Regresión por Procesos Gaussianos

Anthony Huertas

Universidad Nacional de Ingeniería

Técnica de Machine Learning, 2016

Índice

El trabajo va enfocado directamente a la mejora de la técnica de Regresión Lineal mediante *Procesos Gaussianos*, estableciéndonos análisis estadísticos en un espacio de funciones, y otorgándonos un alto porcentaje de certeza sobre un conjunto pequeño de posibles resultados facilitando la toma de decisiones en cuanto a un valor de predicción.

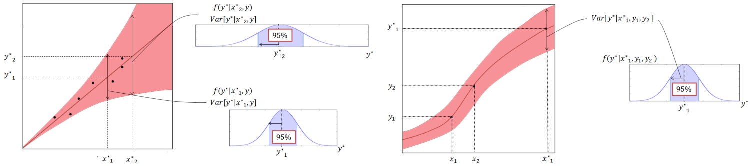


Figure: Regresion Lineal vs Regresión por Procesos Gaussianos

Condicionalidad - Independencia

-Probabilidad Condicional:
Dado dos eventos A y B, la probabilidad de que ocurra A, luego de que B haya ocurrido, se define como

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ dado } P(B) > 0.$$

-Independencia:
Dado dos eventos A y B, se dice que son independientes si

$$P(A|B) = P(A)$$

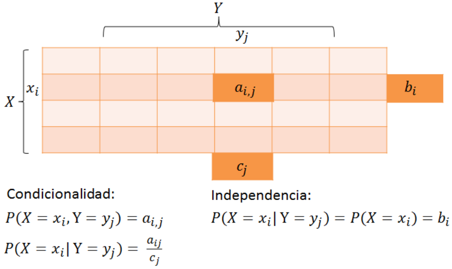


Figure: Condicionalidad, Independencia

Teorema de Bayes

-Teorema de la Probabilidad Total:
Sean $A_1, \ldots, A_n$ sucesos de Ω con $P(A_i) > 0$ para todo $i = 1, \ldots, n$ tales que

- $A_i \cap A_j = \emptyset (i \neq j)$.
- $\Omega = \bigcup_{i=1}^{n} A_i$.

entonces,

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

-Teorema Bayesiano:
Si $P(B) > 0$, con las condiciones del Teorema anterior, entonces

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}.$$

Regla de Bayes

Si X, Y R.V. continuas,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{+\infty} f_{X|Y}(x|y')f_Y(y')dy'}$$

Definición

Función de densidad de probabilidad:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Notación:

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \text{donde } E[X] = \mu,\ Var[X] = \sigma^2.$$

Intervalos de confianza:

$$P(\mu - \sigma < x \leq \mu + \sigma) = \int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(u-\mu)^2\right) du = 0.682 = 68.2\%$$

Descripciones:

- Existe 68.2% de certeza entre 1 desviación estándar de la media.
- Existe 95% de certeza entre 2 desviaciones estándar de la media.
- Existe 99.7% de certeza entre 3 desviaciones estándar de la media.

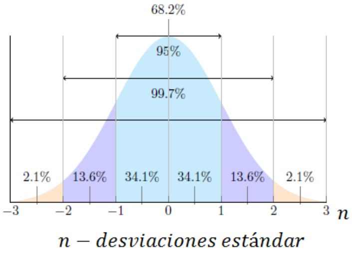


Figure: Porcentaje de certeza entre desviaciones estándar de la media.

Distribución Gaussiana Multivariable

Sea $X = X_1, \ldots, X_n$.
- Independencia:

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) = \frac{1}{(2\pi)^{1/n}\prod_{i=1}^{n}\sigma_i} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T(\mathbf{x}-\mu)\right)$$

-Generalización:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{1/n}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

Σ es la matriz de covarianza, y μ es un vector con elementos μ_i.

Marginalización

Sea una distribución Gaussiana multivariable, con representación

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{21}^T \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

entonces,

$$\begin{aligned} \mathbf{x}_1 &\sim \mathcal{N}(\mu_1, A), \\ \mathbf{x}_2 &\sim \mathcal{N}(\mu_2, B), \\ \mathbf{x}_1|\mathbf{x}_2 &\sim \mathcal{N}(\mu_1 + \Sigma_{21}^T\Sigma_{22}^{-1}(\mathbf{y} - \mu_2), \Sigma_{11} - \Sigma_{21}^T\Sigma_{22}^{-1}\Sigma_{21}), \\ \mathbf{x}_2|\mathbf{x}_1 &\sim \mathcal{N}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{y} - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}^T). \end{aligned}$$

Principio de Verosimilitud

De obtenerse resultados $y_1, \ldots, y_n$, entonces el principio de verosimilitud no permite intuir que estos resultados se rigen bajo un parámetro.

-**Función de verosimilitud**:
Siendo una muestra aleatoria , se define la función de verosimilitud como

$$\mathcal{L}(\theta|\mathbf{y}) = f(\mathbf{y}|\theta), \qquad \text{donde } \mathbf{y} = y_1, \ldots, y_n. \ (y_i \in Y_i)$$

De ser los $Y_i's$ independientes, entonces

$$\mathcal{L}(\theta|\mathbf{y}) \quad = \quad \prod_{i=1}^{n} f(\mathbf{y}_i|\theta).$$

Estimadores de Máxima Verosimilitud (E.V.M.)

El *estimador de máxima verosimilitud* se determina como:

$$\tilde{\theta} = \max_{\theta}\{\mathcal{L}(\theta|\mathbf{y})\}. \tag{1}$$

Debido a que la función ln : $\mathbb{R}^+ \to \mathbb{R}$ es creciente, en muchas ocasiones la estimación de máxima verosimilitud corresponde a determinar

$$\tilde{\theta} = \max_{\theta}\{l(\theta|\mathbf{y})\} \tag{2}$$

donde $l(\theta|\mathbf{y}) = \ln(\mathcal{L}(\theta|\mathbf{y}))$.

Estimador de Máximo a Posteriori (M.A.P.)

De existir la correspondiente distribución para el parámetro θ, entonces el análisis bayesiano implica que la *regla de Bayes* tomaría la siguiente forma

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})} = \frac{\mathcal{L}(\theta|\mathbf{y})\pi(\theta)}{\int_{-\infty}^{+\infty} f(\mathbf{y}|\theta')\pi(\theta')d\theta'}.$$

- Distribución a priori: $\pi(\theta)$.
- Distribución a posteriori: $\pi(\theta|\mathbf{y})$.

Coeficientes de Regresión, Ruidos Gaussianos

Suponiendo *n* datos de entrenamiento que diseñen el modelo.

$$Y = h_\theta(\mathbf{x}) + \epsilon, \qquad \text{con } h_\theta(\mathbf{x}) = \mathbf{x}^T\theta, \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_k \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_k \end{bmatrix}$$

Parámetro: θ, cuyos elementos son los *coeficientes de regresión*.
$\epsilon \sim \mathcal{N}(0, \sigma^2)$ (ruidos Gaussiano), independientes.

Coeficientes de Regresión, Ruidos Gaussianos

$$X = \begin{bmatrix} 1 & \dots & 1 \\ x_{11} & & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{k1} & \dots & x_{kn} \end{bmatrix}, \quad \mathbf{y} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{3}$$

entonces tenemos una representación matricial del modelo de la siguiente forma

$$\mathbf{y} = X^T\theta + \varepsilon, \text{ con } \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_{n\times n}).$$

A causa de la independencía de ϵ_i,

$$f(\varepsilon) = -\frac{1}{(2\pi)^{1/n}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\varepsilon^T\varepsilon\right)$$

Como $\varepsilon = \mathbf{y} - X^T\theta$:

$$\mathcal{L}(\theta|X, \mathbf{y}) = f(\mathbf{y}|X, \theta) = -\frac{1}{(2\pi)^{1/n}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - X^T\theta)^T(\mathbf{y} - X^T\theta)\right).$$

Estimación por Maximá Verosimilitud

Por el *principio de verosimilitud*, estimamos el parámetro θ establecido por la *máxima verosimilitud*

$$\tilde{\theta} = \arg\max_{\theta}\{l(\theta|X, \mathbf{y})\}, \quad \text{donde } l(\theta|X, \mathbf{y}) = \ln(\mathcal{L}(\theta|X, \mathbf{y})).$$

Esto nos conduce a determinar el parámetro que maximice la siguiente función

$$l(\theta|X, \mathbf{y}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - X^T\theta)^T(\mathbf{y} - X^T\theta).$$

Función Costo

$$F(\theta) = \frac{1}{n}(\mathbf{y} - X^T\theta)^T(\mathbf{y} - X^T\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k}\theta x_j\right). \quad (4)$$

Siendo esta función costo una función convexa, se puede aplicar la técnica de Descenso de Gradiente.

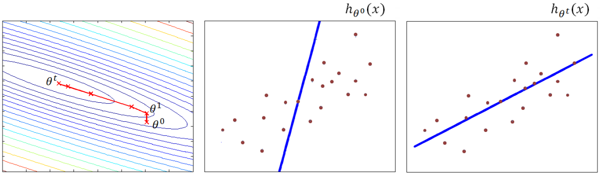


Figure: Descenso de Gradiente para técnicas de regresión lineal simple.

Análisis Bayesiano

$$\mathcal{L}(\theta|X,\mathbf{y}) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{\|\mathbf{y}-X^T\theta\|^2}{2\sigma^2}\right) = \mathcal{N}(X^T\theta, \sigma^2\mathbb{I}_{n\times n}).$$

Regla de Bayes:

$$\pi(\theta|\mathbf{y},X) = \frac{\mathcal{L}(\theta|X,\mathbf{y})\pi(\theta)}{\int_{-\infty}^{+\infty} f(\mathbf{y}|X,\theta')\pi(\theta')d\theta'}$$

Marginalización condicionada:

$$f(\mathbf{y}|X) = \int_{-\infty}^{+\infty} f(\mathbf{y}|X,\theta')\pi(\theta')d\theta'$$

es independiente con el parámetro θ.

Análisis Bayesiano

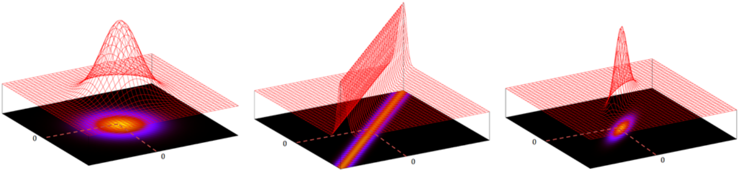
Asumiendo $\pi(\theta) \sim \mathcal{N}(\mathbf{0}, \widehat{\Sigma})$:

$$\begin{aligned}
\pi(\theta|X,\mathbf{y}) &\propto \mathcal{L}(\theta|X,\mathbf{y})\pi(\theta) \\
\pi(\theta|X,\mathbf{y}) &\propto \exp\left(-\frac{\|\mathbf{y}-X^T\theta\|^2}{2\sigma^2}\right)\exp\left(-\frac{\theta^T\widehat{\Sigma}^{-1}\theta}{2}\right) \\
\pi(\theta|X,\mathbf{y}) &\propto \exp\left(\frac{1}{2}(\theta-\overline{\theta})^T A^{-1}(\theta-\overline{\theta})\right)
\end{aligned}$$

por tanto

$$\pi(\theta|\mathbf{y},X) = \mathcal{N}(\overline{\theta}, A^{-1}) \quad \text{donde} \begin{cases} A = \sigma^{-2}XX^T + \widehat{\Sigma}^{-1} \\ \overline{\theta} = \sigma^{-2}A^{-1}X\mathbf{y}. \end{cases}$$

Análisis Bayesiano



$\pi(\theta) \sim N(0, \mathbb{I}_{2\times 2})$ (a) $f(y|X,\theta) \sim N(X^T\theta, \sigma^2\mathbb{I}_{2\times 2})$ (b) $\pi(\theta|X,y) \sim N(\theta, A^{-1})$ (c)

Figure: Sea el modelo de regresión $y = \theta_0 + \theta_1 x + \epsilon$, ($a$) representa la distribución a priori $\pi(\theta) \sim \mathcal{N}(0, \mathbb{I}_{2\times 2})$; ($b$) la densidad de probabilidad de las observaciones dado los parámetros $f(y|X,\theta) \sim \mathcal{N}(X^T\theta, \sigma^2\mathbb{I}_{n\times n})$; (c) representa la distribución a posteriori $\pi(\theta|\mathbf{y}, X) \sim \mathcal{N}(\bar{\theta}, A^{-1})$.

Procesos Gaussianos

Un proceso gaussiano (**GP**), cuyo técnica se basa en distribuciones sobre funciones, representando datos como una muestra de una distribución multivariada. La representación es

$$y = h(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{5}$$

Notación:

$$h(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{6}$$

donde $m(x)$ es la función promedio, y k una función de covarianza sobre funciones.

Distribución Predictiva

-Regresión Lineal:

$$\mathbf{x}^* \text{ (Vector de características en prueba),}$$
$$y^* = h_\theta(\mathbf{x}^*) \text{ (valor predicho).}$$

Luego:

$$f(y^*|\mathbf{x}^*, \mathbf{y}) = \int f(y^*|x^*, \theta')\pi(\theta'|X, \mathbf{y})d\theta'$$

Marginalización Condicionada:

$$f(y^*|\mathbf{x}^*, \mathbf{y}) = \mathcal{N}\left(\sigma^{-2}\mathbf{x}^{*T}A^{-1}X\mathbf{y}, \mathbf{x}^{*T}A^{-1}\mathbf{x}^*\right).$$

Análisis:

$$Var[Y^*|\mathbf{x}^*, \mathbf{y}] \to \infty \quad \text{si } \|\mathbf{x}^*\| \to \infty.$$

Extensión de la Dimensionalidad

El modelo a tratar sería el siguiente

$$y = h_\theta(\mathbf{x}) + \epsilon, \text{ donde } h_\theta(\mathbf{x}) = \phi(\mathbf{x})^T\theta$$
$$\text{y } \ \phi : \mathbb{R}^k \to \mathbb{R}^K (k < K)$$

con $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

$$X = \begin{array}{c} \begin{array}{ccc} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{array} \\ \begin{bmatrix} 1 & \dots & 1 \\ x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{k1} & \dots & x_{kn} \end{bmatrix} \end{array} \xrightarrow{\phi} \begin{array}{c} \begin{array}{ccc} \phi(\mathbf{x_1}) & \cdots & \phi(\mathbf{x_n}) \end{array} \\ \begin{bmatrix} \phi_1(\mathbf{x}_1) & \dots & \phi_1(\mathbf{x_n}) \\ \phi_2(\mathbf{x}_1) & \dots & \phi_2(\mathbf{x_n}) \\ \vdots & \ddots & \vdots \\ \phi_K(\mathbf{x}_1) & \dots & \phi_K(\mathbf{x_n}) \end{bmatrix} \end{array} = \Phi(X).$$

Extensión de la Dimensionalidad

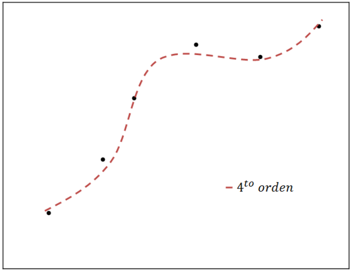


Figure: Ejemplo: Datos unidimensionales, siendo x el valor de la caraterística y $\phi(x) = (1, x, x^2, x^3, x^4)$ entonces el modelo se representaría como $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \epsilon$, el cual se denomima modelo de regresión polinomial de 4to grado.

Extensión de la Dimensionalidad

Ahora,

$$\mathcal{L}(\theta|X,\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|\mathbf{y}-\Phi(X)^T\theta\|^2}{2\sigma^2}\right)$$

$$\pi(\theta|X,\mathbf{y}) = \mathcal{N}(\overline{\theta}, A^{-1}) \quad \text{donde} \begin{cases} A = \sigma^{-2}\Phi(X)\Phi(X)^T + \widehat{\Sigma}^{-1} \\ \overline{\theta} = \sigma^{-2}A^{-1}\Phi(X)\mathbf{y}. \end{cases}$$

Distribución predictiva:

$$y^*|\mathbf{x}^*,\mathbf{y} \quad \sim \quad \mathcal{N}\left(\sigma^{-2}\phi(\mathbf{x}^*)^T A^{-1}\Phi(X)\mathbf{y}, \phi(\mathbf{x}^*)^T A^{-1}\phi(\mathbf{x}^*)\right).$$

Reformulando,

$$\begin{aligned} y^*|\mathbf{x}^*,\mathbf{y} \quad \sim \quad \mathcal{N}\Big(&\phi(\mathbf{x}^*)^T\widehat{\Sigma}\Phi(X)(\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\mathbf{y}, \\ &\phi(\mathbf{x}^*)^T\widehat{\Sigma}\phi(\mathbf{x}^*) - \phi(\mathbf{x}^*)^T\widehat{\Sigma}\Phi(X)(\tilde{\Sigma}+\sigma^2\mathbb{I})^{-1}\Phi(X)^T\widehat{\Sigma}\phi(\mathbf{x}^*)\Big) \end{aligned} \tag{7}$$

donde $\tilde{\Sigma} = \Phi(X)^T\widehat{\Sigma}\Phi(X)$ se denomina Matriz Gram.

Función Kernel

$$\begin{aligned} \kappa(\cdot, \cdot) : \mathbb{R}^k \times \mathbb{R}^k &\rightarrow \mathbb{R} \\ (\mathbf{x}, \mathbf{x}') &\longmapsto \kappa(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \widehat{\Sigma} \phi(\mathbf{x}') \end{aligned} \tag{8}$$

la cual evaluada en los n datos de entrenamiento $\mathbf{x}_i$, en datos de prueba (supongamos m) o de forma conjunta, denotamos las siguiente represetanciones matriciales,

$$\tilde{\Sigma} = \begin{bmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \dots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_n, \mathbf{x}_1) & \dots & \kappa(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \tag{9}$$

$$\Sigma^{**} = \begin{bmatrix} \kappa(\mathbf{x}_1^*, \mathbf{x}_1^*) & \dots & \kappa(\mathbf{x}_1^*, \mathbf{x}_m^*) \\ \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_m^*, \mathbf{x}_1^*) & \dots & \kappa(\mathbf{x}_m^*, \mathbf{x}_m^*) \end{bmatrix} \quad \Sigma^{*} = \begin{bmatrix} \kappa(\mathbf{x}_1^*, \mathbf{x}_1) & \dots & \kappa(\mathbf{x}_1^*, \mathbf{x}_m) \\ \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_m^*, \mathbf{x}_1) & \dots & \kappa(\mathbf{x}_m^*, \mathbf{x}_m) \end{bmatrix} \tag{10}$$

Modelamiento

$$h_\theta \triangleq h_\theta(\mathbf{x}) \sim \mathcal{GP}\left(m(x), \kappa(\mathbf{x}, \mathbf{x}')\right) \tag{11}$$

$m(x)$ y $\kappa(\mathbf{x}, \mathbf{x}')$ se denominan *función media* y *función kernel*.

$$\begin{aligned} E[\mathbf{h}_\theta] &= E[\Phi(X)^T\theta] = \Phi(X)^T E[\theta] = \Phi(X)^T\mathbf{0} = \mathbf{0}. && (12) \\ Cov(\mathbf{h}_\theta) &= E[\Phi(X)^T\theta\theta^T\Phi(X)^T] = \Phi(X)^T E[\theta\theta^T]\Phi(X) \\ &= \Phi(X)^T\widehat{\Sigma}\Phi(X)^T = \tilde{\Sigma}. && (13) \end{aligned}$$

Luego,

$$\mathbf{h}_\theta \sim \mathcal{N}\left(\mathbf{0}, \tilde{\Sigma}\right). \tag{14}$$

Ajuste con Datos de prueba

$$\begin{bmatrix} \mathbf{h}_\theta \\ \mathbf{y}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \tilde{\Sigma} & {\Sigma^*}^T \\ \Sigma^* & \Sigma^{**} \end{bmatrix}\right) \tag{15}$$

Distribución Gaussiana condicionada:

$$y^*|X^*, \mathbf{h}_\theta \sim \mathcal{N}\left(\Sigma^*\tilde{\Sigma}^{-1}\mathbf{h}_\theta, \Sigma^{**} - \Sigma^*\tilde{\Sigma}^{-1}{\Sigma^*}^T\right) \tag{16}$$

Distribución Predictiva conjunta

$$E[\mathbf{y}] = E\left[\Phi(X)^T\theta + \varepsilon\right] = \Phi(X)^T E\left[\theta\theta^T\right] + E[\varepsilon] = \mathbf{0}. \tag{17}$$

$$\begin{aligned} Cov(\mathbf{y}) &= E\left[\mathbf{y}\mathbf{y}^T\right] = E\left[(\Phi(X)^T\theta + \varepsilon)(\Phi(X)^T\theta + \varepsilon)^T\right] \\ &= \Phi(X)^T E\left[\theta\theta^T\right]\Phi(X) + E\left[\varepsilon\varepsilon^T\right] = \Phi(X)^T\widehat{\Sigma}\Phi(X) + \sigma^2\mathbb{I} \\ &= \tilde{\Sigma} + \sigma^2\mathbb{I}. \end{aligned} \tag{18}$$

$$Cov\begin{pmatrix} \mathbf{y} \\ \mathbf{y}^* \end{pmatrix} = \begin{bmatrix} \tilde{\Sigma} + \sigma^2\mathbb{I} & \Sigma^{*T} \\ \Sigma^* & \Sigma^{**} \end{bmatrix} \tag{19}$$

Por lo que se establece la distribución correspondiente

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \tilde{\Sigma} + \sigma^2\mathbb{I} & \Sigma^{*T} \\ \Sigma^* & \Sigma^{**} \end{bmatrix}\right). \tag{20}$$

Con uso de la distribución Gaussiana condicionada, se obtiene

$$\mathbf{y}^*|X^*, \mathbf{y} \sim \mathcal{N}\left(\Sigma^*(\tilde{\Sigma} + \sigma^2\mathbb{I})^{-1}\mathbf{y}, \Sigma^{**} - \Sigma^*(\tilde{\Sigma} + \sigma^2\mathbb{I})^{-1}\Sigma^{*T}\right). \tag{21}$$

Entrenamiento del Modelo

$$\begin{bmatrix} \mathbf{y} \\ y^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} & \tilde{\Sigma} + \sigma^2 \mathbb{I} & & \begin{matrix} \kappa(\mathbf{x}^*, \mathbf{x}_1) \\ \cdots \\ \kappa(\mathbf{x}^*, \mathbf{x}_n) \end{matrix} \\ \kappa(\mathbf{x}^*, \mathbf{x}_1) & \cdots & \kappa(\mathbf{x}^*, \mathbf{x_n}) & \kappa(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix}\right). \tag{22}$$

Denotando,

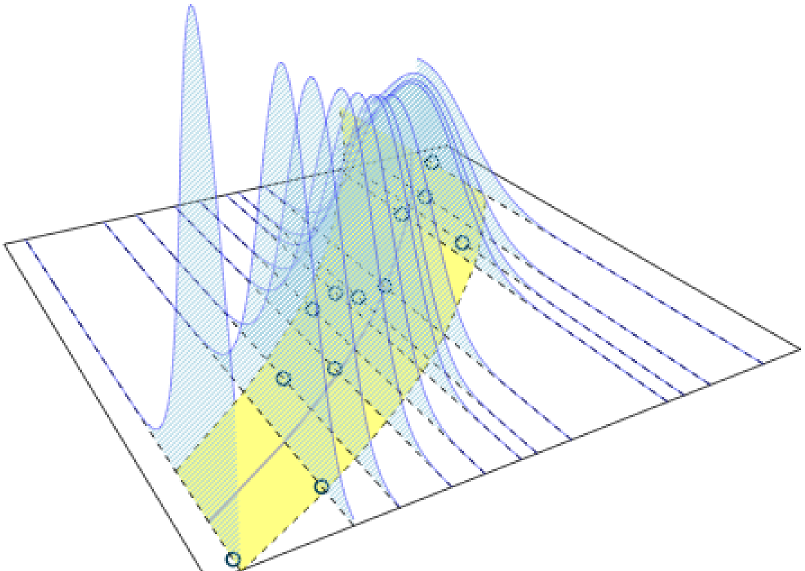
$$\kappa^* = \begin{bmatrix} \kappa(\mathbf{x}^*, \mathbf{x}_1) & \cdot & \kappa(\mathbf{x}^*, \mathbf{x}_n) \end{bmatrix} \quad \kappa^{**} = \kappa(\mathbf{x}^*, \mathbf{x}^*) \tag{23}$$

Tenemos, haciendo uso de la Proposición 2.4, la distribución predictiva siguiente

$$y^* | \mathbf{x}^*, \mathbf{y} \sim \mathcal{N}\left(\kappa^*(\tilde{\Sigma} + \sigma^2 \mathbb{I})^{-1}\mathbf{y}, \kappa^{**} - \kappa^*(\tilde{\Sigma} + \sigma^2 \mathbb{I})^{-1} \kappa^{*T}\right).$$

$$\text{donde} \begin{cases} \overline{y}^* = \kappa^*(\tilde{\Sigma} + \sigma^2 \mathbb{I})^{-1}\mathbf{y} \text{ (valor esperado de predicción).} \\ \sigma^2_{y^*} = \kappa^{**} - \kappa^*(\tilde{\Sigma} + \sigma^2 \mathbb{I})^{-1} \kappa^{*T} \text{ (varianza de predicción)} \end{cases} \tag{24}$$

Entrenamiento del Modelo



Hiperparámetros

$$\begin{aligned} \kappa(\cdot,\cdot) : \mathbb{R}^k \times \mathbb{R}^k & \quad\rightarrow\quad \mathbb{R} \\ (\mathbf{x},\mathbf{x}') & \quad\longmapsto\quad \kappa(\mathbf{x},\mathbf{x}') = \sigma_h^2 \exp\left(-\frac{1}{2l^2}(\mathbf{x}-\mathbf{x}')^2\right) \end{aligned} \tag{25}$$

denominada *función exponencial cuadrática*.

l, σ_h, σ" se denominan **hiperparámetros**. La distribución $y \sim \mathcal{N}\left(0, \tilde{\Sigma} + \sigma^2 I\right)$ vendría implícitamente condicionada a **hiperparámetros**

$$f(y|\theta_{hiper}) = \mathcal{N}\left(0, \tilde{\Sigma} + \sigma^2 I\right) = \frac{1}{(2\pi)^{n/2}\left|\tilde{\Sigma} + \sigma^2\mathbb{I}\right|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{y}^T(\tilde{\Sigma} + \sigma^2\mathbb{I})^{-1}\mathbf{y}\right). \tag{26}$$

donde $\theta_{hiper} = (l, \sigma_h, \sigma)$

Hiperparámetros

$$\begin{aligned}\theta_{hiper,\ estimado} &= \max_{\theta_{hiper}} \{ l(\theta_{hiper}|\mathbf{y}) \} \\ &= \max_{\theta_{hiper}} \left\{ -\frac{1}{2} \ln\left(\left| \tilde{\Sigma} + \sigma^2 \mathbb{I} \right| \right) - \frac{n}{2} \ln(2\pi) + \frac{1}{2} \mathbf{y} (\tilde{\Sigma} + \sigma^2 \mathbb{I})^{-1} \mathbf{y} \right\} \end{aligned} \tag{27}$$

Denotando $B = \tilde{\Sigma} + \sigma^2 \mathbb{I}$,

$$0 = \frac{\partial \left(l(\theta_{hiper}) \, |\mathbf{y} \right)}{\partial a_i} = \frac{1}{2} Traza \left(B^{-1} \frac{\partial B}{\partial a_i} \right) + \frac{1}{2} \mathbf{y}^T \frac{\partial B}{\partial a_i} B^{-1} \frac{\partial B}{\partial a_i} \mathbf{y}. \tag{28}$$

(29)

para $i = 1, 2$ donde $a_1 = \sigma_h, a_2 = l$; y por tanto establecer los hiperparámetros estimados.

Hiperparámetros

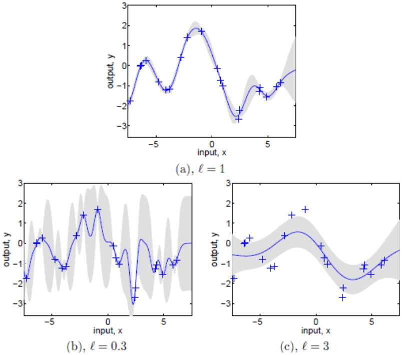


Figure: Visualización de Regresión por Proceso Gaussianos para ciertos datos usando distintos valores en el hiperparámetro l. He aquí la importancia de la estimación sobre los hiperparámetros.