Chap I

M2DS

Méthodes GPR: Gaussian process regression et applications:

Motivation:

On rappelle le problème de régression linéaire:

$$y = f(x) + \varepsilon ,$$

avec $$f(x) = x^T w = \langle x, w \rangle = \sum_i w_i x_i$$

où x est un vecteur, w le vecteur de poids (paramètres) du modèle linéaire. Souvent il y a un biais qu'on peut compenser en augmentant la dimension du vecteur x par élément additionnel toujours égal à 1 (Dans la notation il faut ajouter 1 aux "inputs" d'origine x $X = (1, x)$ (abus de notation)).

exple

$x_0 = 1, x_1 = x_1, \ldots$.

Dans ce modèle, la différence entre y et x est donné par le bruit additif ε, qui peut être considéré comme bruit blanc fort Gaussien (i.i.d)

$$\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$$

Une conséquence directe de ce modèle, on peut expliciter la vraisemblance associée, sachant les paramètres, Grâce à l'indépendance, la densité de probabilité de observation sachant les paramètres :

$$p(Y/X,W) = \prod_{i=1}^{n} p(y_i/x_i,W)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(y_i - x_i^T W)^2}{2\sigma_n^2}\right).$$

X la matrice formée par les colonnes $(x_i)_{1\leq i\leq n}$

$$= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2}|Y - X^T W|^2\right) \quad (*)$$

$$= \mathcal{N}(X^T W, \sigma_n^2 I_d).$$

$|\cdot|$ désigne la norme euclidienne.

Dans l'approche Bayesienne, on doit spécifier une loi à priori pour les paramètres qui reflètent notre intuition sur ce que peuvent être ces paramètres.

Par défaut, on met une loi à priori gaussienne de matrice de covariance Σ_p sur les poids :

$$W \sim \mathcal{N}(0, \Sigma_p)$$

Rappels :

(.) La densité d'une gaussienne multivariée :

$$p(x/m,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}\,|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right)$$

où m : vecteur moyenne (taille D)

Σ : matrice de covariance (symétrique définie positive) de taille $D \times D$

et on écrit $x \sim \mathcal{N}(m,\Sigma)$

(..) Soit x, y de loi jointe gaussienne

$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}\right)$ alors la loi marginale de x conditionnelle à y est :

$$x \sim \mathcal{N}(\mu_x, A), \quad x/y \sim \mathcal{N}\left(\mu_x + CB^{-1}(y-\mu_y),\ A - CB^{-1}C^T\right)$$

(..)
$$p(x/y) = \frac{p(x)\,p(y/x)}{p(y)}$$

(mem. appliqué deux fois la formule de Bayes $p(x|y)$ et $p(y/x)$).

Dans le problème d'inférence Bayesienne, on utilise:

$$\text{La postérieure} = \frac{\text{loi prior} \times \text{vraisemblance}}{\text{loi marginale vraisemblance}}$$

$$\Rightarrow \underbrace{p(w/y,X)}_{p_X(w/y)} = \frac{p(w)\cdot p(y/X,w)}{p(y/X)} = \frac{p(w)\, p_X(y/w)}{p_X(y)} \qquad (**)$$

Notons que, le dénominateur ne dépend pas de w

et on a:

$$p(y/X) = \int p(y/X,w)\, p(w)\, dw$$

ne dépend pas de w

Proposition:

On a:

$$p(w/X,y) = \underbrace{C^{te}} \times \exp\left(-\frac{1}{2}(w-\bar{w})^T\left(\frac{XX^T}{\sigma_n^2}+\Sigma_p^{-1}\right)(w-\bar{w})\right)$$

où

$$\bar{w} = \frac{1}{\sigma_n^2}\left(\frac{XX^T}{\sigma_n^2}+\Sigma_p^{-1}\right)^{-1} X y$$

⚠ -1

indication: Comme $p(y/X)$ ne dépend pas de w on peut le mettre ds C^{te}, donc il suffit de calculer le numérateur de $(**)$ en utilisant $(*)$

Preuve: Par la formule $(**)$:

$$p(w/X,y) = C^{te} \exp\left(-\frac{1}{2}w^T\Sigma_p^{-1}w\right)\exp\left(-\frac{1}{2\sigma_n^2}\overbrace{(y-X^Tw)^T(y-X^Tw)}^{|y-X^Tw|^2}\right)$$

Dans la preuve on fera le terme donné dans la proposition

de sorte à récupérer à nouveau une loi normale au prix de multiplier et diviser par des constantes qui ne dépendent pas de w, ces constantes n'impactent par l'optimisation en w.

C.C : On a :

$$p(w/X,y) \sim \mathcal{N}(\bar{w}, A^{-1})$$

où $\bar{w} = A^{-1}\frac{Xy}{\sigma_n^2}$ et $A^{-1} = \frac{XX^T}{\sigma_n^2} + \Sigma_p^{-1}$

En partant d'une loi à priori $w \sim \mathcal{N}(0, \Sigma_p)$
on obtient une loi à postériori $w/X,y \sim \mathcal{N}(\bar{w}, A^{-1})$

I - Projection des données dans l'espace des caractéristiques

L'inconvénient du modèle précédent est qu'il est très limité puisque ça reste un modèle linéaire.

Une idée intuitive pour améliorer cela est de projeter les "inputs" dans un espace de plus haute dimension d'abord en utilisant un ensemble de bases de fonctions et puis appliquer l'idée de régression linéaire dans cet espace.

Par exemple, un scalaire x peut être projeté dans l'espace des puissances de x : $\phi(x) = (1, x, x^2, \ldots)^T$, une régression linéaire avec $\phi(x)$, on obtient une régression multiple.

Tant que les projections sont des fonctions fixées au préalable (indep. de w) ceci est tjrs dans la classe des modèle paramétriques.

La mise en application de cette idée nécessite de savoir comment choisir cette base de fonctions. Le formalisme des GPR permet de répondre à cette question.

Dans la suite, on suppose que la fonction de base ϕ est donnée.

Plus précisément, on introduit $\phi(x)$ qui est une application d'un espace D-dimensionnel (input $x \in \mathbb{R}^D$) vers un espace N-dimensionnel des caractéristique.

Notation : On note $\Phi(X)$ la matrice formée par les colonnes $\phi(x)$ données dans le "training set".

Maintenant, le modèle devient

$$f(x) = \phi(x)^T w \quad \text{où maintenant } w \in \mathbb{R}^N$$

→ L'analyse de ce modèle est analogue à l'étude de régression

linéaire précédente ; il faut juste remplacer X par $\Phi(x)$

Par analogie, la prédiction f_* est donnée par la loi.

$$f_* / X_*, X, Y \sim \mathcal{N}\left(\frac{\phi(x_*)^T A^{-1} \Phi Y}{\sigma_n^2}, \phi(x_*)^T A^{-1} \phi(x_*) \right)$$

où $\Phi := \Phi(X)$ et $A = \frac{\Phi \Phi^T}{\sigma_n^2} + \Sigma_p^{-1}$

Rq : Faire des prédictions à l'aide de cette équation, nécessite d'inverser A de taille $N \times N$ qui peut être problématique quand N est grand.

Afin de résoudre ce problème, on peut réécrire l'équation précédente comme :

$$f_* / X_*, X, Y \sim \mathcal{N}\left(\phi_*^T \Sigma_p \Phi (K + \sigma_n^2 I)^{-1} Y, \right.$$
$$\left. \phi_*^T \Sigma_p \phi_* - \phi_*^T \Sigma_p \Phi (K + \sigma_n^2 I)^{-1} \Phi^T \Sigma_p \phi_* \right)$$

où $\phi_* = \phi(x_*)$ et $K = \Phi^T \Sigma_p \Phi$

Exercice M.q Le terme d'espérance dans les deux formule précédentes sont égale.
Montrer l'égalité des variances en utilisant le lemme suivant :

Lemme

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

$$Z \in \mathbb{R}^{n \times n}, \quad W \in \mathbb{R}^{m \times m}, \quad U, V \in \mathbb{R}^{n \times m}$$

Il faut appliquer ce lemme avec

$$\left[Z^{-1} = \Sigma_p \, , \; W^{-1} = \sigma_n^2 I \; ; \; V = U = \underline{\Phi} . \right]$$

r.) L'avantage de cette deuxième représentation de la prédiction f_* est qu'on inverse des matrice de taille $n \times n$ ce qui est mieux quand $n < N$.

b.) Dans cette nouvelle expression de la prédiction f_* on voit toujours apparaître :

$\underline{\Phi}^T \Sigma_p \underline{\Phi}$, $\underline{\Phi}_*^T \Sigma_p \underline{\Phi}$ ou $\underline{\Phi}_*^T \Sigma_p \underline{\Phi}_*$, en d'autres termes on voit toujours apparaître la forme $\underline{\Phi}(x)^T \Sigma_p \underline{\Phi}(x')$

où $x, x' \in \{x, x_*\}$.

En partant de cette déf. on peut introduire

$$k(x, x') = \phi(x)^T \Sigma_p \phi(x').$$

Dans la suite, cette fonction k est appelé fonction de Covariance ou tout simplement noyau.

→ Comme Σ_p est symétrique définie positive, on peut trouver $\Sigma_p^{1/2}$ t.q $(\Sigma_p^{1/2})^2 = \Sigma_p$.

On peut alors introduire $\psi(x) = \Sigma_p^{1/2} \phi(x)$, on peut réécrire. $\boxed{k(x, x') = \psi(x)^T \psi(x')}$ Comme produit scalaire.

II - Point de vue espace fonctionnel.

Déf: Un processus gaussien est une famille de V.a. t.q toute sous famille de dimension finie forme un vecteur gaussien.

Un processus gaussien est complétement caractérisé par sa fonction moyenne et sa fonction covariance.

On définit, une fonction moyenne $m(x)$ et une fonction de covariance $k(x, x')$ d'un processus réel $f(x)$ comme :

$$m(x) = \mathbb{E}[f(x)]$$

$$k(x, x') = \mathbb{E}\left[[f(x) - m(x)][f(x') - m(x')] \right]$$

et on écrit le processus gaussien comme

$$f(x) \sim GP(m(x), k(x,x'))$$

Dans la pratique $m(x)=0$, mais on peut proposer d'autres moyennes.

⚠ Généralement, les processus gaussiens sont définis comme v.a dépendant du temps t ($(X_t)_{t\geq 0}$ est un processus gaussien).

Dans l'utilisation GPR ce n'est pas le cas, les v.a sont indéxés par un ensemble $\mathcal{X}$: ensemble des "inputs" possibles qui est généralement $\mathbb{R}^D$.

→ On se permet l'abus de notation

$f_i = f(x_i)$ est la v.a correspondant au couple (x_i, y_i)

$(y_i = f(x_i) + \epsilon)$

⚠ Dans les méthodes GP, on suppose l'hypothèse de <u>Cohérence</u> vérifiée :

Si $(U, V) \sim \mathcal{N}(\mu, \Sigma) \Rightarrow U \sim \mathcal{N}(\mu_1, \Sigma_{11})$ où Σ_{11} est la sous-matrice adéquate de Σ

<u>Exemple</u>: GP par le modèle Bayésien <u>linéaire</u>.

Dans ce cas, $f(x) = \phi(x)^T w$ où $w \sim \mathcal{N}(0, \Sigma_p)$ loi à priori

$$\mathbb{E}[f(x)] = \phi(x)^T \mathbb{E}[w] = 0$$

$$\mathbb{E}[f(x) f(x')] = \phi(x)^T \mathbb{E}[ww^T] \phi(x') = \phi(x)^T \Sigma_p \phi(x')$$

i.e $f(x)$ et $f(x')$ sont jointement gaussien de moyenne 0 et de covariance :

$\phi(x)^T \Sigma_p \phi(x)$

Ainsi, les valeurs des fonctions $f(x_1), \dots, f(x_n)$ correspondent à une loi gaussienne multivariée.

→ Dans la suite, on se focalise sur une fonction de Covariance spécifique donnée par:

$$\text{Cov}\left(f(x_p), f(x_q)\right) = k(x_p, x_q) = \exp\left(-\frac{1}{2}|x_p - x_q|^2\right)$$

Notons que la Covariance entre les outputs $f(x_p)$ et $f(x_q)$ sont écrites comme une fonction des inputs x_p et x_q

§ Prédiction avec de observations sans bruit.

On considère, le cas simple où les observations ne contiennent pas de bruit i.e on observe directement

$$\{(x_i, \underbrace{f_i}_{\text{"}y_i\text{"}}), \text{t.q } i = 1, \dots, n\} \quad \underline{\text{i.e}} \quad \boxed{y = f(x)}$$

la loi jointe des "training outputs" f et des "testing outputs" f_*

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} k(X,X) & k(X,X_*) \\ k(X_*,X) & k(X_*,X_*) \end{bmatrix}\right)$$

S'il y a n : "training points" et n_* : "testing point"

alors $K(X, X_*)$ n'est autre que la matrice $n \times n_*$ des Covariances calculé en toute les paires "training" et "testing"

et donc on peut en déduire que

$$f_* \mid X_*, X, \underbrace{f}_{=y} \sim \mathcal{N}\Big(K(X_*,X)K(X,X)^{-1} f,$$

$$K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*)\Big)$$

$y \equiv f$ observable

Prédiction avec des observations bruitées :

Pour une modélisation plus réaliste, c'est plus naturel de supposer qu'on a pas accès directement aux fonctions valeurs mais des versions bruitées :

$$y = f(x) + \varepsilon .$$

On suppose que le bruit $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$. iid

Dans ce cas,

$$\mathrm{Cov}(y_p, y_q) = K(x_p, x_q) + \sigma_n^2 \delta_{pq}$$

$\delta_{pq} = \mathbb{1}_{p=q}$. Donc

$$\boxed{\mathrm{Cov}(y) = K(X,X) + \sigma_n^2 I .}$$

Par la même procédure qu'avant, on a la loi jointe

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right)$$

Donc

$$f_* \mid X, y, X_* \sim \mathcal{N}\left(\bar{f}_*, \mathrm{Cov}(f_*)\right) \quad \text{avec}$$

$$\bar{f}_* = \mathbb{E}\left[f_* \mid X, y, X_*\right] = K(X_*, X)\left[K(X, X) + \sigma_n^2 I\right]^{-1} y$$

$$\mathrm{Cov}(f_*) = K(X_*, X_*) - K(X_*, X)\left[K(X, X) + \sigma_n^2 I\right]^{-1} K(X, X_*)$$

Dans la suite, pour simplifier ces écritures on introduit

$$K = K(X, X) \; ; \quad K_* = K(X, X_*)$$

et s'il y a un seul point test x_*, on écrit $k(x_*) = k_*$

On obtient,

$$\bar{f}_* = k_*^T \left(K + \sigma_n^2 I\right)^{-1} y \quad (1)$$

$$\mathrm{Var}(f_*) = k(X_*, X_*) - k_*^T \left(K + \sigma_n^2 I\right)^{-1} k_* \quad (2)$$

L'équation précédente de $\bar{f}_*$ équation (1) s'écrit :

$$\bar{f}_* = \sum_{i=1}^{n} \alpha_i \, k(x_i, x_*)$$

où $\alpha = \left(K + \sigma_n^2 I\right)^{-1} y$.

$\bar{f}_*$ s'exprime dans une base de fonction représentée par le noyau k.

§ Mise en place de méthodes GPR :

On rappelle que $f : \mathbb{R}^p \to \mathbb{R}$ suit un modèle GPR noté

$$f \sim GP(\mu, k)$$

Si pour des "inputs points" $x_1, x_2, \dots, x_n \in \mathbb{R}^p$

alors

$$(f(x_1), f(x_2), \dots, f(x_n)) \sim \mathcal{N}(\mu, K_{xx})$$

avec $\boxed{(K_{xx})_{ij} = k(x_i, x_j)}$

→ Les noyaux k peuvent être n'importe quelle fonction symétrique positive i.e. vérifiant :

$$\sum_{i,j=1}^{n} k(x_i, x_j)\, \xi_i \xi_j \geq 0 \quad \forall \; x_k \in \mathbb{R}^p \; ; \; \xi_k \in \mathbb{R}.$$

$$k : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$$

[p : nbre de caractéristiques
n : nbre d'observations.]

Classe des noyaux RBF : "Radial Basis Function"

Les noyaux de type RBF dépendent seulement de $\|x - x'\|$, comme par exple le noyau gaussien ou exponentiel :

$$k(x, x') = \exp\left(-\frac{1}{2\ell^2}\|x - x'\|^2\right)$$

ℓ : "length scale" mesure la distance à avoir entre les données pour être décorrélées.

→ ± ℓ existe d'autres noyaux : noyau Matern (MA)

$$k(x,x') = \frac{\sigma^2 2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\,\frac{\|x-x'\|}{\ell}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\,\frac{\|x-x'\|}{\ell}\right)$$

où Γ : c'est la fonction gamma.

(·) $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$.

(·) K_ν : la fonction modifiée de Bessel de second type

$$K_\nu(z) = \frac{\Gamma(\nu+\frac{1}{2})(2z)^\nu}{\sqrt{\pi}} \int_0^\infty \frac{\cos t}{(t^2+z^2)^{\nu+\frac{1}{2}}}\, dt$$

Étant donné, les données de test : x_*

$f_* / x, x_*, y \sim$ loi gaussienne

d'espérance $\mathbb{E}[f_* | x, x_*, y] = K_{x_*,x}\left[K_{x,x} + \sigma_n^2 I\right]^{-1} y$

et de variance.

$\mathrm{Var}[f_* | x, x_*, y] = K_{x_*,x_*} - K_{x_*,x}\left[K_{x,x} + \sigma_n^2 I\right]^{-1} K_{x,x_*}$

§ optimization de l'hyperparamètre du noyau:

On suppose que le noyau k dépend d'un paramètre λ

ex: $\lambda = (\ell, \sigma_n)$ dans le cas du noyau exponentiel.

pour trouver λ^* : on a: output training

$$\lambda^* = \underset{\lambda}{\text{argmax}} \; \text{Log}\, p(Y/X, \lambda)$$ ← hyperparamètre.

↑ input training

Comme on est dans un modèle gaussien on a:

$$\log p(Y/X,\lambda) = -\frac{1}{2}\left[Y^T (K_{X,X} + \sigma_n^2 I)^{-1} Y + \log \det (K_{X,X} + \sigma_n^2 I) \right] - \frac{n}{2} \log 2\pi$$

§ Calcul de sensibilité:

On rappelle que la prédiction GPR

$$f_* / X, Y, X_* \sim \mathcal{N}\left(\mathbb{E}[f_*/X,Y,X_*], \text{Var}[f_*/X,Y,X_*]\right)$$

avec

$$\mathbb{E}[f_*(X,Y,X_*] = \mathbb{E}[X_*] + K_{X,X_*}\left[K_{X,X} + \sigma_n^2 I\right]^{-1} Y$$

$$\text{Var}[f_*|X,Y,X_*] = K_{X_*,X_*} - K_{X_*,X}\left[K_{X,X} + \sigma_n^2 I\right]^{-1} K_{X,X_*}$$

On rappelle que

$(K_{X,X})_{ij} = k(x_i, x_j)$. Dans le cas du noyau RBF

on a $\boxed{k(x_i, x_j) = \sigma_f \exp\left(-\frac{1}{2\ell^2} \| x_i - x_j \|^2\right)}$

→ Une fois le GPR entraîné, sa prédiction donnée par $\mathbb{E}[f_* / X, y, X_*]$, on peut en déduire directement $\frac{\partial}{\partial X_*}$.

En particulier, dans le cas d'un noyau exponentiel gaussien :

$$\frac{\partial}{\partial X_*} \mathbb{E}[f_* / X, y, X_*] = \frac{\partial}{\partial X_*} \mathbb{E}[X^*] + \underbrace{\frac{\partial}{\partial X_*} K_{X_*,X}} [K_{X,X} + \sigma_n^2 I]^{-1} y$$

$$(K_{X_*,X})_{ij} = \sigma_f \exp\left(-\frac{1}{2\ell^2} \langle X_i^* - X_j, X_i^* - X_j \rangle\right)$$

$$= -\frac{\sigma_f}{2\ell^2} \frac{\partial}{\partial X_*} \langle X_i^* - X_j, X_i^* - X_j \rangle \times \underbrace{\exp\left(-\frac{1}{2\ell^2} \| X_i^* - X_j \|^2\right)}$$

$$= -\frac{1}{2\ell^2} 2 (X_i^* - X_j) (K_{X_*,X})_{ij}$$

$$= \frac{1}{\ell^2} (X_j - X_i^*) (K_{X_*,X})_{ij}$$

Donc en prenant $\mathbb{E}[X_*] = 0$

$$\frac{\partial}{\partial x_0} \mathbb{E}[f_* / X, y, X_*] = \frac{1}{\ell^2} (X - X_*) K_{X_*, X} \left[K_{X,X} + \sigma_n^2 I \right]^{-1} y$$

Donc avec cette formule, il n'y a pas besoin d'entraîner à nouveau le GPR pour calculer la dérivée de la fonction cible initiale qu'on a déjà apprise par la méthode GPR

L'estimée de la dérivée est presque gratuite :