

Parcial 1: Teoría de Aprendizaje de Máquina

2024-I

Profesor: Andrés Marino Álvarez Meza, Ph.D.
Departamento de Ingeniería Eléctrica, Electrónica, y Computación
Universidad Nacional de Colombia - sede Manizales

1. Instrucciones

- Para recibir crédito total por sus respuestas, estas deben estar claramente justificadas e ilustrar sus procedimientos y razonamientos (paso a paso) de forma concreta, clara y completa.
- La componente teórica de cada uno de los puntos deberá entregarse a mano. La componente práctica (programación), debe trabajarse sobre Colaboratory o Kaggle. Enviar link de GitHub con la solución teórica y de simulación al correo electrónico amalvarezme@unal.edu.co antes de las 23:59 del 31 de marzo del 2024.
- Los códigos deben estar debidamente comentados en las celdas de código, y discutidos/explicados en celdas de texto (markdown). Códigos no comentados ni discutidos, no serán contabilizados en la nota final.

2. Preguntas

- 2.1 (Valor 2.5 puntos). Sea el modelo de regresión $t_n = \phi(\mathbf{x}_n)\mathbf{w}^\top + \eta_n$, con $\{t_n \in \mathbb{R}, \mathbf{x}_n \in \mathbb{R}^P\}_{n=1}^N$, $\mathbf{w} \in \mathbb{R}^Q$, $\phi : \mathbb{R}^P \rightarrow \mathbb{R}^Q$, $Q \geq P$, y $\eta_n \sim \mathcal{N}(\eta_n | 0, \sigma_\eta^2)$. Presente el problema de optimización (inferencia) y la solución del mismo para los modelos mínimos cuadrados, mínimos cuadrados regularizados, máxima verosimilitud, máximo a-posteriori, y Bayesiano con modelo lineal Gaussiano. Asuma datos i.i.d. Discuta las diferencias y similitudes entre los modelos estudiados.
- 2.2 (Valor 2.5 puntos) Genere una simulación sobre Python de los regresores por máxima verosimilitud y máximo a-posteriori, discutidos en el punto 2.1, para ajustar la señal: $t_n = \cos[x_n/3] + \cos[x_n/4] + \eta_n$, con $x_n \in [0, 24\pi]$, contaminada con ruido blanco Gaussiano η_n ($SNR_{dB} = 2[dB]$). Asuma mapeo $\phi(\cdot)$ del tipo polinomial de orden Q y prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | 0, \sigma_w^2)$. Simule 500 datos para entrenar los modelos y 200 para predecir. Incluya normalización por `MinMaxScaler()` de `sklearn` después de generar el mapeo no lineal.

* Minimos Cuadrados:

$$t_n = \phi(x_n) w^T + \eta_n, \text{ con } \left\{ t_n \in \mathbb{R}, x_n \in \mathbb{R}^P \right\}_{n=1}^N, w \in \mathbb{R}^Q, \phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q$$

$Q \geq P$, y $\eta_n \sim N(\eta_n | 0, \sigma_{\eta}^2)$

1. Despejamos el η_n :

$$\eta_n = t_n - \phi(x_n) w^T$$

* Problema de Optimización:

$$w^* = \operatorname{argmin}_w \frac{1}{N} \sum_{n=1}^{N-1} \|\eta_n\|_2^2$$

2. Reemplazamos η_n :

$$w^* = \operatorname{argmin}_w \frac{1}{N} \sum_{n=1}^{N-1} \|t_n - \phi(x_n) w^T\|_2^2$$

3. Abrir cuadrados:

$$\frac{\partial}{\partial w} \operatorname{argmin}_w \|t_n - \phi w^T\|_2^2$$

$$\frac{\partial}{\partial w} \operatorname{argmin}_w \langle t_n - \phi w^T, t_n - \phi w^T \rangle$$

$$\frac{\partial}{\partial w} \operatorname{argmin}_w t_n^T t_n - t_n^T \phi w^T - (\phi w^T)^T t_n + (\phi w^T)^T \phi w^T$$

$$\text{Por } (AB^T)^T = BA^T$$

$$\frac{\partial}{\partial w} \operatorname{argmin}_w t_n^T t_n - t_n^T \phi w^T - w \phi^T t_n + w \phi^T \phi w^T$$

$$\text{Por } (A \cdot B)^T = B^T A^T$$

$$\frac{\partial}{\partial w} \operatorname{argmin}_w t_n^T t_n - t_n^T \phi w^T - w \phi^T t_n + w \phi^T \phi w^T$$

$$\text{Por } A^T B = B^T A$$

$$\frac{\partial}{\partial w} \operatorname{argmin}_w (t_n^T t_n - 2 t_n^T \phi w^T + w \phi^T \phi w^T)$$

4. Derivando:

$$(t_n^T t_n - 2 t_n^T \phi w^T + w \phi^T \phi w^T) \quad \text{Se invierte por la regla de la cadena}$$

$$0 - 2 \phi^T t_n + 2 \phi^T \phi w^T = 0$$

$$f'(g(w)) \cdot g'(w)$$

$$2 \phi^T \phi w^T = 2 \phi^T t_n \rightarrow w (\phi^T \phi)^T = t_n^T \phi$$

$$w = t_n^T \Phi (\Phi^T \Phi)^{-1}$$

* Mínimos Cuadrados Regularizados:

$$t_n = \phi(x_n) w^T + \eta_n$$

$$\eta_n = t_n - \phi(x_n) w^T$$

$$W^* = \underset{w}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1} \| t_n - \phi(x_n) w^T \|_2^2$$

$$\text{Sujeto a: } \| w^T \|_2^2 < \alpha$$

Para poder calcular la norma a este vector y que no me de un producto externo, que me coincida con las dimensiones.

* Problema de Optimización:

$$W^* = \underset{w}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1} \| t_n - \phi(x_n) w^T \|_2^2 + \lambda \| w^T \|_2^2 \quad \text{Nota: } w \text{ es un vector fila}$$

2. Abrir cuadrados:

$$\frac{\partial}{\partial w} \underset{w}{\operatorname{argmin}} \langle t_n - \Phi w^T, t_n - \Phi w^T \rangle + \lambda \langle w^T, w^T \rangle$$

$$\frac{\partial}{\partial w} \underset{w}{\operatorname{argmin}} \langle t_n^T t_n - t_n^T \Phi w^T - (\Phi w^T)^T t_n + (\Phi w^T)^T \Phi w^T + \lambda (w^T w^T) \rangle$$

$$\frac{\partial}{\partial w} \underset{w}{\operatorname{argmin}} \langle t_n^T t_n - t_n^T \Phi w^T - w \Phi^T t_n + w \Phi^T \Phi w^T + \lambda (w^T w^T) \rangle$$

$$\frac{\partial}{\partial w} \underset{w}{\operatorname{argmin}} \langle t_n^T t_n - t_n^T \Phi w^T - w \Phi^T t_n + w \Phi^T \Phi w^T + \lambda (w^T w^T) \rangle$$

$$\frac{\partial}{\partial w} \underset{w}{\operatorname{argmin}} \langle t_n^T t_n - 2 t_n^T \Phi w^T + w \Phi^T \Phi w^T + \lambda (w^T w^T) \rangle$$

3. Derivando:

$$(t_n^T t_n - 2 t_n^T \Phi w^T + w \Phi^T \Phi w^T + \lambda (w^T w^T))$$

$$0 - 2 \Phi^T t_n + 2 \Phi^T \Phi w^T + 2 \lambda w^T = 0$$

$$\cancel{2 \Phi^T \Phi w^T} + \cancel{2 \lambda w^T} = \cancel{2 \Phi^T t_n}$$

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

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

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

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

* Máxima Verosimilitud :

Maximizar el Log de la $\prod_{n=1}^N P(t_n | \phi w^\top, \sigma^2)$ → $\sum_{n=1}^N \log P(t_n | \phi w^\top, \sigma^2)$ Datos iid

* Problema de Optimización :

$$W_{ML} = \arg \max_w \sum_{n=1}^N \log P(t_n | \phi w^\top, \sigma^2)$$

1. Despejando :

$$\eta_n = t_n - \phi(x_n) w^\top$$

2. Tenemos lo siguiente :

$$P(x|y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|x - x_n\|_2^2}{2\sigma^2}\right)$$

3. Reemplazando :

$$\prod_{n=1}^N 2 = 2 \cdot 2 \cdot 2 \cdot 2_n = 2^N$$

$$\log \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|t_n - \phi(x_n) w^\top\|_2^2}{2\sigma^2}\right)$$

4. Por Propiedades de los Log :

$$\log\left(\prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\prod_{n=1}^N \exp\left(-\frac{\|t_n - \phi(x_n) w^\top\|_2^2}{2\sigma^2}\right)\right)$$

$$\log\left(\frac{1}{(2\pi\sigma^2)^{N/2}}\right) + \log\left(\exp\left(-\sum_{n=1}^N \frac{\|t_n - \phi(x_n) w^\top\|_2^2}{2\sigma^2}\right)\right)$$

$$\log\left(\frac{1}{(2\pi\sigma^2)^{N/2}}\right) = \cancel{\log 1} - \log(2\pi\sigma^2) - N/2 \log(2\pi\sigma^2)$$

$$\log(2\pi\sigma^2)^{-N/2} - \sum_{n=1}^N \frac{\|t_n - \phi(x_n) w^\top\|_2^2}{2\sigma^2}$$

$$\prod_{n=1}^N \left(\frac{1}{(2\pi\sigma^2)^{1/2}}\right) = \left(\frac{1}{(2\pi\sigma^2)^{1/2}}\right)^N = \left(\frac{1}{(2\pi\sigma^2)^{N/2}}\right)$$

$$\prod_{n=1}^N \exp(x) = \exp^N \exp^{2...} = \exp \sum_{n=1}^N (x_1 + x_2 + x_3) = \exp \sum_{n=1}^N (x)$$

$$-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|t_n - \Phi w^\top\|_2^2$$

$$\frac{\partial}{\partial w} - \frac{N}{2} \left(\log(2\pi) \log(\sigma^2) \right) - \frac{\partial}{\partial w} \left(\frac{1}{2\sigma^2} \|t_n - \Phi w^\top\|_2^2 \right)$$

5. Abrir cuadros :

$$\frac{\partial}{\partial w} \left(\frac{1}{2\sigma^2} \|t_n - \Phi w^\top\|_2^2 \right)$$

$$\frac{\partial}{\partial w} \left(\frac{1}{2\sigma^2} (\langle t_n - \Phi w^\top, t_n - \Phi w^\top \rangle) \right)$$

$$\frac{\partial}{\partial w} \left(\frac{1}{2\sigma^2} (t_n^\top t_n - t_n^\top \Phi w^\top - (\Phi w^\top)^\top t_n + (\Phi w^\top)^\top \Phi w^\top) \right)$$

$$\frac{\partial}{\partial w} \left(\frac{1}{2\sigma^2} (t_n^\top t_n - t_n^\top \Phi w^\top - w \Phi^\top t_n + w \Phi^\top \Phi w^\top) \right)$$

$$\frac{\partial}{\partial w} \left(\frac{1}{2\sigma^2} (t_n^\top t_n - t_n^\top \Phi w^\top - w \Phi^\top t_n + w \Phi^\top \Phi w^\top) \right)$$

$$\frac{\partial}{\partial w} \left(\frac{1}{2\sigma^2} (t_n^\top t_n - 2t_n^\top \Phi w^\top + w \Phi^\top \Phi w^\top) \right)$$

$$\frac{1}{2\sigma^2} (t_n^\top t_n - 2t_n^\top \Phi w^\top + w \Phi^\top \Phi w^\top)$$

$$(-2\Phi^\top t_n + 2\Phi^\top \Phi w^\top = 0) \xrightarrow{\sigma^2} 0$$

$$2\Phi^\top \Phi w^\top = 2\Phi^\top t_n$$

$$(\Phi^\top \Phi w^\top)^\top = (\Phi^\top t_n)^\top \quad (A^\top B)^\top = B A^\top$$

$$w(\Phi^\top \Phi)^\top = t_n^\top \Phi$$

$$w = t_n^\top \Phi (\Phi^\top \Phi)^{-1}$$

* Máximo a Posteriori

Tenemos :

$$P(t_n | \Phi w^\top, \sigma^2) \rightarrow N(t_n | \Phi w^\top, \sigma^2)$$

$$\eta_n = t_n - \Phi(x_n) w^\top$$

Teniendo en cuenta :

$$P(w|t) = \frac{P(t|w)P(w)}{P(t)} \quad \begin{matrix} \xrightarrow{\text{Verosimilitud}} \\ \xrightarrow{\text{Prior}} \\ \xrightarrow{\text{Evidencia}} \end{matrix} \quad w \in \mathbb{R}^Q$$

Asumimos:

$$P(w|t, \phi, \sigma_w^2, \sigma_\eta^2) \propto P(t|\phi, w, \sigma_\eta^2) P(w|\sigma_w^2) \rightarrow \text{iid}$$

* Problema de Optimización:

$$w_{MAP} = \arg \max \log \left(\prod_{n=1}^N N(t_n | \phi w^\top, \sigma^2) \prod_{q=1}^Q N(w_q | 0, \sigma_w^2) \right)$$

Por Propiedades de los Log:

$$w_{MAP} = \arg \max \log \prod_{n=1}^N \left(\frac{1}{\sqrt{2\pi\sigma_\eta^2}} \exp \left(-\frac{\|t_n - \phi(x_n)w^\top\|_2^2}{2\sigma^2} \right) \right) + \log \prod_{q=1}^Q \left(\frac{1}{\sqrt{2\pi\sigma_w^2}} \exp \left(-\frac{\|w_q - 0\|_2^2}{2\sigma_w^2} \right) \right)$$

$$\log \prod_{n=1}^N \left(\frac{1}{\sqrt{2\pi\sigma_\eta^2}} \right) + \log \left(\prod_{n=1}^N \left(\exp \left(-\frac{\|t_n - \phi(x_n)w^\top\|_2^2}{2\sigma^2} \right) \right) \right) + \log \left(\prod_{q=1}^Q \left(\frac{1}{\sqrt{2\pi\sigma_w^2}} \right) \right) + \log \left(\prod_{q=1}^Q \left(\exp \left(-\frac{\|w_q - 0\|_2^2}{2\sigma_w^2} \right) \right) \right)$$

$$\log \left(\frac{1}{(2\pi\sigma_\eta^2)^{N/2}} \right) + \log \left(\exp \left(-\sum_{n=1}^N \frac{\|t_n - \phi(x_n)w^\top\|_2^2}{2\sigma^2} \right) \right) + \log \left(\frac{1}{(2\pi\sigma_w^2)^{Q/2}} \right) + \log \left(\exp \left(-\sum_{q=1}^Q \frac{\|w_q - 0\|_2^2}{2\sigma_w^2} \right) \right)$$

$$-\frac{N}{2} \log (2\pi\sigma_\eta^2) - \sum_{n=1}^N \frac{\|t_n - \phi(x_n)w^\top\|_2^2}{2\sigma^2} - \frac{Q}{2} \log (2\pi\sigma_w^2) - \sum_{q=1}^Q \frac{\|w_q\|_2^2}{2\sigma_w^2}$$

$$\frac{\partial}{\partial w} \left(-\frac{N}{2} \log (2\pi\sigma_\eta^2) - \sum_{n=1}^N \frac{\|t_n - \phi(x_n)w^\top\|_2^2}{2\sigma^2} - \frac{Q}{2} \log (2\pi\sigma_w^2) - \sum_{q=1}^Q \frac{\|w_q\|_2^2}{2\sigma_w^2} \right)$$

5. Abrir cuadrados:

$$\frac{\partial}{\partial w} \left(-\frac{1}{2\sigma^2} \|t_n - \phi w^\top\|_2^2 \right) - \frac{\partial}{\partial w} \left(\frac{1}{2\sigma_w^2} \|w\|_2^2 \right)$$

$$\frac{\partial}{\partial w} \left(-\frac{1}{2\sigma_\eta^2} (\langle t_n - \phi w^\top, t_n - \phi w^\top \rangle) \right) - \frac{\partial}{\partial w} \left(\frac{1}{2\sigma_w^2} \langle w, w \rangle \right)$$

$$\frac{\partial}{\partial w} \left(-\frac{1}{2\sigma_\eta^2} (t_n^\top t_n - t_n^\top \phi w^\top - (\phi w^\top)^\top t_n + (\phi w^\top)^\top \phi w^\top) \right) - \frac{\partial}{\partial w} \left(\frac{1}{2\sigma_w^2} (w^\top w) \right)$$

$$\frac{\partial}{\partial w} \left(-\frac{1}{2\sigma_\eta^2} (t_n^\top t_n - t_n^\top \phi w^\top - w^\top \phi^\top t_n + w^\top \phi^\top \phi w^\top) \right) - \frac{1}{2\sigma_\eta^2} (2\phi^\top w)$$

$$\frac{\partial}{\partial w} \left(-\frac{1}{2\sigma_\eta^2} (t_n^\top t_n - t_n^\top \phi w^\top - w^\top \phi^\top t_n + w^\top \phi^\top \phi w^\top) \right) - \frac{1}{\sigma_w^2} w^\top$$

$$\frac{\partial}{\partial w} \left(-\frac{1}{2\sigma_\eta^2} (t_n^\top t_n - 2t_n^\top \phi w^\top + w^\top \phi^\top \phi w^\top) \right) - \frac{1}{\sigma_w^2} w^\top$$

6. Derivando:

$$\begin{aligned}
 & -\frac{1}{2\sigma_\eta^2} \left(t_n^\top (n - 2t_n^\top \Phi w^\top + w^\top \Phi^\top \Phi w^\top) - \frac{1}{\sigma_w^2} w^\top \right) = 0 \\
 & \cancel{-\frac{1}{2\sigma_\eta^2} (-2\Phi^\top t_n + 2\Phi^\top \Phi w^\top)} - \frac{1}{\sigma_w^2} w^\top = 0 \\
 & \frac{1}{\sigma_\eta^2} \Phi^\top t_n - \frac{1}{\sigma_\eta^2} \Phi^\top \Phi w^\top - \frac{1}{\sigma_w^2} w^\top = 0 \\
 & \frac{1}{\sigma_\eta^2} (\Phi^\top \Phi w^\top) - \frac{1}{\sigma_w^2} w^\top = -\frac{1}{\sigma_\eta^2} \Phi^\top t_n \\
 & \frac{1}{\sigma_\eta^2} \left((\Phi^\top \Phi) + \frac{\sigma_\eta^2}{\sigma_w^2} \right) w^\top = \frac{1}{\sigma_\eta^2} \Phi^\top t_n \quad \text{Siendo: } \lambda = \frac{\sigma_\eta^2}{\sigma_w^2} \\
 & w \left((\Phi^\top \Phi) + \lambda I \right)^\top \left(\Phi^\top t_n \right)^\top \\
 & w = t_n^\top \Phi \left((\Phi^\top \Phi) + \lambda I \right)^{-1}
 \end{aligned}$$

* Bayesiano con modelo lineal Gaussiano

Sea el modelo de regresión

$$t_n = \Phi(x_n) w^\top + \eta_n$$

$$\text{con } \{ t_n \in \mathbb{R}, x_n \in \mathbb{R}^P \}_{n=1}^N$$

$$w \in \mathbb{R}^Q, \Phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q \quad Q \geq P$$

$$\eta_n \sim N(\eta_n | 0, \sigma_\eta^2)$$

$$\eta_n = t_n - \Phi(x_n) w^\top \sim N(t_n - \Phi(x_n) w^\top | 0, \sigma_\eta^2)$$

$$\text{Entonces: } P(t_n | \Phi(x_n) w^\top, \sigma_\eta^2) = N(t_n | \Phi(x_n) w^\top, \sigma_\eta^2) \rightarrow$$

$$P(t_n | w) = N(t_n | \Phi w, \beta' I) \quad \beta' = \sigma_\eta^2$$

$$\text{Ahora: } w \sim P(w) = N(w | m_0, S_0)$$

Sabemos que:

$$M_{x|y} = (\Delta + A^\top L A)^{-1} [A^\top L(y - b) + \Delta u] \quad y$$

$$\sum_{x \mid y} = (\Delta + A^T L A)^{-1}$$

Entonces nos da que:

$$M_N = M_{w \mid t} = (S_0^{-1} + \Phi^T B^{-1} B \Phi)^{-1} [\Phi^T B I(t=0) + S_0^{-1} m_0]$$

$$M_N = M_{w \mid t} = (S_0^{-1} + B \Phi^T \Phi)^{-1} [B \Phi^T t + S_0^{-1} m_0]$$

$$S_n = (S_0^{-1} + B \Phi^T \Phi)^{-1}$$

$$S_n^{-1} = S_0^{-1} + B \Phi^T \Phi$$

$$M_n = M_{w \mid t} = S_n (S_0^{-1} m_0 + B \Phi^T t)$$

$$P(w \mid t) = N(w \mid M_{w \mid t}, S_{w \mid t})$$

$$P(w \mid t) = N(w \mid S_n (S_0^{-1} m_0 + B \Phi^T t), S_n^{-1})$$

Conclusiones:

Después de realizar el problema de optimización y la solución para cada uno de los modelos, se pudo observar que:

* Mínimos Cuadrados:

Objetivo: Minimizar la suma de los cuadrados de la diferencia entre los valores observados (t_n) y los predichos por el modelo ($\phi(x_n) w^T$)

Método: Se encuentra la solución analítica que minimiza la función del error cuadrático.

Regularización: No incorpora términos de regularización para controlar el Sobreajuste.

* Mínimos Cuadrados Regularizados:

Objetivo: Minimizar la suma de los cuadrados de la diferencia entre los valores observados (t_n) y los predichos por el modelo ($\phi(x_n) w^T$)

Método: Se incluye el término de regularización en la función ($\lambda \|w\|_2^2$) que penaliza los coeficientes del modelo.

Regularización: Controla el Sobreajuste al introducir el término de penalización que reduce la magnitud de los coeficientes del modelo.

* Máximo a Posteriori (MAP):

Objetivo: Encontrar los Parámetros del modelo que maximizan la probabilidad a posteriori, dada la distribución previa de los Parámetros.

Método: Combina la Verosimilitud de los datos observados con la distribución previa de los parámetros para obtener la distribución posterior.

Regularización: Incorpora la regularización implícita mediante la elección de una distribución previa para los parámetros del modelo.

* Enfoque Bayesiano con Modelo Lineal Gaussiano:

Objetivo: Estimar la distribución posterior de los parámetros del modelo dada la distribución previa y los datos observados.

Método: Se utiliza el teorema de Bayes para combinar la información previa con la Verosimilitud de los datos observados y obtener la distribución posterior.

Regularización: Incorpora regularización de forma natural a través de la elección de una distribución previa para los parámetros, que actúa como regularización bayesiana.

Diferencias:

- * Como se incorpora la regularización y como se maneja la incertidumbre en los parámetros del modelo.
- * Mínimos cuadrados no incluye regularización, mientras los otros métodos considera algún tipo de regularización
- * ML, MAP y el enfoque Bayesiano toma en cuenta la incertidumbre a través de distribuciones de probabilidad, Mínimos cuadrados no la incorpora directamente.