Simulation basée sur les chaînes de Markov MCMC

Prof. Mohamed El Merouani

2018/2019

Plan

- Rappels sur l'inférence Bayésienne
- Introduction aux Méthodes de Simulation basées sur les chaînes de Markov (MCMC)
- Echantillonnage de Gibbs
- Algorithme de Metropolis-Hastings
- Stratégies hybrides

Rappels sur l'inférence Bayésienne

Dans l'approche Bayésienne de la statistique, le paramètre θ est une v.a. avec une distribution de probabilité. L'information initiale sur θ est repésentée par la distribution a priori $\pi(\theta)$. Étant donnée la distribution $p(x/\theta) = \prod_{i=1}^{n} f(x_i/\theta)$ appelée vraisemblance, qui fournit les probabilités des valeurs d'échantillon sachant θ, on obtient la distribution a posteriori en appliquant le théorème de Bayes :

$$\pi(\theta/x) = \frac{p(x/\theta)\pi(\theta)}{\int p(x/\theta)\pi(\theta)d\theta}$$

Le dénominateur de cette expression, $p(x) = \int p(x/\theta)\pi(\theta)d\theta$, est la distribution marginale des données. La distribution a posteriori contient toute l'information nécessaire pour réaliser des inférences.

Rappels sur l'inférence Bayésienne

Si nous voulons obtenir un estimateur ponctuel, nous pouvons résumer une telle distribution avec son mode, sa médiane, sa moyenne,...Pour le choix, nous introduisons une fonction de perte $\ell : \Theta \times \Theta \longrightarrow \mathbb{R}$, et nous suggérons comme estimateur $\delta(x)$ la solution, si elle existe, de

$$\min_{\delta} \int \ell(\theta, \delta)\pi(\theta/x)d\theta$$

Par exemple, si $\ell(\theta, \delta) = (\theta - \delta)^2$ et l'espérance de θ/x existe, $\delta(x) = E(\theta/x)$.
Si nous voulons une région de probabilité $(1 - \alpha)$, il suffit de choisir un ensemble C_α tel que

$$\int_{C_\alpha} \pi(\theta/x)d\theta \geq 1 - \alpha$$

Rappels sur l'inférence Bayésienne

Pour tester l'hypothèse $H_0 : \theta \in \Theta_0$, contre l'hypothèse $H_1 : \theta \in \Theta_1$, nous acceptons H_0 si

$$\int_{\Theta_0} \pi(\theta/x)d\theta \geq \int_{\Theta_1} \pi(\theta/x)d\theta.$$

Pour réaliser des prédictions, nous utilisons la distribution predictive :

$$f(y/x) = \int f(y/\theta)\pi(\theta/x)d\theta$$

Lorsque la distribution a priori est non informative, les résultats obtenus sont, généralement et formellement, les mêmes que ceux obtenus par l'Inférence Classique, bien que leurs interprétations soient différentes. Aussi, dans de nombreux cas, en augmentant les données disponibles, les résultats obtenus par Inférence Beyésienne sont formellement similaires aux résultats obtenus par Inférence Classique.

Rappels sur l'inférence Bayésienne

- Donc, comme dans la simulation, il est conceptuellement possible d'obtenir de grandes quantités de données pour l'inférence, nous fournissons des estimateurs classiques dans l'analyse des résultats.
- Cependant, dans la construction des métamodèles, nous fournissons des estimateurs Bayésiens. Alors, nous aurons généralement accès à peu de données de telles situations, avec lesquelles l'utilisation de l'information a priori est primordiale.

Introduction aux Méthodes de Simulation basées sur les chaînes de Markov :

- Les méthodes basées sur les chaînes de Markov (MCMC) utilisent la simulation des chaînes de Markov et la statistique Bayésienne et s'appliquent principalement en Intélligence Artificielle.
- L'idée basique des MCM est très simple : On veut générer un échantillon à partir d'une distribution $\pi(x)$ avec $x \in \mathcal{X} \subset \mathbb{R}^n$, mais on ne peut pas le faire directement. Pourtant, on peut construire un processus de Markov $p(\cdot/\cdot)$ dont l'espace d'états est $\mathcal{X}$, à partir duquel c'est facil d'échantillonner, et sa distribution d'équilibre est $\pi(x)$.
- Si nous laissons parcourir le processus une période suffisamment longue, on échantillonnera approximativement à partir de π.

Introduction aux Méthodes de Simulation basées sur les chaînes de Markov :

Nous avons, alors, la stratégie suivante pour générer un échantillon approximatif à partir de π de taille N

$i = 0$, choisir X_0 arbitrairement
Jusqu'à avoir convergence
    Générer $X_{i+1} \sim p(X_i, \cdot)$
    Faire $i = i + 1$
Depuis $j = 1$ jusqu'à N
    Générer $X_{i+j} \sim p(X_{i+j-1}, \cdot)$
    Sortir X_{i+j}
    Faire $j = j + 1$

Entre autres problèmes, nous devons étudier les méthodes de construction de chaînes avec la distribution désirée. Nous allons voir trois d'entre elles par la suite.

Échantillonnage de Gibbs :

- L'intérêt récent pour les méthodes des chaînes de Markov a commencé par le succès que l'échantillonnage de Gibbs a eu dans l'inférence Bayésienne.
- Son nom le plus correct serait l'échantillonnage de substitution, mais dans la première exposition (Geman et Geman, 1984) la distribution de Gibbs est intervenue, d'où son nom.

Comme introduction, considérons l'exemple simple suivant :
Supposons que (X, Y) sont des v.a. de Bernoulli de loi conjointe :

X	Y	$P(X, Y)$
0	0	p_1
1	0	p_2
0	1	p_3
1	1	p_4

avec $p_i > 0$, $\sum_{i=1}^{4} p_i = 1$.

Échantillonnage de Gibbs :

- La loi marginale de X est de Bernoulli de paramètre $p_2 + p_4$, c'est-à-dire, $P(X = 1) = p_2 + p_4$.
- Les distributions de $X/Y = y$ et de $Y/X = x$ sont aussi faciles à calculer. Par exemple, la distribution de $X/Y = 1$ est de Bernoulli avec paramètre $\frac{p_4}{p_3+p_4}$, c'est-à-dire $P(X = 1/Y = 1) = \frac{p_4}{p_3+p_4}$.
- En effet, toutes les distributions conditionnelles peuvent être exprimées par deux matrices :

$$A_{yx} = \begin{pmatrix} P(Y=0/X=0) & P(Y=1/X=0) \\ P(Y=0/X=1) & P(Y=1/X=1) \end{pmatrix} = \begin{pmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{pmatrix}$$

$$A_{xy} = \begin{pmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{pmatrix}$$

Échantillonnage de Gibbs :

Considérons le schéma itératif suivant :

Choisir $Y_0 = y_0$, $j = 1$
  Répéter
    Générer $X_j \sim X/Y = y_{j-1}$
    Générer $Y_j \sim Y/X = X_j$
    $j = j + 1$

La suite $(X_n)_n$ définit une chaîne de Markov de matrice de transition :

$$A = A_{yx}.A_{xy}$$

Comme les probabilités p_i sont positives, cette chaîne est érgodique et elle a comme distribution limite, la marginale de X, alors

$$(p_1 + p_3p_2 + p_4) = (p_1 + p_3p_2 + p_4)A$$

avec lequel

$$X_n \xrightarrow{\mathcal{L}} X$$

Échantillonnage de Gibbs :

Par analogie

$$Y_n \xrightarrow{\mathcal{L}} Y$$

$$(X_n, Y_n) \xrightarrow{\mathcal{L}} (X, Y)$$

- La procédure décrite s'appelle échantillonnage de Gibbs. Elle nous donne, dans ce cas, une chaîne de Markov avec la distribution limite désirée, et elle est, en fait, assez générale.
- En effet, supposons que nous voulons échantillonner à partir d'une v.a. p-variée $X = (X_1, X_2, \cdots, X_p)$ de distribution π. Nous ne pouvons pas le faire directement, mais nous connaissons les densités conditionnelles de $X_s/X_r, r \neq s$ que nous désignons par $\pi(x_s/x_r, r \neq s)$.

L'échantillonnage de Gibbs, dans ce cas est :

Échantillonnage de Gibbs :

L'échantillonnage de Gibbs, dans ce cas est :

Choisir $X_1^0, X_2^0, \cdots, X_p^0$, $j = 1$
Répéter
Générer $X_1^j \sim X_1/X_2^{j-1}, \cdots, X_p^{j-1}$
Générer $X_2^j \sim X_2/X_1^j, X_3^{j-1}, \cdots, X_p^{j-1}$
$\cdots\cdots$
Générer $X_p^j \sim X_p/X_1^j, X_2^j \cdots, X_{p-1}^j$
$j = j + 1$

Remarquons que $X_n = (X_1^n, X_2^n, \cdots, X_p^n)$ définit une chaîne de Markov de transitions

$$P_G(x_n, x_{n+1}) = \prod_{i=1}^{p} \pi(x_i^{n+1}/x_j^n, j > i; x_j^{n+1}, j < i)$$

Sous des conditions suffisamment générales, on obtient la définition correcte et la convergence de l'échantillonnage.

Échantillonnage de Gibbs :

Proposition : (Roberts & Smith, 1994)

Supposons que $D = \{x : \pi(x) > 0\}$ est un ensemble produit $D = \prod_{i=1}^{p} D_i$. Alors :

- $\pi(x_i/x_j, j \neq i)$ est bien définie $\forall x \in D$ et P_G est bien définie.
- P_G est irréductible par rapport à π et elle est apériodique.
- π est invariante par rapport à P_G.
- $X_n \xrightarrow{\mathcal{L}} X$

Échantillonnage de Gibbs :

Exemple :

Supposons que nous voulons échantillonner à partir de la densité

$$\pi(x_1, x_2) = \frac{1}{\pi} \exp\left(-x_1(1 + x_2^2)\right)$$

pour $(x_1, x_2) \in (0, \infty) \times (-\infty, \infty)$. Nous avons

$$\pi(x_1/x_2) = \frac{\pi(x_1, x_2)}{\pi(x_2)} \propto \pi(x_1, x_2) \propto \exp\left(-x_1(1 + x_2^2)\right)$$

$$\pi(x_2/x_1) \propto \pi(x_1, x_2) \propto \exp\left(-x_1 x_2^2\right)$$

avec lequel

$$x_1/x_2 \sim \mathcal{E}xp(1 + x_2^2)$$

$$x_2/x_1 \sim \mathcal{N}(0, \sigma^2 = \frac{1}{2x_1})$$

et l'échantillonnage de Gibbs devient :

Choisir une valeur initiale X_2^0, $j = 1$
Répéter

Générer $X_1^j \sim \mathcal{E}xp\left(1 + (X_2^{j-1})^2\right)$

Générer $X_2^j \sim \mathcal{N}\left(0, \frac{1}{2X_1^j}\right)$

$j = j + 1$

Rappelons que chaque itération ne doit pas être faite dans la séquence naturelle, elle peut être faite, par exemple, dans une séquence aléatoire. De plus, l'échantillonnage peut être effectué par blocs d'indices.

Algorithme de Metropolis-Hastings :

L'algorithme de Metropolis-Hastings (MH) fournit des transitions de X_n à X_{n+1} comme suit. Soit $q(x, y)$ une fonction de transition de probabilités ; alors, en supposant que $X_n = x$, la transition est :

Générer $y \sim q(x, Y)$
Si $\pi(x)q(x, y) > 0$
$$\alpha(x, y) = \min\left\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right\}$$
Sinon
$$\alpha(x, y) = 1$$
Faire $X_{n+1} = y$ avec probabilité $\alpha(x, y)$
Sinon , $X_{n+1} = x$

L'algorithme de Metroplis et al. (1953) correspond au cas précédent avec la distribution de test symétrique, c'est-à-dire $q(x, y) = q(y, x)$, avec les simplifications correspondantes. Remarquons que pour calculer $\alpha(x, y)$, il n'est pas, en fait, nécessaire de connaître π, mais seulement une fonction $\pi_1 \propto \pi$.

Algorithme de Metropolis-Hastings :

Il existe plusieurs autres variantes de l'algorithme MH. Notons que l'échantillonnage de Gibbs peut être vu comme une variante de l'algorithme MH.L'algorithme MH définit un processus de Markov avec probabilité de transition de x à y donnée par :

$$p_{MH}(x,y) = \begin{cases} q(x,y)\alpha(x,y) & \text{si } y \neq x \\ 1 - \int q(x,z)\alpha(x,z)dz & \text{si } y = x \end{cases}$$

Il est facile de voir que π est invariant pour le processus avec transition p_{MH}, donc, il suffit d'assurer l'apériodicité et la π-irréductibilité de p_{MH} pour avoir que $X_n \xrightarrow{\mathcal{L}} X$.

Proposition : (Roberts & Smith, 1994)

1. Si q est apériodique, p_{MH} est apériodique.
2. Si q est π-irréductible et $q(x,y) = 0$ si et seulement si $q(y,x) = 0$, p_{MH} est π-irréductible

Algorithme de Metropolis-Hastings :

- Le problème serait de choisir q, qui est en principe arbitraire, à condition que cela assure la convergence. Habituelllement, on choisit les distributions de test symétriques et de simulation facile.
- Par exemple, dans le cas continu, on choisit des distributions normales de moyenne x et de matrice de covariances Σ.
- On choisit Σ de façon que le taux d'acceptation soit d'environ 25% (Gelman et al., 1996).
- Pour le cas discret, on choisit une distribution uniforme sur un voisinage de x.

Stratégies hybrides :

- Les stratégies hybrides sont particulièrement intéressantes.
- Une d'entre elles, dûe à Müller (1991) consiste en faire des pas de Gibbs toujours lorsque les distributions conditionnelles sont disponibles (et sont de simulation efficiente) et des pas de Metropolis en cas contraire.
- Cette stratégie s'applique à l'inférence et à la prédiction (apprentissage) en des modèles de réseaux de neurones.

Exercices :

Exercice 1 :

Supposons que la densié conjointe de (x_1, x_2) est proportionnelle à $\pi(x_1, x_2) = \frac{1}{\pi} \exp\left(-x_1(1 + x_2^2)\right)$

1. Estimer la moyenne de x_1 à partir d'un échantillon simulé selon l'échantillonnage de Gibbs.
2. Donner un algorithme de Metroplis-Hastings pour ce problème.

Exercice 2 :

Supposons que $f(x, y) \propto xye^{-xy}\mathbb{I}_{[0,B]\times[0,B]}(x, y)$.
Donner l'échantillonnage de Gibbs associé.

Exercice 3 :

Supposons que la distribution conjointe de X et Y est :
$f(x, y) \propto C_n^x y^{x+\alpha-1}(1 - y)^{n-x+\beta-1}$, pour $x = 0, 1, \cdots, n$, $0 \leq y \leq 1$.
Donner l'échantillonnage de Gibbs associé.