Regresión lineal y estimación

Mauricio A. Álvarez

Modelos probabilísticos profundos
AIR Institute

Contenido

Modelo de base lineal

- **Regresión lineal.** El modelo más simple de regresión lineal consiste de una combinación lineal de las variables de entrada

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D.$$

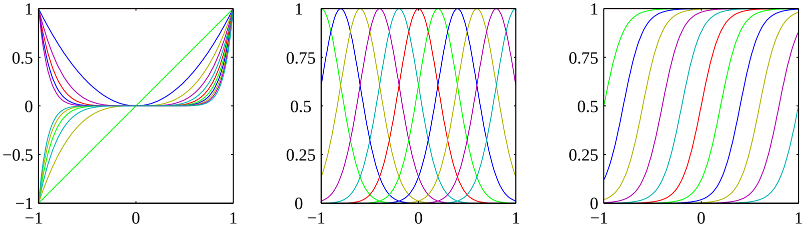
- El modelo anterior se puede extender para combinaciones lineales de funciones no lineales fijas de las variables de entrada,

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{M-1} w_i \phi_i(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}),$$

  donde $\phi_i(\mathbf{x})$ son funciones base, M es el número de parámetros del modelo, y w_0 es el desplazamiento.

- Igualmente, $\mathbf{w} = [w_0 \cdots w_{M-1}]^\top$, $\phi(\mathbf{x}) = [\phi_0(\mathbf{x}) \cdots \phi_{M-1}(\mathbf{x})]^\top$.

Ejemplos funciones base



Polinomial: $\phi_i(x) = x^i$.

Exponencial: $\phi_i(x) = \exp\left\{-\frac{(x-\mu_i)^2}{2s^2}\right\}$

Sigmoidal: $\phi_i(x) = \sigma\left(\frac{x-\mu_i}{s}\right),\ \sigma(a) = 1/(1+\exp(-a))$.

Contenido

Máxima verosimilitud (I)

- Supongamos t dado como

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon,$$

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

- La incertidumbre en t está dada como

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

- Consideremos un conjunto de datos (de entrenamiento)

$$\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\},$$
$$\mathbf{t} = \{t_1, \cdots, t_N\}$$

Máxima verosimilitud (II)

- Suponiendo que los datos son iid

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}).$$

- Tomando el logaritmo de la verosimilitud se tiene

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) &= \sum_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w}),\end{aligned}$$

  donde

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - y(\mathbf{x}_n)\}^2.$$

- Cuáles son los **w**, y el parámetro β que mejor explican los datos.

Máxima verosimilitud (III)

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

$$
\begin{aligned}
E_D(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2 \\
&= \frac{1}{2}\left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\right)^\top \left(\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\right),
\end{aligned}
$$

donde

$$
\boldsymbol{\Phi} = \begin{bmatrix} \phi(\mathbf{x}_1)^\top \\ \phi(\mathbf{x}_2)^\top \\ \vdots \\ \phi(\mathbf{x}_N)^\top \end{bmatrix} = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & & & \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix}
$$

Máxima verosimilitud (IV)

- La verosimilitud logarítmica está dada entonces como

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

- Se tiene entonces,

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

- Recordemos las siguientes derivadas

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

Máxima verosimilitud (V)

- Esto significa que

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

- La solución de máxima verosimilitud para **w** está dada como

$$\mathbf{w}_{ML} = (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \mathbf{t},$$

  donde $\mathbf{\Phi}^\dagger \equiv (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top$ es la pseudo-inversa Moore-Penrose.

- La solución de máxima verosimilitud para β se obtiene de

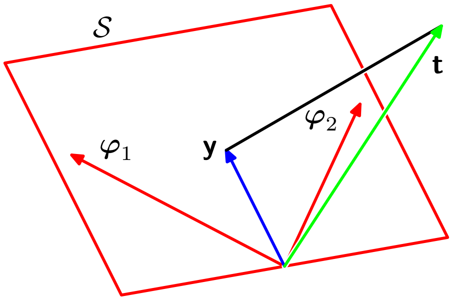
$$\frac{\partial \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta)}{\partial \beta} = \frac{N}{2\beta} - \frac{1}{2} (\mathbf{t} - \mathbf{\Phi} \mathbf{w})^\top (\mathbf{t} - \mathbf{\Phi} \mathbf{w}).$$

- Y así,

$$\frac{1}{\beta_{ML}} = \frac{1}{N} (\mathbf{t} - \mathbf{\Phi} \mathbf{w}_{ML})^\top (\mathbf{t} - \mathbf{\Phi} \mathbf{w}_{ML}) = \frac{1}{N} \sum_{n=1}^{N} \{ t_n - \mathbf{w}_{ML}^\top \phi(\mathbf{x}_n) \}^2.$$

Interpretación geométrica

Se quiere aproximar el vector **t** usando el vector **y**



La aproximación corresponde a la proyección ortogonal del vector **t** en el subespacio de los φ_i, donde φ_i es una columna de Φ.

Contenido

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

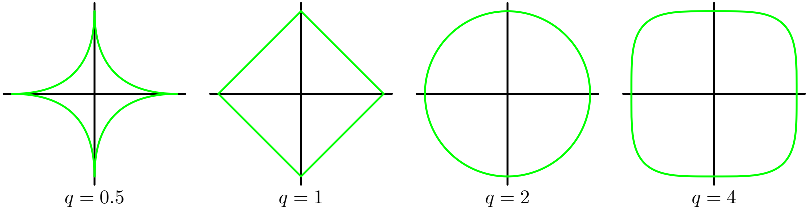
$$\mathbf{w} = (\lambda\mathbf{I} + \mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{t}.$$

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**.



Contenido

Laboratorio I

Python: máxima verosimilitud para regresión.

Contenido

Definiciones

- Una alternativa a la regularización es el tratamiento Bayesiano.

- Como hemos dicho, la verosimilitud del modelo está dada como
$$p(\mathbf{t}|\mathbf{w}, \beta) = \mathcal{N}(\mathbf{t}|\mathbf{\Phi w}, \beta^{-1}\mathbf{I}).$$

- 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

$$p(\mathbf{w}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{t})},$$

  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.

- 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

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

  donde

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

Un paréntesis: propiedades de la Gaussiana

Dadas una distribución Gaussiana marginal para **x**, y una distribución Gaussiana condicional para **y**, dado **x**, de la forma

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

la distribución marginal de **y**, y la distribución condicional de **x** dado **y** están dadas como

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

donde

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\top}\mathbf{L}\mathbf{A})^{-1}.$$

(*Demostración: pgs 90-93, Bishop, C. (2006)*).

Prior más simple

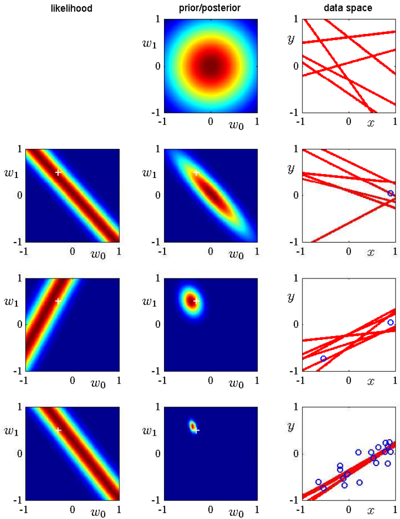
- Un prior más sencillo sigue la forma $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$.
- El posterior está dado como

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

donde

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

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

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

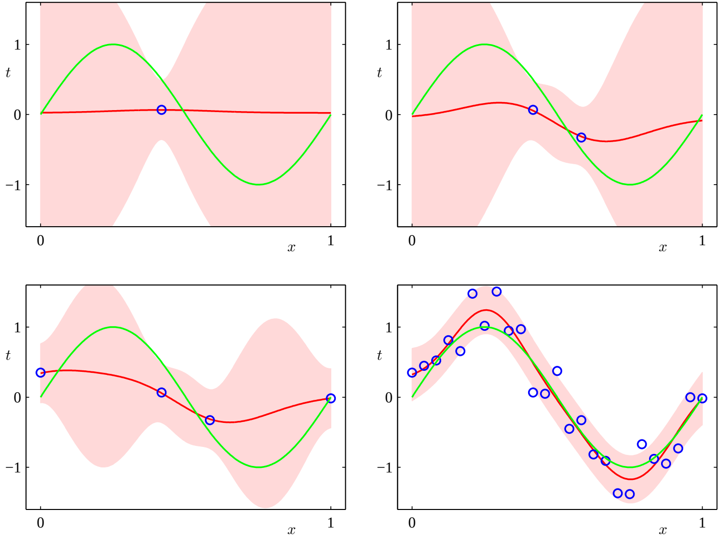
- Usando las propiedades de la Gaussiana (diapositiva anterior) se puede demostrar que

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

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

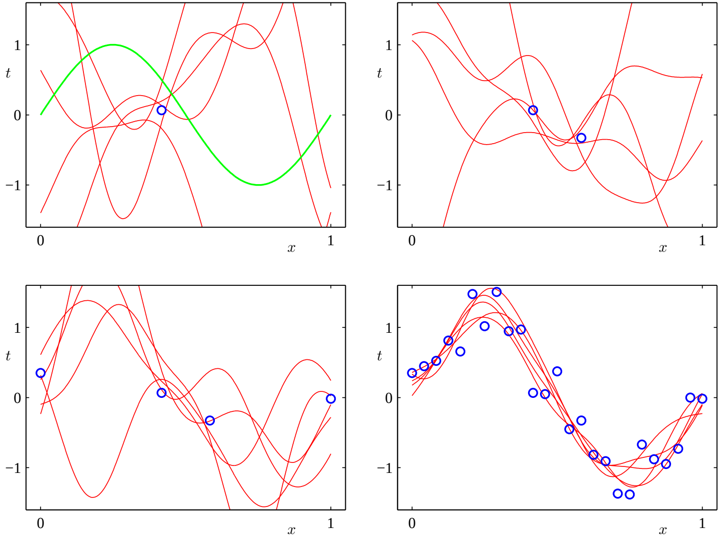
- *Importante*: nótese que se ha asumido que β y α son conocidos.

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

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

$$\text{evidencia} = \int \text{verosimilitud x prior}$$

- 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

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

Aproximación de la evidencia (III)

- Se quiere integrar sobre **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 $\mathbf{m}_N = \beta \mathbf{A}^{-1} \mathbf{\Phi}^\top \mathbf{t}$, $\mathbf{A} = \alpha \mathbf{I} + \beta \mathbf{\Phi}^\top \mathbf{\Phi}$, y

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

- 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

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

- 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

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

- En la expresión anterior

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

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

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

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

Maximización con respecto a α (II)

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

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

- Igualando a cero y despejando α se encuentra

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

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

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

- 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

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

- Igualando a cero y despejando β se obtiene

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

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

Contenido

Laboratorio I

Python: regresión lineal Bayesiana.

Contenido

Solución 1 (a)

- Se comienza con

$$
\begin{aligned}
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \\
p(\mathbf{t}|\mathbf{w}) &= \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \beta^{-1}\mathbf{I}).
\end{aligned}
$$

- Marginalizando **w**, usando propiedades de las Gaussianas multivariadas

$$p(\mathbf{t}|\alpha, \beta) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \beta^{-1}\mathbf{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\top}).$$

- El logaritmo $\ln p(\mathbf{t}|\alpha, \beta)$ está dado como

$$-\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|\beta^{-1}\mathbf{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\top}| - \frac{1}{2}\mathbf{t}^{\top}(\beta^{-1}\mathbf{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\top})^{-1}\mathbf{t}$$

Solución 1 (b)

- Se aplica la identidad de Woodbury, y el lemma del determinante de matrices

$$
\begin{aligned}
(\beta^{-1}\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^\top)^{-1} &= \beta\mathbf{I} - \beta\mathbf{I}\mathbf{\Phi}(\alpha\mathbf{I} + \beta\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top(\beta\mathbf{I}),\\
|\beta^{-1}\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^\top| &= |\alpha\mathbf{I} + \beta\mathbf{\Phi}^\top\mathbf{\Phi}||\alpha^{-1}\mathbf{I}||\beta^{-1}\mathbf{I}|
\end{aligned}
$$

- Reemplazando en la expresión del logaritmo

$$
\begin{aligned}
&-\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{A}| + \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta\\
&-\frac{1}{2}\mathbf{t}^\top\left[\beta\mathbf{I} - \beta\mathbf{I}\mathbf{\Phi}(\alpha\mathbf{I} + \beta\mathbf{\Phi}^\top\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top(\beta\mathbf{I})\right]\mathbf{t}
\end{aligned}
$$

Solución 1 (c)

- La expresión anterior se puede escribir como

$$\frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln 2\pi$$
$$-\frac{\beta}{2}\mathbf{t}^\top\mathbf{t} + \frac{1}{2}\mathbf{t}^\top(\beta\mathbf{I})\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\top(\beta\mathbf{I})\mathbf{t}$$

- Queda por demostrar que

$$\begin{aligned}
-E(\mathbf{m}_N) &= -\frac{\beta}{2}\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}_N\|^2 - \frac{\alpha}{2}\mathbf{m}_N^\top\mathbf{m}_N \\
&= -\frac{\beta}{2}\mathbf{t}^\top\mathbf{t} + \frac{1}{2}\mathbf{t}^\top(\beta\mathbf{I})\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\top(\beta\mathbf{I})\mathbf{t} \\
&= -\frac{\beta}{2}\mathbf{t}^\top\mathbf{t} + \frac{\beta^2}{2}\mathbf{t}^\top\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\top\mathbf{t}
\end{aligned}$$

Solución 1 (d)

- Recordemos que $\mathbf{m}_N = \beta \mathbf{A}^{-1} \boldsymbol{\Phi}^\top \mathbf{t}$.

- De esta forma

$$
\begin{aligned}
& -\frac{\beta}{2}\|\mathbf{t}-\boldsymbol{\Phi}\mathbf{m}_N\|^2-\frac{\alpha}{2}\mathbf{m}_N^\top\mathbf{m}_N \\
& -\frac{\beta}{2}\mathbf{t}^\top\mathbf{t}+\beta\mathbf{t}^\top\boldsymbol{\Phi}\mathbf{m}_N-\frac{\beta}{2}\mathbf{m}_N^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\mathbf{m}_N-\frac{\alpha}{2}\mathbf{m}_N^\top\mathbf{m}_N \\
& -\frac{\beta}{2}\mathbf{t}^\top\mathbf{t}+\beta\mathbf{t}^\top\boldsymbol{\Phi}\mathbf{m}_N-\frac{1}{2}\mathbf{m}_N^\top(\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}+\alpha\mathbf{I})\mathbf{m}_N \\
& -\frac{\beta}{2}\mathbf{t}^\top\mathbf{t}+\beta\mathbf{t}^\top\boldsymbol{\Phi}\beta\mathbf{A}^{-1}\boldsymbol{\Phi}^\top\mathbf{t}-\frac{\beta}{2}\mathbf{t}^\top\boldsymbol{\Phi}\mathbf{A}^{-\top}\mathbf{A}\beta\mathbf{A}^{-1}\boldsymbol{\Phi}^\top\mathbf{t} \\
& -\frac{\beta}{2}\mathbf{t}^\top\mathbf{t}+\beta^2\mathbf{t}^\top\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\top\mathbf{t}-\frac{\beta^2}{2}\mathbf{t}^\top\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\top\mathbf{t} \\
& -\frac{\beta}{2}\mathbf{t}^\top\mathbf{t}+\frac{\beta^2}{2}\mathbf{t}^\top\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^\top\mathbf{t}
\end{aligned}
$$