

Modèles de Markov et modèles de Markov cachés

Hélène Touzet

`helene.touzet@univ-lille.fr`

CNRS, Bonsai, CRIStAL

Exemple : Doudou le hamster [[modifier](#) | [modifier le code](#)]



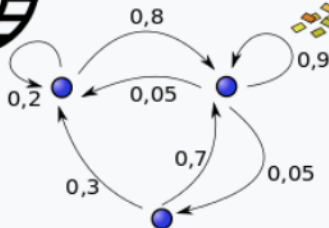
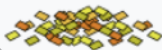
- Doudou le hamster ne connaît que trois endroits dans sa cage
 - les copeaux où il dort
 - la mangeoire où il mange
 - la roue où il fait de l'exercice
- toutes les minutes, il peut soit changer d'activité, soit continuer celle qu'il était en train de faire.

- quand il dort, il a 9 chances sur 10 de ne pas se réveiller la minute suivante
- quand il se réveille, il y a 1 chance sur 2 qu'il aille manger et 1 chance sur 2 qu'il parte faire de l'exercice
- le repas ne dure qu'une minute, après il fait autre chose
- après avoir mangé, il y a 3 chances sur 10 qu'il parte courir dans sa roue, mais surtout 7 chances sur 10 qu'il retourne dormir
- courir est fatigant pour Doudou ; il y a 8 chances sur 10 qu'il retourne dormir au bout d'une minute. Sinon il continue en oubliant qu'il est déjà un peu fatigué



Roue

Copeaux



Mangeoire

au départ	dort (copeaux)
minute 1	dort (copeaux)
minute 2	dort (copeaux)
minute 3	mange (mangeoire)
minute 4	court (roue)
minute 5	dort (copeaux)

- quelle est la probabilité de cette suite d'états ?

au départ	dort (copeaux)
minute 1	dort (copeaux)
minute 2	dort (copeaux)
minute 3	mange (mangeoire)
minute 4	court (roue)
minute 5	dort (copeaux)

- quelle est la probabilité de cette suite d'états ?

$$0,9 \times 0,9 \times 0,05 \times 0,3 \times 0,8$$

- un nombre fini d'états
- des probabilités qui décrivent le passage d'un état à l'autre
- processus sans mémoire: connaissance uniquement de l'état actuel

	copeaux	mangeoire	roue
copeaux	0,9	0,05	0,05
mangeoire	0,7	0	0,3
roue	0,8	0	0,2

matrice de transition

- état initial: le hamster dort $[1 \ 0 \ 0]$
- que fera-t-il au bout d'une minute

$$[1 \ 0 \ 0] \begin{bmatrix} 0,9 & 0,05 & 0,05 \\ 0,7 & 0 & 0,3 \\ 0,8 & 0 & 0,2 \end{bmatrix} = [0,9 \ 0,05 \ 0,05]$$

- que fera-t-il au bout de deux minutes

$$[1 \ 0 \ 0] \begin{bmatrix} 0,9 & 0,05 & 0,05 \\ 0,7 & 0 & 0,3 \\ 0,8 & 0 & 0,2 \end{bmatrix}^2 = [0,885 \ 0,045 \ 0,07]$$

- cas général: $X_n = X_0 P^n$

X_0 vecteur initial, P matrice de transition, X_n vecteur à l'étape n

Le casino malhonnête



- 1 dé normal : probabilité de $1/6$ par face
- 1 dé pipé : le 6 a une probabilité de 0.5 et les autres faces ont une probabilité de 0.1
- passage du dé normal au dé pipé : probabilité 0.05
- passage du dé pipé au dé normal : probabilité 0.1

4 6 4 3 2 3 5 6 1 4 6

- pas de changement de dé
- quelle est la probabilité de cette observation avec le dé normal ?

$$\frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6}$$

$\frac{1}{6}$: probabilités d'émission 0,95: probabilités de transition

- quelle est la probabilité de cette observation avec le dé pipé ?

$$0,1 \times 0,9 \times 0,5 \times 0,9 \times 0,1 \times 0,9 \times 0,1 \times 0,9 \times 0,1 \times 0,9 \times 0,1 \times 0,9 \times 0,5$$

0,1 et 0,5: probabilités d'émission 0,9: probabilités de transition

4 6 4 3 2 5 6 1 4 6 5 6 3 1 2 4 2 1 5 6 2

dé normal

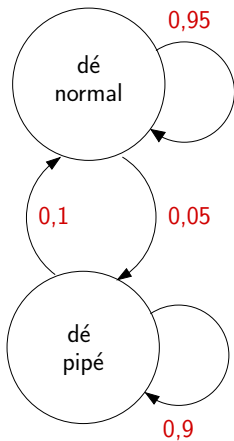
dé pipé

- quelle est la probabilité de cette observation ?

$$\begin{aligned} & \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \\ & \times 0,05 \times \\ & 0,1 \times 0,9 \times 0,5 \times 0,9 \times 0,1 \times 0,9 \times 0,1 \times 0,9 \times 0,5 \times 0,9 \times 0,1 \times 0,9 \times 0,5 \times 0,9 \times 0,1 \\ & \times 0,1 \times \\ & \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \times 0,95 \times \frac{1}{6} \end{aligned}$$

4 6 4 3 2 3 5 6 1 4 6 5 2 6 3 1 1 2 4 2 1
5 6 2 4 3 5 1 1 3 6 2 4 2 6 1 3 4 5 1

- changement de dé possible, un nombre arbitraire de fois
- quelle est l'alternance de dés qui rend cette observation la plus probable ?



un	deux	trois	quatre	cinq	six
$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

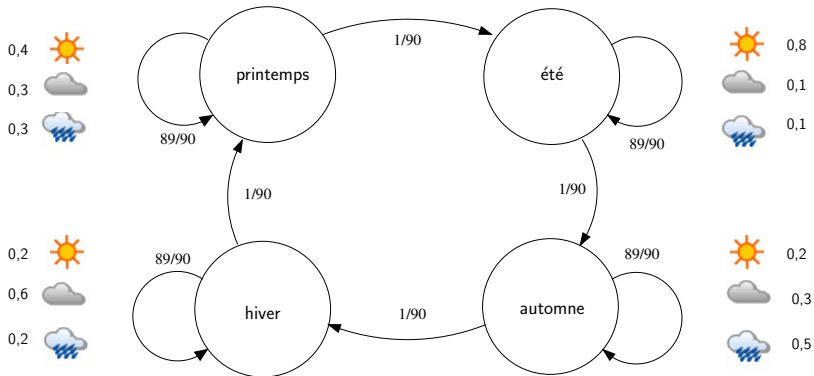
un	deux	trois	quatre	cinq	six
0,1	0,1	0,1	0,1	0,1	0,5

chemin de probabilité maximale dans le graphe/automate

La météo des saisons

	printemps	été	automne	hiver
soleil	0.4	0.8	0.2	0.2
nuages	0.3	0.1	0.3	0.6
pluie	0.3	0.1	0.5	0.2





Processus markovien, chaîne de Markov



Andreï Markov
(1856-1922)

- modélisation de processus stochastiques
- applications en finance, en économie, en analyse du signal, pour la modélisation de processus industriels. . .

Modèle de Markov (homogène)

- un ensemble fini d'états: π_1, \dots, π_m
- des probabilités de transition (matrice)
 $a_{k\ell}$: probabilité d'accéder à l'état ℓ alors que l'on est dans l'état k (en un pas)
- propriétés
 - **sans mémoire**: la probabilité d'être dans un état à l'instant i ne dépend que de l'état à l'instant $i - 1$
 - à partir d'un état initial, on peut **prédire** les états au bout de n itérations
 - à chaque suite d'états, on peut associer une **probabilité** dans le modèle

Modèle de Markov caché – HMM

- un ensemble fini d'états, π_1, \dots, π_m avec probabilités de transition $a_{k\ell}$
- un ensemble fini d'observations: $x_1, \dots, x_j \leftarrow$ nouveau
- des probabilités d'émissions \leftarrow nouveau
 $e_k(b)$: probabilité d'observer b alors que l'on est dans l'état k
- perte d'information entre le modèle et l'observation

Application à l'analyse de séquences biologiques

- chaque résidu (nucléotide ou acide aminé) est une observation
- les états correspondent à des fonctions ou des éléments structuraux des séquences
- applications à l'analyse de protéines, la prédiction de gènes, la modélisation de motifs biologiques. . .

Exemple 1: prédiction des hélices α dans les séquences protéiques



- les hélices α présentent un biais de composition en acides aminés qui les différencie des autres région de la protéine
- p_i : fréquence d'apparition de l'acide aminé i dans une hélice α
- q_i : fréquence d'apparition de l'acide aminé i dans un peptide hors hélice α
- m : longueur moyenne d'une hélice α (10 résidus en général)
- f : fréquence des hélices α

Observations : une séquence d'acides aminés

Question : où sont les hélices α ?

Modélisation sous forme de HMM

- deux états

hélice alpha

autre

- probabilités de transition

hélice alpha \rightarrow autre : $1/m$

hélice alpha \rightarrow hélice alpha: $(m-1)/m$

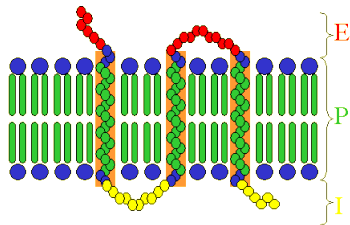
autre \rightarrow hélice alpha: f

autre \rightarrow autre: $1-f$

- probabilités d'émission de l'état hélice alpha: p_i
- probabilités d'émission de l'état autre: q_i

Exemple 2: Protéines transmembranaires

- protéines fichées dans la membrane d'une cellule qui permettent à la cellule de recevoir des informations extérieures
- le domaine transmembranaire est structuré en hélice α , avec un fort biais de composition en acides aminés hydrophobes
- la protéine contient souvent une succession de domaines transmembranaires



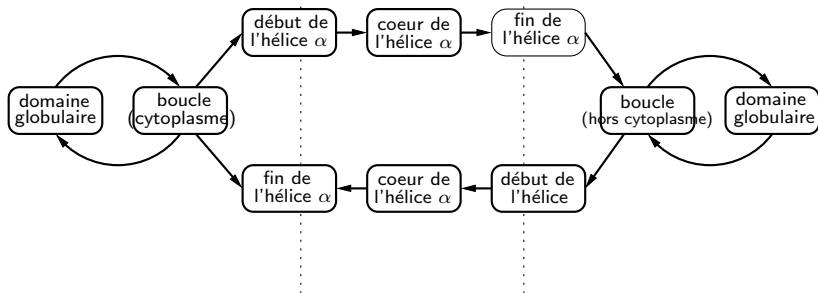
Observations : une séquence d'acides aminés

Question : où sont les domaines transmembranaires ?

cytoplasme


membrane

*extérieur de la
cellule*



Exemple 3: Les îlots CpG

- méthylation : dans le dinucléotide ...CG..., le C mute souvent en T
- la distribution des nucléotides n'est donc pas indépendante, et la fréquence d'apparition d'un nucléotide dépend du nucléotide précédent
- modèle de Markov simple, avec quatre états: A, C, G et T
- matrice de transition

	A	C	G	T
A	0.30	0.21	0.28	0.21
C	0.32	0.30	0.08	0.30
G	0.25	0.25	0.30	0.20
T	0.17	0.23	0.30	0.30

- îlot CpG: dans les zones du génome précédant un gène, le phénomène de méthylation disparaît, et la proportion en CpG est donc plus importante.
- modèle de Markov simple
- probabilités de transition dans un îlot CpG

	A	C	G	T
A	0.18	0.27	0.43	0.12
C	0.17	0.37	0.27	0.19
G	0.16	0.34	0.37	0.13
T	0.08	0.36	0.38	0.18

- deux sous-modèles : modèle + (îlots) / modèle - (hors îlots)
- articulation des deux modèles ?

8 états : A-, C-, G-, T-
A+, C+, G+, T+

4 observations : A, C, G, T

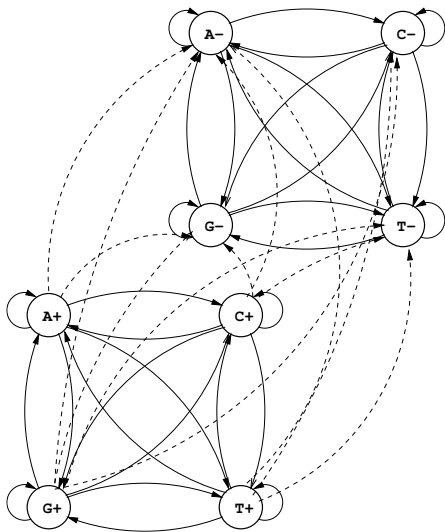
probabilités de transition entre les états
A-, C-, G-, T-: données par les
transitions du modèle de Markov -

probabilités de transition entre les états
A+, C+, G+, T+: données par les
transitions du modèle de Markov +

probabilités de transition des états +
vers les états -: dépendent de la taille
d'un îlot Cpg

probabilités de transition des états -
vers les états +: dépendent de la taille
des régions entre les îlots Cpg

probabilités d'émission ?



- étant donnée une protéine, où sont les domaines transmembranaires ?
- étant donnée une séquence ADN, où sont les îlots CpG ?
- étant donnée une suite d'observations, quelle est la suite d'états qui maximise ces observations ?

décodage: algorithme de Viterbi

- étant donnée une suite d'observations, quelle est sa probabilité ?

vraisemblance: algorithme Forward (ou Backward)

- à partir de données d'entraînement, comment déterminer les paramètres du modèle HMM ?

algorithme de Baum-Welch

Algorithme de Viterbi, 1967

- une suite d'observations $S = S_1, \dots, S_n$
- objectif: trouver la suite d'états qui maximise la probabilité de cette suite
- $v_\ell(i)$: probabilité du chemin le plus probable entre S_1 et S_i , terminant sur l'état ℓ

$$\begin{cases} v_\ell(1) &= 1 \\ v_\ell(i+1) &= e_\ell(S_{i+1}) \max_k \{v_k(i) a_{k\ell}\} \end{cases}$$

- le chemin cherché est reconstruit à partir de $\max_\ell \{v_\ell(n)\}$

- mise en œuvre par programmation dynamique
- table de $n \times m$ (longueur de la suite observée \times nombre d'états)
- complexité en temps : $O(n)$
- variante: calculer $V_\ell(i) = \log_2 v_\ell(i)$ (intérêt numérique)

$$V_\ell(i+1) = \log_2 e_\ell(x_{i+1}) + \max_k \{ V_k(i) + \log_2 a_{k\ell} \}$$

Algorithme Forward

- une suite d'observations $S = S_1, \dots, S_n$
- objectif : trouver la probabilité totale de S (pour toutes les suites d'états possibles)
- $f_\ell(i)$: probabilité de l'observation entre S_1 et S_i , le dernier état étant ℓ

$$f_\ell(i+1) = e_\ell(S_{i+1}) \sum_k f_k(i) a_{k\ell}$$

- implémentation : programmation dynamique
- complexité : $O(n)$
- version symétrique : algorithme Backward

$b_\ell(i)$: probabilité de l'observation entre S_{i+1} et S_n
en partant de l'état ℓ

Comment déterminer les paramètres d'un modèle de Markov ?

- ce dont on dispose
 - un ensemble d'états π_1, \dots, π_m
 - un ensemble d'observations x_1, \dots, x_j
 - un échantillon d'observations qui constitue un ensemble d'apprentissage
- ce que l'on cherche: déterminer les probabilités de transition $a_{k\ell}$ et les probabilités d'émission $e_\ell(b)$

Cas 1 : observations = états (Modèle de Markov simple)

Cas 2 : observations \neq états,
mais les états pour chaque observation sont connus

Cas général : observations \neq états
→ algorithme de Baum-Welch

Cas 1: Modèle de Markov simple

- dénombrement des transitions

$A_{k\ell}$: nombre de transitions de l'état k vers l'état ℓ

- problème de *sur-adaptation* : introduction de *pseudo-poids* $r_{k\ell}$

$$A_{k\ell} \leftarrow A_{k\ell} + r_{k\ell}$$

- normalisation pour avoir une probabilité

$$a_{k\ell} = \frac{A_{k\ell}}{\sum_{\ell'} A_{k\ell'}}$$

- pas de probabilités d'émission

Cas 2: observations \neq états, mais les états sont connus

- le calcul des probabilités de transition ne change pas

$$a_{k\ell} = \frac{A_{k\ell}}{\sum A_{k\ell'}}$$

- calcul des probabilités d'émission
 - dénombrement des observations

$E_k(b)$: nombre de fois où l'état k donne l'observation b

- correction optionnelle avec des pseudo-poids

$$E_k(b) \leftarrow E_k(b) + r_k(b)$$

- normalisation, pour avoir une probabilité

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

Algorithme de Baum-Welch, 1972

- cas général : observations \neq états
- approximations itératives des paramètres
- fonction d'objectif: vraisemblance
 - probabilité globale de l'échantillon
 - calculée avec l'algorithme Forward

- probabilité que $a_{k\ell}$ soit utilisé à la position i dans la séquence $S = S_1 \dots, S_n$

$$\frac{f_k(i) a_{k\ell} e_\ell(S_{i+1}) b_\ell(i+1)}{P(S)}$$

$f_k(i)$: algorithme Forward, probabilité de la suite d'observations S_1, \dots, S_i , le dernier état étant k

$b_\ell(i+1)$: algorithme Backward, probabilité de la suite d'observations S_{i+1}, \dots, S_n , le premier état étant ℓ

- espérance de $a_{k\ell}$

$$A_{k\ell} = \sum_S \sum_i \frac{f_k(i) a_{k\ell} e_\ell(S_{i+1}) b_\ell(i+1)}{P(S)} \quad (1)$$

on somme sur toutes les positions de toutes les séquences de l'échantillon

- pour les paramètres d'émission

$$E_k(b) = \sum_S \sum_{S_i=b} \frac{f_k(i) b_k(i)}{P(S)} \quad (2)$$

1. initialisation: choix de valeurs initiales arbitraires pour les paramètres a et e
2. itération:
 - calcul des valeurs f_k et b_k pour toutes les séquences de l'échantillon
 - détermination des valeurs pour E et A avec les équations (1) et (2)
 - nouvelles valeurs pour e et $a \rightarrow$ cas 2
3. critère d'arrêt: recommencer 2. jusqu'à avoir convergence de la probabilité de l'échantillon

convergence vers un maximum **local** de la probabilité de l'échantillon.