Nom :

Id :

Année 2022-2022 **Unsupervised Learning - M2 Data Science**

Examen du cours « Unsupervised Learning »

Exercice 1 : ADN

Soit une séquence d'ADN dont le premier nucléotide est a : pour tout $n \geq 1$, on note X_n la n-ième base composant la séquence d'ADN en partant d'une de ses extrémités. On suppose que (X_n) est une chaîne de Markov homogène d'espace d'états $\mathcal{A} = \{a, c, g, t\}$ et d'état initial a (on identifiera a à l' état 1, c à l'état 2, g à l'état 3 et t à l'état 4). On note Q sa matrice de transition.

1. Exprimer à l'aide de la matrice Q, la probabilité que la séquence commence par le motif $aacg$.
2. Montrer que la 4-ième base est indépendante de la seconde sachant la 3-ième base.
3. Expliquer comment simuler les n premières bases d'une séquence pour ce modèle.

Exercice 2 : Loi stationnaire d'une chaîne à deux états

Soit $(X_n)_n$ une chaîne de Markov homogène d'espace d'états $\{1, 2\}$ et de matrice de transition $Q = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$ avec $p, q \in [0, 1]$. Si $p + q > 0$, quelle est la loi de probabilité stationnaire de la chaîne ?

Exercice 3 : Chaîne de Markov cachée

Soit $\mathcal{MC}(\boldsymbol{\pi}, A)$ une chaîne de Markov cachée à deux états cachés $\{1, 2\}$ de matrice de transition $A = \begin{pmatrix} 0.5 & 0.5 \\ 0.7 & 0.3 \end{pmatrix}$ de probabilité d'état initial $\boldsymbol{\pi} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$ et de probabilité d'émission $B = \begin{pmatrix} 0.2 & 0.9 \\ 0.8 & 0.1 \end{pmatrix}$ sur les symbôles $\{\circ, \times\}$ avec $B_{1k} = P(\circ|Z = k)$ et $B_{2k} = P(\times|Z = k)$.

Attention : la convention pour B utilisée ici numérote les symbôles en lignes et pas en colonnes comme vu en TD.

1. Écrire le pseudo-code de simulation d'une chaîne de Markov cachée de longueur n.

2. Écrire la log-vraisemblance complète (c'est-à-dire la loi jointe des observations et des états cachés) des paramètres d'une chaîne de Markov cachée.
3. Écrire l'espérance de cette log-vraisemblance par rapport à la loi des états cachés sachant les observations.
4. Quel algorithme est classiquement utilisé pour estimer les paramètres d'une chaîne de Markov cachée ? Décrire succintement les étapes principales de cet algorithme.

Exercice 4 : Modèle à blocs stochastiques

Soit un modèle modèle à blocs stochastiques à deux classes :

$$X_{ij}|Z_i = k, Z_j = \ell \sim \mathcal{B}(\gamma_{k\ell})$$

avec $\pi_1 = P(Z_i = 1) = \frac{1}{2} = P(Z_i = 2) = \pi_2$ et $\gamma = \begin{pmatrix} 1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$.

1. Écrire la formule théorique de la vraisemblance complète de ce type de modèle.
2. Pourquoi un algorithme EM classique n'est pas applicable pour estimer les paramètres de ce modèle ?
3. Quelle solution peut être adoptée pour estimer les paramères ?