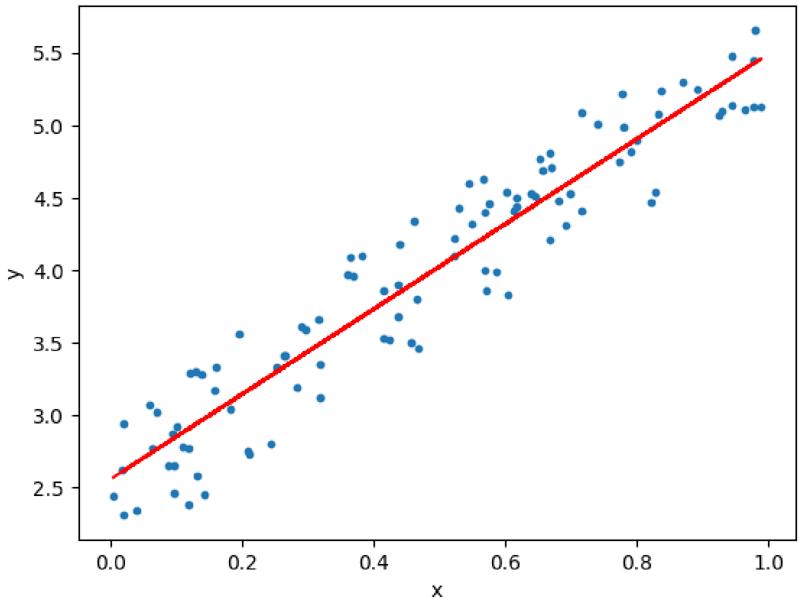


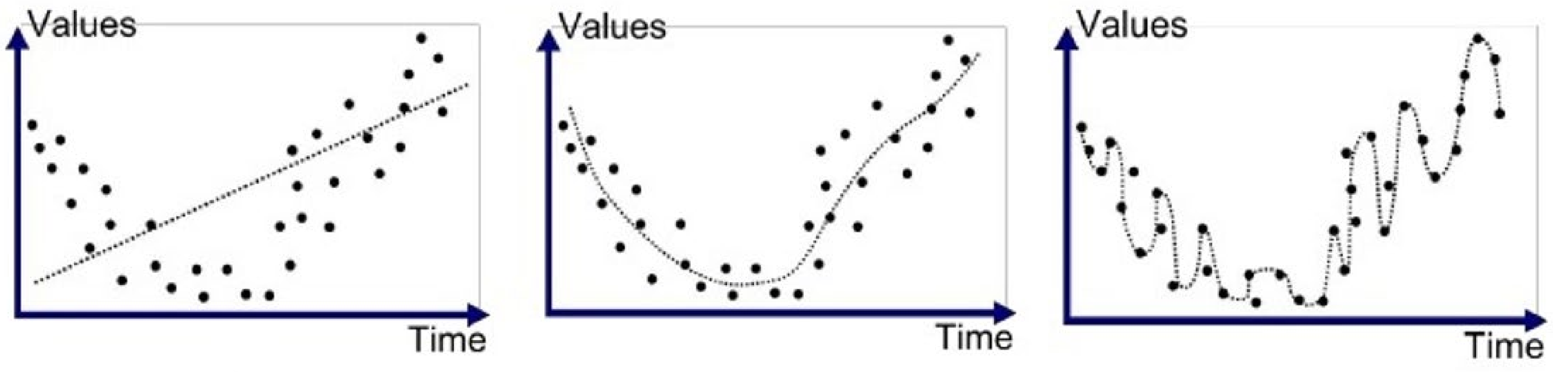
Regresión II

Jonnatan Arias Garcia
jonnatan.arias@utp.edu.co
jariasg@uniquindio.edu.co

David Cardenas peña - dcardenasp@utp.edu.co
Hernán Felipe Garcia - hernanf.garcia@udea.edu.co

Índice

1. Regularización
2. Regresión Lineal Bayesiana
    1. Definiciones
    2. Prior
    3. Posterior
    4. Evidencia
3. Gradiente Descendiente (estocástico)
4. Regresión lineal múltiple
5. Extra: Demostraciones



Underfitted Good Fit/Robust Overfitted

I. Regularización

La regularización se utiliza para evitar el sobreajuste del modelo y mejorar su generalización

Definición

- Controlar el sobre entrenamiento.
- La función de error toma la forma

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w},$$

donde $E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{w}$.

- El valor de $\mathbf{w}$ que minimiza $E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$ está dado por

$$\mathbf{w} = (\lambda \mathbf{I} + \mathbf{\Phi}^\top \mathbf{\Phi})^{-1}\mathbf{\Phi}^\top \mathbf{t}.$$

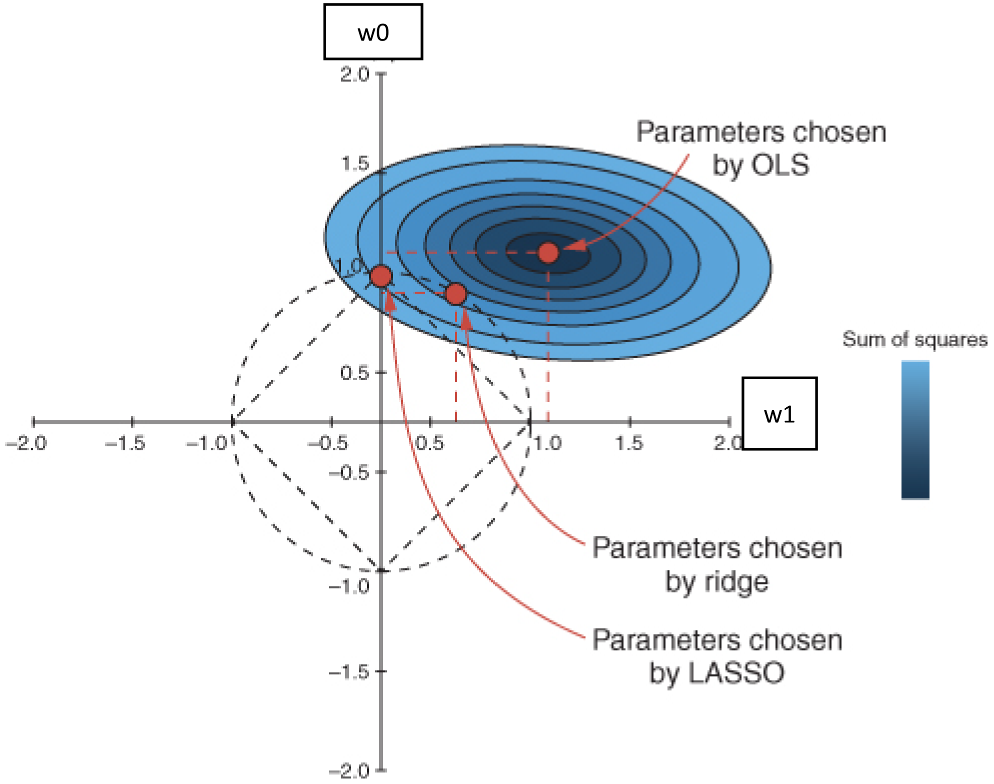
Demostración 1

Alternativas de regularización

- En general, la función de error toma la forma

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=0}^{M-1}|w_j|^q.$$

- El caso $q = 2$ es el regularizador cuadrático anterior.
- El caso $q = 1$ se conoce como la regresión **lasso**.



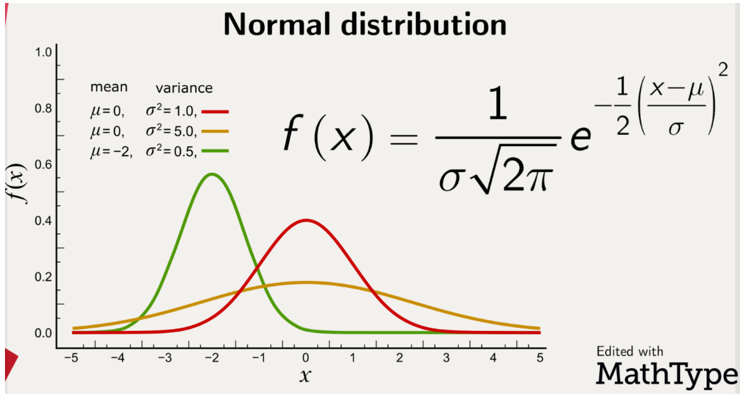
II. Regresión lineal Bayesiana

Definiciones

- Una alternativa a la regularización es el tratamiento Bayesiano.

- Como hemos dicho, la verosimilitud del modelo está dada como

$$p(\mathbf{t}|\mathbf{w},\beta) = \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w},\beta^{-1}\mathbf{I}).$$



- Lo que se hizo en máxima verosimilitud fue realizar una estimación puntual para **w**, que denotamos como $\mathbf{w}_{ML}$.

- En estimación Bayesiana, asumimos un prior para **w** y calculamos la probabilidad a posteriori de **w** dados los datos **t**.

- El posterior sobre **w** se usa para hacer predicciones.

Teorema de Bayes

- Para calcular el posterior sobre **w** usamos el teorema de Bayes

$$\underbrace{p(\mathbf{w}|\mathbf{t})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{t}|\mathbf{w})}^{\text{verosimilitud}}\,\overbrace{p(\mathbf{w})}^{\text{prior}}}{\underbrace{p(\mathbf{t})}_{\text{evidencia o prob total}}},$$

  donde $p(\mathbf{t})$ es la evidencia, $p(\mathbf{t}|\mathbf{w})$ es la verosimilitud y $p(\mathbf{w})$ es el prior.

- Usando el modelo $t = y(\mathbf{w}, \mathbf{x}) + \epsilon$ (con $\epsilon \sim \mathcal{N}(0, \beta^{-1})$), la verosimilitud es conocida.

- Dependiendo del prior que se escoja para **w**, es posible calcular analíticamente el posterior.

- Se dice que un prior es conjugado a una verosimilitud, si el posterior tiene la misma forma del prior.

Prior y posterior

- Asumiendo que el prior es Gaussiano, el posterior es igualmente Gaussiano.

Prior Gaussiano

- En particular, supongamos que $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$.

- Usando propiedades de la Gaussiana, se puede demostrar que

$$p(\mathbf{w}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{t})} = \frac{\overbrace{\mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I})}^{\text{Verosimilitud}}\mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)}{p(\mathbf{t})} = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N),$$

donde

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^\top\mathbf{t}) \quad \text{Media del post.}$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^\top\mathbf{\Phi}. \quad \text{Cov. del post.}$$

Demostración 2: Gaussianas auto conjugadas

Prior más simple

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

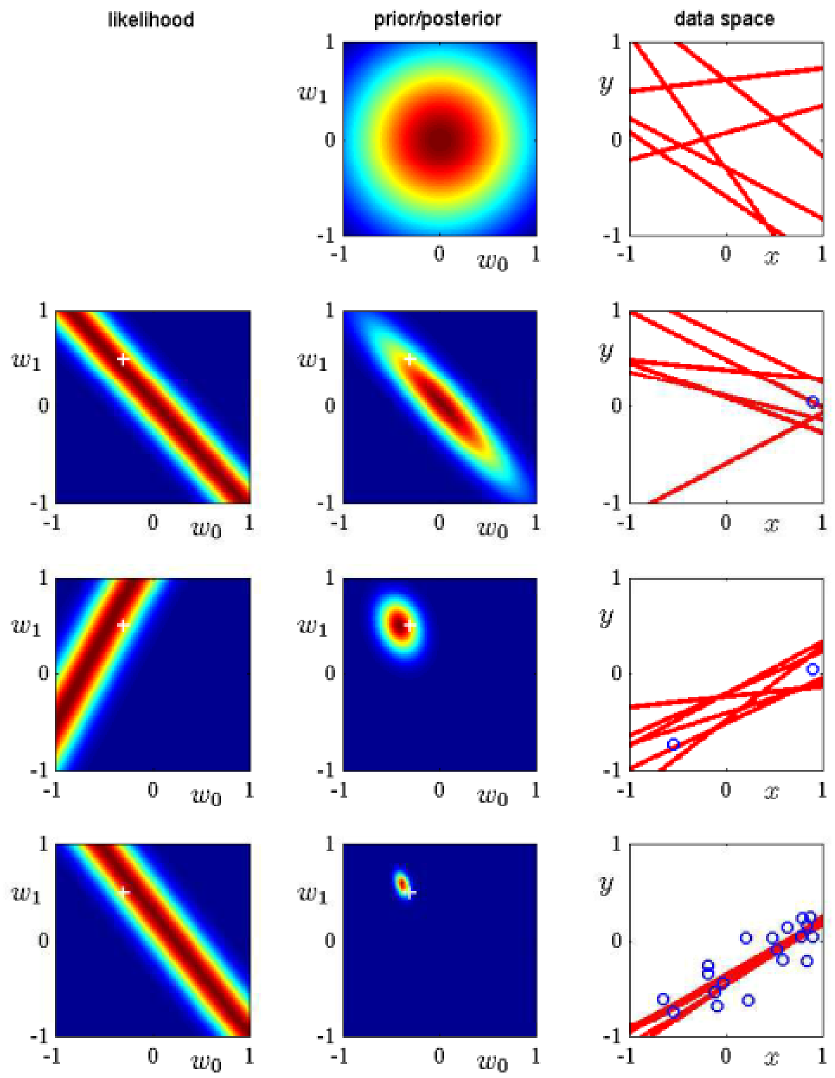
$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N),$$

donde

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

Demostración 3: Gaussianas auto conjugadas simplificada

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 estimación Maximum A Posteriori (MAP).

- El logaritmo del posterior es una función de **w**

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top\phi(\mathbf{x}_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} + \text{const.}$$

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

Distribución predictiva

- **Objetivo**: hacer predicciones de t para nuevos valores $\mathbf{x}$.

- Denotemos ese nuevo valor de entrada como $\mathbf{x}_*$, y la predicción resultante como t_*.

- La distribución predictiva para t_* está dada como

$$p(t_*|\mathbf{t}, \alpha, \beta, \mathbf{x}_*) = \int p(t_*|\mathbf{w}, \beta, \mathbf{x}_*)p(\mathbf{w}|\mathbf{t}, \alpha, \beta)d\mathbf{w}.$$

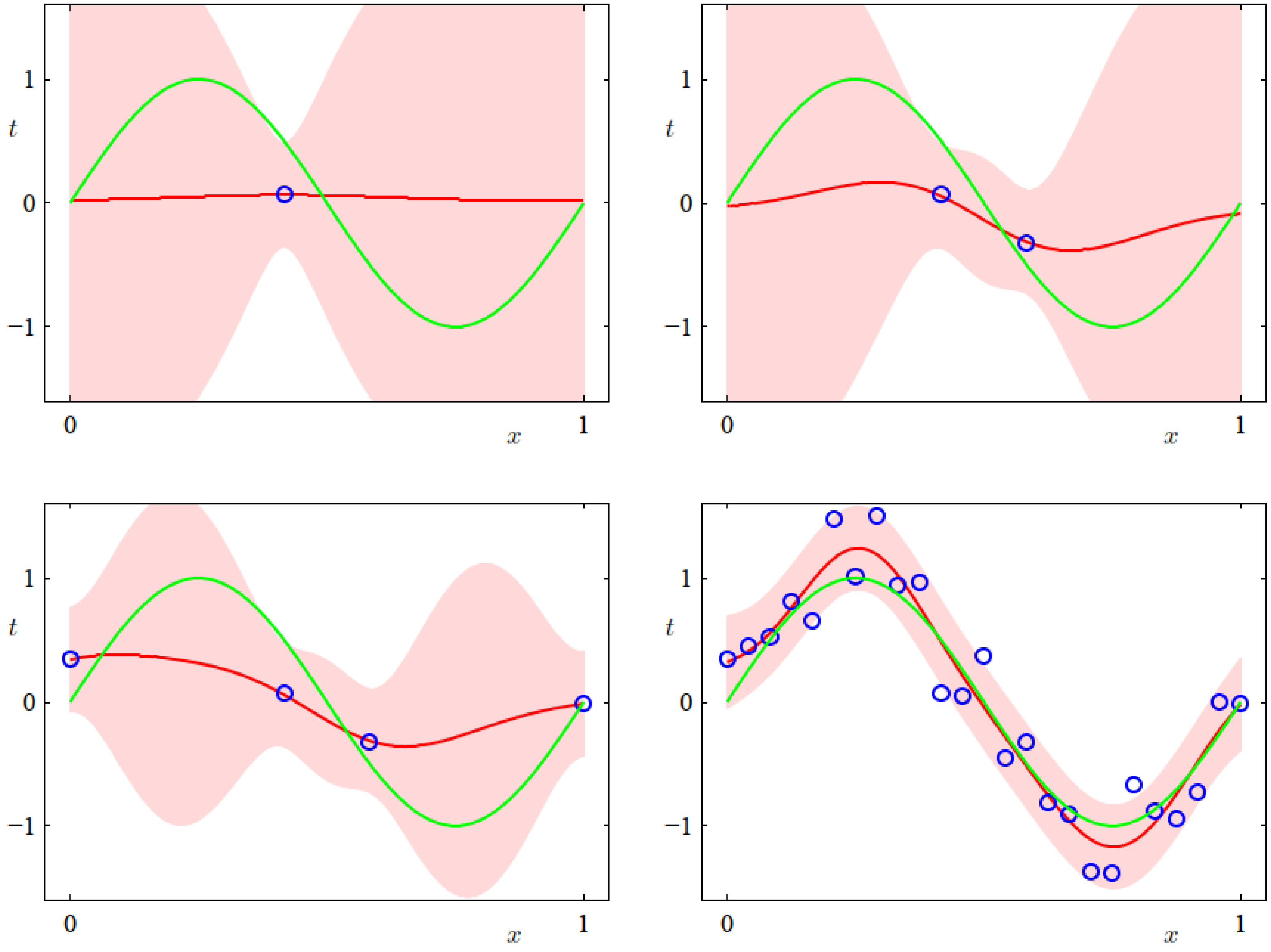
- Usando las propiedades de la Gaussiana (diapositiva anterior) se puede demostrar que

$$p(t_*|\mathbf{t}, \alpha, \beta, \mathbf{x}_*) = \mathcal{N}(t_*|\mathbf{m}_N^\top \phi(\mathbf{x}_*), \sigma_N^2(\mathbf{x}_*)),$$

donde $\sigma_N^2(\mathbf{x}_*) = \beta^{-1} + \phi(\mathbf{x}_*)^\top \mathbf{S}_N \phi(\mathbf{x}_*)$. Demostración 4: predictiva

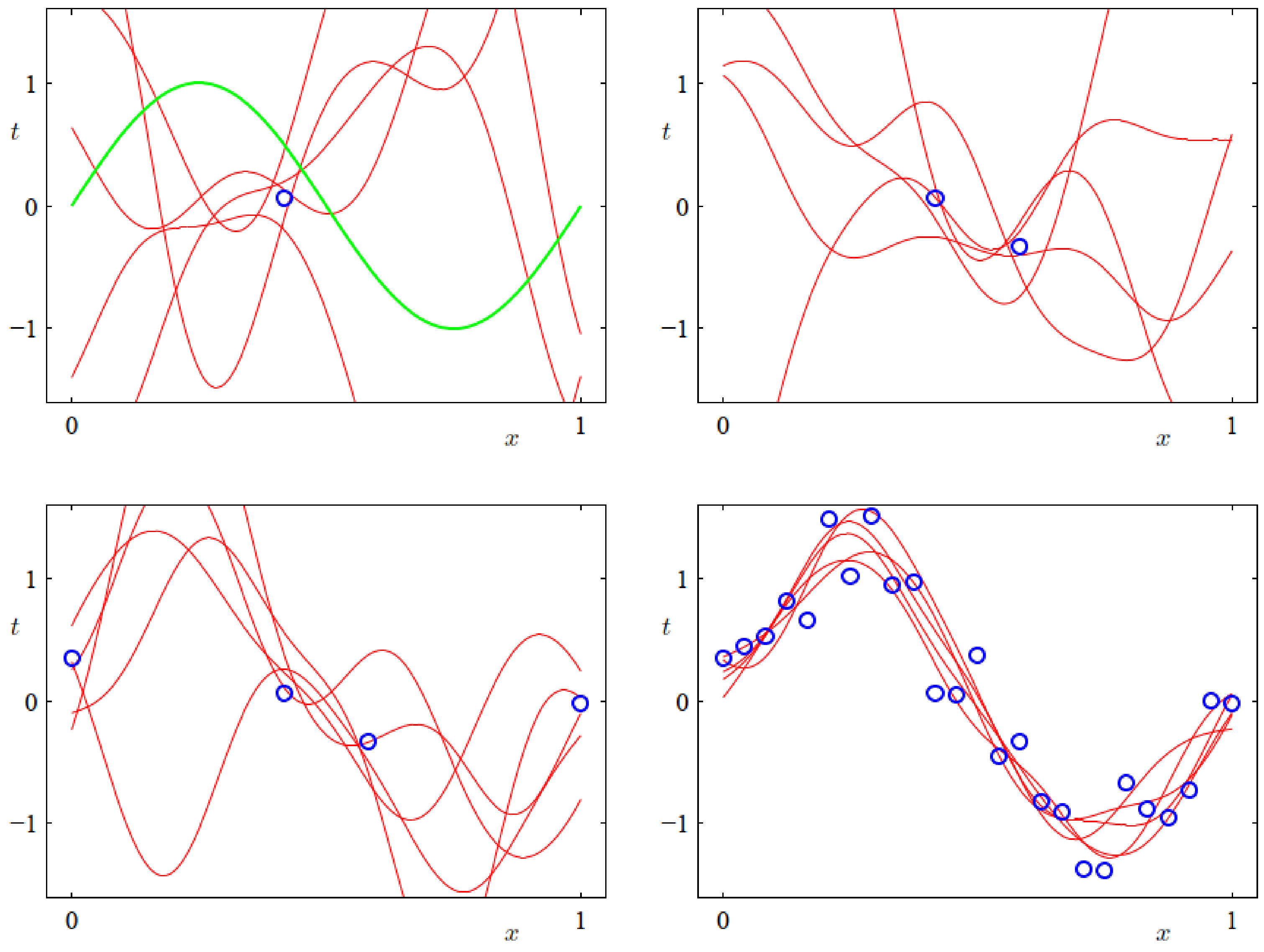
- *Importante*: nótese que se ha asumido que β y α son conocidos.

Ejemplo: distribución predictiva



Ejemplo: otra representación

Se muestrea el posterior $p(\mathbf{w}|\mathbf{t})$, y luego se grafica $y(\mathbf{x}, \mathbf{w})$.



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 $p(\mathbf{t}|\alpha, \beta)$.

- 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

$$
\begin{aligned}
p(\mathbf{t}|\alpha,\beta) &= \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}, \\
\text{evidencia} &= \int \text{verosimilitud x prior}
\end{aligned}
$$

- Reemplazando en la integral $p(\mathbf{t}|\mathbf{w},\beta)$, y $p(\mathbf{w}|\alpha)$ se obtiene

$$
p(\mathbf{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\{-E(\mathbf{w})\}d\mathbf{w},
$$

donde

$$
\begin{aligned}
E(\mathbf{w}) &= \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) \\
&= \frac{\beta}{2}\|\mathbf{t} - \mathbf{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}.
\end{aligned}
$$

Aproximación de la evidencia (III)

- Se quiere integrar sobre $\mathbf{w}$. Para eso se completa el cuadrado obteniéndose

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top \mathbf{A}(\mathbf{w}-\mathbf{m}_N),$$

donde $\mathbf{m}_N = \beta \mathbf{A}^{-1}\boldsymbol{\Phi}^\top \mathbf{t}$, $\mathbf{A} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$, y

$$\boxed{E(\mathbf{m}_N) = \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^\top \mathbf{m}_N.}$$

Demostración 5: Evidencia

- Nótese que

$$\nabla\nabla E(\mathbf{w}) = \mathbf{A},$$

es la matriz Hessiana.

Aproximación de la evidencia (IV)

- Para calcular la integral se tiene entonces

$$\int \exp\{-E(\mathbf{w})\}d\mathbf{w} = \exp\{-E(\mathbf{m}_N)\}\times$$
$$\int \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^\top \mathbf{A}(\mathbf{w}-\mathbf{m}_N)\right\} d\mathbf{w}$$
$$= \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}.$$

- La evidencia logarítmica es entonces igual a

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln(\alpha) + \frac{N}{2}\ln(\beta) - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi).$$

- α y β se estiman maximizando la expresión anterior e igualando a cero.

Maximización con respecto a α (I)

- Recordemos que el determinante de una matriz cuadrada **P** se puede calcular como

$$|\mathbf{P}| = \prod_i p_i, \qquad p_i = \text{eig}(\mathbf{P}).$$

- En la expresión anterior

$$|\mathbf{A}| = |\alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}| = \prod_i (\alpha + \lambda_i),$$

donde λ_i es el i-ésimo valor propio de la matriz $\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$.

- El valor propio λ_i se puede calcular resolviendo la siguiente ecuación espectral

$$(\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}) \mathbf{u}_i = \lambda_i \mathbf{u}_i.$$

Maximización con respecto a α (II)

- Usando el resultado anterior, la derivada de $\ln p(\mathbf{t}|\alpha, \beta)$ con respecto a α sigue

$$\ln p(\mathbf{t}|\alpha, \beta) = \frac{M}{2}\ln(\alpha) + \frac{N}{2}\ln(\beta) - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi).$$

$$\frac{\partial \ln p(\mathbf{t}|\alpha, \beta)}{\partial \alpha} = \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^\top \mathbf{m}_N - \frac{1}{2}\sum_i \frac{1}{\alpha + \lambda_i}.$$

- Igualando a cero y despejando α se encuentra

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

donde $\gamma = \sum_i \frac{\lambda_i}{\alpha+\lambda_i}$.

- Nótese que esta es una solución implícita para α, porque γ y $\mathbf{m}_N$ dependen de α. La solución es iterativa.

Maximización con respecto a β (I)

- Los valores propios λ_i dependen de β a través de la ecuación espectral

$$\lambda_i \mathbf{u}_i = (\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}) \mathbf{u}_i.$$

- Derivando a ambos lados la expresión anterior con respecto a β

$$\frac{d\lambda_i}{d\beta} = \frac{\lambda_i}{\beta}.$$

Maximización con respecto a β (II)

- Usando el resultado anterior, la derivada de $\ln p(\mathbf{t}|\alpha, \beta)$ con respecto a β sigue

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln(\alpha) + \frac{N}{2}\ln(\beta) - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi).$$

$$\frac{\partial \ln p(\mathbf{t}|\alpha,\beta)}{\partial\beta} = \frac{N}{2\beta} - \frac{1}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 - \frac{\gamma}{2\beta}.$$

- Igualando a cero y despejando β se obtiene

$$\boxed{\frac{1}{\beta} = \frac{1}{N-\gamma}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2.}$$

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

Algoritmo Completo

1. Inicializar a_0, β_0
2. Calcular los parámetros del posterior $\boldsymbol{w}, \boldsymbol{m}_N, \boldsymbol{S}_N$
3. Calcular γ
4. Calcular a_k, β_k
5. Hasta la convergencia del paso 2

III. Gradiente descendiente estocástico

Stochastic Gradient Descent I

- El algoritmo Stochastic Gradient Descent (SGD) ajusta un modelo (en este caso lineal) minimizando una función de costo (en este caso el MSE) posiblemente regularizada.
- El gradiente del costo se calcula para cada muestra en cada iteración y el modelo se actualiza en la dirección contraria del gradiente según una tasa de aprendizaje (learning rate).
- Para el modelo lineal, la optimización por gradiente descendiente evita el cálculo de inversas, x e:

$$\boldsymbol{w} = \left(\boldsymbol{\phi}^{\boldsymbol{T}}\boldsymbol{\phi} + \lambda \boldsymbol{I}_{\boldsymbol{M}}\right)^{-\mathbf{1}} \phi^T t$$

- Dada una función de costo J$(\theta) \in \mathbb{R}$ a minimizar respecto de $\theta \in \mathbb{R}^D$, la regla de actualización del gradiente descendiente está dada por:

$$\boldsymbol{\theta}_{\boldsymbol{k}} = \boldsymbol{\theta}_{\boldsymbol{k}-\mathbf{1}} - \boldsymbol{\mu}\nabla_{\boldsymbol{\theta}}\mathbf{J}(\boldsymbol{\theta})$$

- donde $\mu \in \mathbb{R}$ es la tasa de aprendizaje y $\nabla_{\boldsymbol{\theta}}\boldsymbol{J}$ es el vector gradiente que reúne las derivadas del costo respecto a cada parámetro θ_d

$$\nabla_\theta J(\boldsymbol{\theta}) = \begin{bmatrix} \frac{dJ}{d\theta_1} \\ \frac{dJ}{d\theta_2} \\ \vdots \\ \frac{dJ}{d\theta_d} \\ \vdots \\ \frac{dJ}{d\theta_D} \end{bmatrix}$$

Stochastic Gradient Descent II

Para el caso de regresión lineal la función de costo es

$$L(w) = E_D(w) = \sum_{n=1}^{N} l_n(w) = \sum_{n=1}^{N} \frac{1}{2}\{t_n - y(X_n, w)\}^2$$

Usando la regla de la suma, la derivada del costo será la suma de las derivadas respecto a los parámetros del error de cada muestra:

$$\nabla L(w) = \frac{dL}{dw} = \frac{d}{dw}\left\{\sum_{n=1}^{N} l_n(w)\right\} = \sum_{n=1}^{N} \frac{d}{dw} l_n(w)$$

Stochastic Gradient Descent III

*Usando la regla de la cadena, la derivada del costo respecto a los parámetros será:

$$\nabla J(w) = \sum_{n=1}^{N} \frac{d}{dw} l_n(w) = \sum_{n=1}^{N} \frac{dl_{n(y_n)}}{dy_n} \frac{dy_{n(w)}}{dw}$$

*La primera derivada corresponde a la derivada del costo (pérdida) respecto a la salida del modelo:

$$\frac{dl_{n(y_n)}}{dy_n} = \frac{d}{dy_n}\left\{\frac{1}{2}(t_n - y_n)^2\right\} = -(t_n - y_n)$$

*La segunda derivada corresponde a la derivada de la salida del modelo respecto a los parámetros:

$$\frac{dy_n(w)}{dw} = \frac{d}{dw}\{w^T \varphi(x_n)\} = \varphi_n$$

*La derivada del costo respecto a los parámetros será entonces:

$$\nabla J(w) = -\sum_{n=1}^{N} (t_n - y_n)\varphi_n = -\sum_{n=1}^{N} e_n \varphi_n = -e^T \phi$$

Donde $e \in \mathbb{R}^N$ es el vector de errores y $\phi \in \mathbb{R}^{NxM}$

*si el error de predicción es cero, el gradiente se hace cero y el algoritmo converge

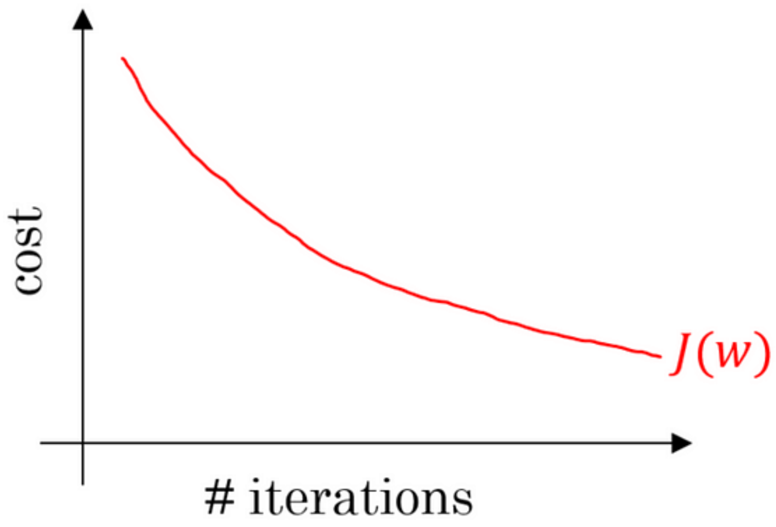
Stochastic Gradient Descent III

El algoritmo de optimización por gradiente descendiente queda:

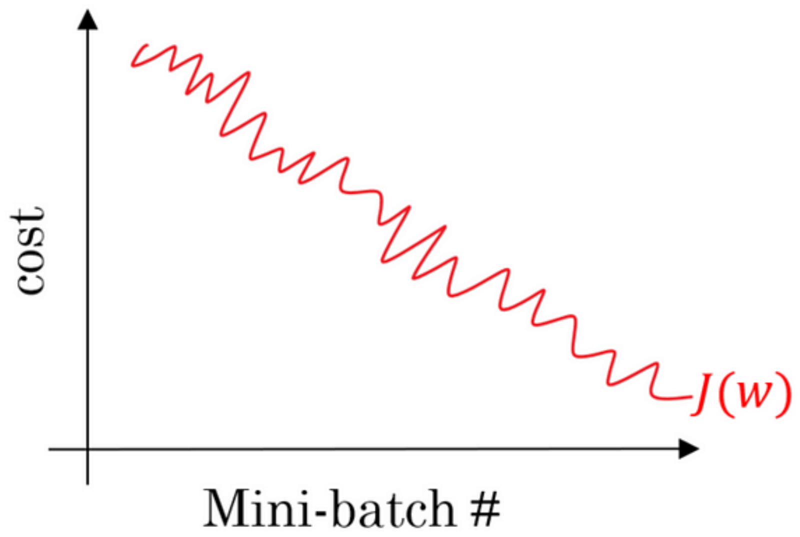
1. Inicial el modelo con $\boldsymbol{w_0}$ y fijar una tasa de aprendizaje $\boldsymbol{\mu}$
2. Realizar las predicciones con los parámetros de $\boldsymbol{y} = \boldsymbol{\phi w_k}$
3. Calcular los errores de predicción $e = t - y$
4. Calcular la función de costo total para los parámetros actuales $L(w_k) = \frac{1}{2} e^T e$
5. Actualizar los parámetros para la siguiente iteración
$$\boldsymbol{w_{k+1}} = \boldsymbol{w_k} - \boldsymbol{\mu \nabla J(w_k)}$$
$$\boldsymbol{w_{k+1}} = \boldsymbol{w_k} + \boldsymbol{\mu \phi^T e}$$
6. Hasta la convergencia volver al paso 2

Stochastic Gradient Descent IV

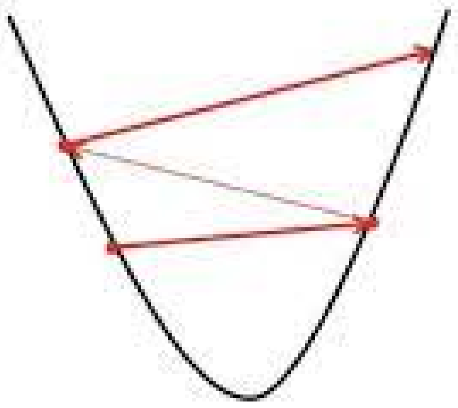
Batch gradient descent



Mini-batch gradient descent



Big learning rate



Small learning rate

IV. Regresor lineal múltiple

Diferencias

Lineal simple

Variables Predictoras:

Única variable independiente para predecir la dependiente.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Interpretación de Coeficientes:

En la regresión lineal simple, el coeficiente representa el cambio promedio en la variable dependiente por unidad de cambio en la variable independiente.

Gráficos:

Puede ser visualizado fácilmente en un gráfico bidimensional.

Complejidad:

Más simple y fácil de interpretar, pero puede no capturar la complejidad de relaciones más intrincadas.

Lineal Múltiple

Variables Predictoras:

Dos o más variables independientes para predecir la dependiente.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon.$$

Interpretación de Coeficientes:

Cada coeficiente representa el cambio promedio en la variable dependiente por unidad de cambio en la correspondiente variable independiente, manteniendo las otras variables constantes.

Gráficos:

Más desafiante visualmente, ya que implica múltiples dimensiones.

Complejidad:

Puede modelar relaciones más complejas entre las variables predictoras y la variable dependiente.

V. Otros posibles regresores

- Regresión por Mínimos cuadrados parciales (método aproximación de cuadrados, visto anteriormente)
- Por arboles de decisión y bosques aleatorios (Deep learning)
- SVM (lo veremos en clúster)
- Por KNN (lo veremos en redes)

Extra: Demostraciones

Demostración 1: w para ridge

Bishop 3.2a $\quad q=2 \quad$ Jonnatan Arias G.

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^2$$

$$= \frac{1}{2}\{(t-\Phi w)^T(t-\Phi w) + \lambda w^T w\}$$

Derivando e igualanda a 0

$$\frac{\partial}{\partial w}\{(t-\Phi w)^T(t-\Phi w) + \lambda w^T w\} = 0$$

$$-2\Phi^T(t-\Phi w) + 2\lambda w = 0$$

$$-\Phi^T t + \Phi^T\Phi w + \lambda w = 0$$

$$(\Phi^T\Phi + \lambda I)w = \Phi^T t$$

$$w = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T t$$

Demostración 2: Gaussianas auto conjugadas

$$P(w|t) = \frac{P(t|w)P(w)}{P(t)} = \frac{\mathcal{N}(t|\Phi w, \beta^{-1} I_N)\,\mathcal{N}(w|m_0, S_0)}{P(t)}$$

$$P(w|t) = \mathcal{N}(w|m_N, S_N)$$

$$P(t) = \frac{1}{Z_L} e\left(-\frac{1}{2}(t-\Phi w)^T \Sigma^{-1}(t-\Phi w)\right) \frac{1}{Z} e\left(-\frac{1}{2}(w-m_0)^T S_0^{-1}(w-m_0)\right)$$

$$\propto e\left(-\frac{1}{2}(w-m_0)^T S_0^{-1}(w-m_0) - \frac{1}{2}(t-\Phi w)^T \Sigma^{-1}(t-\Phi w)\right)$$

Omitiendo $\left(-\frac{1}{2}\right)$

$$(w-m_0)^T S_0^{-1}(w-m_0) + (t-\Phi w)^T \Sigma^{-1}(t-\Phi w)$$

$$= \underbrace{w^T S_0^{-1} w} \underbrace{- 2 w^T S_0^{-1} m_0} + \overset{cte}{\cancel{m_0 S_0 m_0}} + \overset{cte}{\cancel{t \Sigma^{-1} t}} + \underbrace{(w\Phi)^T \Sigma^{-1} \Phi w}$$

$$\underbrace{-2 w^T \Phi t}$$

$$= w^T\left(S_0^{-1} + \Sigma^{-1}\Phi^T\Phi\right) w - 2 w^T\left(S_0^{-1} m_0 + \Sigma^{-1}\Phi^T t\right) + cte$$

Donde

$$S_N^{-1} = \left(S_0^{-1} + \Sigma^{-1}\Phi^T\Phi\right)$$

$$m_N = \overset{\#}{\cancel{S_0^{-1}}} m_0 + \Sigma^{-1}\Phi^T t$$

Demostración 3: Gaussianas auto conjugadas simplificada

$$p(t^*|x, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t^* | m_N^T \phi^*, \sigma_N^2)$$

Con $\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x)$

Distribución Predictiva

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|w, \beta)\, p(w|\mathbf{t}, \alpha, \beta)\, dw$$

teniendo para el posterior que:

$$y(x, w) = w^T \phi(x)$$

$$p(t|x, w, \beta) = \mathcal{N}(t | y(x, w), \beta^{-1})$$

$$p(t|x, w, \beta) = \mathcal{N}(t | w^T \phi(x), \beta^{-1} \mathbb{I})$$

Además

$$p(w|m_N, S_N) \rightarrow p(w|\mathbf{t}, \alpha, \beta)$$

Ya que

$$m_N = S_N (\cancel{S_0^{-1} m_0} + \beta \Phi^T \mathbf{t}) = \boxed{\beta S_N \Phi^T \mathbf{t}}$$

(cte)

$$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi \quad \text{con} \quad S_0^{-1} = \alpha \mathbb{I}$$

$$\boxed{S_N^{-1} = \alpha \mathbb{I} + \beta \Phi^T \Phi}$$

Demostración 4: Predictiva. Partiendo de la Demost.3

Ahora, usando propiedades tenemos que

$$\mathcal{N}(t \mid \mathbf{w}^T \phi(x), \beta^{-1}) \text{ —— } \mathcal{N}(y \mid Ax+b, L^{-1})$$

$$t \to y$$
$$\phi(x) \to Ax$$
$$0 \to b$$
$$\beta^{-1} \to L^{-1}$$

$$\mathcal{N}(w \mid m_N, S_N) \text{ —— } \mathcal{N}(x \mid \mu, \Lambda^{-1})$$

$$w \to x$$
$$m_N \to \mu$$
$$S_N \to \Lambda^{-1}$$

$$p(y) = \mathcal{N}(y \mid A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$

Reemplazando

$$p(y) = \mathcal{N}(t \mid m_N^T \phi(x), \beta^{-1} + \phi(x)^T S_N \phi(x))$$

Donde

$$\beta^{-1} + \phi(x)^T S_N \phi(x) = \sigma^2$$

Demostración 5: Evidencia

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

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

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

- $p(\boldsymbol{t}|\boldsymbol{w},\beta)$ es la verosimilitud de la base de datos $\boldsymbol{X}, \boldsymbol{t}$:

$$p(\boldsymbol{t}|\boldsymbol{w},\beta) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{\Phi}\boldsymbol{w}, \beta^{-1}\boldsymbol{I}_N)$$

- $p(\boldsymbol{w}|\alpha)$ es el prior de $\boldsymbol{w}$:

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}_M)$$

- Reemplazando:

$$p(\boldsymbol{t}|\alpha,\beta) = \int \mathcal{N}(\boldsymbol{t}|\boldsymbol{\Phi}\boldsymbol{w}, \beta^{-1}\boldsymbol{I}_N)\,\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}_M)\,d\boldsymbol{w}$$

$$p(\boldsymbol{t}|\alpha,\beta) = \int \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left(-\frac{\beta}{2}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{w})^\top(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{w})\right)\left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left(-\frac{\alpha}{2}\boldsymbol{w}^\top\boldsymbol{w}\right)d\boldsymbol{w}$$

$$p(\boldsymbol{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\left(-\frac{\beta}{2}\underbrace{(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{w})^\top(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{w})}_{(t-y)^2} - \frac{\alpha}{2}\boldsymbol{w}^\top\boldsymbol{w}\right)d\boldsymbol{w}$$

$$p(\boldsymbol{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\left(-\frac{\beta}{2}(\boldsymbol{t}^\top\boldsymbol{t} - 2\boldsymbol{t}^\top\boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{w}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{w}) - \frac{\alpha}{2}\boldsymbol{w}^\top\boldsymbol{w}\right)d\boldsymbol{w}$$

$$p(\boldsymbol{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\left(-\frac{\beta}{2}\boldsymbol{t}^\top\boldsymbol{t} + \beta\boldsymbol{t}^\top\boldsymbol{\Phi}\boldsymbol{w} - \frac{\beta}{2}\boldsymbol{w}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{w} - \frac{\alpha}{2}\boldsymbol{w}^\top\boldsymbol{w}\right)d\boldsymbol{w}$$

$$p(\boldsymbol{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\left(-\frac{\beta}{2}\boldsymbol{t}^\top\boldsymbol{t} + \beta\boldsymbol{t}^\top\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{w}^\top\underbrace{\left(\frac{\beta}{2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{\alpha}{2}\boldsymbol{I}_M\right)}_{\frac{S_N^{-1}}{2}}\boldsymbol{w}\right)d\boldsymbol{w}$$

$x^\top A x + b^\top x + c$

$\int e\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)dx = Z$

Sp

- Para resolver la integral se completa el cuadrado en la exponencial:

- Para resolver la integral se completa el cuadrado en la exponencial:

$$\boldsymbol{S}_N^{-1} = \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \alpha\boldsymbol{I}_M$$ Cov inv del post

$(w-m_N)^\top S_N^{-1}(w-m_N) = w^\top S_N^{-1} w - 2m_N^\top S_N^{-1} w + m_N^\top S_N^{-1} m_N$

- Si $\beta\mathbf{t}^\top\boldsymbol{\Phi}\boldsymbol{w} = \boldsymbol{m}_N^\top\boldsymbol{S}_N^{-1}\boldsymbol{w}$, entonces: $\boldsymbol{m}_N = \beta\boldsymbol{S}_N\boldsymbol{\Phi}^\top\boldsymbol{t}$ media del post.
- Se completa entonces sumando y restando con el término $\boldsymbol{m}_N^\top\boldsymbol{S}_N^{-1}\boldsymbol{m}_N$

$$p(\boldsymbol{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)^\top\boldsymbol{S}_N^{-1}(\boldsymbol{w}-\boldsymbol{m}_N) - E(\boldsymbol{m}_N)\right)d\boldsymbol{w}$$

$$E(\boldsymbol{m}_N) = \frac{\beta}{2}(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{m}_N)^\top(\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{m}_N) + \frac{\alpha}{2}\boldsymbol{m}_N^\top\boldsymbol{m}_N$$