Exercices sur les modèles de Markov et HMM

Exercice 1: Le génome de *Simplicimum bestiolus*

On considère un organisme dont le génome suit les règles suivantes:

1. un `A` est suivi soit d'un `C` avec une probabilité de 0.1, soit d'un `G`, soit d'un `T` avec la même probabilité,
2. un `T` a deux fois plus de chance d'être suivi par un `A` que par tout autre nucléotide,
3. les `C` et les `G` sont suivis de n'importe quel nucléotide, avec la même probabilité.

Donnez un modèle de Markov qui décrive ce génome.

Exercice 2: Le génome de *Bizarrum organismus*

On considère des séquences nucléotidiques d'un organisme étrange *Bizarrum organismus*:

1. Le nucléotide actuel est un `A`, un `C`, un `G` ou un `T` avec une probabilité de 25% si les deux nucléotides précédents sont identiques.
2. Le nucléotide actuel a 40% de chance d'être un `C`, 40% de chance d'être un `G`, 10% de chances d'être un `A` et 10% de chance d'être un `T`, si les deux nucléotides précédents sont différents.

Montrez que l'on peut modéliser les séquences de *Bizarrum organismus* par un modèle de Markov, dont vous donnerez les états et les probabilités de transition.

Exercice 3: TATAAT

Dans le génome, les sites d'initiation de la transcription de l'ADN en ARN sont précédés d'un site *promoteur* : c'est une région de composition particulière sur laquelle peut se fixer l'ARN polymérase. Dans les organismes procaryotes, ce promoteur est représenté par le motif approché `TATTAT`.

Le tableau ci-dessous donne les fréquences à chaque position pour `TATAAT` relevées dans le génome de la bactérie *E. coli*:

	T	A	T	A	A	T
A	0.04	0.88	0.26	0.58	0.48	0.03
C	0.09	0.03	0.11	0.13	0.22	0.06
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89

Question 1: Donnez un modèle de Markov caché (sans pseudo-poids) pour représenter ce motif.

On considère maintenant que les séquences à analyser peuvent contenir des erreurs de séquençage, sous forme d'insertion d'un nucléotide ou de délétion d'un nucléotide. La probabilité de chaque type d'erreur à une position donnée est 0,02. Qu'il s'agisse d'une insertion ou d'une délétion, il n'y a à chaque fois qu'un seul nucléotide ajouté ou supprimé, et dans le cas d'une insertion, il peut s'agir d'un `A`, d'un `C`, d'un `G` ou d'un `T` avec la même probabilité.

Question 2: Modifiez le modèle de Markov caché de la question précédente pour tenir compte des erreurs de séquençage.

Exercice 4: Les magiciens

Deux magiciens A et B donnent un spectacle devant un public de mathématiciens. Chacun a devant lui un chapeau qui contient pour A trois mouchoirs bleus et un mouchoir vert, et pour B un mouchoir bleu et deux mouchoirs verts.

La règle est la suivante:

- pour savoir qui commence, les deux magiciens lancent le dé: si le nombre est pair, A commence et prend le dé, sinon B commence et prend le dé;
- à chaque étape:
  - le magicien qui a le dé tire un mouchoir au hasard dans son chapeau, et le passe à un assistant qui le montre au public, puis remet le mouchoir dans le chapeau;
  - le magicien lance le dé: si le nombre est compris entre 1 et 4, il garde le dé. Sinon il le passe à l'autre magicien.

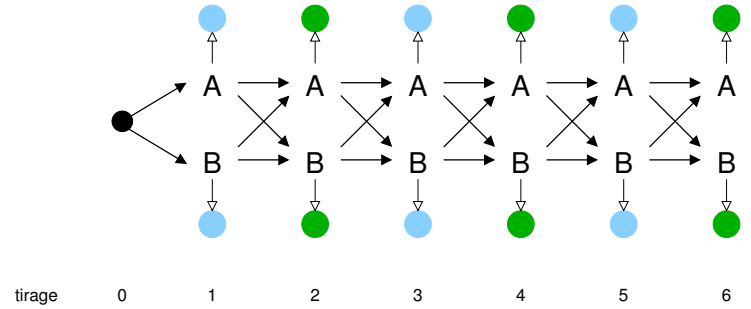
  Bien sûr, le public ne voit ni le dé, ni les chapeaux, ni les magiciens. Il ne voit que le mouchoir qui a été choisi à chaque étape.

Question 1: Modélisez le problème sous forme de modèle de Markov à états cachés. Quelles sont les probabilités de transition, d'émission ?

On considère la suite de couleur suivante

```
Bleu Vert Bleu Vert Bleu Vert
```

On peut représenter tous les suites d'états possibles associés à ce tirage sous forme de graphe.



Question 2: Ajoutez les probabilités de transition et d'émission sur ce graphe.

Question 3: Déduisez-en le chemin de plus grande probabilité. Pour cela, vous pouvez calculer sur chaque nœud A ou B la probabilité depuis le début des tirages.

Question 4: Ecrivez des formules de récurrence pour $P(i, A)$, la probabilité du meilleur chemin jusqu'au $i$ème tirage inclus se terminant sur l'état A et pour $P(i, B)$, la probabilité du meilleur chemin jusqu'au $i$ème tirage inclus se terminant sur l'état B. A quoi vous font penser ces formules ?

Question 5: Toujours à partir du graphe, calculez la probabilité cumulée de tous les chemins compatibles avec l'observation `Bleu Vert Bleu Vert Bleu Vert`. Ecrivez ensuite les formules de récurrence. De nouveau, à quoi cela vous fait-il penser ?

Exercice 5: Les deux dés

Un joueur a deux dés à six faces, identiques et non truqués. À chaque tour, il lance les deux dés (que vous ne voyez pas), et annonce comme résultat la somme des deux faces. Entre deux tours, il peut également décider de mettre un dé de côté, et de continuer à jouer avec un seul dé. Dans ce cas, le nombre annoncé est simplement la valeur de la face du dé. La probabilité l'événement *passer de deux dés à un dé* est égale à 1/10. Quand le joueur joue avec un seul dé, la probabilité de repasser à deux dés est de 1/5. Ce jeu peut être modélisé par un modèle de Markov à états cachés qui comprend deux états:

- α: le joueur joue avec un seul dé
- β: le joueur joue avec deux dés

Les probabilités d'émission de cet HMM sont données par le tableau ci-dessous. La première ligne indique les douze observations possibles (nombres de 1 à 12), la deuxième ligne les probabilités d'émission pour l'état α et la troisième ligne les probabilités d'émission pour l'état β.

	`un`	`deux`	`trois`	`quatre`	`cinq`	`six`	`sept`	`huit`	`neuf`	`dix`	`onze`	`douze`
α	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	0	0	0	0	0
β	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Question 1. Comment ont été calculées les probabilités d'émission pour l'état α ? Pour l'état β ? Pour l'état β, vous pouvez étayer votre explication en détaillant le calcul de la probabilité d'émission pour `huit`.

Question 2. Dessinez le modèle de Markov caché: états, probabilités de transition, probabilités d'émission

Question 3. Pour la suite d'observations `quatre onze six cinq`, quelle est la suite d'états la plus probable parmi les trois suites $s_1 = \alpha\alpha\alpha\alpha$, $s_2 = \beta\beta\beta\beta$ et $s_3 = \alpha\beta\beta\beta$?

Question 4. On change la règle du jeu: au lieu d'annoncer le résultat de la somme obtenue, le joueur indique simplement si le résultat est compris entre 1 et 4, ou compris entre 5 et 12. Que doit-on modifier dans le modèle de Markov caché ? Donnez ce nouveau modèle.

Question 5. La règle du jeu change de nouveau. Cette fois, le joueur indique si le résultat obtenu est pair ou impair. Quelles sont les probabilités d'émission associées à cette règle ? Qu'en pensez-vous ?

Exercice 6: Le Génie

Un génie dispose de trois urnes, remplies respectivement de

- urne 1: 2 boules oranges et 3 boules vertes
- urne 2: 2 boules oranges et 1 boule verte
- urne 3: 3 boules vertes

A chaque étape, le génie tire une boule au hasard dans une des trois urnes, et montre la boule, sans dire de quelle urne elle provient. Pour le choix de l'urne, il a préférence pour l'urne 1, qu'il choisit systématiquement au départ puis 1 fois sur 2 quand il est sur l'urne 2 ou 3, et 3 fois sur 4 quand il est déjà sur l'urne 1. Les deux autres urnes, 2 et 3, sont choisies de manière équiprobables quand on est sur l'urne 1, 2 ou 3. L'objectif est de retrouver la suite des urnes pour une succession de boules donnée.

Question 1. Montrez que le problème peut être modélisé par un modèles de Markov à états cachés (HMM), dont vous préciserez les états, les probabilités d'émission et les probabilités de transition.

Question 2. Pour la suite d'observations `orange vert vert orange`, quelle est la suite d'états la plus probable parmi les quatre suites $s_1 = 1123$, $s_2 = 2113$, $s_3 = 1132$ et $s_4 = 1322$?

Question 3. Le génie change de règle du jeu: à chaque tirage, il choisit deux urnes, et deux boules (une dans chaque urne). Le choix des deux urnes est donné par les règles suivantes:

- on reste sur le même couple d'urnes avec une probabilité de 1/2;
- quand on change de couple d'urnes, on garde l'urne de plus petit numéro avec une probabilité de 3/4.

Montrez que ce nouveau problème peut également être modélisé par un modèle de Markov à états cachés. Quels sont les états, les probabilités de transition et les probabilités d'émission ?