MAPSI – Examen final

Durée : 2 heures

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

Exercice 1 – Caisses de super-marché (3.5 pts)

Dans une grande surface, les caisses sont regroupées en 5 îlots distincts de tailles similaires. Pourtant la distribution des clients est loin d'être uniforme. En moyenne sur une journée, nous observons la répartition suivante :

Îlot	1	2	3	4	5
Nb de clients	1000	300	500	450	750

La direction du magasin décide d'embaucher un statisticien pour résoudre son problème...

Q 1.1 Explication. Après réflexion, le statisticien pense que l'affluence est liée aux types de rayon près des caisses : il distingue 3 types de rayon –(1) nourriture fraîche, (2) autre nourriture, (3) ce qui ne se mange pas– et calcule le tableau de probabilités suivant :

Rayon\ Îlot	1	2	3	4	5
(1)	0.8	0	0	0	0.1
(2)	0.2	0.25	0.9	1	0.8
(3)	0	0.75	0.1	0	0.1

Q 1.1.1 (0.5 pt) Que modélise ce tableau ?

R : rayon (proche de la caisse)
C : Caisse
$p(R|C)$

Q 1.1.2 (0.5 pt) Est-il possible de retrouver la distribution marginale des rayons à partir des données fournies ? Comment ?

Oui,

$$p(R) = \sum_C p(R, C) = \sum_C p(R|C) * p(C)$$

En estimant p(C) à partir du tableau fourni prcédemment

Q 1.1.3 (1 pt) Les variables Rayon et Ilot sont-elles indépendantes ? Démontrer votre réponse et proposer une courte analyse.

0.5 pour l'indép
0.5 pour l'analyse

```
P(R,C) =
[[ 0.26666667,  0.        ,  0.        ,  0.        ,  0.025     ],
 [ 0.06666667,  0.025     ,  0.15      ,  0.15      ,  0.2       ],
 [ 0.        ,  0.075     ,  0.01666667,  0.        ,  0.025     ]]
```

Impossible de retrouver des dépendances linéaires entre lignes ou entre colonne
Donc pas d'indépendance. On peut émettre l'hypothèse que le type de rayon proche influence fortement le fait d'aller vers un ilot... En particulier que les personnes finissent leurs courses au rayon frais et vont ensuite à la caisse la plus proche.

Q 1.2 (1.5 pt) Plan et efficacité du plan. Suite à cette première analyse, la direction du magasin et le statisticien proposent de mettre une place une solution très simple de fléchage des îlots. Après une semaine, la répartition quotidienne des clients est devenue la suivante :

Îlot	1	2	3	4	5
Nb de clients	970	350	500	430	750

Le statisticien est satisfait du changement mais la direction pas du tout... D'après vous, ce changement est-il dérisoire ou significatif ? D'où provient la différence d'interprétation ?

1pt pour le chi2 + DDL + limite
0.5 pt analyse 95% VS 99%
Il faut faire un test du chi2 en prenant comme référence le premier tableau (C)

$$D = \sum_i \frac{(C_i - C2_i)^2}{C_i} = 10.12$$

DDL = 4
à 95% de confiance, la limite est 9.49, le changement est alors significatif... Mais à 99% de confiance, la limite est à 13.3 et le changement n'en est pas un !
On est à la limite d'un changement significatif...
... Et dans la pratique, on a l'impression que rien n'a changé :)

Exercice 2 – Programmation dynamique & MMC : rétro-propagation des β (3.5 pts)

Dans le cadre de la manipulation des modèles de Markov Cachés, nous nous limitons à un modèle à observations discrètes. Nous nous intéressons au problème 1 : l'évaluation de la vraisemblance d'une séquence d'observations, les états n'étant pas connus. Nous considérons un modèle de paramètres $\lambda = \{\Pi, A, B\}$, une séquence de longueur T avec des observations x_1^T, et des états cachés s_1^T (notations du cours). Les états sont discrets et directement représentés par leur indice : $s_t \in \{1, \ldots, i, \ldots, N\}$ – Rappel : $\pi_i = p(s_1 = i \mid \lambda), A_{ij} = p(s_{t+1} = j \mid s_t = i, \lambda), B_j(x_t) = p(x_t \mid s_t = j, \lambda)$

Q 2.1 (1 pt) Rappeler la formulation de ce problème et indiquer rapidement la principale difficulté associée.

0.5 pt formulation
0.5 pt explication de la combinatoire
Définition de la vraisemblance : $p(x_1^T|\lambda)$
Principale difficulté : pour calculer la vraisemblance, il faut repasser par la loi jointe sur les états et observations :

$$p(x_1^T|\lambda) = \sum_{s_1^T} p(x_1^T, s_1^T|\lambda)$$

A ce moment là, la combinatoire dans la somme est ingérable en direct : il faut factoriser des termes, c'est le principe de la programmation dynamique.

Q 2.2 Pour pallier ce problème, nous allons introduire des termes $\beta_t(i) = p(x_{t+1}^T|s_t = i, \lambda)$, avec une initialisation à $\forall i, \beta_T(i) = 1$.

Q 2.2.1 (0.5 pt) Terminaison. Exprimer la vraisemblance de la séquence d'observation étant donné λ en fonction des β.

$$p(x_1^T|\lambda) = \sum_i \beta_1(i)\pi_i$$

La relation de récursion est la suivante :

$$\beta_t(i) = \sum_{j=1}^{N} A_{ij} B_j(x_{t+1}) \beta_{t+1}(j)$$

Q 2.2.2 (2 pts) Démontrer la validité de cette formule à l'aide des théorèmes de base et des hypothèses d'indépendance introduites lors de la définition des chaînes de Markov cachées. Vous procéderez de la même manière que lors du cours, en détaillant toutes les étapes.

Pour la lisibilité, nous enlevons les λ.

$$\beta_t(i) = p(x_{t+1}^T|s_t = i, \lambda) = p(x_{t+1}^T|s_t = i) = p(x_{t+1}, x_{t+2}^T|s_t = i)$$

Introduction de s_{t+1} :

$$p(x_{t+1}, x_{t+2}^T|s_t = i) = \sum_j p(x_{t+1}, x_{t+2}^T, s_{t+1} = j|s_t = i)$$

En utilisant : $p(A, B|C) = p(A|B, C)p(B|C)$

$$\sum_j p(x_{t+1}, x_{t+2}^T, s_{t+1} = j|s_t = i) = \sum_j p(x_{t+1}, x_{t+2}^T|s_{t+1} = j, s_t = i)p(s_{t+1} = j|s_t = i)$$

Si on connait s_{t+1}, alors s_t n'a plus d'influence sur les observations postérieures. De plus connaissant s_{t+1}, l'observation x_{t+1} est indépendante des observations postérieures.

$$= \sum_j p(x_{t+1}, x_{t+2}^T|s_{t+1} = j)A_{ij} = \sum_j p(x_{t+1}|s_{t+1} = j)p(x_{t+2}^T|s_{t+1} = j)A_{ij}$$

$$= \sum_j B_j(x_{t+1})\beta_{t+1}(j)A_{ij}$$

Exercice 3 – Modélisation des étudiants (5.5 pts)

On considère un examen, pour lequel un étudiant s peut répondre à la question q correctement, ce qui est noté $x_{qs} = 1$ ou incorrectement, $x_{qs} = 0$. On suppose que la chance de succès dépend de la capacité de l'étudiant $\alpha_s \in \mathbb{R}$ et de la difficulté de la question $\delta_q \in \mathbb{R}$. On postule le modèle de réponse suivant :

$$p(x_{qs} = 1|\alpha_s, \delta_q) = \sigma(\alpha_s - \delta_q), \quad \text{avec : } \sigma(x) = \frac{1}{1 + \exp(-x)}$$

Q 3.1 (0.5 pt) Préliminaire

Q 3.1.1 Etudier rapidement la fonction σ et conclure qu'elle permet effectivement de modéliser une probabilité.

si $x \to +\infty$ alors $\sigma(x) = 1$
si $x \to -\infty$ alors $\sigma(x) = 0$

Q 3.1.2 A quelle condition un étudiant s a-t-il autant de chance de répondre correctement que de se tromper à une question q ?

$p(x_{qs} = 1|\alpha_s, \delta_q) = 0.5 \iff x = 0 \iff \alpha_s = \delta_q$

Q 3.2 (0.5 pt) Quelle loi permet de modéliser la variable x_{qs} ? Montrer que :

$$p(x_{qs}|\alpha_s, \delta_q) = \sigma(\alpha_s - \delta_q)^{x_{qs}}(1 - \sigma(\alpha_s - \delta_q))^{(1-x_{qs})}$$

où x_{qs} peut prendre les valeurs 0 ou 1.

Modélisation classique pour Bernoulli de paramètre p, il suffit de vérifier $p(x) = x^p(1 - x)^{1-p}$

Q 3.3 Les données binaires x_{qs} sont dans une matrice X de taille $Q \times N$ pour Q questions et N étudiants.

Q 3.3.1 (1.5 pt) Donner la vraisemblance $p(X|\alpha,\delta)$ et la log vraisemblance $\mathcal{L} = \log p(X|\alpha,\delta)$? Rappeler les hypothèses que vous faites pour avancer dans le calcul.
Indice : $\mathcal{L}$ est de la forme : $\mathcal{L} = \sum_q \sum_s \beta(\alpha_s - \delta_q) + \gamma \log \sigma(\alpha_s - \delta_q)$

Questions indep. + Etudiants indep.

$$p(X|\alpha,\delta) = \prod_q \prod_s p(x_{qs}|\alpha_s,\delta_q) = \prod_q \prod_s \sigma(\alpha_s - \delta_q)^{x_{qs}} (1 - \sigma(\alpha_s - \delta_q))^{(1-x_{qs})}$$

$$\mathcal{L} = \log p(X|\alpha,\delta) = \sum_q \sum_s x_{qs} \log(\sigma(\alpha_s - \delta_q)) + (1 - x_{qs}) \log(1 - \sigma(\alpha_s - \delta_q))$$

$$1 - \sigma(\alpha_s - \delta_q) = \frac{\exp(-(\alpha_s - \delta_q))}{1 + \exp(-(\alpha_s - \delta_q))}$$

$$\log(\sigma(\alpha_s - \delta_q)) = -\log(1 + \exp(-(\alpha_s - \delta_q))),$$

$$\log(1 - \sigma(\alpha_s - \delta_q)) = -(\alpha_s - \delta_q) - \log(1 + \exp(-(\alpha_s - \delta_q)))$$

$$\mathcal{L} = \sum_q \sum_s -x_{qs} \log(1 + \exp(-(\alpha_s - \delta_q))) + (1 - x_{qs})(-(\alpha_s - \delta_q) - \log(1 + \exp(-(\alpha_s - \delta_q))))$$

$$\mathcal{L} = \sum_q \sum_s -(1 - x_{qs})(\alpha_s - \delta_q) - \log(1 + \exp(-(\alpha_s - \delta_q))) = \sum_q \sum_s -(1 - x_{qs})(\alpha_s - \delta_q) + \log \sigma(\alpha_s - \delta_q)$$

Q 3.3.2 (1.5 pt) Montrer que $\sigma'(x) = \sigma(x)(1 - \sigma(x))$, en déduire $(\log \sigma(x))'$ puis exprimer $\frac{\partial \mathcal{L}}{\partial \alpha_s}$ et $\frac{\partial \mathcal{L}}{\partial \delta_q}$.

$$\sigma(x)' = -u'/u^2 = \frac{-\exp(-x)}{(1 + \exp(-x))^2} = \sigma(x)(1 - \sigma(x))$$

$$(\log \sigma(x))' = \frac{\sigma(x)'}{\sigma(x)} = 1 - \sigma(x)$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_s} = \sum_q -(1 - x_{qs}) + (1 - \sigma(\alpha_s - \delta_q))$$

Attention au signe pour δ

$$\frac{\partial \mathcal{L}}{\partial \delta_q} = \sum_s (1 - x_{qs}) - (1 - \sigma(\alpha_s - \delta_q))$$

Q 3.3.3 (0.5 pt) Proposer un algorithme de gradient pour optimiser cette vraisemblance.

1. Soit ε, X
2. Tant que convergence non atteinte (ou nb itération fixée...)
3. calculer $\frac{\partial \mathcal{L}}{\partial \alpha_s}$ et $\frac{\partial \mathcal{L}}{\partial \delta_q}$
4. MAJ : $\alpha_s \leftarrow \alpha_s + \varepsilon \frac{\partial \mathcal{L}}{\partial \alpha_s}$ ...

Attention à faire les MAJ en parallèle (calcul préalable des dérivées) Attention au signe (montée de gradient)

Q 3.4 (1 pt) On voudrait se servir de ce type de formulation pour construire un classement des joueurs aux échecs. Comment procéder ?

C'est une épreuve de Bernoulli. Pour une partie p, soit i gagne, soit i perd face à j

$$p(i,j,p) = p(i>j)^{x_{ij}^p} p(j>i)^{1-x_{ij}^p}$$

La formule proposée contient un double produit sur i et j qui permet d'envisager les deux solutions pour chaque partie.

$$p(X|\alpha) = \prod_{p=1}^{P} \prod_{i,j=1}^{N} \sigma(\alpha_i - \alpha_j)^{x_{ij}^p}$$

Exercice 4 – Est-ce que la publicité marche ? (3 pts)

Pendant une journée de travail, le service client d'une grande entreprise opérant à niveau national reçoit un très grand nombre de messages. Dans la dernière semaine, ils ont reçu le nombre suivant de messages par jour –exprimé en dizaine de milliers– :

3	6	5	16	10

Q 4.1 (1 pt) Dans un premier temps, le nombre de messages par jour est modélisé par une variable aléatoire X suivant une loi de Poisson :

$$P(X=n) = \frac{\lambda^n}{n!} e^{-\lambda}$$

Trouver le paramètre λ de ce modèle en utilisant l'échantillon donné par *maximum de vraisemblance*.

On cherche le paramètre $\theta = \lambda$ qui maximise la vraisemblance :

$$L(\mathbf{x},\theta) = L(x_1,x_2,x_3,x_4,x_5;\theta) = \prod_{i=1}^{5} \frac{\theta^{x_i}}{x_i!} e^{-\theta}$$

En faisant le logarithme :

$$\ln L(\mathbf{x},\theta) = \sum_{i=1}^{5} \ln \frac{\theta^{x_i}}{x_i!} e^{-\theta} = \sum_{i=1}^{5} \ln \theta^{x_i} + \sum_{i=1}^{5} \ln e^{-\theta} - \sum_{i=1}^{5} \ln x_i! = \ln\theta \cdot \sum_{i=1}^{5} x_i - 5\theta - \sum_{i=1}^{5} \ln x_i!$$

En cherchant le maximum :

$$\frac{\partial \ln L(\mathbf{x},\theta)}{\partial\theta} = 0 \qquad \Leftrightarrow \qquad \frac{\partial \ln\theta \cdot \sum_{i=1}^{5} x_i - 5\theta - \sum_{i=1}^{5} \ln x_i!}{\partial\theta} = 0$$

$$\frac{1}{\theta} \sum_{i=1}^{5} x_i - 5 = 0 \qquad \Leftrightarrow \qquad \theta = \frac{\sum_{i=1}^{5} x_i}{5} = \overline{x}$$

Donc $\theta = 8$.

Q 4.2 (1 pt) Un expert a proposé une amélioration du modèle, en supposant que le nombre de messages est proportionnel au nombre des passages publicitaires programmés sur le média. En notant le nombre d'affichages publicitaires dans la journée i avec k_i, on suppose que le nombre de message X_i suit une loi de Poisson de paramètre : $\lambda_i = k_i \cdot \lambda$ Trouver le paramètre λ maximisant la vraisemblance de ce nouveau modèle, en sachant que les nombres d'affichages k_i correspondant à chaque jour sont les suivants :

10	30	20	70	70

On suit la même méthode qu'avant, en modifiant quelques équations.

$$\ln L(\mathbf{x},\theta) = \cdot \sum_{i=1}^{5} x_i \ln(k_i) + \ln(\theta) \sum_{i=1}^{5} x_i - \theta \sum_{i=1}^{5} k_i - \sum_{i=1}^{5} \ln x_i!$$

En cherchant le maximum :

$$\frac{\partial \ln L(\mathbf{x},\theta)}{\partial \theta} = 0 \quad \Leftrightarrow \quad \frac{1}{\theta}\sum_{i=1}^{5} x_i - \sum_{i=1}^{5} k_i = 0 \quad \Leftrightarrow \quad \theta = \frac{\sum_{i=1}^{5} x_i}{\sum_{i=1}^{5} k_i} = \frac{\overline{x}}{\overline{k}}$$

Donc $\theta = \frac{40}{200} = 0.2$.

Q 4.3 (1 pt) Après ces tests initiaux, un nombre d'affichage publicitaire rentable $k^\star = 42$ a été calculé et est utilisé tous les jours. Au bout de 5 semaines (de 5 jours ouvrés), le service client a enregistré les nombres de messages et d'affichages publicitaires par jour et veut faire un bilan de la politique publicitaire mise en place. Ils ont peur que l'impact publicitaire n'ait baissé.
Pour rappel : selon notre modèle, un jour X_i correspond en moyenne à $k^\star\lambda$ dizaines de milliers de messages (avec un écart-type de $k^\star\lambda$).
Pouvez-vous tester si l'impact de la publicité est resté le même en utilisant un test de confiance ? Comment ?

L'effet de la publicité est capturé par le λ calculé comme à la question 4.2. Le test d'hypothèse est donc $H_0 : \lambda = 0.2$ vs $H_1 : \lambda < 0.2$. Pour construire un test, pour le cas standard, il faut qu'on considère un seuil c pour laquelle H_1 peut être accepté.
Soit X_i la variable aléatoire "nombre de messages" au jour i. Puisque le $k^\star$ reste constante, la distribution est à priori la même pour tous les jours, et on peut donc tester si l'extraction confirme ça.
Soit $\overline{X}$ le poids moyen de l'échantillon (on a $s \times 5 = 25$ jours). Pour le théorème central limite, on sait que $\frac{\overline{X}-\mu}{\sigma/\sqrt{25}} \sim \mathcal{N}(0,1)$.
On cherche alors c tels que :

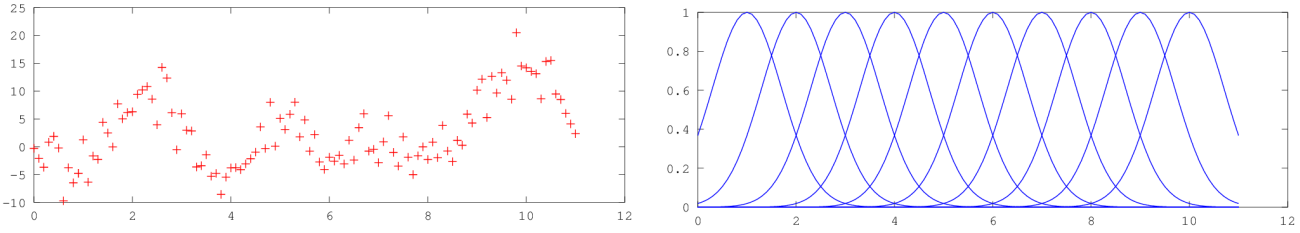
$$P(\frac{\overline{X} - k^\star\lambda}{k^\star\lambda/5} < \frac{c_i - k^\star\lambda}{k^\star\lambda/5}) = P(Z < \frac{c_i - k^\star\lambda}{k^\star\lambda/5}) = \alpha = 1 - \text{niveau de confiance}$$

On comparant les moyennes empiriques $\overline{x}$ avec c, on peut déduire si la moyenne a baissé, et donc si l'impact de la publicité a finalement changé ou non.

Exercice 5 – Régression par mixture de gaussiennes (4.5 pts)

Soit le problème de régression unidimensionnelle (non-linéaire) présenté dans la figure ci-dessous (figure de gauche). On propose d'utiliser un modèle à base d'une mixture pondérée de n_G gaussiennes (présentées en figure de droite) :

$$f(x) = \sum_{k=1}^{n_G} w_k g_k(x), \qquad g_k(x) = \exp(-\|x - \mu_k\|^2)$$



Les μ_k sont données et les fonctions g_k sont représentées sur la figure de droite. Le problème consiste à trouver les coefficients w_k de la mixture. On dispose d'une base de données composée de N couples : $(\mathbf{x},\mathbf{y}) = \{(x_j, y_j)\}_{j=1,\ldots,N}$

Q 5.1 (0.5 pt) Exprimer le coût de régression au sens des moindres carrés $C(\mathbf{x},\mathbf{y},\mathbf{w})$.

$$C(\mathbf{x},\mathbf{y},\mathbf{w}) = \sum_j (\sum_i w_i g_i(x_j) - y_j)^2$$

Q 5.2 (1 pt) Quel problème doit-on résoudre pour trouver $\mathbf{w}^\star$, le vecteur des poids optimaux au sens des moindres carrés ? Proposer une solution approximative, uniquement à partir des figures ci-dessus.

0.5 formulation
0.5 estimation à la louche

$$\mathbf{w}^\star = \arg\max_{\mathbf{w}} \sum_j (\sum_i w_i g_i(x_j) - y_j)^2$$

Dimension = 10
[-5 0 15 -5 10 0 5 0 0 15]

Q 5.3 (0.5 pt) Soit la matrice G, avec $G \in \mathbb{R}^{n_G \times N}$ composée de $g_{ij} = g_i(x_j)$. Exprimer $\hat{\mathbf{y}}$ en fonction de G et $\mathbf{w}$ de manière matricielle.

$\hat{\mathbf{y}} = G\mathbf{w}$

Q 5.4 (2 pts) Calculer la dérivée de la fonction coût par rapport à w_i. Expliquer en détail comment annuler ces dérivées, soit de manière matricielle – en donnant l'expression de $\mathbf{w}^\star$ en fonction de la matrice G et de $\mathbf{y}$– soit en conservant les notations indicielles.

En fait, ça devient exactement le problème traité en cours de semestre :

$$C = (G^T\mathbf{w} - \mathbf{y})^2$$

$$\nabla_{\mathbf{w}} C = 2G(G^T\mathbf{w} - \mathbf{y})$$

$$\nabla_{\mathbf{w}} C = 0 \Longleftrightarrow \mathbf{w}^\star = (GG^T)^{-1}G\mathbf{y}$$

Q 5.5 (0.5 pt) Donner un exemple d'application où ce type de modèle est utile.