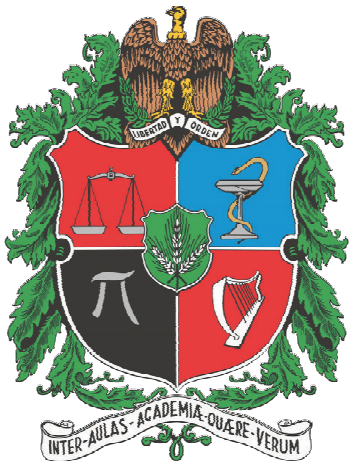
Modelo lineal Bayesiano

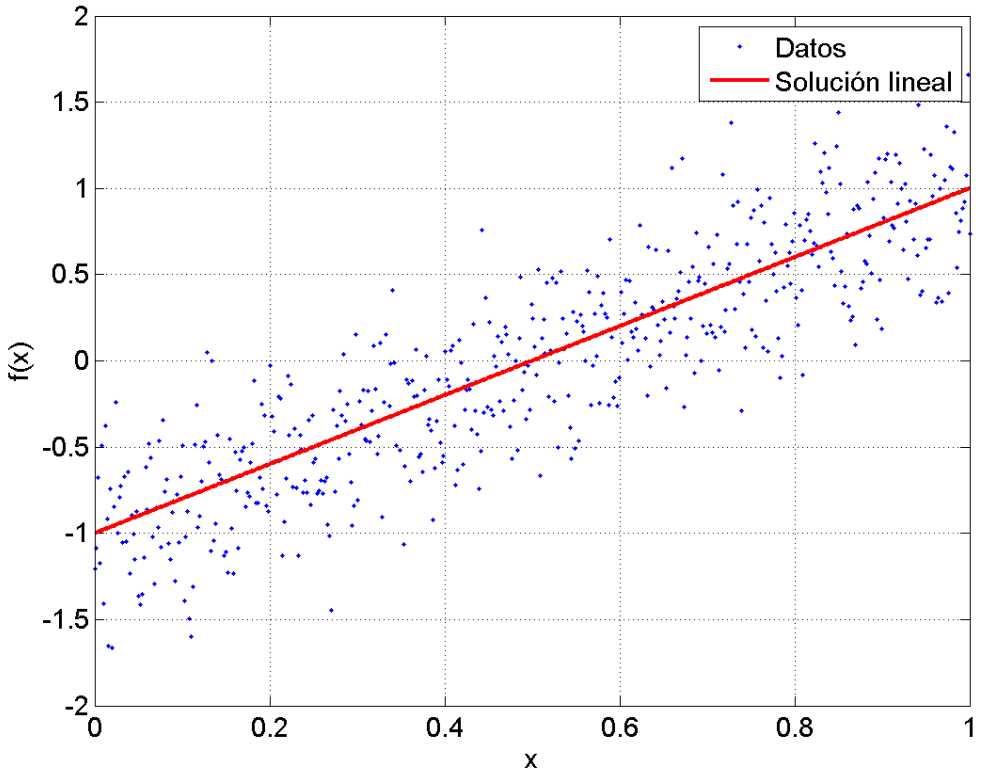
A. M. Alvarez-Meza, Ph.D.
amalvarezme@unal.edu.co

Departamento de ingeniería eléctrica, electrónica y computación
Universidad Nacional de Colombia-sede Manizales



Contenido

Modelo lineal



- $y = f(x) = mx + b \Rightarrow$ noción lineal desde algebra básica.
- $\boldsymbol{y} = f(\boldsymbol{x}) = \boldsymbol{xw} + \boldsymbol{b} \Rightarrow$ extensión algebra vectorial.
- Los datos no siempre llegan limpios y no siempre comparten relaciones lineales!

Modelo lineal: extensión matricial

$$\hat{\boldsymbol{y}} = f(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{w} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1P} \\ x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NP} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_P \end{bmatrix}$$

- $\hat{\boldsymbol{y}} \in \mathbb{R}^N$, $\boldsymbol{X} \in \mathbb{R}^{N \times P}$, $\boldsymbol{w} \in \mathbb{R}^P$
- N : # muestras.
- P : # características.

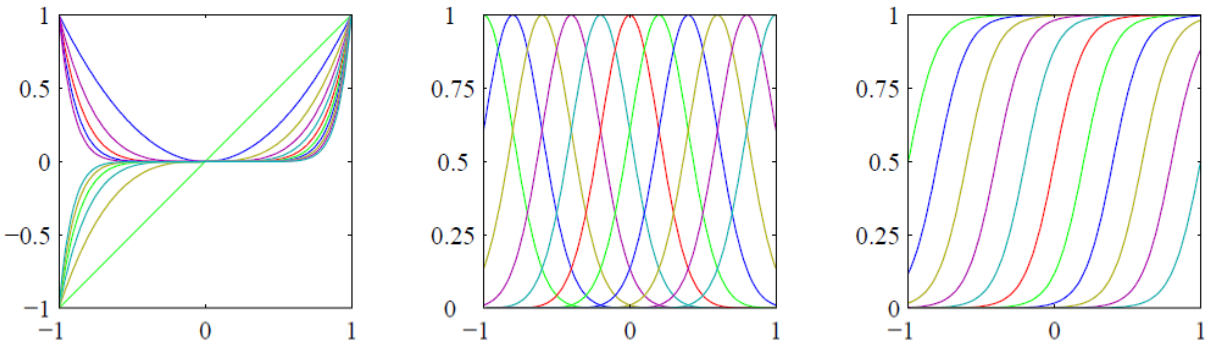
Cómo encontrar los parámetros del modelo lineal ($\boldsymbol{w}$)?

Modelo lineal regularizado

$$\epsilon(\boldsymbol{w}, \lambda) = \frac{1}{2}\|\boldsymbol{e}\|_2^2 + \frac{1}{2}\lambda\|\boldsymbol{w}\|_2^2$$

- $\|\boldsymbol{e}\|_2^2$: cuantifica el desajuste entre los datos y el modelo lineal de aproximación.
- $\|\boldsymbol{w}\|_2^2$: cuantifica el sobreajuste (complejidad) de la solución.
- $\lambda \in \mathbb{R}^+$: parámetro de balance ("trade-off").
- Se necesita un λ que garantice una solución simple pero "exacta".

Regresión desde representación no lineal



- Polinomial: $\phi(\boldsymbol{x}) = \left[\boldsymbol{x}^j\right]_{j=1}^{D}$, D : grado del polinomio.
- Exponencial: $\phi\left(\boldsymbol{x}|\{\boldsymbol{\mu}_j\}_{j=1}^{Q}, \sigma\right) = \left[\exp\left(\frac{-\|\boldsymbol{x}-\boldsymbol{\mu}_j\|_2^2}{2\sigma^2}\right)\right]_{j=1}^{Q}$.
- Sigmoidal:
  $\phi\left(\boldsymbol{x}|\{\boldsymbol{\mu}_j\}_{j=1}^{Q}, \sigma\right) = \left[1/\left(1+\exp\left(\|\boldsymbol{x}-\boldsymbol{\mu}_j\|_2^2/(2\sigma^2)\right)\right)\right]_{j=1}^{Q}$.

Solución del modelo lineal sobre representaciones no lineales

- Solución sobre $\boldsymbol{X}$:

$$\boldsymbol{w}^* = \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

$$y_i = \boldsymbol{x}_i \boldsymbol{w}$$

- Solución sobre $\boldsymbol{\Phi}$:

$$\boldsymbol{w}^* = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y}$$

$\boldsymbol{X} \in \mathbb{R}^{N \times P}$, $\boldsymbol{w} \in \mathbb{R}^{P}$, $\boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{P \times P}$.

$$y_i = \phi\left(\boldsymbol{x}_i\right) \boldsymbol{w}$$

$\boldsymbol{\Phi} \in \mathbb{R}^{N \times Q}$, $\boldsymbol{w} \in \mathbb{R}^{Q}$, $\boldsymbol{\Phi}^\top \boldsymbol{\Phi} \in \mathbb{R}^{Q \times Q}$.

Thomas Bayes

Thomas Bayes
1701–1761

Thomas Bayes was born in Tunbridge Wells and was a clergyman as well as an amateur scientist and a mathematician. He studied logic and theology at Edinburgh University and was elected Fellow of the Royal Society in 1742. During the 18^{th} century, issues regarding probability arose in connection with gambling and with the new concept of insurance. One particularly important problem concerned so-called inverse probability. A solution was proposed by Thomas Bayes in his paper 'Essay towards solving a problem in the doctrine of chances', which was published in 1764, some three years after his death, in the *Philosophical Transactions of the Royal Society*. In fact, Bayes only formulated his theory for the case of a uniform prior, and it was Pierre-Simon Laplace who independently rediscovered the theory in general form and who demonstrated its broad applicability.

Máxima verosimilitud I

- Supongamos y dado como

$$y = f(\boldsymbol{x}, \boldsymbol{w}) + \epsilon$$

donde $\epsilon \sim \mathcal{N}(0, \beta^{-1})$.

- La incertidumbre en y está dada como

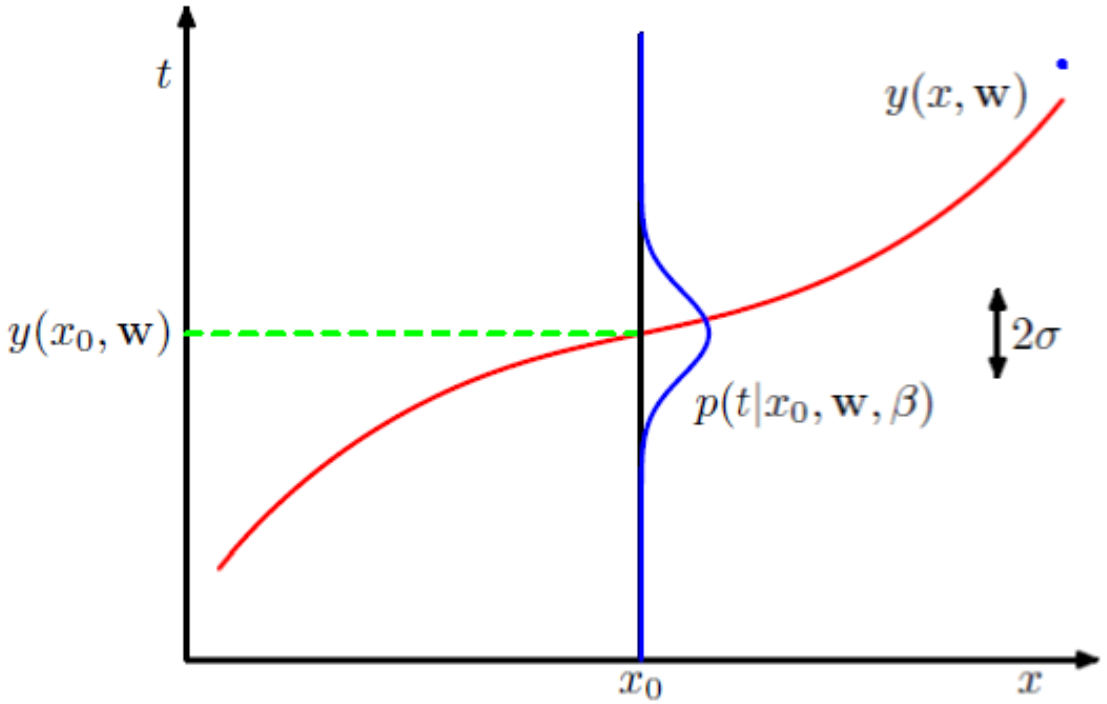
$$p\left(y|\boldsymbol{x}, \boldsymbol{w}, \beta^{-1}\right) = \mathcal{N}\left(y|f(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}\right).$$

- Consideremos un conjunto de datos (de entrenamiento)

$$\boldsymbol{X} = \left[\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top\right]^\top,$$
$$\boldsymbol{y} = [y_1, \ldots, y_N]^\top,$$

con $\boldsymbol{x}_i \in \mathbb{R}^P$, $y \in \mathbb{R}$.

Ejemplo:



Máxima verosimilitud II

- Suponiendo que los datos son independientes e idénticamente distribuidos (iid) y $f(\boldsymbol{x}, \boldsymbol{w}) = \phi(\boldsymbol{x})\boldsymbol{w}$:

$$p\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \beta\right) = \prod_{i=}^{N} \mathcal{N}\left(y_i|\phi(\boldsymbol{x}_i)\boldsymbol{w}, \beta^{-1}\right)$$

- Tomando el logaritmo de la verosimilitud se tiene

$$\begin{aligned}\log\left(p\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \beta\right)\right) &= \sum_{i=1}^{N} \log\left(\mathcal{N}\left(y_i|\phi(\boldsymbol{x}_i)\boldsymbol{w}, \beta^{-1}\right)\right) \\ &= \frac{N}{2}\log\left(\beta\right) - \frac{N}{2}\log\left(2\pi\right) - \beta E_D\left(\boldsymbol{w}\right)\end{aligned}$$

$$E_D\left(\boldsymbol{w}\right) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \phi(\boldsymbol{x}_i)\boldsymbol{w}\right)^2$$

- Cuáles son los $\boldsymbol{w}$ y el parámetro β que mejor explican los datos?

Máxima verosimilitud III

- Maximizar la verosimilitud es equivalente a minimizar $-\beta E_D(\boldsymbol{w})$
- De nuevo,

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \phi(\boldsymbol{x}_i)\boldsymbol{w})^2$$
$$= \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})^\top(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})$$

donde

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi(\boldsymbol{x}_1) \\ \phi(\boldsymbol{x}_2) \\ \vdots \\ \phi(\boldsymbol{x}_N) \end{bmatrix}$$

$\boldsymbol{\Phi} \in \mathbb{R}^{N \times Q}$, $\phi : \mathbb{R}^P \to \mathbb{R}^Q$.

Máxima verosimilitud IV

- La verosimilitud logarítmica está dada entonces como:

$$\log\left(p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w},\beta\right)\right)=\frac{N}{2}\log(\beta)-\frac{N}{2}\log\left(2\pi\right)-\frac{\beta}{2}\left(\boldsymbol{y}-\boldsymbol{\Phi w}\right)^{\top}\left(\boldsymbol{y}-\boldsymbol{\Phi w}\right)$$

- Se tiene entonces

$$\begin{aligned}\frac{\partial\log\left(p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w},\beta\right)\right)}{\partial\boldsymbol{w}} &= -\frac{\beta}{2}\frac{\partial}{\partial\boldsymbol{w}}\left[\left(\boldsymbol{y}-\boldsymbol{\Phi w}\right)^{\top}\left(\boldsymbol{y}-\boldsymbol{\Phi w}\right)\right]\\ &= -\frac{\beta}{2}\frac{\partial}{\partial\boldsymbol{w}}[\boldsymbol{y}^{\top}\boldsymbol{y}-2\boldsymbol{y}^{\top}\boldsymbol{\Phi w}+\boldsymbol{w}^{\top}\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi w}]\end{aligned}$$

- Note que

$$\frac{\partial\left(\boldsymbol{a}^{\top}\boldsymbol{x}\right)}{\partial\boldsymbol{x}}=\boldsymbol{a},\qquad\frac{\partial\left(\boldsymbol{x}^{\top}\boldsymbol{A}\boldsymbol{x}\right)}{\partial\boldsymbol{x}}=\left(\boldsymbol{A}+\boldsymbol{A}^{\top}\right)\boldsymbol{x}.$$

Máxima verosimilitud V

- Por ende

$$\frac{\partial \log\left(p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w},\beta\right)\right)}{\partial \boldsymbol{w}} = \beta\left[\boldsymbol{\Phi}^{\top}\boldsymbol{y}\boldsymbol{w} - \boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\boldsymbol{w}\right].$$

- La solución de máxima verosimilitud para $\boldsymbol{w}$ está dada como:

$$\boldsymbol{w}_{ML} = \left(\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{y},$$

- La solución de máxima verosimilitud para β se obtiene de:

$$\frac{\partial \log\left(p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w},\beta\right)\right)}{\partial \beta} = \frac{N}{2\beta} - \frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{w}\right)^{\top}\left(\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{w}\right).$$

- Y así,

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\left(\boldsymbol{y}-\boldsymbol{\phi}\boldsymbol{w}_{ML}\right)^{\top}\left(\boldsymbol{y}-\boldsymbol{\phi}\boldsymbol{w}_{ML}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \phi\left(\boldsymbol{x}_i\right)\right)^2.$$

Máxima verosimilitud VI

- Al igual que en mínimos cuadrados, el modelo de máxima verosimilitud puede regularizarse.
- Se pretende controlar el sobre entrenamiento.
- La función de error toma la forma

$$E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \phi(\boldsymbol{x}_i)\boldsymbol{w})^2 + \frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w},$$

  donde $E_W(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w}$.
- El valor de $\boldsymbol{w}$ que minimiza $E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w})$ está dado por:

$$\boldsymbol{w} = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y}.$$

Modelo Bayesiano como alternativa a la regularización

- Modelos lineales basados en mínimos cuadrados y máxima verosimilitud pueden regularizarse.
- El valor del parámetro de regularización se impone o asume.
- Una alternativa a la regularización es el tratamiento Bayesiano.

Algunas consideraciones

- La verosimilitud del modelo lineal esta dado por:

$$p\left(\boldsymbol{y}|\boldsymbol{\phi}, \boldsymbol{w}, \beta\right) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{\phi}\boldsymbol{w}, \beta^{-1}\boldsymbol{I}\right).$$

- En máxima verosimilitud se calculó una solución puntual para $\boldsymbol{w} \rightarrow \boldsymbol{w}_{ML}$.
- En estimación Bayesiana se asume un prior para $\boldsymbol{w}$ y se calcula la probabilidad a posteriori de $\boldsymbol{w}$ dados los datos $\boldsymbol{y}$.
- El posteriori sobre $\boldsymbol{w}$ se usa para hacer predicciones.

Teorema de Bayes

- Para calcular el posterior sobre $\boldsymbol{w}$ se usa el teorema de Bayes:

$$p\left(\boldsymbol{w}|\boldsymbol{y}\right) = \frac{p\left(\boldsymbol{y}|\boldsymbol{w}\right)p\left(\boldsymbol{w}\right)}{p(\boldsymbol{y})},$$

  donde $p(\boldsymbol{y})$ es la evidencia, $p(\boldsymbol{y}|\boldsymbol{w})$ es la verosimilitud y $p(\boldsymbol{w})$ es el prior.
- Usando el modelo $\boldsymbol{y} = f\left(\boldsymbol{w}, \boldsymbol{x}\right) + \epsilon$; con $\epsilon \sim \mathcal{N}\left(0, \beta^{-1}\right)$, la verosimilitud es conocida.
- Dependiendo del prior que se escoja para $\boldsymbol{w}$, es posible calcular analíticamente el posterior.

Prior y posterior

- Asumiendo que el prior es Gaussiano, el posterior es igualmente Gaussiano.
- En particular, suponga que $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$.
- Usando propiedades de la Gaussiana, se puede demostrar que:

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})\, p(\boldsymbol{w})}{p(\boldsymbol{y})} = \frac{\mathcal{N}\left(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \beta^{-1}\boldsymbol{I}\right)\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)}{p(\boldsymbol{y})}$$
$$= \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N).$$

donde

$$\boldsymbol{m}_N = \boldsymbol{S}_N\left(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{\Phi}^\top\boldsymbol{y}\right)$$
$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}.$$

Tarea: Demuestre las igualdades anteiores.

Propiedades de la Gaussiana

- Dadas una distribución Gaussiana marginal para $\boldsymbol{x}$, y una distribución Gaussiana condicional para $\boldsymbol{y}$, dado $\boldsymbol{x}$, de la forma:

$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Delta}^{-1}\right)$$
$$p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}\right)$$

la distribución marginal de $\boldsymbol{y}$, y la distribución condicional de $\boldsymbol{x}$ dado $\boldsymbol{y}$ están dadas por:

$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Delta}^{-1}\boldsymbol{A}^{\top}\right)$$
$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\Sigma}\{\boldsymbol{A}^{\top}\boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Delta}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}\right)$$

donde

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Delta} + \boldsymbol{A}^{\top}\boldsymbol{L}\boldsymbol{A}\right)^{-1}$$

Prior más simple

- Un prior más sencillo sigue la forma $p(\boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}\right)$.
- El posterior está dado como

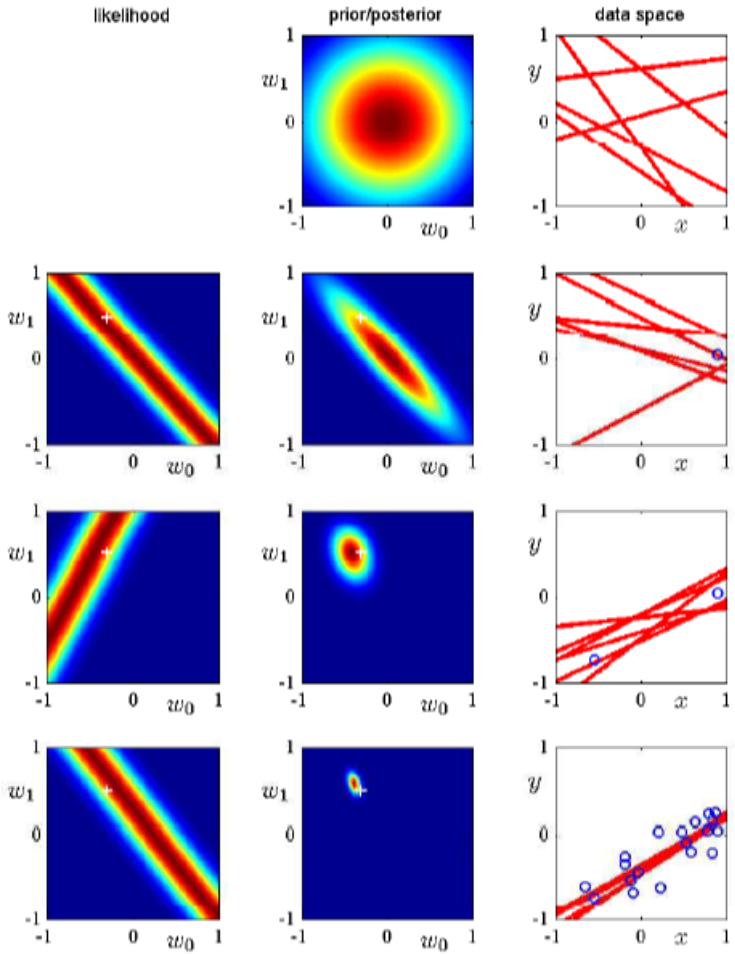
$$p\left(\boldsymbol{w}|\boldsymbol{y}\right) = \mathcal{N}\left(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N\right).$$

donde

$$\boldsymbol{m}_N = \beta \boldsymbol{S}_N \boldsymbol{\Phi}^\top \boldsymbol{y}$$
$$\boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}.$$

Tarea: Demuestre las igualdades anteiores.

Ejemplo: posterior



$\beta^{-1} = 0.04,\ \alpha = 2,\ w_0 = -0.3,\ w_1 = 0.5.$

Maximum a posteriori (MAP)

- La regularización se puede ver como la estimación del Maximum A Posteriori (MAP).
- El logaritmo del posterior es una función de $\boldsymbol{w}$

$$\log\left(p(\boldsymbol{w}|\boldsymbol{y})\right) = -\frac{\beta}{2}\sum_{i=1}^{N}(y_i - \phi(\boldsymbol{x}_i)\boldsymbol{w})^2 - \frac{\alpha}{2}\boldsymbol{w}^\top\boldsymbol{w} + \text{cte}$$

- Equivalente a la regularización si $\lambda = \alpha/\beta$.

Tarea: Demuestre la igualdad anteior.

Distribución predictiva

- **Objetivo**: hacer predicciones de y para nuevos valores $\boldsymbol{x}$.
- Denotemos ese nuevo valor de entrada como $\boldsymbol{x}_*$ y la predicción resultante como y_*.
- La distribución predictiva para y_* está dada como:

$$p\left(y_*|\boldsymbol{y}, \alpha, \beta, \boldsymbol{x}_*\right) = \int p\left(y_*|\boldsymbol{w}, \beta, \boldsymbol{x}_*\right) p\left(\boldsymbol{w}|\boldsymbol{y}, \alpha, \beta\right) d\boldsymbol{w}$$

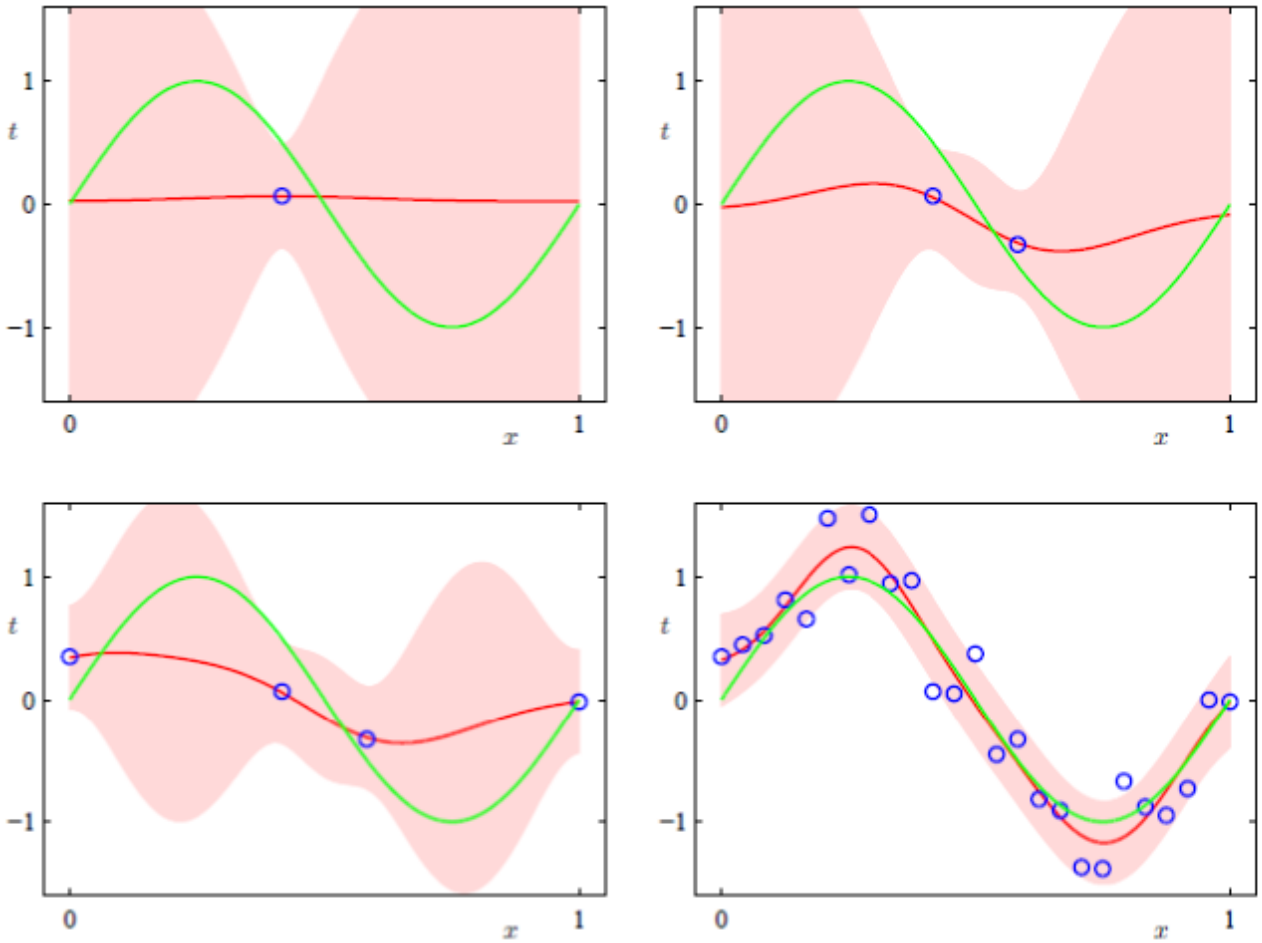
- Usando las propiedades de la Gaussiana se puede demostrar que:

$$p\left(y_*|\boldsymbol{y}, \alpha, \beta, \boldsymbol{x}_*\right) = \mathcal{N}\left(y_*|\phi(\boldsymbol{x}_*)\boldsymbol{m}_N, \sigma_N^2\left(\boldsymbol{x}_*\right)\right),$$

donde $\sigma_N^2\left(\boldsymbol{x}_*\right) = \beta^{-1} + \phi\left(\boldsymbol{x}_*\right)\boldsymbol{S}_N\phi\left(\boldsymbol{x}_*\right)^\top$

- β y α se han asumido conocidos.

Ejemplo: distribución predictiva



Aproximación de la evidencia I

- Si no se conocen α y β, cómo se pueden estimar a partir del conjunto de entrenamiento?
- En un tratamiento Bayesiano general, se ponen priors sobre α y β y se calculan los posteriores.
- Alternativamente, se puede estimar como los parámetros que maximizan la evidencia $p(\boldsymbol{y}|\alpha, \beta)$.
- Este método se conoce como máxima verosimilitud tipo II, aproximación de la evidencia, Bayes empírico.

Aproximación de la evidencia II

- La evidencia está dada como

$$p\left(\boldsymbol{y}|\alpha,\beta\right) = \int p\left(\boldsymbol{y}|\boldsymbol{w},\beta\right) p\left(\boldsymbol{w}|\alpha\right) d\boldsymbol{w}$$

$$\text{evidencia} = \int \text{verosimilitud} \times \textit{prior}$$

- Maximizando la expresión anterior en función de α y β:

$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^\top \boldsymbol{m}_N},$$

donde $\gamma = \sum_i \frac{\lambda_i}{\alpha+\lambda_i}$, $\left(\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right) \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$,
$\boldsymbol{m}_N = \beta \left(\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \alpha \boldsymbol{I}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y}$.

- Note que es una solución iterativa para α, porque γ y $\boldsymbol{m}_N$ dependen de α.

Aproximación de la evidencia III

- Derivando $\log\left(p(\boldsymbol{y}|\alpha,\beta)\right)$ con respecto a β

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\left|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{m}_N\right|^2 .$$

- De nuevo esta es una solución implícita para β, porque $\boldsymbol{m}_N$ depende de β, la solución es iterativa.

Ejercicios Laboratorio

En un cuaderno (notebook) responda a las siguientes preguntas con ejemplos concretos de implementación sobre Python. Envie/comparta su notebook al correo `amalvarezme@unal.edu.co`.

- Consultar el funcionamiento (modelo matemático, función de costo y optimización) de los algoritmos de regresión Bayesiana `sklearn.linear_model.BayesianRidge` y `sklearn.linear_model.ARDRegression` según sus implementaciones en el paquete `Scikit-Learn` de Python.
- Consulte cómo podría incluir funciones de representación no lineal en los algoritmos de regresión del punto anterior.
- Utilizando las bases de datos estudiadas en la sesión de mínimos cuadrados, realice un análisis comparitivo de los métodos de regressión mediante validación cruzada. Recuerde sintonizar los parámetros libres de cada algoritmo.

Referencias I

Hansen, P. C. (1998).

Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion. (Vol. 4). Siam.

Bishop, C. (2006).

Pattern recognition.
Machine Learning, 128.