Víctor Quintero T. 9976089.

1. Sea el modelo de regresión:

$$t_n = \phi(x_n)\, w + n_n$$

con el conjunto de datos

$$\{(t_n \in \mathbb{R}, x_n \in \mathbb{R}^P)\}_{n=1}^{N} \quad \text{Donde:}$$

- t_n es la variable objetivo para la muestra n
- x_n es el vector de características de entrada para la muestra n.
- $w \in \mathbb{R}^Q$ vector de pesos (parámetros)
- $\phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q$ función que mapea el espacio de entrada a un espacio de características de Q dimensiones. $\phi(x_n)$ es un vector columna.
- $Q \geq P$.
- n_n es el ruido, Gaussiano con media $\emptyset$ y varianza σ_n^2: $n_n \sim \mathcal{N}(n_n \mid 0, \sigma_n^2)$.
- los datos son i.i.d

$$t = [t_1, t_2, \dots, t_N]^T$$

Φ = Matriz de $N \times Q$:

$$\Phi = \begin{pmatrix} \phi(x_n)[1]^T \\ \phi(x_n)[2]^T \\ \vdots \\ \phi(x_n)[N]^T \end{pmatrix}$$

$$n = [n_1, n_2, \dots n_N]^T$$

$$t = \Phi w + n$$

Powered by CamScanner

Mínimos Cuadrados.

Víctor Quintero T.

$$L(w) = \sum_{n=1}^{N} (t_n - \phi(x_n)^T w)^2$$

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

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

$$= t^T t - 2t^T \Phi w + w^T \Phi^T \Phi w$$

Se deriva con respecto a w e igualamos a 0:

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

$$\Phi^T \Phi w = \Phi^T t.$$

Si $\Phi^T \Phi$ tiene inversa

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

Mínimos Cuadrados Regularizados

$$L(w) = \sum_{n=1}^{N} (t_n - \phi(x_n)^T w)^2 + \lambda \|w\|_2^2$$

$$= (t - \Phi w)^T (t - \Phi w) + \lambda w^T w.$$

Con $\lambda \geq 0$ tenemos:

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

$$(\Phi^T \Phi + \lambda I_Q) w = \Phi^T t.$$

Donde I_Q es la matriz Identidad $Q \times Q$.

$$= (\Phi^T \Phi + \lambda I_Q)^{-1} \Phi^T t.$$

Máxima Verosimilitud

$\eta_n = t_n - \phi(x_n)^T w$ y $\eta_n \sim \mathcal{N}(0, \sigma_n^2)$, entonces:

$$t_n \mid x_n, w, \sigma_n^2 \sim \mathcal{N}(t_n \mid \phi(x_n)^T w, \sigma_n^2)$$

la verosimilitud es:

$$p(t_n \mid x_n, w, \sigma_n^2) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(t_n - \phi(x_n)^T w)^2}{2\sigma_n^2}\right)$$

Aplicando log en base natural tenemos:

$$\ln p(t \mid X, w, \sigma_n^2) = \sum_{n=1}^{N} \ln p(t_n \mid x_n, w, \sigma_n^2)$$

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

$$= \frac{N}{2}\ln(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\sum_{n=1}^{N}(t_n - \phi(x_n)^T w)^2$$

Problema de optimización: Maximizar $\ln p(t \mid X, w, \sigma_n^2)$ con respecto a w (y σ_n^2). Esto es equivalente a minimizar $\sum (t_n - \phi(x_n)^T w)^2$ por lo tanto

$$= (\Phi^T \Phi)^{-1} \Phi^T t.$$

Es igual que mínimos cuadrados.

Para σ_n^2 derivando el logaritmo de la verosimilitud con respecto a σ_n^2 (o $\beta = 1/\sigma_n^2$) e igualando a 0

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

Máximo a-posteriori Víctor Quintero T.

Sea p una distribución gaussiana:

$$p(w|\alpha) = \mathcal{N}(w|0,\alpha^{-1}I_Q) = \left(\frac{\alpha}{2\pi}\right)^{Q/2} \exp\left(-\frac{\alpha}{2} w^T w\right).$$

Aplicando Bayes:

$$p(w|t, X, \sigma_n^2, \alpha) \propto p(t|X, w, \sigma_n^2)\, p(w|\alpha)$$

Problema de Optimización:

Maximizar $\ln p(w|\ldots)$ que es equivalente a minimizar:

$$-\ln p(w|\ldots) \propto \frac{1}{2\sigma_n^2} \sum_{n=1}^{N} (t_n - \phi(x_n)^T w)^2 + \frac{\alpha}{2} w^T w.$$

multiplicando por $2\sigma_n^2$:

$$\sum_{n=1}^{N} (t_n - \phi(x_n)^T w)^2 + \alpha\sigma_n^2 w^T w$$

si $\lambda = \alpha\sigma_n^2$ tenemos:

$$= (\Phi^T\Phi + \alpha\sigma_n^2 I_Q)^{-1}\Phi^T t.$$

Bayesiano con modelo lineal Gaussiano

Con una distribución $p(w|t, X, \alpha, \beta)$ donde $\beta = \frac{1}{\sigma_n^2}$

- Verosimilitud: $p(t|X, w, \beta) = \mathcal{N}(t|\Phi w, \beta^{-1}I_N)$.
- Prior: $p(w|\alpha) = \mathcal{N}(w|\mu_m, S_m)$ $(\mu_m = 0,\ S_m = \alpha^{-1}I_Q)$
- Posterior: $p(w|t, X, \alpha, \beta)$ es Gaussiana $\mathcal{N}(w|m_N, S_N)$

$$S_N = (S_m^{-1} + \beta\Phi^T\Phi)^{-1}$$

$$m_N = S_N(S_m^{-1}\mu_m + \beta\Phi^T t)$$

Víctor Quintero.

Para el Prior común $M_m = 0$ y $S_0 = \alpha^{-1} I_Q$:

$$S_N = (\alpha I_Q + \beta \Phi^T \Phi)^{-1}$$

$$m_N = \beta S_N \Phi^T t = (\alpha \beta^{-1} I_Q + \Phi^T \Phi)^{-1} \Phi^T t.$$

Si $\alpha \beta^{-1} = \alpha \sigma_n^2$ se vuelve máximo a posteriori.

Regresión Rígida Kernel

Problema de optimización:

$$L(w) = (t - \Phi w)^T (t - \Phi w) + \lambda w^T w.$$

$w = \Phi^T \alpha$ con $\alpha \in \mathbb{R}^N$, tenemos:

$$L(\alpha) = (t - \Phi \Phi^T \alpha)^T (t - \Phi \Phi^T \alpha) + \lambda (\Phi^T \alpha)^T (\Phi^T \alpha)$$

con $K = \Phi \Phi^T$, donde $K_{ij} = \phi(x_i)^T \phi(x_j) = k(x_i, x_j)$

$$L(\alpha) = (t - K\alpha)^T (t - K\alpha) + \lambda \alpha^T \Phi \Phi^T \alpha =$$

$$(t - K\alpha)^T (t - K\alpha) + \lambda \alpha K \alpha.$$

Derivando $L(\alpha)$ con respecto a α e igualando a 0 =

$$\frac{\partial L(\alpha)}{\partial \alpha} = -2K^T (t - K\alpha) + 2\lambda K \alpha = 0.$$

Si K es simétrico $(K^T = K)$ tenemos;

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

$$\alpha = (K + \lambda I_N)^{-1} t$$

Donde I_N es la identidad $N \times N$.

GP - Procesos Gaussianos

Víctor Quintero T.

Distribución a priori sobre funciones $f(x) \sim GP(m(x), k(x,x'))$

$m(x) = 0$. $\quad t_n = f(x_n) + \eta_n$, con $\eta_n \sim \mathcal{N}(0, \sigma_n^2)$

$$p(t|X) = \mathcal{N}(t|0, K_N + \sigma_n^2 I_n)$$

donde $(K_n)_{ij} := k(x_i, x_j)$ se tiene:

$$\begin{pmatrix} t \\ t_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K_N + \sigma_n^2 I_n & k_* \\ k_*^T & k(x_*, x_*') \end{pmatrix}\right)$$

k_* es un vector $N \times 1$ con $(k_*)_n = k(x_n, x_*)$

Media predictiva: $\mu_* = k_*^T (K_N + \sigma_n^2 I_N)^{-1} t$

Varianza Predictiva: $\varepsilon^2_{f,*} = k(x_*, x_*) - k(K_N + \sigma_n^2 I_N)^{-1} k_*$

Varianza (para t_*): $\varepsilon^2_{t_*} = \varepsilon^2_{f,*} + \sigma_n^2$.

$$\ln p(t|X,\theta) = -\frac{1}{2} t^T (K_N(\theta) + \sigma_n^2 I_N)^{-1} t - \frac{1}{2} \ln |K_N(\theta) + \sigma_n^2 I_N| - \frac{N}{2} \ln(2\pi).$$

donde θ son los hiperparámetros del kernel.

Diferencias y Similitudes

Víctor Quintero Toro

- Paramétrico vs. No paramétrico:
  - Paramétricos: Mínimos Cuadrados, Mínimos Cuadrados Regularizados, Máxima Verosimilitud, Máximo a posteriori, Bayesiano Lineal (con ϕ fijo).
  - No paramétricos: Regresión rígida Kernel, GP. La complejidad puede crecer con N.
- Estimación puntual vs. distribución:
  - Estimación puntual de w: Mínimos Cuadrados, Mínimos cuadrados regularizados, Máxima verosimilitud, Máximo a posteriori.
  - Distribución sobre w: Bayesiano Lineal.
  - Predicción Puntual: Mínimos Cuadrados, Mínimos Cuadrados Regularizados, Máxima Verosimilitud, Máximo a posteriori, Regresión Rígida Kernel.
  - Distribución Predictiva: Bayesiano Lineal, GP
- Regularización
  - Sin regularización (directa): Mínimos Cuadrados, Máxima Verosimilitud.
  - Con regularización:
    - Mínimos cuadrados con regularización: penaliza $\|w\|_2^2$
    - Máximo a posteriori: $p(w)$ regula penalización L_2
    - Bayesiano Lineal: el prior $p(w)$ regulariza.
    - KRR: el término λI_N en $(K + \lambda I_N)^{-1}$
    - GP: El kernel y σ_n^2 controlan la suavidad.

Diferencias y Similitudes

Víctor Quintero T.

Conexiones clave:

- Mínimos Cuadrados y Máxima Verosimilitud para w son idénticos bajo ruido gaussiano.
- Mínimos cuadrados regularizados y máximo a posteriori con prior Gaussiano para w son idénticos si $\lambda = \alpha\sigma_n^2$
- La media posterior m_N del Bayesiano lineal es W de máximo a posteriori con hiperparámetros consistentes.
- La media predictiva de GP μ_* es idéntica a la de predicción de la Regresión Rígida kernel $y(x_*)$ si $\lambda = \sigma_n^2$ y se usa el mismo kernel. GP también da varianza.
- El Bayesiano Lineal con prior $\mathcal{N}(w|0, \alpha^{-1} I_Q)$ y kernel $k(x, x') = \alpha^{-1}\phi(x)^T\phi(x')$ es un caso de GP.