Examen Final - MAPSI

Durée : 2 heures

Seuls documents autorisés : Calculatrice, antisèche recto-verso, tables de lois.
– Barême indicatif –

Exercice 1 (3 pts) – Indépendence

Soit quatre variables aléatoires X, Y, Z, W, de modalités respectives $\{x_1, x_2\}$, $\{y_1, y_2\}$, $\{z_1, z_2\}$ et $\{w_1, w_2\}$. La probabilité jointe de ces quatre variables est fournie par le tableau suivant :

	y_1				y_2			
	z_1		z_2		z_1		z_1	
	w_1	w_2	w_1	w_2	w_1	w_2	w_1	w_2
x_1	0,048	0,032	0,048	0,032	0,012	0,008	0,012	0,008
x_2	0,192	0,128	0,192	0,128	0,048	0,032	0,048	0,032

Q 1.1 (1 pt) Indépendence conditionnelle Calculer le tableau $P(X, W|Y, Z)$ et justifier le fait que X et W sont indépendantes du couple (Y, Z).

Le tableau $P(X, W|Y, Z)$ est le suivant :

	y_1				y_2			
	z_1		z_2		z_1		z_1	
	w_1	w_2	w_1	w_2	w_1	w_2	w_1	w_2
x_1	0,12	0,08	0,12	0,08	0,12	0,08	0,12	0,08
x_2	0,48	0,32	0,48	0,32	0,48	0,32	0,48	0,32

Depuis le tableau, il est évident que $P(X, W|Y, Z) = P(X, W)$. [De plus, car la proportion entre les $p(X, W)$ est constante à changer de W ou à changer de X, X et W sont indépendantes aussi.]

Q 1.2 (1 pt) Indépendence. Déterminer si Z est indépendante du couple (X, Y).

On calcule le tableau $P(X, Y, Z)$ en marginalisant W :

	y_1		y_2	
	z_1	z_2	z_1	z_2
x_1	0,08	0,08	0,02	0,02
x_2	0,32	0,32	0,08	0,08

Dans ce cas aussi la proportion de la valeur de probabilité est constante à changer de X ou à changer de Y (il y a indépendance soit vers X, soit vers Y). Différemment, on peut constater que pour tout $x \in \{x_1, x_2\}$ et $y \in \{y_1, y_2\}$, on a $P(x, y, z_1) = P(x, y, z_2)$.

Q 1.3 (1 pt) Indépendence et conjonction Soit trois variables aléatoires X, Y, Z. Montrer que si X est indépendante du couple (Y, Z), et Y est indépendante de Z, alors Z est indépendante du couple (X, Y).

Si X est indépendant de (Y, Z) :

$$P(X, Y, Z) = P(X)P(Y, Z)$$

puisque Y est indépendant de Z :

$$P(Y, Z) = P(Y)P(Z)$$

on a donc

$$P(X, Y, Z) = P(X)P(Y, Z) = P(X)P(Y)P(Z) = P(X, Y|Z)P(Z)$$

Alors :

$$P(X, Y|Z) = P(X)P(Y)$$

soit la probabilité conditionnelle en Z du couple (X, Y) ne dépend pas de Z, i.e. ils sont indépendants.

Exercice 2 (4 pts) – À l'attaque !

L'autorité maritime d'un certain pays souhaite évaluer un logiciel qui analyse les données de navigation de vaisseaux (satellitaires, enregistrées dans les ports, etc.) pour détecter des attaques de piraterie. D'après son constructeur, une attaque est détectée dans 90% des cas. Malheureusement, il y a aussi 20% de chances que le logiciel identifie une attaque lorsqu'il n'y en a pas. Pour l'évaluation, l'autorité se concentre sur un trajet en particulier qui, dans la dernière année, a enregistré 200 attaques des pirates sur 4000 passages de vaisseaux. On note L la variable aléatoire pour la prédiction du logiciel et A pour l'attaque.

Q 2.1 (1 pt) Probabilité à posteriori Le logiciel signale une attaque en ce moment. En utilisant le nombre d'attaques observé dans la dernière année pour estimer la probabilité a priori, calculer la probabilité qu'il y ait effectivement une attaque. Écrire la formule correspondant à cette probabilité, puis calculer sa valeur.

Les caractéristiques diagnostiques du logiciel sont résumés par ce tableau :

	A	$\neg A$
$+_L$	0.9	0.2
$-_L$	0.1	0.8

La probabilité a posteriori est donnée par la formule de Bayes :

$$P(A|+_L) = \frac{P(+_L|A)p(A)}{P(+_L|A)P(A) + P(+_L|\neg A)P(\neg A)}$$

Si on considère la statistique donnée pour la probabilité a priori on a : $P(A) = \frac{200}{4000} = 0.05$, $P(\neg A) = 1 - P(A) = 0.95$. Et depuis le tableau on a : $P(+_L|A) = 0.9$, $P(+_L|\neg A) = 0.2$. Après le calcul, $P(A|+_L) = 0.1915$.

Q 2.2 (1.5 pt) Information complémentaire L'autorité a à disposition des outils supplémentaires : depuis quelques années elle a mis en place un réseau d'observateurs permanents (choisi parmi des pécheurs et d'autres navires civils), dotés d'un appareil spécial pour notifier en temps réels l'occurrence de mouvements suspects. Selon une statistique interne, ce réseau a permis de reconnaitre 40% des attaques en avance. Malheureusement, cette méthode donne aussi 30% de faux positifs (mouvements suspects sans attaque). En supposant les deux notifications (logiciel et réseaux d'observateurs) indépendantes, calculer la probabilité qu'il y ait une attaque lorsque les deux notifications sont actives (utiliser R pour la variable aléatoire de prédiction venant du réseau).

Les caractéristiques diagnostiques du réseau d'observateurs sont résumés par ce tableau :

	A	$\neg A$
$+_R$	0.4	0.3
$-_R$	0.6	0.7

La probabilité a posteriori est donnée par la formule de Bayes, en utilisant les deux données :

$$P(A|+_L,+_R) = \frac{P(+_L,+_R|A)P(A)}{P(+_L,+_R|A)P(A) + P(+_L,+_R|\neg A)P(\neg A)}$$

Par indépendance on a :

$$P(+_L,+_R|A) = P(+_L|A)P(+_R|A) = 0.36$$
$$P(+_L,+_R|\neg A) = P(+_L|\neg A)P(+_R|\neg A) = 0.06$$

Après le calcul, $P(A|+_L,+_R) = 0.24$.

Q 2.3 (1.5 pt) Recalibration ? Après une recherche qualitative, nous nous sommes aperçus que la statistique des attaques utilisée pour estimer la valeur des probabilités a priori était non optimale. En effet, les pirates n'attaquent pas tous les navires, mais principalement ceux qui ont un certain tonnage. Supposons que les navires se distribuent en 3 classes (I, II, III). D'après les statistiques historiques, la probabilité d'attaque en fonction de la classe est la suivante : $P(\text{I}) = 0.1$, $P(\text{II}) = 0.5$ $P(\text{III}) = 0.4$. D'un autre coté, les données détaillées de l'année passée donne :

Classe de navire	I	II	III
Nombre d'attaques	40	120	40

Faire un test d'ajustement avec un niveau de confiance de 90% pour déterminer si les observations correspondent toujours à la distribution historique sur les classes de navires.

La population est repartie en 3 classes. Chaque individu est supposé d'avoir une probabilité p_r d'appartenir à une classe $r \in \{\text{I}, \text{II}, \text{III}\}$.
Notre hypothèse est donc que les probabilités données décrivent l'ensemble population. Si le test d'ajustement est satisfait, l'hypothèse est correcte.
Si on appelle N_r le nombre d'individus d'un classe r, pour la théorème de la limite centrale, N_r suit une loi normal centré en $n \cdot p_r$.
On sait que la somme des écarts carrés parmi ce qui est observé et le modèle suit la loi du Chi :

$$D^2_{(n)} = \sum_{r \in \{\text{I,II,III}\}} \frac{(N_r - n \cdot p_r)^2}{n \cdot p_r} \sim \chi^2_{(2)}$$

Si on calcule la valeur effective des écarts carrés d^2 on a :

$$d^2 = \frac{(40 - 0.1 \cdot 200)^2}{0.1 \cdot 200} + \frac{(120 - 0.5 \cdot 200)^2}{0.5 \cdot 200} + \frac{(40 - 0.4 \cdot 200)^2}{0.4 \cdot 200} = 44$$

Puisque on veut $1 - \alpha = 0.9$ on considère $\alpha = 0.1$.
En lisant sur la table de la loi du χ^2, on voit que $P(\chi^2_{(2)} > d^2_\alpha) = 0.1$ pour $d^2_\alpha = 4,61 < 44$
Le test n'est pas satisfait, et donc l'hypothèse n'est pas supporté.

Exercice 3 (4 pts) – Rejection sampling

Supposons qu'un phénomène réel peut être modélisé par une variable aléatoire $X \in [0, 1]$ qui suit une loi normale tronquée proportionnelle à $\mathcal{N}(\frac{3}{4}, 1)$. La fonction de densité d'une loi normale tronquée proportionnelle à $\mathcal{N}(\mu, \sigma^2)$ définie sur $[a, b]$ peut être écrite comme :

$$f(x) = \begin{cases} C \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} & \text{pour } \ a \leq x \leq b \\ 0 \quad \text{pour } \ x < a \ \text{ et } \ x > b \end{cases}$$

où C est un facteur de normalisation.

Q 3.1 (1.5 pt) Déterminer la fonction de densité $f(x)$ de X en calculant C.
Note : utiliser la table de la loi normale.

Puisque la loi est tronquée, on doit considérer un facteur de normalisation C :

$$f(x) = C \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\frac{3}{4})^2}$$

Sur le table de la loi normale centrée réduite on lit $P(Z > 0,25) \approx 0,4013$ et $P(Z > 0,75) = P(Z < -0.75) \approx 0,2266$. La surface tronquée est donc de 0.6279, celle qui reste est $1 - 0.6279 = 0.3721$. Pour normaliser, le produit avec C de cette surface doit être 1, et donc :

$$C \approx \frac{1}{0.3721} = 2.6874$$

Q 3.2 (1.5 pt) Imaginons que la fonction $f(x)$ est difficile à échantillonner. Nous allons donc utiliser la méthode Monte Carlo appelée *rejection sampling*.
Rappel de cours : Choisir une distribution $q(\cdot)$, facile à échantillonner, telle qu'il existe un facteur k satisfaisant $\forall x, k \cdot q(x) \geq f(x)$. L'algorithme d'échantillonnage est constitué de quatre étapes : (1) tirer un nombre z selon $q(\cdot)$ (*pre-échantillonnage*) ; (2) calculer $m_q = k \cdot q(z)$; (3) tirer un nombre u selon la distribution uniforme sur $[0, m_q]$; (4) accepter z comme échantillon si $u \leq f(x)$.

Calculez le *taux d'acceptation* (la proportion de pre-échantillons acceptés) lorsque $q(\cdot)$ est une loi uniforme sur $[0, 1]$, et $k \cdot q(\cdot) = \max_{x\in[0,1]} f(x)$.

Le taux d'acceptation est donnée par :

$$P(acceptation) = \int_0^1 q(x) \cdot \frac{f(x)}{k \cdot q(x)} = \int_0^1 q(x) \cdot \frac{f(x)}{k \cdot q(x)} = \frac{1}{k}$$

Car $q(x) = 1$ et $k \cdot q(x) = \max_{x\in[0,1]} f(x)$, on a :

$$k = \max_{x\in[0,1]} f(x) = \frac{C}{\sqrt{2\pi}} \approx \frac{2.6874}{2.50663} = 1.0721$$

Donc $P(acceptation) \approx 0.93$

Q 3.3 (1 pt) Supposons que le calcul de $f(x)$ est relativement coûteux. Une méthode pour augmenter l'efficacité (appelée *compression*) est de proposer une fonction $r(x)$ simple—e.g. une droite—qui est une limite inférieure de $f(x)$. L'algorithme modifié prend en compte un *pré-filtrage* des éléments qui sont certainement acceptés, quand $u \leq r(z)$. Calculez le taux des préfiltrage si on prend $r(x) = \min_{x\in[0,1]} f(x)$.

$$P(prefiltrage) = \int_0^1 q(x) \cdot \frac{r(x)}{k \cdot q(x)} = \int_0^1 \frac{r(x)}{k}$$

En dessinant $f(x)$ il est évident que le min est pour $x = 0$.

$$r(x) = \min_{x\in[0,1]} f(x) = f(0) = C \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(0-\frac{3}{4})^2} = k \cdot e^{-\frac{9}{32}}$$

Donc $P(prefiltrage) \approx e^{-\frac{9}{32}} = 0.75$

Exercice 4 (4 pts) – Modélisation markovienne

Nous souhaitons construire un générateur automatique de morceaux de piano (séquence de notes), en s'inspirant du travail d'un compositeur contemporain n'utilisant que 20 notes. Nous disposons donc de plusieurs morceaux $\mathbf{x} = \{n_1, \ldots, n_T\}, n_t \in \{1, \ldots, 20\}$.

Q 4.1 (0.5 pt) Proposer un système simple capable de générer des notes en s'inspirant des 3 morceaux dont nous disposons $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$.
Donner les dimensions des matrices de paramètres et détailler les étapes successives à suivre.

1. Initialiser un modèle de markov simple à 20 états : $\Pi \in \mathbb{R}^{20}, A \in \mathbb{R}^{20\times 20}$
2. Optimiser les paramètres au sens du max de vraisemblance par rapport à X :
   (a) Compter les premières notes pour définir Π. (normaliser le vecteur de comptage pour sommer à 1)
   (b) Compter toutes les transitions de notes puis normaliser chaque ligne de A pour sommer à 1.
3. Tirer un morceau selon Π, A :
   (a) Tirer une note initiale selon Π
   (b) Tirer les notes suivantes selon la bonne ligne de A
   (c) S'arrêter autour de la longueur moyenne des morceaux

Q 4.2 (1 pt) Le compositeur n'est pas du tout satisfait du résultat. Il vous explique que chacun de ses morceaux est structuré de la manière suivante : introduction (style A), couplet (style B), couplet (style C), conclusion (style A de nouveau). Il y a des variations dans chaque partie, mais globalement, les portions A, B, C sont plutôt homogènes. Toutes les parties ont des longueurs assez proche.
Proposer un système tenant compte de ces contraintes ; pour simplifier l'apprentissage, nous faisons l'hypothèse que les parties sont *exactement* de la même longueur. (Détailler les étapes et citer les algorithmes utilisés et donner une idée de la forme des matrices quand c'est possible).

Il faut un HMM. Il y a 20 observation et 3 états

$$\Pi = [1, 0, 0]$$

$$A = \begin{bmatrix} \alpha & 1-\alpha & 0 \\ 0 & \beta & 1-\beta \\ 1-\gamma & 0 & \gamma \end{bmatrix}$$

$$B \in \mathbb{R}^{4\times 20}$$

$\alpha = \beta = \gamma = \frac{n-1}{n}$ avec n la longueur moyenne des parties
Le plus simple : simplement donner les dimensions des matrices puis :

1. Initialiser des états (entre 1 et 3) selon la technique gauche droite dans X
2. Une fois les couples états/observations connus ; calculer les matrices A, Π, B par comptage + normalisation
3. Utiliser ce modèle en mode génératif : tirer un état initial, tirer une observation, tirer un état suivant...

Q 4.3 (0.5 pt) Quels ajustements sont nécessaires pour tenir compte des variations de longueurs possibles dans les différentes parties ?

Faire un Viterbi, une fois connus A, Π, B pour avoir une meilleure estimation des états (c'est à dire des durées de chaque partie). Puis ré-estimer les paramètres. Ré-itérer si besoin jusqu'à convergence ;
C'est l'algorithme de Baum Welch simplifié...

Q 4.4 (1 pt) Est-il facile d'insérer un refrain R entre A et B, entre B et C et entre C et A ? Pourquoi ? Donner une idée pour insérer ce refrain.

Non, c'est difficile, car il chaque refrain doit déboucher sur un couplet différent ! Il faut de la mémoire.
Une idée : faire des états différents pour chaque refrain (6 états), mais apprendre la même distribution d'observation pour les refrains (comptage fusionné des états refrain).

Q 4.5 (1 pt) Un statisticien pense que les différentes parties ne sont pas très importante ; pour lui, tout se joue sur la mémoire : il pense qu'il faudrait prendre en compte entre les 4 et 6 dernières notes pour prédire efficacement la suivante.
Comment construire et apprendre un tel système ? Cela vous semble-t-il raisonnable avec les 3 morceaux de la base d'apprentissage ?

On repart donc sur un système de markov simple (non caché).
Si on veut une mémoire M, il faut une matrice de transition de taille 20^{M+1}. $A = p(n_t|n_{t-1}, \ldots, n_{t-M})$. Pour bien modéliser le système, on pourrait se passer de Π mais introduire une 21ème note virtuelle correspondant à avant le morceau.
Malheureusement, $20^5 = 3.2$ millions de cases... IL y a peu de chance que nous ayons des partitions suffisamment longues pour estimer correctement A.

Exercice 5 (6.5 pts) – Recommandation par filtrage collaboratif

Nous voulons construire un système de recommandation de livres ; nous avons des utilisateurs $U = \{\ldots, \mathbf{u}_i, \ldots\}$ et des livres $X = \{\ldots, \mathbf{x}_j, \ldots\}$, nous proposons de modéliser l'intérêt de $\mathbf{u}_i$ pour $\mathbf{x}_j$ avec une variable de Bernoulli Y : $y_{ij} = 1$ si l'utilisateur aime le livre et 0 sinon. Nous utilisons la fonction logistique pour modéliser Y : $p(Y_{ij} = 1|\mathbf{u}_i, \mathbf{x}_j) = f(\mathbf{u}_i, \mathbf{x}_j) = \frac{1}{1+\exp(-\mathbf{u}_i \cdot \mathbf{x}_j)}$, où $\cdot$ désigne le produit scalaire. Dans l'approche que nous envisageons, chaque utilisateur $\mathbf{u}_i$ (respectivement livre $\mathbf{x}_j$) est un vecteur *abstrait* de $\mathbb{R}^d$. Le concepteur du système fixe d et initialise aléatoirement U et X. L'idée est de concevoir un algorithme d'optimisation qui modifient les $\mathbf{u}_i$ et $\mathbf{x}_j$ pour rendre la fonction d'évaluation de l'intérêt pertinente.

Q 5.1 (1 pt) En imaginant que les normes des $\mathbf{x}_j$ et $\mathbf{u}_i$ sont contraintes à une valeur α, dans quels cas particuliers obtenons-nous un interêt minimum/maximum ?

Si les normes de tous les vecteurs sont égales, les valeurs mini et maxi de f correspondent à $\mathbf{x}_j = \mathbf{u}_i$ et $\mathbf{x}_j = -\mathbf{u}_i$.
On a alors $f(\mathbf{u}_i, \mathbf{x}_j) = \frac{1}{1+\exp(-\mathbf{u}_i \cdot \mathbf{x}_j)} = \frac{1}{1+\exp(-\alpha^2)} \approx 1$ ou $f(\mathbf{u}_i, \mathbf{x}_j) = \frac{1}{1+\exp(\alpha^2)} \approx 0$(si α assez grand)

Q 5.2 (1.5 pt) Exprimer la probabilité de la variable de Bernoulli $P(Y = y_{ij}|\mathbf{u}_i, \mathbf{x}_j)$ en fonction de $y_{ij}, \mathbf{u}_i, \mathbf{x}_j$. Puis exprimer la vraisemblance conditionnelle $\mathcal{L}$ d'une base de donnée contenant N couples $\mathbf{u}_i, \mathbf{x}_j$ associés à des valeurs d'intérêt binaires $y_{ij} \in \{0, 1\}$ (comme pour la régression logistique). Vous ferez les hypothèses d'indépendance usuelles.
Montrer que la log-vraisemblance est de la forme $\sum_{couples(i,j)} \beta(\mathbf{u}_i \cdot \mathbf{x}_j) + \gamma \log(f(\mathbf{u}_i, \mathbf{x}_j))$

$$P(Y = y_{ij}|\mathbf{u}_i, \mathbf{x}_j) = p(Y_{ij} = 1|\mathbf{u}_i, \mathbf{x}_j)^{y_{ij}}(1 - p(Y_{ij} = 1|\mathbf{u}_i, \mathbf{x}_j))^{1-y_{ij}}$$
$$= (\frac{1}{1 + \exp(-\mathbf{u}_i \cdot \mathbf{x}_j)})^{y_{ij}}(1 - \frac{1}{1 + \exp(-\mathbf{u}_i \cdot \mathbf{x}_j)})^{1-y_{ij}}$$

$$\mathcal{L} = \prod_{couples(i,j)} P(Y = y_{ij}|\mathbf{u}_i, \mathbf{x}_j)$$

$$\log \mathcal{L} = \sum_{couples(i,j)} y_{ij} \log(\frac{1}{1 + \exp(-\mathbf{u}_i \cdot \mathbf{x}_j)}) + (1 - y_{ij}) \log(1 - \frac{1}{1 + \exp(-\mathbf{u}_i \cdot \mathbf{x}_j)})$$

$$= \sum_{couples(i,j)} y_{ij} \log(\frac{1}{1 + \exp(-\mathbf{u}_i \cdot \mathbf{x}_j)}) + (1 - y_{ij}) \underbrace{\log(1 - \frac{1}{1 + \exp(-\mathbf{u}_i \cdot \mathbf{x}_j)})}_{\log \frac{\exp(-\mathbf{u}_i \cdot \mathbf{x}_j)}{1+\exp(-\mathbf{u}_i \cdot \mathbf{x}_j)} = -\mathbf{u}_i \cdot \mathbf{x}_j + \log(\frac{1}{1+\exp(-\mathbf{u}_i \cdot \mathbf{x}_j)})}$$

$$= \sum_{couples(i,j)} (1 - y_{ij})(-\mathbf{u}_i \cdot \mathbf{x}_j) + \log(\frac{1}{1 + \exp(-\mathbf{u}_i \cdot \mathbf{x}_j)})$$

Q 5.3 (1.5 pt) Afin de simplifier la suite, nous nous intéressons à un seul couple $(\mathbf{u}_i, \mathbf{x}_j)$, ce qui simplifie l'expression de la vraisemblance $\log \mathcal{L}_{ij}$. Donner la dimension de $\nabla_{\mathbf{u}_i} \log \mathcal{L}_{ij}$ puis montrer que le gradient $\nabla_{\mathbf{u}_i} \log \mathcal{L}_{ij}$ est de la forme $\mathbf{x}_j\,(y_{ij} - f(\mathbf{u}_i, \mathbf{x}_j))$.

$$\log \mathcal{L}_{ij} = (1 - y_{ij})(-\mathbf{u}_i \cdot \mathbf{x}_j) + \log(\frac{1}{1 + \exp(-\mathbf{u}_i \cdot \mathbf{x}_j)}) = (1 - y_{ij})(-\mathbf{u}_i \cdot \mathbf{x}_j) - \log(1 + \exp(-\mathbf{u}_i \cdot \mathbf{x}_j))$$

$$\nabla_{\mathbf{u}_i} \log \mathcal{L} = \begin{bmatrix} \vdots \\ \frac{\partial \log \mathcal{L}}{\partial u_{ik}} \\ \vdots \end{bmatrix} \in \mathbb{R}^d$$

$$\frac{\partial \log \mathcal{L}}{\partial u_{ik}} = -(1 - y_{ij})x_{jk} - \frac{-x_{jk} \exp(-\mathbf{u}_i \cdot \mathbf{x}_j)}{1 + \exp(-\mathbf{u}_i \cdot \mathbf{x}_j)}$$

$$= x_{jk}\left(y_{ij} - 1 + \frac{exp}{1+exp}\right) = x_{jk}\left(y_{ij} - 1 + 1 - \frac{1}{1+exp}\right) = x_{jk}(y_{ij} - f(\mathbf{u}_i, \mathbf{x}_j))$$

En vectorisant pour toutes les dimensions k :

$$\nabla_{\mathbf{u}_i} \log \mathcal{L} = \mathbf{x}_j \left(y_{ij} - f(\mathbf{u}_i, \mathbf{x}_j)\right)$$

Q 5.4 (0.5 pt) En déduire $\nabla_{\mathbf{x}_j} \log \mathcal{L}_{ij}$.

La forme de la vraisemblance est parfaitement symétrique :

$$\nabla_{\mathbf{x}_j} \log \mathcal{L} = \mathbf{u}_i \left(y_{ij} - f(\mathbf{u}_i, \mathbf{x}_j)\right)$$

Q 5.5 (1.5 pt) Pour un couple ($\mathbf{x}_j$, $\mathbf{u}_i$) correspondant à $y_{ij} = 1$. Rappeler les formules de mise à jour des paramètres pour une montée de gradient. Démontrer que la mise à jour des paramètres par montée de gradient provoque une hausse de la fonction d'intérêt f.

Formules de mise à jour pour une montée de gradient :

$$\mathbf{u}_i^{t+1} \leftarrow \mathbf{u}_i^t + \varepsilon \nabla_{\mathbf{u}_i} \log \mathcal{L}$$

$$\mathbf{x}_j^{t+1} \leftarrow \mathbf{x}_j^t + \varepsilon \nabla_{\mathbf{x}_j} \log \mathcal{L}$$

On note :

$$\mathbf{u}_i^{t+1} \leftarrow \mathbf{u}_i^t + \varepsilon \nabla_{\mathbf{u}_i} \log \mathcal{L} = \mathbf{u}_i^t + \varepsilon \mathbf{x}_j \left(y_{ij} - f(\mathbf{u}_i, \mathbf{x}_j)\right) = \mathbf{u}_i^t + \varepsilon^\dagger \mathbf{x}_j, \qquad (y_{ij} - f(\mathbf{u}_i, \mathbf{x}_j)) > 0$$

Pour un couple intéressé : $y_{ij} = 1$:

$$f(\mathbf{u}_i^{t+1}, \mathbf{x}_j^{t+1}) = \frac{1}{1 + \exp(-\mathbf{u}_i^{t+1} \cdot \mathbf{x}_j^{t+1})} = \frac{1}{1 + \exp(-(\mathbf{u}_i^t + \varepsilon \mathbf{x}_j^t) \cdot (\mathbf{x}_j^t + \varepsilon \mathbf{u}_i^t))}$$

$$f(\mathbf{u}_i^{t+1}, \mathbf{x}_j^{t+1}) = \frac{1}{1 + \exp(-(\mathbf{u}_i^t \mathbf{x}_j^t (1+\varepsilon) + \varepsilon (\mathbf{x}_j^t)^2 + \varepsilon (\mathbf{u}_i^t)^2))} > f(\mathbf{u}_i^t, \mathbf{x}_j^t)$$

Q 5.6 (0.5 pt) Expliquer le titre de l'exercice (notamment l'aspect collaboratif de la recommandation).

En considérant tous les couples d'une population, les individus modèlent les profiles de livres pour toute la population, puis les livres *contaminent* les autres utilisateurs. On prend donc en compte les gouts de toute la population pour définir l'ensemble des profiles : le processus est collaboratif.