PARCIAL 1

Modelos de Regresión:

$w \in \mathbb{R}^Q$; donde $\phi: \mathbb{R}^P \to \mathbb{R}^Q$

target Datos

$t_n = \underbrace{\phi(x_n)}_{\text{Vector de funciones}} w^T + \eta_n$; $\{t_n \in \mathbb{R}, x_n \in \mathbb{R}^P\}_{n=1}^N$: $\eta_n \sim N(\eta_n | 0, \sigma_n^2)$

Mínimos Cuadrados:

$\mathcal{E} = \arg\min_w \|t - \phi w^T\|^2$

Se minimiza

$\frac{\partial \mathcal{E}}{\partial w} = \langle t - \phi w^T, t - \phi w^T \rangle \longrightarrow = (t - \phi w^T)^T (t - \phi w^T)$

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

$\frac{\partial}{\partial w}(\|t\| - (2t^T \phi w^T)^T + \phi^T (w^T)^T (\phi w^T)) = 0$

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

$\phi^T \phi w = \phi^T t \longrightarrow \underbrace{w^*}_{\text{Pesos}} = (\phi^T \phi)^{-1} \phi^T t$

Mínimos Cuadrados Regularizados: (Ridge Regression)

Son los mín. Cuadrados con regularización ℓ_2

Penaliza pesos grandes.

$\mathcal{E} = \arg\min_w \|t - \phi w^T\|_2^2 + \lambda \|w\|_2^2$ → Regularización ($2 \to \ell_2$)

$\frac{\partial \mathcal{E}}{\partial w} = (t - \phi w^T)^T (t - \phi w^T) + \lambda w w^T$

$\frac{\partial}{\partial w}(\|t\| - 2t\phi^T w + \phi^T w \phi w^T + \lambda w w^T) = 0$

$0 - 2t\phi^T + 2\phi^T \phi w + 2\lambda w^T = 0$

$\phi^T \phi w + \lambda w = \phi^T t \longrightarrow \underbrace{w^*}_{\text{Pesos grandes}} = (\phi^T \phi + \lambda \mathbb{I})^{-1} \phi^T t$

Máxima Verosimilitud (MLE):

Es un modelo probabilístico, describe t_n según la entrada x_n

W se describe según el ruido η_n

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

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

distribución Normal para t_n

$P(t_n | \phi(x_n) w^T, \sigma_n^2) = N(t_n | \underbrace{\phi(x_n) w^T}_{\mu_x}, \underbrace{\sigma_n^2}_{\sigma_x^2})$

$$\underbrace{P(t_n | \phi(x_n) w^T, \sigma_n^2) = \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{\left(-\frac{(t_n - \phi(x_n) w^T)^2}{2\sigma_n^2}\right)}}_{\text{func. de densidad para cada } t_n}$$

$$W_{ML} = \arg\max_{w, \sigma_n^2} \log \left(\prod_{n=1}^N N(t_n | \phi(x_n) w^T, \sigma_n^2) \right)$$

$$L(w, \sigma_n^2) = \prod_{n=1}^N \left(\frac{1}{\sqrt{2\pi\sigma_n^2}} e^{\left(-\frac{(t_n - \phi(x_n) w^T)^2}{2\sigma_n^2}\right)} \right)$$

Observaciones independientes → iid, por eso $\prod$

$$L(w, \sigma_n^2) = \log \left(\prod_{n=1}^N \left(\frac{1}{\sqrt{2\pi\sigma_n^2}} e^{\left(-\frac{(t_n - \phi(x_n) w^T)^2}{2\sigma_n^2}\right)} \right) \right)$$

Con $\log \prod_{n=1}^N = \sum_{n=1}^N$

. . . . .

. . . . .

$$L(w, \sigma_n^2) = \sum_{n=1}^N \left(-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_n^2) - \frac{(t_n - \phi(x_n) w^T)^2}{2\sigma_n^2} \right)$$

$$L(w, \sigma_n^2) = -\frac{N}{2} \log(2\pi) - \underbrace{\frac{N}{2} \log(\sigma_n^2)}_{\text{Para } \sigma_n^2 \text{ grandes}} - \frac{1}{2\sigma_n^2} \underbrace{\sum_{n=1}^N (t_n - \phi(x_n) w^T)^2}_{\text{Error Cuadrático respecto } w_n, \sigma_n^2}$$

- Maximización respecto W_{ML} (Pesos óptimos):

$$\frac{\partial L(w, \sigma_n^2)}{\partial w} = 0 - 0 - \frac{\partial}{\partial w} \left(\underbrace{\frac{1}{2\sigma_n^2} \|t_n - \phi(x_n) w^T\|_2^2}_{= \text{ Mínimos Cuadrados}} \right)$$

$$= -\frac{1}{2\sigma_n^2} (t_n - \phi(x_n) w^T)^T (t_n - \phi(x_n) w^T)$$

$$= -\frac{1}{2\sigma_n^2} \left(t^T t - t^T \phi(x_n) w^T - t_n \phi(x_n)^T w + \phi(x_n)^T (w^T)^T \phi(x_n) w^T \right)$$

$$= \frac{\partial}{\partial w} \left(-\frac{1}{2\sigma_n^2} \left(\|t\|^2 - 2t_n \phi(x_n)^T w + \phi(x_n)^T w \phi(x_n) w^T \right) \right) = 0$$

$0 - 2\phi(x_n)^T t + 2\phi(x_n)^T \phi(x_n) w = 0$

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

Pesos óptimos: $W_{ML} = (\phi(x_n)^T \phi(x_n))^{-1} \phi(x_n)^T t$

- Maximizamos respecto σ_{ML}^2 (Varianza óptima)

$$\frac{\partial L(w, \sigma_n^2)}{\partial \sigma_n^2} = 0 - \frac{N}{2\sigma_n^2} + \frac{1}{2(\sigma_n^2)^2} \|t_n - \phi(x_n w^T)\|_2^2 = 0$$

$$\frac{\|t_n - \phi(x_n) w^T\|^2}{2(\sigma_n^2)^2} = \frac{N}{2\sigma_n^2}$$

$\sigma_n^2 = \frac{1}{N} \|t_n - \phi(x_n) w^T\|^2$ → Promedio Er residual

Varianza óptima: $\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N \underbrace{(t_n - \phi(x_n) w^T)^2}$ → Distancia o diferencia entre predicciones y target

Máximo a-posteriori: (MAP)

Con prior $P(w)$

Maximiza distribución a-posteriori $P(w|t)$

$P(t_n | \phi(x_n) w^T, \sigma_n^2) = \overbrace{N(t_n | \phi(x_n) w^T, \sigma_n^2)}^{\text{Gaussiana}}$

$P(t_n|w)$

$$P(t_n | \phi(x_n) w^T, \sigma_n^2) = \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{\left(-\frac{(t_n - \phi(x_n) w^T)^2}{2\sigma_n^2}\right)}$$

$P(w) = N(w | 0, \underbrace{\sigma_w^2 \mathbb{I}}_{})$ → Pesos indep. con var. σ_w^2 Controla regularización

$$P(w) = \frac{1}{\sqrt{2\pi\sigma_w^2}} e^{\left(-\frac{\|w\|_2^2}{2\sigma_w^2}\right)}$$

El posteriori $P(w|t_n)$ se maximiza con verosimilitud y prior del teorema de Bayes.

$$P(w|t_n) = \frac{P(t_n|w)P(w)}{P(t_n)}$$ → Maximizamos sólo numerador (Maximo a-posteriori)

Para pasar a Σ

$$W_{MAP} = \arg\max_w \log\left(\prod_{n=1}^{N} N(t_n|\phi(x_n)W^T, \sigma_n^2)\prod_{q=1}^{Q} N(w_q|0,\sigma_w^2)\right)$$

$$\log P(w|t_n) = \underbrace{\log P(t_n|w)}_{\text{Verosimilitud}} + \underbrace{\log P(w)}_{\text{Prior}}$$ → Procede verosimilitud de MLE

$$\frac{d\log P(w|t_n)}{dw} = \frac{d}{dw}\left(-\frac{1}{2\sigma_n^2}\|t-\phi(x_n)w^T\|_2^2 + cte\right) + \frac{d\log P(w)}{dw}$$

$$\frac{d\log P(w)}{dw} = \frac{d}{dw}\left(\sum_{q=1}^{Q}\left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma_w^2 - \frac{\|w\|_2^2}{2\sigma_w^2}\right)\right)$$

Asumimos iid:

Maximizar $\log P(w|t)$ = min.

$$W_{MAP} = \arg\max_w\left(-\frac{1}{2\sigma_n^2}\|t-\phi(x_n)w^T\|_2^2 - \frac{1}{2\sigma_w^2}\|w\|_2^2\right)$$

$$W_{MAP} = \arg\max_w\left(\|t_n-\phi(x_n)W^T\|_2^2 + \frac{2\sigma^2}{2\sigma_w^2}\|w\|_2^2\right)$$

$$W_{MAP} = \arg\max_w\left(\|t_n-\phi(x_n)W^T\|_2^2 + \lambda\|w\|_2^2\right)$$ → Mismo result. de min. cuadrados regularizados

$$W^*_{MAP} = (\phi(x_n)^T\phi(x_n)+\lambda I)^{-1}\phi(x_n)^T t_n$$

Bayesiano con modelo lineal Gaussiano:

No trata de encontrar un único valor para W_{ML} como MAP, sino que se calcula toda la distribución.

$$P(w|t_n,x_n) = N(w|\underbrace{W_{MAP}}_{\text{Punto de MAP}}, \underbrace{\Sigma_{MAP}}_{\text{Covarianza a-posteriori}})$$ → Distribución Posteriori de W en Bayesiano Gaussiano

$$\Sigma_{MAP} = S_N = \left(S_0^{-1} + \frac{1}{\sigma_n^2}\phi^T\phi\right)^{-1}$$ → Captura la Incertidumbre

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

Se asume $m_0=0$ → Punto de inicio

Matriz de Covarianza } S_0 → Icertidumbre inicial sobre W; Correlaciones entre pesos

$P(w) = N(w|m_0,S_0)$ → Prior sobre pesos (W)

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

$P(w|t_n,x_n) = P(t_n|w)\cdot P(w)$

$$= \frac{1}{\sqrt{2\pi\sigma^2_{MAP}}}\, e^{\left(\underbrace{\underbrace{-\frac{1}{2\sigma_n^2}\|t_n-\phi w^T\|_2^2}_{\text{Min. Cuadrados}} - \frac{1}{2}\|w-w_0\|_2^2}_{\text{Min. Cuadrados Regularizados}}\right)}$$

$$P(w|t_n,x_n) = -\frac{1}{2\sigma_n^2}\left(\|t_n\|_2^2 - 2t_n\phi(x_n)^T W + \phi(x_n)^T W\phi(x_n)W^T\right)$$

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

• • • • •

• • • • •

$$= -\frac{1}{2} S_0^{-1}(w-m_0)^T(w-m_0) = -\frac{1}{2}\left(W^TS_0^{-1}W - (W^T)S_0^{-1}(m_0)^T - m_0^T S_0^{-1}W + S_0^{-1}m_0^Tm_0\right)$$

$$= -\frac{1}{2}\left(\underbrace{S_0^{-1}W^TW}_{\text{Cuadrático}} - \underbrace{2S_0^{-1}m_0^TW}_{\text{Lineal}} + \underbrace{S_0^{-1}m_0^Tm_0}_{\text{cte}}\right)$$

Optimizamos $P(w|t_n,x_n)$ respecto W:

$$\frac{dP(w|t_n,x_n)}{dw} = -\frac{1}{2\sigma_n^2}\left(0 - 2t_n\phi(x_n)^T + 2\phi(x_n)^T\phi(x_n)W\right) - \frac{1}{2}\left(2S_0^{-1}W - 2S_0^{-1}m_0^T + 0\right)$$

$$-\frac{1}{2}\left(-\frac{2}{\sigma_n^2}t_n\phi(x_n)^T + \frac{2\phi(x_n)^T\phi(x_n)W}{\sigma_n^2} - 2S_0^{-1}W - 2S_0^{-1}m_0^T\right) = 0$$

$$\underbrace{\frac{\phi(x_n)^T\phi(x_n)W}{\sigma_n^2} + S_0^{-1}W}_{\text{Cuadrático}} = \underbrace{\frac{t_n\phi(x_n)^T}{\sigma_n^2} + S_0^{-1}m_0^T}_{\text{Lineal}}$$

$$W\left(S_0^{-1}I + \frac{\phi(x_n)^T\phi(x_n)}{\sigma_n^2}\right) = \frac{t_n\phi(x_n)^T}{\sigma_n^2} + S_0^{-1}m_0^T$$

$$W = \underbrace{\left(S_0^{-1}I + \frac{\phi(x_n)^T\phi(x_n)}{\sigma_n^2}\right)^{-1}}_{S_N}\underbrace{\left(S_0^{-1}m_0^T + \frac{t_n\phi(x_n)^T}{\sigma_n^2}\right)}_{M_N}$$ → Pesos óptimos

$$P(w) = N(w|0,\sigma_w^2) \longrightarrow \frac{P(w|t_n,x_n)}{P(w|t_n)} = N(w|\bar{m}_N, \bar{S}_N)$$

$$W = \underbrace{\left(\frac{1}{\sigma_w^2}I + \frac{\phi(x_n)^T\phi(x_n)}{\sigma_n^2}\right)^{-1}}_{\bar{S}_N}\frac{\phi(x_n)^Tt_n}{\sigma_n^2} = \bar{m}_N$$

$$\bar{S}_N = \left(\frac{1}{\sigma_n^2}\right)^{-1}\left(\frac{\sigma_n^2}{\sigma_w^2}I + \frac{\phi(x_n)^T\phi(x_n)}{\sigma_n^2}\right)^{-1}\phi(x_n)^Tt_n$$

$$W = \left(\frac{\phi(x_n)^T\phi(x_n)}{\sigma_n^2} + \lambda I\right)^{-1}\phi(x_n)^Tt_n$$ → ~Min. cuadrados Reg. $\lambda = \frac{\sigma_n^2}{\sigma_w^2}$

Kernel Ridge: (KRR)

Para No linealidades

Mapeo x_n a un espacio de dimensión mayor con $\phi(x_n)$

Mejora:

$$K(x,x') = \phi(x)^T\phi(x')$$

Partimos de Bayesiano completo:

$P(w|t_n) = \|t_n-\phi(x_n)W^T\|_2^2 + \lambda\|w\|_2^2$

$W^* = \arg\min_w \|t_n - \phi(x_n)W^T\|_2^2 + \lambda\|w\|_2^2$

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

$W^* = \underbrace{(\underbrace{\phi^T(x_n)\phi(x_n)}_{K(x,x')}+\lambda I)^{-1}t_n}_{\alpha_i}\underbrace{\phi^T(x_n)}_{\phi(x_i)^T}$

$W^* = \sum_i \alpha_i\phi(x_i)^T$

Necitamos predicción para un nuevo punto: Proyectando sobre solución W^*

$$t_* = \phi(X_*)^T W^*$$

$$t_* = \underbrace{\phi(X_*)^T \phi^T(X_n)}_{K(*)^T} \underbrace{(\overbrace{\phi^T(X_n)\phi(X_n)}^{K(X,X')} + \lambda \mathbb{I})^{-1} t_n}_{\alpha_i}$$

$$t_* = (K + \lambda \mathbb{I})^{-1} t_n K(*)^T \quad K(*)^T = [K(X_*, X_1), \ldots, K(X_*, X_N)]$$

Procesos Gaussianos: (GPs / GPR)

Generaliza la regresión Kernel modelando distribuciones sobre func.

Define por → func. media $m(X) = 0$
→ func. covarianza $K(X, X')$

$$GP: \; f(X) \sim GP(m(X), K(X, X'))$$

$$f(X) = \phi(X_n)^T W \;:\; W \sim N(0, \underbrace{\Sigma_{MAP}}_{S_N}) \rightarrow \text{Sin ruido}$$

$$m(X) = E\{f(X)\} = E\{\phi(X_n)^T W\} = \phi(X_n)^T E\{W\} = 0 \rightarrow \text{Media}$$

$$K(X, X') = Cov(f(X), f(X'))$$

$$K(X, X') = E\{f(X), f(X')\} = E\{\phi(X)^T W \phi^T(X') W^T\}$$

$$= \phi(X)^T E\{W W^T\} \phi(X') = \phi(X)^T \Sigma_W \phi(X') \rightarrow \text{Cov}$$

$$f(X) \sim GP(f(X)|0, K) \;:\; K \in \mathbb{R}^{N \times N} \sim K_{ij} = K(X_i, X_j)$$

$$t_n = f(X) + \eta_n \quad \eta_n \rightarrow \text{Ruido},\; \eta_n \rightarrow N(0, \sigma_n^2)$$

$$P(f(X)) = N(f(X)|0, K) \rightarrow \text{Prior sobre } f(X)$$

$$p(t_n|f(X)) = N(t_n|f(X), \sigma_n^2 \mathbb{I}_N) \rightarrow \text{Verosimilitud}$$

Hallamos distribución marginal de t_n:

$$P(t_n) = \int p(t|f(X))\, p(f(X))\, df(X) = N(t|0, K + \sigma_n^2 \mathbb{I}_N)$$

$$P(t_n) = \int N(t_n|f(X), \sigma_n^2 \mathbb{I}_N) \underbrace{(\sigma_n^2 \mathbb{I}_N)^{-1}}_{S_0} (t_n - f(X))$$

$$- \frac{1}{2} f(X)^T K^{-1} f(X)) \, df(X)$$

$$P(t_n) = \int e^{\left(-\frac{1}{2}\left((\sigma_n^2 \mathbb{I}_N)^{-1} (\|t_n\|^2 - 2 t_n^T f(X) + f(X)^T f(X)) \right.\right.}$$

$$\left.\left. + f(X)^T f(X) K^{-1}\right)\right) df(X)$$

$$P(t_n) = \int e^{\left(-\frac{1}{2} \underbrace{((\sigma_n^2 \mathbb{I}_N)^{-1} + K^{-1})}_{a} f(X)^T f(X) + \underbrace{(\sigma_n^2 \mathbb{I}_N)^{-1}}_{b} t_n^T f(X) \right.}$$

$$\left. \underbrace{- \frac{(\sigma_n^2 \mathbb{I}_N)^{-1}}{2} t_n^T t_n}_{c} \right) df(X)$$

$$P(t_n) = \int e^{\left(-\frac{1}{2} a f(X)^T f(X) + b f(X) + c\right)} df(X) \rightarrow M_f = a^{-1} b$$

$$P(t_n) = \int e^{\left(-\frac{1}{2}(f(X) - M_f)^T a (f(X) - M_f) + \frac{1}{2} b^T a^{-1} b + c\right)} df(X)$$

. . . . .

. . . . .

$$P(t_n) = e^{\left(\frac{1}{2} b^T a^{-1} b + c\right)}$$

$$P(t_n) = e^{\left(\frac{1}{2}(\sigma_n^4 \mathbb{I}_N)^{-1} t_n^T a^{-1} t_n - \frac{(\sigma_n^2 \mathbb{I}_N)}{2} t_n^T t_n\right)}$$

$$P(t_n) = e^{\left(-\frac{1}{2} t_n^T \left(-(\sigma_n^2 \mathbb{I}_N)^{-1} (K + \sigma_n^2 \mathbb{I}_N)^{-1} K + (\sigma_n^2 \mathbb{I}_N)^{-1}\right) t_n\right)}$$

$$P(t_n) = e^{\left(-\frac{1}{2} t_n^T \underbrace{(K + \sigma_n^2 \mathbb{I}_N)^{-1}}_{\text{Varianza}} t_n\right)}$$

$$P(t_n) = N(t_n|0, K + \sigma_n^2 \mathbb{I}_N)$$