

1.)

Modelo de regresión

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

$$\{t_n \in \mathbb{R}; x_n \in \mathbb{R}^{p \times N} \}_{n=1}^N; w \in \mathbb{R}^q; \phi \in \mathbb{R}^p \rightarrow \mathbb{R}^q$$

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

$$q \leq p$$

• Mínimos cuadrados

Pasamos el modelo a su forma matricial:

$$t = \phi w + \eta$$

$$\phi \in \mathbb{R}^{N \times q} = \text{Datos}; t \in \mathbb{R}^N = \text{Vector de salida}$$

$$w \in \mathbb{R}^q = \text{Modelo}$$

$$\text{Optimización } w^* = \arg \min_w \frac{1}{N} \|t - \phi w\|_2^2$$

Para encontrar el valor mínimo se deriva y se iguala a cero:

$$\|t - \phi w\|_2^2 = \langle t - \phi w, t - \phi w \rangle =$$

$$= (t - \phi w)^T (t - \phi w)$$

$$= t^T t - t^T \phi w - (\phi w)^T t + (\phi w)^T (\phi w)$$

$$\frac{\partial}{\partial w} \left\{ \frac{1}{N} [t^T t - 2t^T \phi w + (\phi w)^T (\phi w)] \right\} = 0$$

$$N \times \frac{1}{N} \left[\cancel{\frac{\partial}{\partial w} t^T t} - (2t^T \phi)^T + 2\phi^T \phi w \right] = 0 \times N^0$$

$$\frac{1}{N} \left[-2\phi^T t + 2\phi^T \phi w \right] = 0 \times \frac{1}{2}$$

$$-\phi^T t + \phi^T \phi w = 0$$

$$\phi^T \phi w = \phi^T t$$

$$= (\phi^T \phi)^{-1} (\phi^T t) w = (\phi^T \phi)^{-1} \phi^T t$$

$$w^* = (\phi^T \phi)^{-1} \phi^T t$$

• Mínimos cuadrados regularizados

$$W^* = \operatorname{argmin} \|f - \phi W\|_2^2 + \lambda \|W\|_2^2$$

$$W^* = \frac{\partial}{\partial W} \left\{ \|f - \phi W\|_2^2 + \lambda \|W\|_2^2 \right\} = 0$$

$$\Rightarrow -2 \phi^T f + 2 \phi^T \phi W + 2\lambda W = 0$$

$$(2 \phi^T \phi + 2\lambda I) W = 2 \phi^T f$$

$$W^* = (\phi^T \phi + \lambda I)^{-1} \phi^T f$$

• Máxima verosimilitud

$$t_n = \phi(x_n)^T W + n : n \sim N(0, \sigma_n^2)$$

$$P(t_n | x_n, W) = N(t_n | \phi(x_n)^T W, \sigma_n^2)$$

Dado que los datos son i.i.d

$$\prod_{n=1}^N P(t_n | \phi(x_n)^T W, \sigma_n^2)$$

$$\begin{aligned} \log \left(\prod_{n=1}^N P(t_n | \phi(x_n)^T W, \sigma_n^2) \right) &= \log \left(\prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \left[\frac{|t_n - \phi(x_n)^T W|^2}{2\sigma_n^2} \right] \right) \\ &= N \log \left(\frac{1}{\sqrt{2\pi\sigma_n^2}} \right) - \frac{1}{2\sigma_n^2} \sum_{n=1}^N |t_n - \phi(x_n)^T W|^2 \\ &= -\frac{N}{2} \log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2} \sum_{n=1}^N |t_n - \phi(x_n)^T W|^2 \end{aligned}$$

Se derivan y se iguala a cero, pero

W termina siendo igual que en mínimos cuadrados

$$W^* = \operatorname{argmin} \left(-\frac{1}{2\sigma_n^2} \|f - \phi W\|_2^2 - \underbrace{\frac{N}{2} \log(2\pi\sigma_n^2)}_{\text{No aporta}} \right)$$

En este caso bajo el supuesto ruido Gaussiano, maximizar la verosimilitud es equivalente a minimizar la suma de los errores al cuadrado.

$$W^* = (\phi^T \phi)^{-1} \phi^T f$$

• Máximo a-posteriori

Se asume un prior sobre \mathbf{W} y se busca maximizar la posterior, entonces se debe encontrar un valor que realice esto.

$$\mathbf{W}^* = \operatorname{argmax}_{\mathbf{W}} p(\mathbf{W} | \mathbf{t})$$

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

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

$p(\mathbf{t}) = 1$: Esto con el fin de simplificar

$$\begin{aligned} p(\mathbf{t} | \mathbf{w}) &= \prod_{n=1}^N N(t_n | \phi(x_n) \mathbf{w}, \sigma_n^2) & p(\mathbf{w}) &= N(\mathbf{w} | \mathbf{0}, \mathbf{J}_w^{-1}) \\ &= \log \left(\prod_{n=1}^N p(t_n | \phi_n \mathbf{w}, \sigma_n^2) \right) \prod_{q=1}^Q p(w_q | 0, \mathbf{J}_w^{q-1}) \\ &= \log \left(\prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \left(-\frac{|t_n - \phi_n \mathbf{w}|^2}{2\sigma_n^2} \right) \right) + \log \left(\prod_{q=1}^Q \frac{1}{\sqrt{2\pi\sigma_q^2}} \right) \\ &\quad + \log \left(\prod_{q=1}^Q \exp \left(-\frac{|w_q|^2}{2\sigma_q^2} \right) \right) \\ &= \underbrace{-\frac{N}{2} \log(2\pi\sigma_n^2)}_{\text{No aporta}} - \underbrace{\frac{Q}{2} \log(2\pi\sigma_q^2)}_{\text{No aporta}} - \frac{1}{2\sigma_w^2} \sum_{n=1}^N |t_n - \phi_n \mathbf{w}|^2 \\ &\quad - \frac{1}{2\sigma_w^2} \sum_{q=1}^Q |w_q|^2 \end{aligned}$$

$$\mathbf{W}^* = \min_{\mathbf{W}} \left[\| \mathbf{t} - \phi \mathbf{W} \|^2 + \frac{2\sigma_w^2}{2\sigma_w^2} \|\mathbf{W}\|^2 \right] \quad \text{Min. cuadrados regularizados}$$

$$\mathbf{W}^* = (\phi^T \phi + \frac{\sigma_w^2}{2\sigma_w^2} \mathbb{I})^{-1} \phi^T \mathbf{t}$$

• Bayesiano con modelo lineal Gaussiano

En este modelo además de buscar un vector \mathbf{w} , se busca también la distribución completa a posteriori sobre \mathbf{w} dados los datos observados

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

$$\mathbf{E}_N = \text{Matriz de covarianza} \quad \mathbf{E}_N = \left(\frac{1}{\sigma_w^2} \mathbb{I} + \frac{1}{\sigma_w^2} \phi^T \phi \right)$$

$$m_N = \frac{1}{g_n^2} \sum N \phi^T \phi$$

Optimización: Maximizar la evidencia

$$p(t) = \int p(t|w) p(w) dw$$

Esto se demuestra:

Sea un vector $x \in \mathbb{R}^q$ con prior Gaussian

$$p(x) = N(x|\mu, \Sigma^{-1})$$

Sea el modelo $y = Ax + b$; $p(y|x) = N(y|Ax+b, L^{-1})$

$$p(x,y) = p(x)p(y|x); \quad p(x|y) = N(x|\mu_{x|y}, \Sigma_{x|y})$$

$$\Sigma = (\Delta + A^T L A)^{-1}; \quad \mu_{x|y} = \sum_{x|y} (A^T L (y - b) + \Delta \mu)$$

Para este caso

$$t_n = \phi(x_n) w^T + h_n \quad t = \phi w + h; \quad p(t|w) = N(t|\phi w, g_n^2)$$

Se asume $p(w) = N(w|m_0, S_0)$

$$\text{Se sabe que } \log p(t|w) = -\frac{1}{2g_n^2} \|t - \phi w\|_2^2 + Cte.$$

$$\log(N(x|\mu, \Sigma)) = -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) - \frac{1}{2} \log |\Sigma| - \frac{n}{2} \log(2\pi)$$

$$\begin{aligned} \text{Por tanto } \log(p(w)) &= -\frac{1}{2} (w-m_0)^T S_0^{-1} (w-m_0) + Cte \\ \log(p(t|w) p(w)) &= \log(p(t|w)) + \log(p(w)) \\ &= -\frac{1}{2} \left[\frac{1}{g_n^2} (t - \phi w)^T (t - \phi w) + (w-m_0)^T S_0^{-1} (w-m_0) \right] + Cte \end{aligned}$$

Se agrupan términos

$$w^T \left(\frac{1}{g_n^2} \phi^T \phi + S_0^{-1} \right) w - 2w^T \left(\frac{1}{g_n^2} \phi^T t + S_0^{-1} m_0 \right)$$

$$\text{Se sabe que } \log(p(w)) = -\frac{1}{2} (w-m_0)^T S_0^{-1} (w-m_0)$$

$$= -\frac{1}{2} (w^T S_0^{-1} w - 2w^T S_0^{-1} m_0 + m_0^T S_0^{-1} m_0) + Cte$$

$$S_N = \left(\frac{1}{g_n^2} \phi^T \phi + S_0^{-1} \right)^{-1} \quad S_N^{-1} m_N = \left(\frac{1}{g_n^2} \phi^T t + S_0^{-1} m_0 \right)$$

$$m_N = S_N (S_0^{-1} m_0 + \frac{1}{g_n^2} \phi^T t)$$

Como $P(w) = N(w | m_0, S_0) = N(w | 0, g_n^2 I)$

$$S_N = \left(\frac{1}{g_n^2} I + \frac{1}{g_n^2} \phi^T \phi \right)^{-1} \quad m_N = \frac{1}{g_n^2} S_N \phi^T t$$

$$m_N = \left(\frac{1}{g_n^2} I \right) \left(\frac{1}{g_n^2} \right)^{-1} \left(\frac{g_n^2}{g_n^2} I + \phi^T \phi \right)^{-1} \phi^T t = \left(\frac{g_n^2}{g_n^2} I + \phi^T \phi \right)^{-1} \phi^T t$$

• Regresión rígida Kernel

Se extiende a espacios de funciones no lineales

$$\hat{y} = \phi^T w \quad q \in \mathbb{R}^{N \times Q} \quad Q \rightarrow \infty \quad (\text{RKHS})$$

$$y = RQ \rightarrow R$$

$$w^* = \arg \min \frac{1}{N} \|t - \phi^T w\|_2^2 + \lambda \|w\|_F^2$$

por mínimos cuadrados $w^* = (\phi^T \phi + \lambda I)^{-1} \phi^T t$

sin embargo $\phi^T \phi \in \mathbb{R}^{Q \times Q} \quad Q \rightarrow \infty$

$$(\phi^T \phi + \lambda I)^{-1} \phi^T = (\lambda (I + \phi^T \phi)^{-1})^{-1} \phi^T = \frac{1}{\lambda} (I + \phi^T \phi)^{-1} \phi^T$$

$$\phi^T (I + \phi^T \phi)^{-1} = \phi^T (\lambda I + \phi \phi^T)^{-1}$$

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

se hace la predicción para nuevos puntos

$$f(x_*) = \phi^T (x_*)^T w^*$$

$$f(x_*) = \phi^T (x_*)^T \phi^T (\lambda I + \phi \phi^T)^{-1} t$$

$$K = \phi \phi^T; \quad K_* = \phi^T (x_*)^T \phi^T$$

$$f(x_*) = [K_* (t)^T] (K + \lambda I)^{-1} t$$

• Gaussian process

Extiende el método paramétrico para definir la incertidumbre de los parámetros del regresor al imponer un prior sobre fracciones directamente en RKHS

El GP se va a definir por su media y covarianza

$$f(x) \in \mathbb{R} ; f(x) = \phi(x)^T w ; p(w) = N(w|0, \Sigma_w) \\ \Sigma_w \in \mathbb{R}^{R \times Q}$$

$$m(x) = E\{f(x)\} = E\{\phi(x)^T w\} = \phi(x)^T E\{w\} = 0$$

$$\text{Cov}(f(x), f(x')) = k(x, x') = \phi(x)^T \Sigma_w \phi(x')$$

$$f \sim \text{GP}(f|0, K) ; K = [k(x_i, x_j)] \in \mathbb{R}^{N \times N}$$

Este modelo busca maximizar $\alpha_p(t|\theta)$, donde θ es un hiperparámetro de $K(t^*, \cdot | \theta)$

Se obtiene que

$$m(x_*) = K_*^T (K + \sigma_e^2 I)^{-1} t$$

$$[K]_{\bullet} = [k(x_*, x_1), k(x_*, x_2), \dots, k(x_*, x_N)]^T$$

$$\text{Cov}(f(x), f(x')) = k(x_*, x_*) + \sigma_e^2 - K_*^T (K + \sigma_e^2 I)^{-1} K_*$$

Diferencias claves

- Respuesta única vs incertidumbre:

- Los modelos frequentistas (LS, Ridge, Kernel Ridge) dan una sola respuesta ("el mejor ajuste")
- Los modelos Bayesianos (Bayesiano lineal, GP) dan un rango de respuestas posibles con sus probabilidades, lo que permite medir la incertidumbre de las predicciones.

- Flexibilidad rígida vs. adaptable:

- Los modelos paramétricos (LS, Ridge, Bayesiano lineal) tienen una complejidad rígida.
- Los modelos no-paramétricos (Kernel Ridge, GP) usan el "truco del kernel" para adaptar su complejidad a los datos, haciéndolos mucho más flexibles para patrones complejos

- Pares relacionados
 - Ridge es la versión regularizada de mínimos cuadrados.
 - MAP es la versión simplificada y no probabilística de un modelo Bayesiano, que termina siendo igual a Ridge.
 - Un Proceso Gaussiano es, en esencia, la versión Bayesiana y con incertidumbre de Kernel Ridge.

Regresor	Linear Regression	Lasso	Elastic Net	Kernel Ridge	SGDRegressor
Modelo matemático	$y = w^T x + b$	$y = w^T x + b$	$y = w^T x + b$	$f(x) = \sum_{n=1}^N \alpha_n K(x_n, x)$	$y = w^T x + b$
Función de costo	$J(w) = \ Xw - y\ _2^2$	$J(w) = \ Xw - y\ _2^2 + \alpha \ w\ $	$J(w) = \ Xw - y\ _2^2 + \alpha p \ w\ _1 + \frac{\alpha(1-p)}{2} \ w\ _2^2$	$J(\alpha) = \ y - K\alpha\ _2^2 + \lambda \ \alpha\ _2^2$	$J(w) = L(y_n, \hat{y}_n)$ $\alpha \propto B(w)$
Estrategia de optimización	Usa mínimos cuadrados ordinarios (OLS). Resuelve la ecuación normal $w = (X^T X)^{-1} X^T y$	Usa descenso por coordenadas (coordenadas). Igual que en ElasticNet	Usa descenso por coordenadas. Esta estrategia optimiza un coeficiente a la vez, manteniendo los demás fijos, y repite.	Tiene una solución analítica, pero en el espacio dual. Implica invertir la matriz del kernel para encontrar α .	Descenso del gradiente estocástico. SGD usa un solo dato o un pequeño lote a la vez.
Relación con esquemas básicos	Mínimos cuadrados	Mínimos cuadrados regularizados	Mínimos cuadrados regularizados	Regresión rígida Kernel. $\alpha = \operatorname{argmin}_{\alpha} (\ y - K\alpha\ _2^2 + \lambda \ \alpha\ _2^2)$	Se adapta para minimizar la función de costo. $w \leftarrow w - \eta \frac{\partial L}{\partial w} + \frac{\partial L(y_n, w^T x_n)}{\partial w}$
Complejidad	$O(NP)$	$O(NP \cdot \text{Iteraciones})$ Escala bien para datasets grandes en ambas dimensiones	$O(NP \cdot \text{Iteraciones})$	$O(N^2 P + N^3)$ inviable para datasets con más de unos pocos miles de muestras. Su principal cuello de botella es la creación e inversión de la matriz del Kernel.	$O(NP \cdot \text{épocas})$ Es la mejor opción para escalar a un número muy grande de muestras (N)
Escalabilidad (N: Muestras, P: Características)	$O(NP^2)$ Se vuelve muy lento con millones de muestras pero pocas características.				

Bayesian Ridge	Gaussian Process	SVR	RandomForest	Gradient Boosting / XG Boost
$y = \mathbf{w}^T \mathbf{x} + b$	$f(x) \sim GP(m(x), k(x, x'))$	$f(x) = \mathbf{w}^T \mathbf{q}(x) + b$	Ensemble de árboles de decisión (Promedio)	Ensamble de árboles de decisión (construidos secuencialmente)
$\max_{\theta} \log P(y X, \theta)$. Encontrar los hiperparámetros	$\max_{\theta} \log P(f X, \theta)$, encontrar los hiperparámetros	$J(\mathbf{w}, \mathbf{\xi}) = \frac{1}{2} \ \mathbf{w}\ _2^2 + C \sum_{n=1}^N (\mathbf{\xi}_n + \mathbf{\xi}_n^*)$	$MSE = \frac{1}{N} \sum_n (y_n - \hat{y}_n)^2$ Criterio de división	$Obj = \sum_{n=1}^N \ell(y_n, \hat{y}_n) + \sum_{k=1}^K \Omega(f_k)$
Su estrategia es la inferencia Bayesiana. Utiliza la regla de Bayes para calcular la probabilidad posterior completa.	Su estrategia es ajustar los hiperparámetros de su kernel para maximizar la Verosimilitud Marginal logarítmica.	Su estrategia es resolver un problema de optimización convexa mediante programación cuadrática.	Su método es, la construcción independiente de un ensamble de Árboles. Cada árbol se entrena de forma aislada.	Su estrategia es el impulso del gradiente. Es un método secuencial y aditivo que realiza un descenso de gradiente en el espacio de las funciones.
Regla de Bayes $P(w y, X) \propto P(y X, w) \cdot P(w)$ Prior posterior	Media predictiva $E[f_*] = k_*^T (K + g_n g_n^T)^{-1} y$ Variancia verosimilitud	$Obj = \min_{\mathbf{w}} \frac{1}{2} \ \mathbf{y} - \mathbf{k}^T \mathbf{w}\ ^2 + \lambda \ \mathbf{w}\ _2^2$	Pertenece a una familia de algoritmos completamente distintos.	Pertenece a una familia de algoritmos completamente distintos.
$O(p^2 N + p^3)$ Escala bien con el número de muestras (N), pero su rendimiento se degrada rápidamente con un número alto de características (p)	$O(N^3 + N^2 p)$ Escala muy mal con el número de muestras (N) debido a la inversión de la matriz del Kernel.	$O(N^2 p)$ y $O(N^3)$ Escala muy mal con el número de muestras (N).	$O(B \cdot N \log(N) \cdot p)$ Escala excelentemente tanto con muestras (N) como con características (p).	$O(B \cdot N \cdot R \cdot d_{max})$ Escala aunque su entrenamiento es secuencial, sus implementaciones están optimizadas. A menudo es más rápido y eficiente para el manejo de datos.