Preuves Cours 8 RLD

1 Processus Gaussien

Soit $f(x) = x^T w$, avec w un vecteur de paramètre de régression linéaire dans cet espace. On considère $w \sim \mathcal{N}(0, \Sigma_p)$.

On a alors un processus gaussien pour lequel on veut estimer les postérieures de sortie espérée $p(f^*|x^*, X, y)$ pour chaque nouveau point x^* connaissant un ensemble de points X pour lesquels on a observé les sorties y, avec y le vecteur dont chaque composante $y_i = f(x_i) + \epsilon_i$, avec $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (on n'observe pas directement f, uniquement les y bruités).

On a : $p(y|X, w) = \prod_{i=1}^{n} p(y_i|x_i, w) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2}(y - X^T w)^T(y - X^T w)) = \mathcal{N}(X^T w, \sigma^2 I)$

On a alors :

$$\begin{aligned} p(w|X, y) &\propto p(y|X, w)p(w) \\ &\propto \exp(-\frac{1}{2\sigma^2}(y - X^T w)^T(y - X^T w))exp(-\frac{1}{2}w^T\Sigma_p^{-1}w) \\ &\propto exp(-\frac{1}{2}w^T(\frac{1}{\sigma^2}XX^T + \Sigma_p^{-1})w + \sigma^{-2}X^T yw) \\ &\propto exp(-\frac{1}{2}w^T Aw + \bar{w}Aw) \end{aligned}$$

avec $A = \frac{1}{\sigma^2}XX^T + \Sigma_p^{-1}$ et $\bar{w} = A^{-1}\sigma^{-2}X^T y$.

On a donc : $p(w|X, y) = \mathcal{N}(\bar{w}, A^{-1})$
Soit $f^* = x^{*T} w$. Puisque x^* est une constante et que $p(w|X, y)$ est normale, alors $p(f^*|x^*, X, y)$ est également normale : $p(f^*|x^*, X, y) = \mathcal{N}(\mu_{f^*}, \Sigma_{f^*})$ avec :

$$\mu_{f^*} = \mathbb{E}[x^{*T} w|X, y, x^*] = x^{*T}\mathbb{E}[w|X, y] = x^{*T}\bar{w}$$

$$\begin{aligned} \Sigma_{f^*} &= Var(x^{*T} w|X, y, x^*) \\ &= \mathbb{E}[(x^{*T} w - x^{*T}\bar{w})^2|X, y, x^*] \\ &= \mathbb{E}[x^{*T}(w - \bar{w})(w - \bar{w})^T x^*|X, y, x^*] \\ &= x^{*T}\mathbb{E}[(w - \bar{w})(w - \bar{w})^T|X, y]x^* \\ &= x^{*T} Var(w|X, y)x^* \\ &= x^{*T} A^{-1} x^* \end{aligned}$$

Soit le lemme d'inversion matricielle suivant, qui va nous permettre d'introduire des noyaux pour traiter des cas non linéaires :

$$(Z + UWV^T)^{-1} = Z^{-1} - Z^{-1}U(W^{-1} + V^T Z^{-1}U)^{-1}V^T Z^{-1}$$

En posant $Z^{-1} = \Sigma_p$, $W^{-1} = \sigma^2 I$ et $V = U = X$, on peut ré-écrire la variance Σ_{f^*} selon :

$$\begin{aligned}\Sigma_{f^*} &= x^{*T} A^{-1} x^* \\ &= x^{*T}(\Sigma_p^{-1} + X\sigma^{-2} I X^T)^{-1} x^* \\ &= x^{*T}(\Sigma_p - \Sigma_p X(\sigma^2 I + X^T \Sigma_p X)^{-1} X^T \Sigma_p) x^* \\ &= x^{*T}\Sigma_p x^* - x^{*T}\Sigma_p X(\sigma^2 I + X^T \Sigma_p X)^{-1} X^T \Sigma_p x^* \\ &= k(x^*, x^*) - k(x^*, X)(k(X, X) + \sigma^2 I)^{-1} k(x^*, X)\end{aligned}$$

avec $k(x, y) = x^T \Sigma_p y$.

Pour la moyenne on a :

$$\begin{aligned}x^{*T}\bar{w} &= x^{*T} A^{-1} \sigma^{-2} X y \\ &= x^{*T}(\sigma^{-2} X X^T + \Sigma_p^{-1})^{-1} \sigma^{-2} X y \\ &= x^{*T} X (X X^T + \sigma^2 \Sigma_p^{-1})^{-1} y \\ &= x^{*T} \Sigma_p X (X^T \Sigma_p X + \sigma^2 I)^{-1} y \\ &= k(x^*, X)(k(X, X) + \sigma^2 I)^{-1} y\end{aligned}$$

Cette écriture permet d'exprimer les quantités uniquement en terme de produits scalaires entre points d'entrée x. Cela donne l'occasion d'introduire des noyaux reproductibles k, correspondant à des produits scalaires dans l'espace de Hilbert (Kernel trick comme dans SVM). Si on considère $f(x) = \phi(x)^T w$ plutôt que $f(x) = x^T w$, avec ϕ une projection de x dans un espace de Hilbert, on peut utiliser toutes sortes de noyaux (e.g. noyau RBF) $k(x, y) = \phi(x)^T \Sigma_p \phi(y)$ pour travailler sur des problèmes non linéaires (sans avoir à exprimer ϕ de manière explicite). C'est la force des processus gaussiens. Leur limite est cependant leur complexité croissante avec les nombre N de points observés (calcul et inversion d'une matrice $k(X, X)$ de taille $N \times N$).

2 Maximum d'Entropie

On montre que la distribution $p(\tau) \propto exp(-c(\tau))$, avec $c(\tau)$ un coût attribué à la trajectoire τ est la solution du problème de minimisation :

$$\min_q E_q\left[c(\tau)\right] - \mathcal{H}_q(\tau)$$

Où $\mathcal{H}_q(\tau)$ est l'entropie de la distribution q.

Soit $p(\tau) = \frac{exp(-c(\tau))}{Z}$, avec $Z = \int exp(-c(\tau))d\tau$ la partition de la distribution.

Soit la quantité à minimiser $E_q\left[c(\tau)\right] - \mathcal{H}_q(\tau)$.

En prenant $q = p$ on a :

$$\begin{aligned}E_q\left[c(\tau)\right] - \mathcal{H}_q(\tau) =& E_q\left[c(\tau) + \log q(\tau)\right] \\ =& E_q\left[c(\tau) - c(\tau) - \log Z\right] \\ =& -\log Z E_q[1] \\ =& -\log Z\end{aligned}$$

Or on a :

$$\begin{aligned}\log Z =& \log \int exp(-c(\tau))d\tau \\ =& \log \int \frac{q(\tau)}{q(\tau)} exp(-c(\tau))d\tau \\ \geq& \int q(\tau) \left[\log exp(-c(\tau)) - \log q(\tau)\right] \\ =& E_q[-c(\tau)] + H_q(\tau)\end{aligned}$$

On a donc $E_q[c(\tau)] - H_q(\tau) \geq -\log Z$

$-\log Z$ est donc un minorant de la quantité à minimiser, que l'on atteint lorsque $q(\tau) \propto exp(-c(\tau))$.