hat{\sigma}^2$

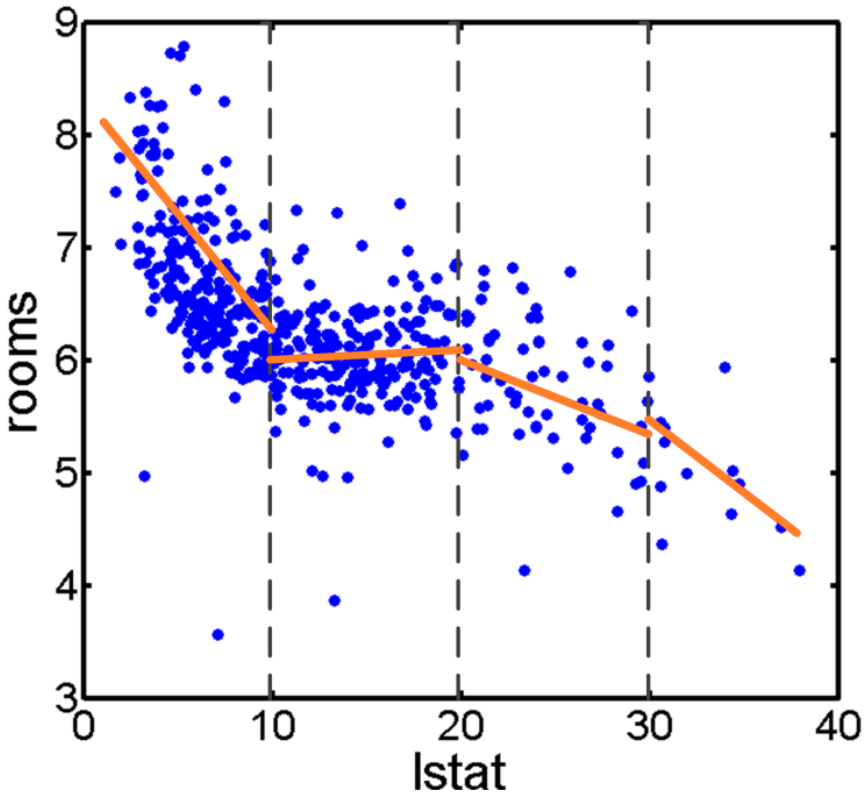
Ajuste local



Ajuste local



Ajuste local por tramos



una primera idea….

dividir el rango de la variable explicativa en intervalos y aproximar linealmente en cada tramo

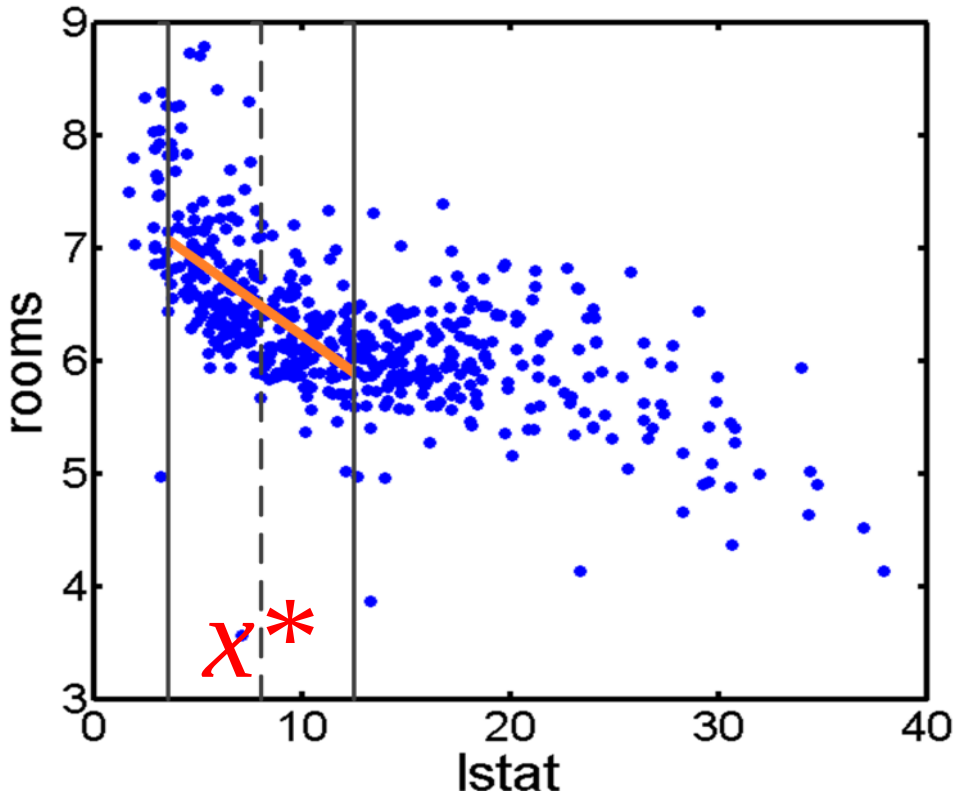
desventajas

función de discontinua

estimación sesgada

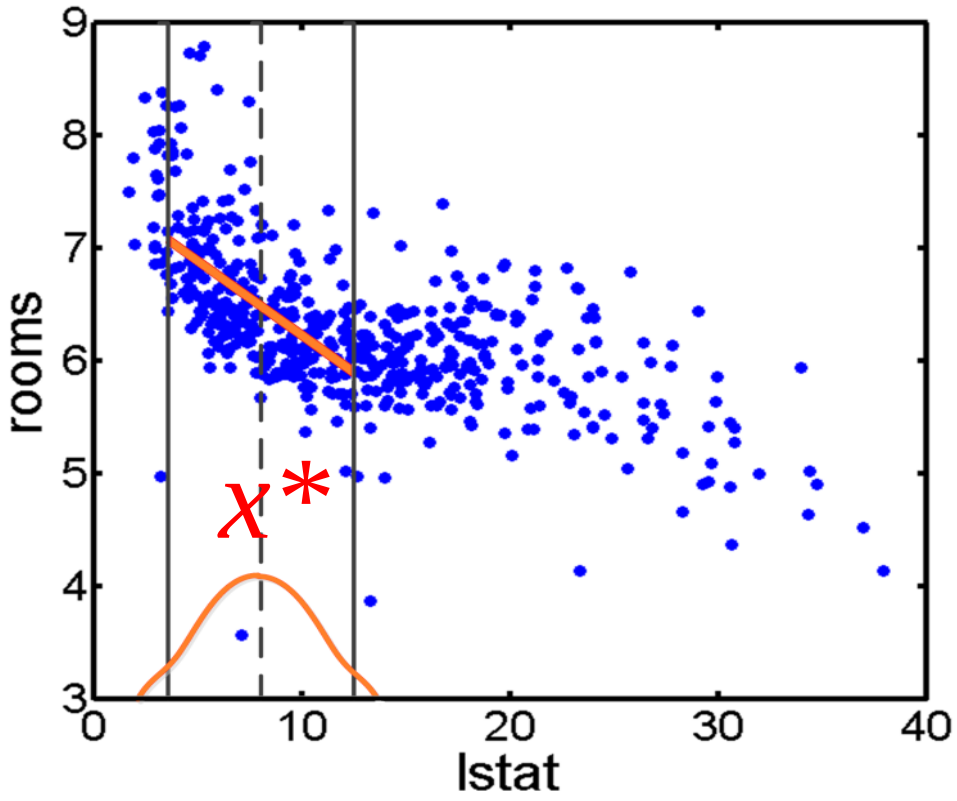
Ajuste local

Núcleo lineal uniforme



Ajuste local

Kernel Gaussiano

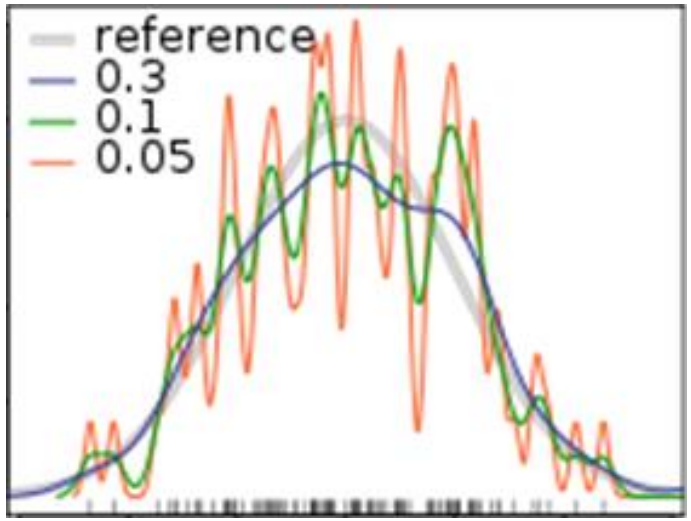


Función núcleo (kernel)

$$w_i = w\ \ x^*, x_i\ \ = \frac{\frac{x_i - x^*}{h}}{\sum_{j=1}^{n} \frac{x_j - x^*}{h}}$$

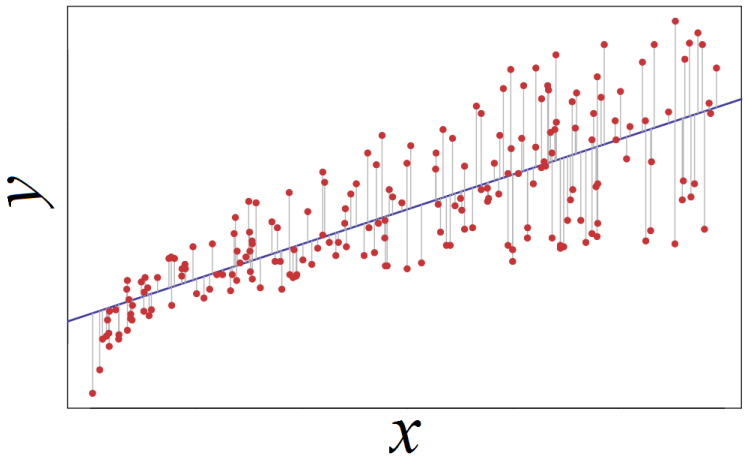
h<< solo las observaciones cercanas tendrán peso, mayor flexibilidad, problemas de sobreajuste.

h>> observaciones alejadas también tendrán peso, falta de ajuste, mejor generalización.

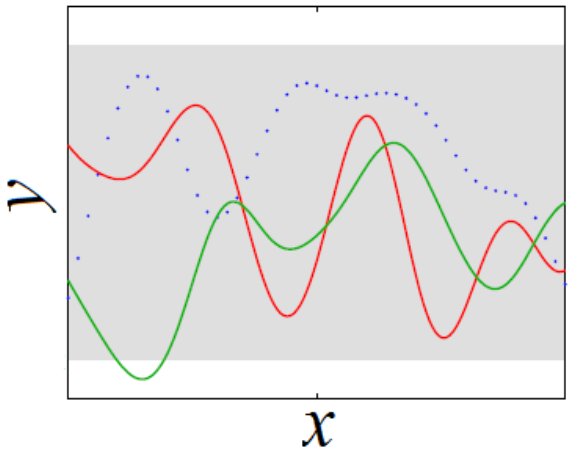


Dos enfoques

Restringir la clase de funciones que consideramos que se van a ajustar mejor a los datos.



Definir múltiples relaciones funcionales y especificar una probabilidad a priori para cada función de aproximación posible.

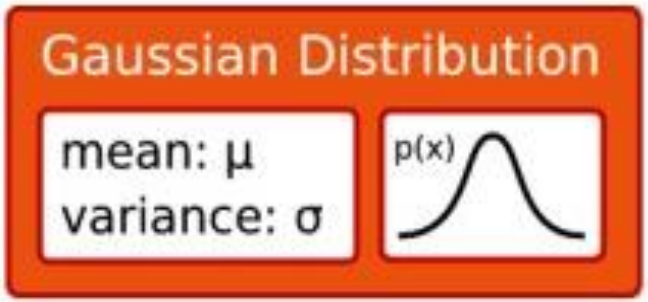


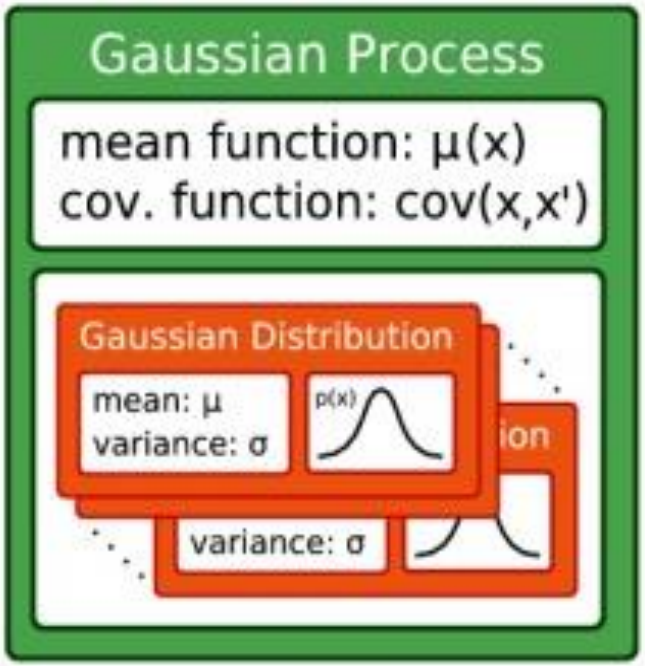
Dudas con el método?

Aprender a mapear valores de $y = f\ x$ de observaciones $x_i, y_i\ _{i=1}^{N}$

- **Ajuste del modelo:**
  - como ajusto los parámetros?
  - flexibilidad vs generalización
- **Selección del modelo:**
  - que modelo uso?
  - es el modelo correcto?
- **Interpretación:**
  - cual es la precisión de las estimaciones?
  - puedo confiar en ellos incluso si:
    - no estoy seguro de los parámetros
    - no estoy seguro de la estructura del modelo
    - no estoy seguro del nivel de ruido en los datos

Proceso Gaussiano





$$\mathbf{f} = f_1, \ldots, f_n \,^T \sim \mathcal{N}\ \mu, \Sigma \longrightarrow f\ x = \mathcal{GP}\ m\ x\ , k\ x, x'$$

Definición de GP

Dado un conjunto $\mathcal{H}$ de todas las posibles funciones que mapean un vector $\mathbf{x}=\{x_1, x_2, x_3,\ldots,x_m\}$.

Ejemplos de evaluar una función particular h_0 podrían ser:

$$h_0(x_1)=5, h_0(x_2)=2.3, h_0(x_3)=\pi,\ldots, h_0(x_m)=-7$$

Como el dominio de $h_0 \in \mathcal{H}$ tiene solo m elementos, podemos representarlo como un vector

$$\vec{h_0} = h_0(x_1), h_0(x_2), h_0(x_3),\ldots, h_0(x_m)^T$$

Especificando una distribución de probabilidad para cada $h \in \mathcal{H}$, le asociamos una probabilidad a priori a cada función.

Definición de GP

Si en particular especificamos $\vec{h} \sim \mathcal{N}\ \mu, \sigma^2 I$ esto implica una distribución sobre funciones

$$p\ \mathcal{H}\ = \prod_{i=1}^{m} \frac{2}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}\ h\ x_i\ - \mu_i\ ^2 \right)$$

…que estamos haciendo?

Asociamos densidades de probabilidades sobre funciones de dominios infinitos, usando una distribución Gaussiana multivariable en un número finito de puntos de entrada $x_1, \ldots, x_m$

Definición de GP

Un proceso Gaussiano es una colección de variables aleatorias, tal que cualquier subconjunto finito de ellas tiene una distribución Gaussiana multivariable.

$$h\begin{pmatrix} x_1 \\ \vdots \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m\ x_1 \\ \vdots \\ m\ x_m \end{bmatrix}, \begin{bmatrix} k\ x_1, x_1 & \cdots & k\ x_1, x_m \\ \vdots & \ddots & \vdots \\ k\ x_m, x_1 & \cdots & k\ x_m, x_m \end{bmatrix}\right)$$

$$h\ x\ \sim \mathcal{GP}\ m\ \cdot\ , k\ \cdot,\cdot$$

Procesos Gaussianos

Para $m()$ podemos usar cualquier función real, pero para $k(;)$ debe cumplirse que para cualquier conjunto $x_1,\ldots, x_m$, la matriz resultante

$$K = \begin{bmatrix} k\ x_1, x_1 & \cdots & k\ x_1, x_m \\ \vdots & \ddots & \vdots \\ k\ x_m, x_1 & \cdots & k\ x_m, x_m \end{bmatrix}$$

sea válida para una distribución Gaussiana multivariable, lo cual son las mismas condiciones que para los *kernels.*

Covarianza exponencial

$$h(\cdot) \sim \mathcal{GP}(0, k(\cdot,\cdot)) \qquad k_{SE}(x, x') = \exp\left(-\frac{1}{2\tau^2}\|x - x'\|^2\right)$$

$h(x)$ y $h(x')$ tienen alta covarianza cuando x y x' están cercanos en el espacio

$$\|x - x'\| \approx 0 \quad \text{y} \quad \exp\left(-\frac{1}{2\tau^2}\|x - x'\|^2\right) \approx 1$$

$h(x)$ y $h(x')$ tienen baja covarianza cuando x y x' están alejados en el espacio

$$\|x - x'\| \gg 0 \quad \text{y} \quad \exp\left(-\frac{1}{2\tau^2}\|x - x'\|^2\right) \approx 0$$

Regresión con GPs

Sea $S = \{x_i, y_i\}_{i=1}^{m}$ el conjunto de entrenamiento

Dado un modelo de regresión no paramétrico

$$y_i = h(x_i) + \varepsilon_i$$

Donde son ϵ_i son variables de ruido con distribuciones $\mathcal{N}(0; \sigma^2)$

Regresión con GPs

Sea $T = \left\{ x_i^* \right\}_{i=1}^{m}$ un conjunto de datos de pruebas

Queremos hacer predicciones y^* en los puntos x^*

Estimación

prior condicionada

La suma de variables Gaussianas también es una Gaussiana, y por lo tanto

$$p\left(\begin{bmatrix}\vec{y}\\ \vec{y}^*\end{bmatrix} \mid X, X^*\right) = p\left(\begin{bmatrix}\vec{h}\\ \vec{h}^*\end{bmatrix}\right) + p\left(\begin{bmatrix}\vec{\varepsilon}\\ \vec{\varepsilon}^*\end{bmatrix}\right) \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} k\ X, X\ + \sigma^2 I & k\ X, x^* \\ k\ x^*, X & k\ x^*, x^*\ + \sigma^2 I \end{bmatrix}\right)$$

Usando las reglas de condicionamiento Gaussianas, se sigue que:

$$p\ \vec{y}^* \mid \vec{y}, X, x^*\ = \mathcal{N}\ \mu^*, \Sigma^*$$

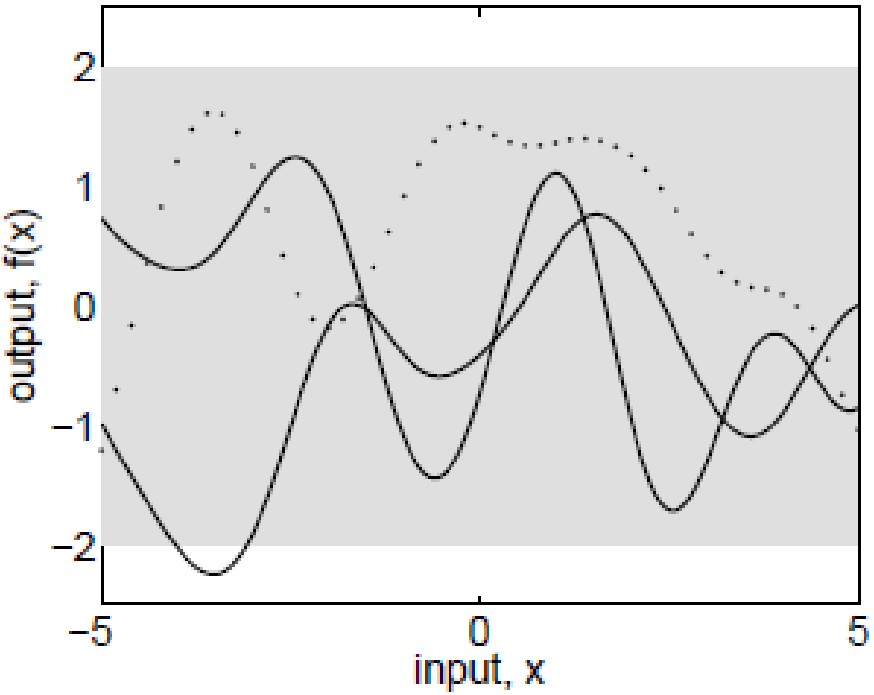
posterior

donde

$$\mu^* = k\ x^*, X\ \ k\ X, X\ + \sigma^2 I\ ^{-1} \vec{y}$$

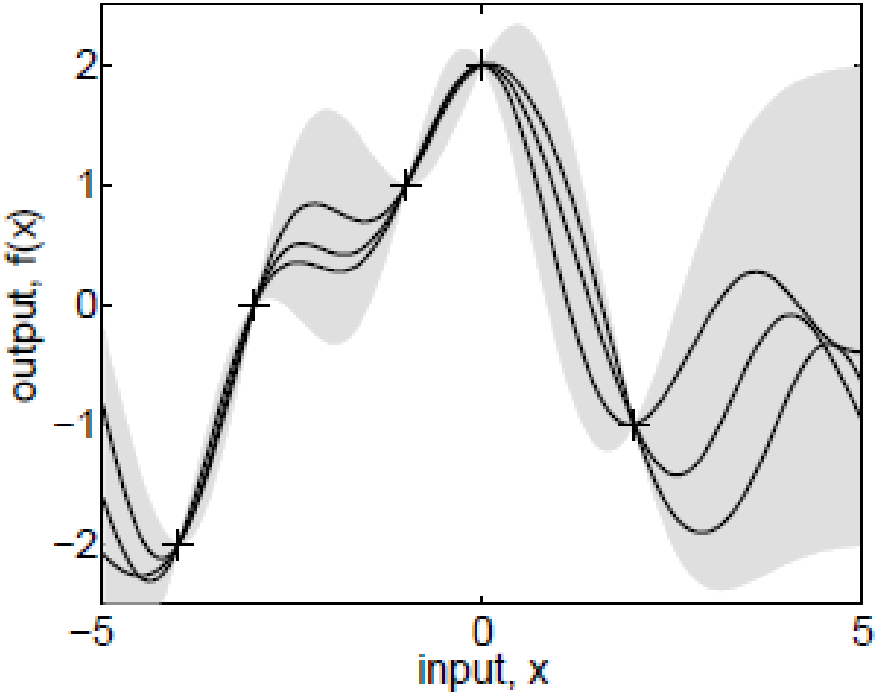
$$\Sigma^* = \underbrace{k\ x^*, x^*\ + \sigma^2 I}_{\text{varianza a priori}} \underbrace{- k\ x^*, X\ \ k\ X, X\ + \sigma^2 I\ ^{-1} k\ X, x^*}_{\text{datos de testeo}}$$

Prior y posterior



(a), prior

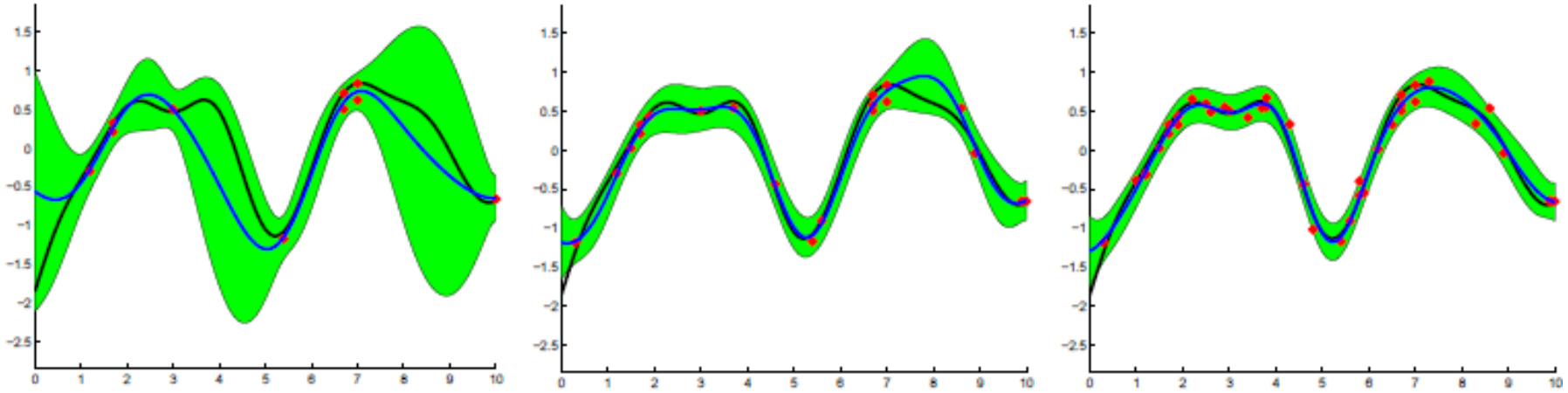
distribución a priori con distintas parametrizaciones iniciales



(b), posterior

prior condicionada a cinco observaciones

Región de confianza



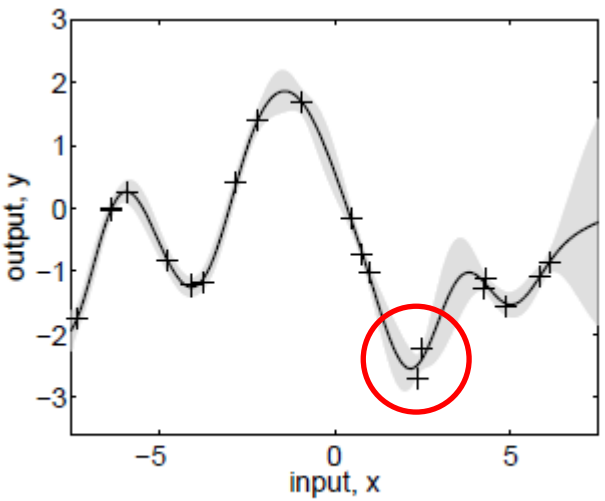
Regresión con media cero y función de covarianza $k_{SE}(\cdot,\cdot)$ con m = 10, m = 20, y m = 40 ejemplos de entrenamiento

Los hiperparámetros

$$k\ \ x, x' \ = \sigma_s^2 \exp\left(-\frac{1}{2}\ \ x - x' \ ^T C\ \ x - x' \ \right)$$

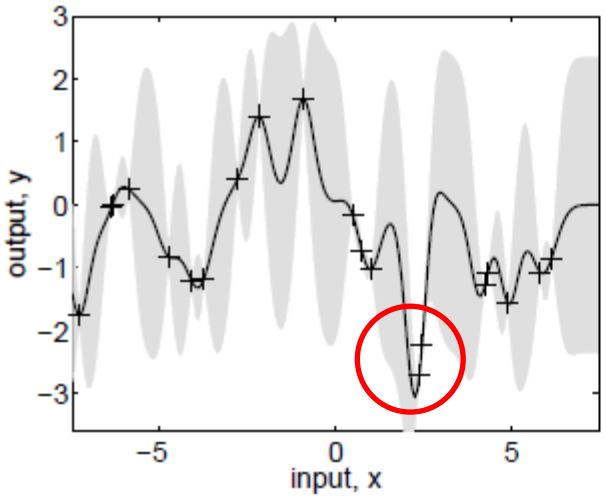
- σ_s es la varianza o factor de escala del ruido
- C es una matriz simétrica que puede tener distintas parametrizaciones
- $C = \mathrm{diag}(\ell^{-2})$ con $\ell = (\ell_1,\ldots,\ell_D)$ Automatic Relevance Determination (ARD)
- Particularmente ℓ se conoce como factor de escala característico

Los hiperparámetros



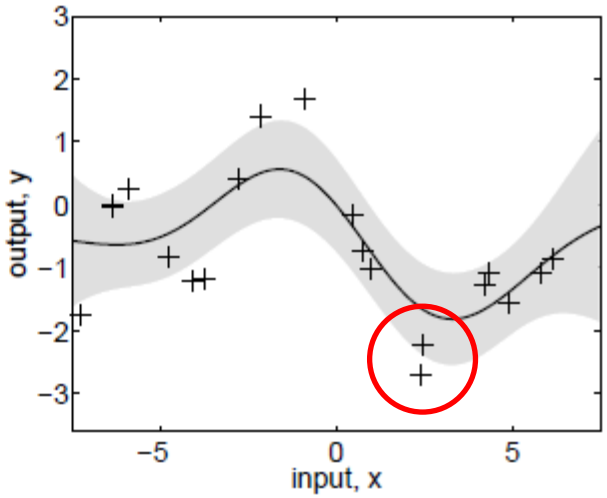
(a), $\ell = 1$

Función relativamente suave con algo de ruido



(b), $\ell = 0.3$

Función más flexible con muy poco ruido



(c), $\ell = 3$

Función muy suave con mucho ruido

Máxima verosimilitud

La función de verosimilitud es una función de lo parámetros del modelo en relación a los datos observados

$$p\ \theta \mid X, y$$

Podemos tomar como estimación de los parámetros estudiados el valor que haga máxima la probabilidad de obtener la muestra observada

$$p\ \theta \mid X, y\ \propto p\ \ y \mid X, \theta$$

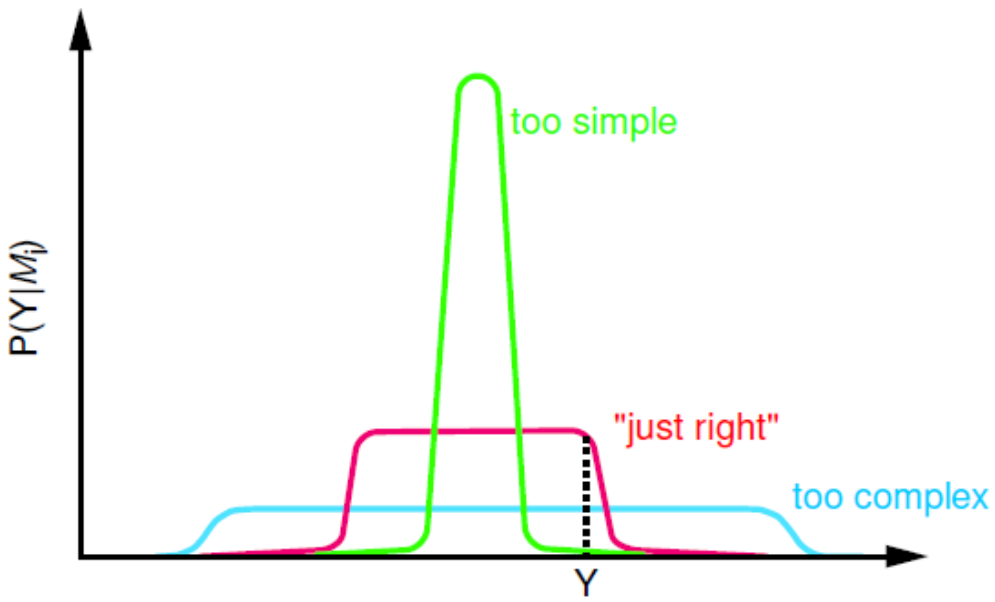
Dada una función de covarianza exponencial cuadráticas seleccionamos σ_n tal que se optimice el logaritmo

$$\mathcal{L} = \log \mathrm{p}(\mathbf{y}|\mathbf{X}, \theta) = \underbrace{-\frac{1}{2}\mathbf{y}^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y}}_{\text{data-fit}} - \underbrace{\frac{1}{2}\log|\mathbf{K} + \sigma_n^2\mathbf{I}|}_{\text{complexity}} - \underbrace{\frac{\mathrm{N}}{2}\log 2\pi}_{\text{normaliz.}}$$

Navaja de Ockham

De un conjunto de variables explicativas que forman parte del modelo, debe seleccionarse la combinación más reducida y simple posible, teniendo en cuenta la varianza residual

Función de máxima verosimilitud



todos los conjuntos de entrenamiento posibles