Introduction aux Processus Gaussiens

Olivier Delalleau

LISA

19 août 2003

Contexte

- Dataset $D = \{x_n, t_n\}_{n=1\ldots N}$
- $t_n = y(x_n) + \epsilon_n$
- But = prédire la **distribution** de t^*, étant donné un nouveau point x^*

$$P(t^*|x^*, D)$$

- $\Rightarrow$ perspective **bayésienne** $\neq$ fréquentiste
- Utilisation typique : régression

$$\hat{y}(x^*) = E[t^*|x^*, D]$$

Principe

- Un GP place une distribution gaussienne *a priori* sur les fonctions :

$$P(y) = \mathcal{N}(f, K) = \frac{1}{Z} exp\left[-\frac{1}{2}(y-f)^T K^{-1}(y-f)\right]$$

- Formellement : $P(y)$ est un GP si pour tous x_1, ..., x_n, $P(y(x_1), ..., y(x_n))$ est une gaussienne.
- En pratique : $f = 0$
- $K(x, x') = Cov[y(x), y(x')]$

Origine

Les GP sont connus depuis longtemps en statistiques, mais ne sont utilisés que depuis peu en machine learning.

[*Bayesian learning for Neural Networks*, Neal, 1996] : un réseau de neurones place une une probabilité a priori sur l'espace de fonctions, et lorsque la couche cachée est infinie, il s'agit d'un GP.

Prédiction

- Loi jointe des t_n :

$$\begin{aligned} & P(y(x_1), ..., y(x_N)) = \mathcal{N}(0, K_N) \\ et \quad & P(\epsilon_n) = \mathcal{N}(0, \sigma^2) \\ \Rightarrow \quad & P(t_1, ..., t_N) = \mathcal{N}(0, K_N + \sigma^2.I) \end{aligned}$$

- Prédiction :

$$\begin{aligned} P(t^*|t_n) &= \frac{P(t^*, t_n)}{P(t_n)} \\ &= \mathcal{N}(\hat{t}^*, \sigma^{*2}) \end{aligned}$$

La Matrice de Covariance K

C'est elle qui définit la solution :

$$(K_N)_{ij} = K(x_i, x_j)$$

$\Rightarrow$ similarité avec les méthodes à noyau comme les SVM. Une fonction de covariance classique est donc bien sûr le noyau gaussien :

$$K(x, x') = \frac{1}{U} exp\left[-\frac{1}{2}\frac{||x - x'||^2}{\sigma_K^2}\right]$$

Estimation des Hyperparamètres

Hyperparamètres θ = {bruit σ^2, variance du noyau σ_K^2}

Optimisation "facile" dans un cadre bayésien :

$$P(\theta|t_n) = \frac{P(t_n|\theta)P(\theta)}{P(t_n)}$$

$\Rightarrow$ minimisation de

$$C(\theta) = -ln(P(t_n|\theta)) - ln(P(\theta))$$

Considérations Pratiques

- **Entraînement** : il faut calculer $K_N^{-1} \Rightarrow O(N^3)$
- **Prédiction** en $O(N)$
- **Bornes de confiance** en $O(N^2)$
- $\Rightarrow$ **impraticable** pour de grands ensembles de données

Méthodes d'Accélération

Principe = utiliser un sous-ensemble de $m \ll N$ points : "Sparse"

$\Rightarrow$ entraînement en $O(m^2 N)$, prédiction en $O(m)$, bornes en $O(mN)$.

Exemples :

- Sélection aléatoire
- *Sparse greedy Gaussian Process regression* (Smola, Bartlett, 2001) : critère de sélection des points assez coûteux
- *Fast sparse Gaussian Process Methods: the informative Vector Machine* (Lawrence, Seeger, Herbrich, 2002) : critère basé sur la théorie de l'information, application à la classification $\Rightarrow$ résultats comparables à une SVM, avec un temps d'entraînement plus rapide
- *Fast forward selection to speed up sparse gaussian process regression* (Seeger, Williams, Lawrence, 2003) : même critère, aussi rapide qu'une sélection aléatoire, mais résutats à peine meilleurs (intérêt = optimisation des hyperparamètres)