Parcial # 1

Teoria de Aprendizaje de Maquina

Yonnier Alexander Munoz Salazar

2.1. Sea el modelo de regresión $t_n = \phi(x_n)W^T + \eta_n$
Con: $\{t_n \in \mathbb{R},\ x_n \in \mathbb{R}^p\}_{n=1}^N$, $W \in \mathbb{R}^Q$, $\phi: \mathbb{R}^p \to \mathbb{R}^Q$
$Q \geq P$, $\eta_n \sim N(\eta_n | 0, \sigma_n^2)$
Presentar el problema de optimización y solución.

◆ Minimos Cuadrados

Se tiene el problema de optimización:

$$\hat{W} = \underset{W}{\operatorname{argmin}} \| t_n - f(\phi | W) \|_2^2$$

Se resuelve para Hallar W^T. De la norma se tiene:

$$\| t_n - f(\phi | W^T) \|_2^2 = \langle t_n - f(\phi | W^T),\ t_n - f(\phi | W^T) \rangle$$

Se supone ϕ como datos transformados con regresion no lineal.
Se distribuye el producto interno

$$\langle t_n - f(\phi | W^T),\ t_n - f(\phi | W^T) \rangle$$

$$= \langle t_n - \phi W^T,\ t_n - \phi W^T \rangle$$

$$= t_n^T t_n - t_n^T \phi W^T - (\phi W^T)^T t_n + (\phi W^T)^T \phi W^T$$

$$= t_n^T t_n - 2 t_n^T \phi W^T + W \phi^T \phi W^T$$

Se deriva respecto a W^T y Se iguala a cero

$$\frac{d}{dW}\left(t_n^T t_n - 2 t_n^T \phi W^T + W \phi^T \phi W^T \right)$$

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

Se despeja W^T

$$2 \phi^T \phi W^T = 2 t_n^T \phi$$

$$\boxed{W^T = t_n^T (\phi^T \phi)^{-1} \phi}$$

Finalmente, Se tiene el predict para MC como:

$$\boxed{t_{new} = \phi(X_{new})\,\hat{W}^T}$$

◆ Mínimos Cuadrados Regularizados

Se tiene el problema de optimización

$$\hat{W} = \underset{W^T}{\operatorname{argmin}} \frac{1}{N} \|t_n - f(\phi(x_n)|W\|_2^2 + \lambda \|W\|_2^2$$

Donde $\lambda \in R$, λ es un hiperparametro de regularización

Se distribuye el producto interno del primer termino

$$\|t_n - f(\phi_n|W^T)\|_2^2 = [t_n^T t - 2 t_n^T \Phi W + W^T \phi^T \phi W]\frac{1}{N}$$

Del segundo termino se tiene:

$$\lambda\|W\|_2^2 = \lambda \langle W, W\rangle = \lambda W^T W$$

Se deriva ambos terminos respecto a W y Se iguala a 0

$$\frac{d}{dw}\|t_n - \phi W\|_2^2 = [-2\phi^T t_n + 2\phi^T \phi W]\,1/N$$

$$\frac{d}{dw}\lambda\|W\|_2^2 = 2\lambda W$$

$$\Rightarrow -\frac{2}{N}\phi^T t_n + \frac{2}{N}\phi^T \phi W + 2\lambda W = 0$$

Se multiplica a ambos lados por $N/2$ y $(\phi^T\phi + \lambda N I)^{-1}$

$$\frac{N}{2}\left[-\frac{2}{N}\phi^T t_n + \frac{2}{N}\phi^T \phi W + 2\lambda W\right] = 0\left[\frac{N}{2}\right]$$

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

Finalmente Se tiene:

$$\boxed{\begin{aligned} W &= (\phi^T\phi + \lambda N I)^{-1}\phi^T t_n \\ \hat{W}^T &= t_n^T(\phi^T\phi + \lambda N I)^{-1}\phi \end{aligned}}$$

Predict para MCR: $\boxed{t_{new} = \phi(X_{new})\,\hat{W}^T}$

Maxima Verosimilitud

En el problema de optimización se agrega WGN = η y se utiliza la función log para simplificar cálculos

$$\hat{W} = \underset{w}{\operatorname{argmax}} \log\left(P(t_n \mid F(X|W); \eta)\right)$$

La información del ruido se toma como:

$$\hat{W} = \underset{w}{\operatorname{argmax}} \log\left(\prod_{n=1}^{N} N(t_n \mid F(X, W), \sigma_\eta^2)\right)$$

Con Φ como los datos transformados, se tiene la forma de la probabilidad como:

$$\log\left(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \exp\left(-\frac{\|t_n - \Phi W\|_2^2}{2\sigma_\eta^2}\right)\right.$$

Se utilizan propiedades de log para simplificar la expresión

$$\log\left(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_\eta^2}}\right) + \log\left(\prod_{n=1}^{N} \exp\left(-\frac{\|t_n - \Phi W\|^2}{2\sigma_\eta^2}\right)\right)$$

$$= \log_\eta\left(\frac{1}{(2\pi\sigma_\eta^2)^{N/2}}\right) + \log_\eta\left(\exp\left(-\sum_{n=1}^{N} \frac{\|t_n - \Phi W\|^2}{2\sigma_\eta^2}\right)\right)$$

$$= \log_\eta\left((2\pi\sigma_\eta^2)^{N/2}\right) - \sum_{n=1}^{N} \frac{\|t_n - \Phi W\|^2}{2\sigma_n^2}$$

Se obtiene

$$-\frac{N}{2}\left[\log(2\pi) + \log(\sigma_n^2)\right] - \frac{1}{2\sigma_n^2}\|t_n - \Phi W\|^2$$

Se deriva respecto a W y se iguala a cero, de MC se tiene la derivada del segundo termino y la derivada del primer termino es igual a cero Finalmente, se tiene:

$$\frac{d}{dw}\left(\frac{N}{2}\left[\log(2\pi) + \log(\sigma_n^2)\right] - \frac{1}{2\sigma_n^2}\|t_n - \Phi W\|^2\right.$$

$$= 0 - \frac{1}{2\sigma_n^2}\left[-2\Phi^T t_n + 2\Phi^T \Phi W\right]$$

Despejando W, se tiene:

$$\boxed{\hat{W} = (\Phi^T \Phi)^{-1} \Phi^T t_n}$$

Para encontrar σ_η^2, Se repite el procedimiento anterior, derivando y despejando para σ_η^2

$$\frac{d}{d\sigma_\eta^2}\left[-\frac{N}{2}\left(\log(2\pi)+\log(\sigma_\eta^2)\right)\right]-\frac{d}{d\sigma_\eta^2}\left[\frac{1}{2\sigma_\eta^2}\|t_n-\Phi w\|_2^2\right]$$

$$=-\frac{N}{2\sigma_\eta^2}+\frac{1}{2(\sigma_\eta^2)^2}\|t_n-\Phi w\|_2^2=0$$

Se multiplica a ambos lados por $(\sigma_\eta^2)^2$

$$=-\frac{N}{2\sigma_\eta^2}(\sigma_\eta^2)^2+\frac{1}{2(\sigma_\eta^2)^2}(\sigma_\eta^2)^2\|t_n-\Phi w\|_2^2=0(\sigma_\eta^2)^2$$

$$=-\frac{N}{2}\sigma_\eta^2+\frac{1}{2}\left[-2\Phi^T t_n+2\Phi^T\Phi w\right]=0$$

Finalmente, Se tiene:

$$\boxed{\sigma_\eta^2=\frac{1}{N}\left[-2\Phi^T t_n+2\Phi^T\Phi w\right]}$$

- Maximo A-posteriori

Se agrega un prior con probabilidad

$$P(w)=N(w\mid 0,\sigma_w^2)\rightarrow w\in\mathbb{R}^a$$

Se tiene la optimización como:

$$\hat{w}=\text{Argmax}_w\ \log(P(w)\,t_n,\Phi,\sigma_\eta^2)$$

Se desarrolla la expresión

$$\log(P(w)\,t_n,\Phi,\sigma_\eta^2)=\log\left(\prod_{n=1}^{N}(t_n\mid\phi(x_n)w,\sigma_\eta^2)\prod_{q=1}^{a}N(w_q\mid 0,\sigma_w^2)\right.$$

$$=\left\{-\frac{N}{2}\left[\log(2\pi)+\log(\sigma_\eta^2)\right]-\frac{1}{2\sigma_\eta^2}\|t_n-\Phi w\|_2^2\right\}\Rightarrow A$$

$$+\log\left(\prod_{q=1}^{a}\left(\frac{1}{\sqrt{2\pi\sigma_w^2}}\right)\exp\left(-\|w_q-0\|_2^2\right)\right)$$

$$=A+\log_2\left(\prod_{q=1}^{a}\frac{1}{(2\pi\sigma_w)^{a/2}}+\log\left(\exp\left(-\sum_{q=1}^{a}\frac{|w|}{2\sigma_w^2}\right)\right)\right.$$

$$=A-\frac{a}{2}\left[\log(2\pi)+\log(\sigma_w^2)\right]-\sum\frac{|w|}{2\sigma_w^2}$$

$$= -\frac{N}{2}\left[\log(2\pi) + \log(\sigma_\eta^2)\right] - \frac{1}{2\sigma_\eta^2}\|t - \Phi w\|_2^2 - \frac{Q}{2}\left[\log(2\pi) + \log(\sigma_w^2)\right] - \frac{1}{2\sigma_w^2}\|W\|_2^2$$

Ahora, se tiene que:

$$= Max - \left[\frac{2\sigma_\eta^2}{2\sigma_\eta^2}\|t_n - \Phi w\|_2^2 + \frac{2\sigma_\eta^2}{2\sigma_w^2}\|W\|_2^2\right] + cte$$

$$= min\left[t_n - \Phi w\|_2^2 + \frac{\sigma_\eta^2}{\sigma_w^2}\|W\|_2^2 + cte\right.$$

Finalmente $\hat{W}$ esta dado por:

$$\hat{W} = (\Phi^T\Phi + \frac{\sigma_\eta^2}{\sigma_w^2}I)^{-1}\Phi^T t_n$$

Bayesiano lineal Gaussiano

Se define el prior con probabilidad de la forma:

$$P(W) = N(W|m_0, S_0)$$

Donde, $P(W)$ incluye una matriz de varianza anisotropica

La probabilidad de WGN esta dada como:

$$P(\eta) = N(\eta|0, \beta^{-1})$$

Donde, $\beta^{-1} = \sigma I$, isotropica $\rightarrow \beta \in \mathbb{R}^+$

El problema de optimización por Bayes es:

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

Donde: $P(w|t) \Rightarrow$ Posterior

$P(t|w) \Rightarrow$ Verosimilitud

$P(w) \Rightarrow$ Prior

$P(t) \Rightarrow$ Evidencia

Despejando el ruido de la verosimilitud $P(t|w,\beta)$

Se tiene: $\eta = t_n - w^T \phi(x_n) \sim N(t_n - w^T \phi(x_n) | 0, B^{-1})$

El termino $t_n - w^T \phi(x_n)$ traslada la media del termino así:

$$P(t|w) = N(t | \phi w, B^{-1} I)$$

Para el posterior, se obtiene $\mu_{w|t}$ y $\Sigma_{w|t}$ a partir de la matriz de precisión y completando cuadrados, se tiene:

$$\mu_{w|t} = M_N = (S_0^{-1} + \phi^T B I \phi)^{-1} [\phi^T B I (t-0) + S_0^{-1} M_0]$$
$$= S_N (S_0^{-1} M_0 + B \phi^T t)$$

Y,

$$\Sigma_{w|t} = S_N$$

$$S_N = (S_0^{-1} + B \phi^T \phi)^{-1} \; ; \; S_N^{-1} = S_0^{-1} + B \phi^T \phi$$

Finalmente se tiene la predictiva del modelo de la forma: como,

$$P(t_{new}|w) = N(t_{new} | w^T \phi(x_n), B^{-1})$$

Y,

$$P(w|t) = N(w | m_N, S_N)$$

Con el operador convolución, se tiene:

$$P(t_{new}|t, w) = \int P(t_{new}|w) \, P(w|t) \, dw$$

Queda:

$$\boxed{P(t_{new}|t, w) = N(t_{new} | M_N^T \phi(x_n) B^{-1} + \phi^T(x_*) S_N \phi(x_*))}$$

Sabiendo que:

$$M_N^T \phi(x_*) = \phi^T(x_*) M_N$$

La predictiva final es la predictiva de cada dato nuevo ponderada respecto a la esperanza del posterior

$$\boxed{\hat{P}(t_{new}|w) = E_{w|t}\{P(t_{new}|w)\}}$$