

1. Sea el modelo de regresión $t_n = \phi(x_n)w^+ + \eta_n$

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

$\phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q$, $Q \geq P$ y $\eta_n \sim N(\eta_n | 0, \sigma_n^2)$

Presente el problema de optimización y la solución del mismo para los modelos de:

- a) mínimos cuadrados
- b) mínimos cuadrados regularizados
- c) máxima verosimilitud
- d) máximo a-posteriori
- e) Bayesiano con modelo Gaussiano
- f) Regresión rígida Kernel
- g) Procesos Gaussiano

Asuma datos i.i.d. Discuta las diferencias y similitudes entre los modelos estudiados

Planteamiento

$t_n \in \mathbb{R}$ target \rightarrow variable objetivo

$x_n \in \mathbb{R}^P$ features \rightarrow vector de características

$w \in \mathbb{R}^Q$ pesos \rightarrow vector de parámetros

$\phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q$ función base que transforma el espacio de entrada

Normas

Otro de mayor dimensión ($d \geq p$) lo que permite manejar relaciones no lineales.

$$\eta_n \sim \mathcal{N}(0, \sigma_n^2) \text{ ruido Gaussiano iid con varianza } \sigma_n^2$$

Mínimos cuadrados

El problema de optimización es encontrar w que minimice el error cuadrático medio (MSE) entre las predicciones ϕw y los targets t .

$$\text{Problema de optimización: } w^* = \arg \min_w \frac{1}{n} \|t - \phi w\|^2$$

$$\begin{aligned}\frac{\partial}{\partial w} \|t - \phi w\|^2 &= \langle t - \phi w, t - \phi w \rangle = (t - \phi w)^+ (t - \phi w) \\&= (t + (\phi w)^+) (t - \phi w) \\&= t^T t - t^T \phi w - (\phi w)^+ t + (\phi w)^+ (\phi w) \\&= t^T t - 2t^T \phi w + (\phi w)^+ (\phi w) \\&= \|t\|^2 - 2\langle t, \phi w \rangle + \langle \phi w, \phi w \rangle \\&= \|t\|^2 - 2\langle t, \phi w \rangle + \|y\|^2\end{aligned}$$

$$\frac{\partial}{\partial w} \left\{ \|t\|^2 - 2\langle t, \phi w \rangle + \langle \phi w, \phi w \rangle \right\}$$

$$= 0 - 2t^T \phi + \frac{\partial}{\partial w} \left\{ w^T \phi + \langle \phi w, \phi w \rangle \right\}$$

$$= \frac{1}{2} (0 - 2\phi^+ (t^T)^+ + 2\phi^+ \phi w) = 0$$

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

$$2\phi^+ \phi w = 2\phi^+ t$$

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

$$w^* = (\phi^T \phi)^{-1} \phi^T t \quad | \text{ solución}$$

Mínimos cuadrados regularizado

Se añade un hiperparámetro de regularización λ

Problema de optimización: $w^* = \arg \min_w \|t - \phi w\|^2 + \lambda \|w\|^2$

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

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

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

$$2(\phi^T \phi + \lambda I) w - 2 \phi^T \phi w = 0$$

$$(\phi^T \phi + \lambda I) w = \phi^T \phi w$$

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

Mayima verosimilitud

Problema de optimización: $w^*, \sigma^2 = \arg \max_{w, \sigma^2} p(t| \phi, w, \sigma^2)$

$$\log p(t| \phi, w, \sigma^2) = \sum_{n=1}^N \log \mathcal{N}(t_n | \phi(x_n)^T w, \sigma^2)$$

$$\begin{aligned} &= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \phi(x_n)^T w)^2 \\ &= \frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|t - \phi w\|^2 \end{aligned}$$

para simplificar

Para minimizar respecto a w

$$\frac{\partial}{\partial w} \left\{ -\frac{N}{2} \log(\pi(t|G^0)) - \frac{1}{2\sigma^2} \|t - \phi w\|^2 \right\} = -\frac{1}{\sigma^2} (\phi^T (\phi w - t)) = 0$$

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

$$w^* = (\phi^T \phi)^{-1} \phi^T t \rightarrow \text{igual que minimos cuadrados}$$

Para minimizar respecto a σ^2

$$\frac{\partial}{\partial \sigma^2} \left\{ \log p(t|\phi, w^*, \sigma^2) \right\} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \|t - \phi w^*\|^2 = 0$$

$$-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \|t - \phi w^*\|^2 = 0$$

$$\sigma^{2*} = \frac{1}{N} \|t - \phi w^*\|^2$$

Máximo a posteriori

Problema de optimización $w^* = \arg \max_w p(w|t, \phi)$

se busca maximizar la probabilidad de los proyectos dados los datos, o decir la distribución posterior.

$$p(w|t, \phi) = \frac{p(t|\phi, w)p(w)}{p(t|\phi)} \rightarrow \text{se maximiza el numerador}$$

\hookrightarrow no depende de w \uparrow

$$w^* = \arg \max_w p(t|\phi, w)p(w)$$

$$\log p(w|t, \phi) = \log p(t|\phi, w) + \log p(w)$$

$$\log p(t|\phi, w) = -\frac{1}{2\sigma^2} \|t - \phi w\|^2$$

$$p(w) = N(w|0, \alpha^{-1}I) \rightarrow \log p(w) = -\frac{\alpha}{2} \|w\|^2$$

(prior gaussiano no trivial)

$$\log p(w|t, \phi) = -\frac{1}{2\sigma^2} \|t - \phi w\|^2 - \frac{\alpha}{2} \|w\|^2,$$

se busca maximizar, que equivale a minimizar la función de coste, que es el problema de mínimos cuadrados regularizados.

$$\text{Entonces la solución } w^* = (\phi^\top \phi + \lambda I)^{-1} \phi^\top t, \quad \lambda = \sigma^2$$

Bayesiano con modelo lineal Gaussiano

$$\text{Se quiere calcular } p(t_* | x_*, D) = \int p(t_* | x_*, w) p(w | D) dw$$

distribución predictiva · condición completa

↓
predicción
condicional

↓
distribución
posterior
sobre los
parámetros

se tiene que $p(z|w) = N(z|\phi w, \Sigma)$ respecto a la verosimilitud

$$S' = (\phi^\top \phi + \Sigma^{-1})^{-1} \quad \text{covarianza posterior}$$

$$y = S' \phi^\top \Sigma^{-1} t \quad \text{medida posterior}$$

se quiere predecir la salida t_* en un nuevo punto x_* , o decir:

$$p(t_* | x_*, D) = \int p(t_* | x_*, w) p(w | D) dw$$

$$\text{se sabe que: } z_* | w \sim N(\phi(x_*)^\top w, \Sigma)$$

$$w | D \sim N(\mu, S)$$

Entonces la integral termina en otra Gaussiana

$$p(t_{\phi} | X_*, D) = N(t_{\phi} | \phi(x_*)^T \mu, \phi(x_*)^T S \phi(x_*) + \sigma^2)$$

Regresión rigida Kernel

Se basa sobre un Kernel de activo compuesto (kernel) del espacio de características $k(x, x') = \phi(x)^T \phi(x')$

Sea $w = \phi^T \alpha$ entonces la predicción:

$$\hat{E}(x) = \phi(x)^T w = \phi(x)^T \phi^T \alpha = k_x^T \alpha$$

$$\text{donde } k_x = [k(x_1, x), \dots, k(x_N, x)]^T$$

Entonces se recibe la fórmula de coste

$$\frac{\partial}{\partial \alpha} \left\{ \|K\alpha - t\|^2 + \lambda \alpha^T K \alpha \right\}$$

$K = \phi \phi^T$ matriz kernel

$$= 2K(\alpha - t) + 2\lambda K\alpha = 0$$

$$K(K + 2I)\alpha = Kt$$

$$\alpha = (K + 2I)^{-1}t$$

Para predecir un nuevo punto x_* se usa:

$$\hat{E}(x_*) = \sum_{n=1}^N \alpha_n k(x_n, x_*) = k_*^T \alpha$$

$$\text{entonces } \hat{E}(x_*) = k_*^T (K + 2I)^{-1}t$$

Procesos Gaussianos

En GPR se consideran las salidas t_n generadas por una función latente $f(x)$ con ruido gaussiano:

$$t_n = f(x_n) + \eta_n$$

Se pone una distribución a priori sobre las funciones

$$f(x) \sim GP(0, K(x, x'))$$

la distribución de los datos observados es $t \sim N(0, K + G^2 I)$

Se quiere el valor predicho $\hat{t}_* = E[t_* | t]$ y su variancia

La distribución conjunta de los datos observados t y el nuevo

entro t_* es $\begin{bmatrix} t \\ t_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K + G^2 I & K_x \\ K_x^T & K_{xx} + G^2 \end{bmatrix}\right)$

$$K_{xx} = K(x_*, x_*)$$

Por propiedades de la distribución Gaussiana conjugada la distribución predictiva de t_* es: $t_* | t, x_* \sim N(\mu_*, \sigma_*^2)$

donde $\mu_* = K_x^T (K + G^2 I)^{-1} t$ es la medida

$$\sigma_*^2 = K_{xx} + G^2 - K_x^T (K + G^2 I)^{-1} K_x$$
 es la varianza