Chaînes de Markov

Gabriel Lang

Université Paris-Saclay, AgroParisTech

MAEE, master BEE

Origine des modèles

Deux grandes familles de modèles à l'origine.

Modèles déterministes : mécanique

Ex : loi de chute des corps

$z(t) = z_0 - \frac{1}{2} g t^2$

Modèles stochastiques : jeux de hasard

Ex : roulette pari sur (Noir, Rouge), recherche de martingale

$\mathbb{P}$(R,N,R) $= 1/8$

$\mathbb{P}$(N sort 2 fois sur 3 tirages) $= 3/8$

$\mathbb{P}$(Nombre N>Nombre R indéfiniment) $= 0$

Séries chronologiques

On observe une série de données mesurées à intervalle de temps fixé $x_0, x_1, \dots x_n$.
Ex : séries de températures, comptage journalier d'individus d'une espèce, données de présence absence d'une espèce...
Les données sont des résultats de tirage aléatoire de variables X_i, $i = 0, \dots, n$.

Echantillon d'une loi

- Les variables sont de même loi de probabilité.
- Les variables sont indépendantes.
- L'ordre des variables n'a pas importance.

Série chronologique

- Les variables ne sont pas nécessairement de même loi de probabilité.
- Les variables ne sont pas indépendantes.
- La dépendance entre variables est plus forte lorsque l'écart de temps est petit.

Définition d'un modèle de série chronologique

Modélisation stochastique

Modèle = Loi jointe des variables.

- Echantillon : Définition de la loi d'une des variables et indépendance.
- Série chronologique : Définition de la loi jointe de l'ensemble des variables.

On peut définir ces lois jointes de manière récurrente par l'utilisation des lois conditionnelles

Lois conditionnelles

On définit la loi de chaque variable conditionnellement aux précédentes : $\mathcal{L}(X_i|X_0,\dots,X_{i-1})$.
On recalcule les lois jointes par déconditionnement :

$$\mathcal{L}(X_0,\dots,X_i) = \mathcal{L}(X_i|X_0,\dots,X_{i-1}) \times \mathcal{L}(X_0,\dots,X_{i-1})$$

Chaîne de Markov

On considère que la loi conditionnelle ne dépend réellement que de la dernière valeur.

Propriéte de Markov

$$\mathcal{L}(X_i|X_0,\dots,X_{i-1}) = \mathcal{L}(X_i|X_{i-1})$$

Exemple :
Dans un modèle de population à générations séparées, chaque génération ne dépend que de la génération précédente.

Dépendance

La propriété de Markov n'implique pas l'indépendance entre variables non consécutives.

Exemple :
Effectif d'une génération par rapport à l'effectif des grands parents.

Usages

Nous allons présenter les deux principaux usages des chaines de Markov avec deux exemples.

- *Simulation* d'une loi de probabilité ayant un nombre d'éventualités supérieur à la capacité machine.
  Ex : Mélange automatisé d'un jeu de carte $\Longrightarrow$ Caractéristiques du modèle
- *Modélisation* d'un phénomène se déroulant dans le temps mesuré à date régulière.
  Ex : Modèle de Wright-Fisher $\Longrightarrow$ Référence en génétique des populations

Mélange d'un paquet de carte

On souhaite proposer des paquets de cartes mélangés pour un jeu en ligne.

- Quel est le but du mélange ?
- Comment caractériser un mélange idéal ?
- Comment juger de la qualité ?
- Quelles sont les méthodes informatiques pour tirer une variable aléatoire ?
- Combien d'éventualités pour un mélange de 52 cartes ?

Un modèle explicite

On restreint notre paquet à 3 cartes : R, D, et V.
L'ordre RDV désigne un paquet où la carte de dessus est le roi, la suivante la dame, la dernière le valet. On répète une procédure identique un certain nombre de fois
Premier mélange
A chaque étape, on place la carte du dessus au dessous.

Un modèle explicite

On restreint notre paquet à 3 cartes : R, D, et V.
L'ordre RDV désigne un paquet où la carte de dessus est le roi, la suivante la dame, la dernière le valet. On répète une procédure identique un certain nombre de fois
Premier mélange
A chaque étape, on place la carte du dessus au dessous.
On obtient

$$RDV \Longrightarrow DVR \Longrightarrow VRD \Longrightarrow RDV$$

ou

$$RVD \Longrightarrow VDR \Longrightarrow DRV \Longrightarrow RVD$$

Quelles sont les caractéristiques de ce mélange ?

Un modèle explicite

On restreint notre paquet à 3 cartes : R, D, et V.
L'ordre RDV désigne un paquet où la carte de dessus est le roi, la suivante la dame, la dernière le valet. On répète une procédure identique un certain nombre de fois
Deuxième mélange
A chaque étape, on place la carte du dessus au dessous avec probabilité 0.5, entre les deux autres cartes sinon.
Exo : Partant de RDV, décrire les probabilité des differents ordres pour les trois premiers tours de mélange

Paramètre d'une chaîne de Markov à nombre d'état fini

Une chaîne de Markov à nombre d'état k est un processus aléatoire entièrement déterminée par

- Une matrice de transition P carrée de taille k stochastique (= dont la somme des lignes vaut 1)
- La loi initiale au temps L_0 (vecteur ligne de dimension k).

La loi L_n de la chaîne à l'instant n est $L_n = L_0 P^n$.

Loi stationnaire

- Une loi stationnaire est un vecteur propre à gauche de la matrice P pour la valeur propre 1 ($LP = L$). Ce vecteur a des coefficients positifs sommant à 1.
- Il existe toujours au moins une loi stationnaire (1 est valeur propre et il y a un vecteur propre à coefficients positifs).

Cette loi n'est pas nécessairement unique.

Irréductibilité

Une chaîne est irréductible si et seulement si la transition de chaque état i vers chaque état j est possible au cours de l'évolution de la chaîne.

- L'irréductibilité dépend de la matrice P
- En pratique on représente la matrice par un graphe orienté sur les états et on vérifie s'il existe un chemin entre i et j pour chaque couple d'état (i, j).

La chaîne n'est pas irréductible

- s'il existe un élément i qui reçoit des flêches mais n'en émet pas (i état absorbant)
- s'il existe un chemin de i à j mais pas de retour (i état transient)
- s'il n'existe aucun chemin de i à j, ni de j à i (ensemble d'état scindé en sous-ensembles)

Périodicité

Une chaîne est périodique de période $p > 1$ si et seulement si on peut découper l'ensemble des états en p sous-ensembles disjoints telle que l'évolution de la chaîne choisit cycliquement les états dans les p sous-ensembles.

- La périodicité dépend de la matrice P.
- Si elle existe, la période est la longueur du plus petit chemin fermé.
- En pratique, on choisit dans le graphe un état et on numérote les états jusqu'au retour sur l'état de départ. On recommence avec les états non numérotés en conservant une numérotation cohérente. Si cette numérotation cohérente est impossible, la chaîne est apériodique.

La chaîne est apériodique

- s'il existe un état i tel que l'élément diagonal $p_{i,i} \neq 0$.
- s'il existe un couple (i, j) avec deux chemins de i à j dont les longueurs sont premières entre elles.

Chaîne de Markov à nombre fini d'états

Chaîne de Markov irréductible et apériodique

- Une seule loi stationnaire
- La loi de X_n converge toujours vers la loi stationnaire
- La convergence est très rapide (l'écart décroit comme une suite géométrique).
- L'histogramme calculé avec les variables successives converge vers la loi stationnaire (ergodicité).

Chaîne de Markov non irréductible et apériodique

- S'il y a deux espaces séparés, il peut y avoir une infinité de lois stationnaires
- La loi de X_n converge vers une des lois stationnaires. Le choix dépend de la loi de départ L_0.

Chaîne de Markov irréductible et périodique

- Une seule loi stationnaire.
- La loi de X_n ne converge pas vers la loi stationnaire (cycle).
- L'histogramme calculé avec les variables successives converge vers la loi stationnaire.

On considère une population de gènes qui se reproduit de manière non sexuée.

Définition

- On considère des générations séparées d'effectif fixe $2N$ (diploidie)
- Chaque individu de la génération $G+1$ a un père choisi avec indépendance et équiprobabilité $p = 1/2N$ dans la génération G.

Propriétés

- Certains pères n'ont pas de fils.
- Au bout d'un certain nombre de générations, tous les individus de la dernière génération proviennent d'un unique individu de la génération 1.
- Il est possible que tous les individus de la dernière génération soient issus d'un ancêtre commun dans un génération $G > 1$.

Modèle de Wright Fisher : nombre de fils

Soit i un père; soit v_i son nombre de fils; l'événement {le fils j a choisi i pour père} est une variable de Bernoulli de paramètre $p_i = 1/2N$. Il y a $2N$ fils qui choisissent indépendamment.

Effectifs des fils

- v_i suit une loi binomiale $B(2N, 1/2N)$, d'espérance 1, de variance $1 - 1/2N$.
- Pour deux pères i et i', v_i et $v_{i'}$ sont (faiblement) dépendants : $\mathrm{Cov}\,(v_i, v_{i'}) = -1/2N$

Grande population

Si N tend vers l'infini,

- la loi de v_i tend vers la loi de Poisson de paramètre 1.
- La probabilité d'être sans enfants $\mathbb{P}(v_i = 0) \to e^{-1} = 0.37$
- La probabilité d'avoir j enfants $\mathbb{P}(v_i = j) \to e^{-1}/j!$.
- Si t est la date d'extinction de la descendance d'un individu i, $\mathbb{P}(t = 1) = 0.37$ et par récurrence $\mathbb{P}(d = t) = exp(-\mathbb{P}(d = t - 1) - 1)$

Probabilité d'extinction en 10 générations =0.84, en 20 générations =0.91.
Si on s'intéresse à l'origine d'une population présente, la construction de la généalogie en partant de la génération 0 est inefficace car la plupart des individus disparaissent sans descendance.

Construction de l'arbre en remontant le temps

On part de la génération présente (dernière génération observée), on considère un sous-ensemble de k individus présents et on remonte les générations. On ne construit que l'arbre en ligne directe de ces k individus.
Si tous les pères d'une génération sont distincts les branches généalogiques restent séparées. Si on considère les $2N$ individus cette probabilité est très faible ($= (2N)!/(2N)^{2N} \approx e^{-2N}$). Mais pour une sous-population de taille k petite, cette probabilité est proche de 1.

Probabilités de branchement de l'arbre en présence de k branches distinctes

Si $2N$ est très grand par rapport à k :

- Avec une probabilité équivalente à $1 - k(k-1)/4N$,tous les individus ont un père distinct.
- Avec une probabilité équivalente à $k(k-1)/4N$, seuls deux individus ont le même père.
- Avec une probabilité d'ordre $1/N^2$, plus de deux individus ont un père commun.

Lorsque deux individus ont le même père on dit qu'il y a coalescence de la généalogie.

Calcul de la probabilité d'avoir k pères distinct

Une généalogie avec k pères distincts correspond au choix d'un k-uplet parmi les $2N$ pères possibles. Il y a $(2N)!/(2N-k)!$ k-uplets possibles. Il y a $(2N)^k$ choix possibles des pères. La probabilité P_0 d'avoir k pères différents pour k individus est

$$P_0 = \frac{(2N)!}{(2N-k)!(2N)^k} = \frac{(2N)\cdots(2N-k+1)}{(2N)^k} = \left(1-\frac{1}{2N}\right)\cdots\left(1-\frac{k-1}{2N}\right)$$

Pour k petit devant $2N$, $P_0 = 1 - \sum_{i=1}^{k-1} \frac{i}{2N} + O(N^{-2}) = 1 - \frac{k(k-1)}{4N} + O(N^{-2})$.
Si deux pères sont identiques, on choisit un couple de fils ($\frac{k(k-1)}{2}$ possibles), puis un père pour ces deux fils ($2N$ possibles) et un $k-1$-uplet de pères autres ($\frac{(2N-1)!}{(2N-k+1)!}$ possibles) :

$$P_1 = \frac{k(k-1)}{2(2N)^k}\frac{(2N)!}{(2N-k+1)!} = \frac{k(k-1)}{4N}\left(1-\frac{1}{2N}\right)\cdot\left(1-\frac{k-2}{2N}\right) = \frac{k(k-1)}{4N} + O(N^{-2})$$

La probabilité qu'il y ait plus de deux individus ayant un père commun est $O(1/N^2)$.