MAPSI – Examen réparti 1

Durée : 2 heures

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

Exercice 1 (3 pts) – Indépendance

Deux logiciels X et Y implémentent deux fonctions structurellement indépendantes de l'infrastructure d'une grande organisation de vente en ligne. Soit trois variables aléatoires :

- A la variable associée à la proposition "les soldes sont en cours",
- B la variable "X est en marche",
- C la variable "Y est en marche".

	a_1		a_2	
	b_1	b_2	b_1	b_2
c_1	0,288	0,192	0,016	0,024
c_2	0,192	0,128	0,064	0,096

Les trois variables ont modalités $\{a_1, a_2\}$, $\{b_1, b_2\}$, $\{c_1, c_2\}$, correspondant à des évaluations vrai et faux de chaque proposition. La probabilité jointe de ces variables est fournie par le tableau ci-dessus.

Q 1.1 (1 pt) Indépendence. Déterminer si B et C sont indépendantes.

Q 1.2 (1 pt) Indépendence conditionnelle Déterminer le tableau $P(B, C|A)$ et vérifier que B et C sont indépendantes conditionnellement à A.

Q 1.3 (1 pt) Donner une explication plausible au fait que B et C ne soient pas indépendantes malgré l'indépendance structurelle de X et Y.

Exercice 2 (4.75 pts) – Tirage de roulette numérique et hypothèse fumeuse

Dans un casino, face à une roulette électronique, un client statisticien voudrait mieux comprendre le système pour maximiser ses chances... Il a lu sur le site du fabricant que le tirage de chaque numéro émis par la roulette électronique était effectué selon une binomiale de paramètres $p, N = 16$.

Q 2.1 (0.25 pt) Combien de tirages différents sont possibles selon cette loi ?

Q 2.2 (1 pt) Nous observons les tirages suivants : $10, 10, 5, 8, 12, 8, 7, 12, 6$. En faisant l'hypothèse que les tirages sont indépendants, donner une estimation du paramètre p au sens du maximum de vraisemblance.
Rappel $X \sim \mathcal{B}(p, N)$, $\quad p(X = k) = C_N^k p^k (1-p)^{N-k} = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}$

Q 2.3 (1.5 pt) Le résultat ne tombe pas rond... Et notre statisticien est convaincu que le créateur de la machine a dû choisir une probabilité simple. Il pense que c'est 0.5 : cela vous semble-t-il raisonnable ? (à un niveau de confiance 0.95 et en prenant $\sigma^2 = 4$ et en se rappelant que l'espérance de la binomiale est Np)

Q 2.4 (0.5 pt) Calculer : $\dfrac{p(X = k+1)}{p(X = k)}$

Q 2.5 (1.5 pt) Sur quel chiffre faut-il miser ? Formuler le problème, et en vous appuyant sur la question précédente, donner votre réponse.

Exercice 3 (4.5 pts) – **Review Spam**

Les revues d'utilisateurs sur Internet constituent une ressource importante comme déclencheur d'achat. En conséquence, ce système est fréquemment attaqué, notamment par des rédactions massives de fausses revues visant à sur-valoriser ou au contraire dénigrer un produit. Les experts estiment à 2% le taux de spam parmi les revues. Des chercheurs ont mis au point un détecteur de fausses revues, dûment évalué par des experts, présentant les caractéristiques suivantes :

	Spam (s)	Non spam ($\bar{s}$)
Alarme (a)	95%	10%
Pas d'alarme ($\bar{a}$)	5%	90%

Q 3.1 (0.5 pt) Étant donné les deux variables aléatoires en présence, à savoir A (alarme du système de détection) et S (spam selon les experts), que représente le tableau ci-dessus ?

Q 3.2 (1.5 pt) Ce type de système s'évalue en précision et rappel : la précision correspond au taux de vrai spam parmi les alarmes et le rappel (aussi appelé couverture) qui mesure le taux de spam détecté.
Donner la formulation de la précision et du rappel puis effectuer l'application numérique.
En s'appuyant sur le rappel et la précision calculés, comment interpréter les performances d'un système qui retireraient toutes les revues qui ont levé une alarme ?

Q 3.3 Une autre étude portant sur 1 million de compte du site nozamA pointe l'importance de la prise en compte de la taille des comptes (le nombre de messages écrits par les utilisateurs). Les auteurs considèrent qu'un compte compromis n'émet que du spam (et respectivement qu'un compte sain n'en émet pas). Après pré-traitements et regroupement, l'étude propose de regrouper les utilisateurs en 6 catégories, et calcule les statistiques suivantes :

Catégorie de compte utilisateur :	1	2	3	4	5	6	Tot
Nombre moyen de messages par compte :	100	50	20	10	5	1	
Nombre de comptes :	10k	20k	50k	80k	100k	740k	1M
Taux de spam sur la catégorie :	0.01	0.004	0.003	0.001	0.0003	0.0001	

Q 3.3.1 (1 pt) En moyenne, combien de revues sont écrites par chaque utilisateur ? En moyenne toujours, combien de spam sont émis par chaque utilisateur (vous pourrez d'abord déterminer le nombre moyen de spam par compte dans chaque catégorie) ? Cela est-il cohérent avec la question précédente ?

Q 3.3.2 (1.5 pt) Un expert pense que le taux de spameurs est proportionnel au nombre de messages émis. Selon cette hypothèse, le nombre de spameurs attendu par catégorie serait le suivant :

Catégorie de compte utilisateur :	1	2	3	4	5	6	Tot
Nombre de spameurs dans la catégorie :	102	102	102	82	51	75	514

Tester cette hypothèse à un niveau de confiance 95% en détaillant le protocole utilisé.

Exercice 4 (16.5 pts) – Modèle de langue

En recherche d'information, le but est de trouver les documents qui sont pertinents pour une question q formée de la suite de mots $\{m_{q1}, \ldots, m_{qn}\}$. Pour déterminer l'intérêt d'un document d, les moteurs calculent :

$$p(d \text{ pertinent pour } q) = p(q|\mathcal{M}_d) \tag{1}$$

où $\mathcal{M}_d$ est un modèle de langue défini par le document d. En pratique, un moteur de recherche présentera en premier les documents pour lesquels la probabilité $p(q|\mathcal{M}_d)$ est la plus haute. Dans un premier temps, un modèle de langue est simplement une distribution multinomiale donnant les probabilités des mots (et donc des documents en faisant une hypothèse d'indépendance naïve) :

$$\mathcal{M}_d = \{\theta_1, \ldots, \theta_T\}, \qquad \theta_i = p(m_i|\mathcal{M}_d) = \text{ probabilité d'observation du mot } i \tag{2}$$

Un modèle de langue permet de calculer la distribution de probabilité sur des séquences de mots. De manière formelle, un modèle de langue $\mathcal{M}$ permet d'estimer la probabilité $p\,(t|\mathcal{M})$ qu'un texte t soit généré suivant le modèle de langue $\mathcal{M}$, où un texte t est représenté comme une séquence de mots $m_{t1}, \ldots, m_{tl_t}$, où l_t est la longueur du texte t. En pratique, un texte t peut être un document, un ensemble de document ou une question. Nous utiliserons les notations suivantes :

T	nombre de mots dans le vocabulaire
$C = \{d_1, \ldots, d_N\}$	Corpus : ensemble des N documents de notre univers.
f_{dm}	nombre d'occurrences du mot m dans le document d
f_m	nombre d'occurrences du mot m dans tout le corpus C ($f_m = \sum_{k=1}^{N} f_{d_i} m$)
ℓ_d	longueur du document d (nombre de mots : $\ell = \sum_{k=1}^{N} \ell_{d_k}$)
ℓ	nombre de mots dans l'ensemble des documents (le corpus C)

Le processus général est le suivant :

0. $\forall d \in C$, calcul des $\mathcal{M}_d$ au sens du maximum de vraisemblance,
1. une requête $q = \{m_{q1}, \ldots, m_{qn}\}$ arrive,
2. calculs des $p(q|\mathcal{M}_d)$, $\forall d \in C$,
3. retour des documents maximisant la vraisemblance de la requête q.

Q 4.1 Lissage de Jelinek-Mercer

Q 4.1.1 (1.5 pt) Donner la log-vraisemblance d'un document $d = \{m_{d1}, \ldots, m_{d\ell_d}\}$ en fonction des paramètres d'un modèle $\mathcal{M}$ multinomial

Q 4.1.2 (1.5 pt) L'optimisation du modèle multinomial pour un document d, au sens du maximum de vraisemblance aboutit à :

$$\theta_{dm}^{MV} = p\left(m|\mathcal{M}_d^{MV}\right) = \frac{f_{dm}}{l_d}$$

Calculer la pertinence de la question "sorties à Paris" pour les deux documents suivants – on enlèvera la ponctuation et la casse des caractères (distinction majuscule/minuscule) – :

(d1) Sorties de Paris, Ginette et Paulette tournèrent à gauche

(d2) Sorties (Paris)

Quel problème se pose lorsqu'un terme de la question n'apparaît pas dans le document, en particulier un terme peu important (comme "à" dans "sorties à Paris") ? Pourquoi est-ce un problème dans le cadre de la recherche documentaire ?

Q 4.1.3 (1.5 pt) Afin de palier à ce problème, on utilise un modèle de lissage comme celui de Jelinek-Mercer, ce qui permet de définir le modèle de langue $\mathcal{M}_d^{JM}$ pour un document d de la manière suivante :

$$p(m|\mathcal{M}_d^{JM}) = \lambda p(m|\mathcal{M}_d^{MV}) + (1-\lambda)p(m|\mathcal{M}_C^{MV})$$

où le paramètre $\lambda \in [0,1]$, et $p(m|\mathcal{M}_C^{MV})$ est le modèle de langue du corpus C, à savoir le modèle qui maximise la probabilité d'observer l'ensemble des documents si on les mettait bout à bout. On utilise $\mathcal{M}_d^{JM}$ dans l'équation (1).

(a) Que vaut $\theta_m^{MV} = p\left(m|\mathcal{M}_C^{MV}\right)$?

(b) À quoi sert λ ? En particulier, comment l'interpréter de manière probabiliste ?

(c) En quoi cela permet-il de régler le problème lié aux mots absents ?

Q 4.2 Jelinek-Mercer : approfondissement

Q 4.2.1 (0.5 pt) Dire pourquoi le modèle $p\left(m|\mathcal{M}_d^{JM}\right)$ est multinomial

Q 4.2.2 (0.5 pt) Expliquer pourquoi ce modèle ne maximise plus forcément la probabilité d'observer le document d.

Q 4.2.3 (0.5 pt) Pour éviter ce problème, on définit un nouveau modèle de langue $\mathcal{M}_d^{JM^*}$ pour un document d :

$$p(m|\mathcal{M}_d^{JM^*}) = \lambda p(m|\mathcal{M}_d^{JV}) + (1-\lambda)p(m|\mathcal{M}_C^{JV})$$

où $\theta_{dm}^{JV} = p\left(m|\mathcal{M}_d^{JV}\right)$, $\theta_m^{JV} = p\left(m|\mathcal{M}_C^{JV}\right)$ et λ sont les paramètres à apprendre.
En vous souvenant de l'interprétation probabiliste 4.1.3(b), il est possible de considérer λ comme la probabilité d'une variable Z non observée. Quelle est-elle ?

Q 4.2.4 (0.5 pts) La vraisemblance à maximiser correspond à la probabilité d'observer l'ensemble des documents, i.e.

$$\mathcal{L} = p(d_1, \ldots, d_n|\mathcal{M}^{JM^*})$$

En supposant les documents indépendants, on peut l'écrire comme

$$\mathcal{L} = \prod_d p(d|\mathcal{M}_d^{JM^*})$$

Écrire $\log \mathcal{L}$ en fonction de f_{dm}

Q 4.2.5 (3.5 pts) Proposer un algorithme pour estimer les paramètres. Vous utiliserez explicitement la variable Z définie précédemment pour l'estimation.

Q 4.2.6 (1 pt) Lorsqu'on cherche une solution à ce problème, λ converge vers 0 ou 1. Dire ce que valent les autres paramètres dans ce cas, et expliquer pourquoi cela pose problème. Proposez une solution (simple) pour éviter ce problème.

Q 4.3 Lissage de Dirichlet
Un autre modèle de lissage est celui basé sur le lissage de Dirichlet. Pour cela, on suppose que la probabilité a priori des termes est donné par une loi de probabilité de Dirichlet

$$p(\theta_1, \ldots, \theta_T; \boldsymbol{\alpha}) = Dir\left(\theta_1, \ldots, \theta_T; \boldsymbol{\alpha}\right) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{T} \theta_i^{\alpha_i - 1}$$

où $\boldsymbol{\alpha}$ est un vecteur de dimension T ($\boldsymbol{\alpha} \in \mathbb{R}^{+T}$) et $B(\boldsymbol{\alpha})$ permet d'obtenir une distribution de probabilité valide, i.e.

$$B(\mathbf{alpha}) = \int_{[0,1]^T} \prod_{i=1}^{T} \theta_i^{\alpha_i - 1} d\theta$$

Q 4.3.1 (0.5 pt) Exprimer le principe de maximum a posteriori pour un document d (formulation).

Q 4.3.2 (2 pts) Donner le θ^{MAP} qui maximise le maximum a posteriori pour chaque mot m

Q 4.3.3 (3 pts) On suppose que $\alpha_m = \mu p\left(m|\mathcal{M}_C^{MV}\right)$.

(a) Ré-écrire la solution θ_{dm}^{MAP}

(b) En analysant la solution, comment peut-on interpréter μ ?

(c) On suppose que μ est de l'ordre de 1000. Quelle est la valeur de θ_{dm}^{MAP} si le document est très court ($l_d \ll \mu$) ? Si le document est long ($l_d \gg \mu$) ? En déduire l'intérêt du lissage de Dirichlet par rapport au lissage de Jelinek-Mercer.