Examen de 1ère session de MAPSI

Durée : 2 heures

Seuls documents autorisés : Calculatrices et votre antisèche recto-verso
– Barême indicatif –

Exercice 1 – Attente au distributeur de billets (4pt)

En supposant que le traitement d'une demande d'un client au distributeur est réalisé en temps constant (égal à 1 unité de temps) et que la distribution des arrivées des clients au distributeur suit une loi de Poisson de paramètre μ (unités de temps), la distribution de probabilité des durées X pendant lesquelles le distributeur est occupé suit une distribution de Borel de paramètre μ définie par :

$$P(X = n \text{ unités de temps}) = \frac{e^{-\mu n}(\mu n)^{n-1}}{n!}$$

où le paramètre μ appartient à l'intervalle $[0, 1]$. Sur une période de quelques jours, on a observé les durées d'occupation suivantes :

X	2	3	1	5	2	7

Q 1.1 (2pt) Déterminez par maximum de vraisemblance la valeur du paramètre μ.

La vraisemblance est égale à :

$$L(\mathbf{x}, \mu) = \prod_{i=1}^{6} \frac{e^{-\mu x_i}(\mu x_i)^{x_i - 1}}{x_i!},$$

où les x_i sont les observations de la variable X. Par conséquent,

$$\log L(\mathbf{x}, \mu) = \sum_{i=1}^{6} [-\mu x_i + (x_i - 1)\log \mu + (x_i - 1)\log x_i - \log(x_i!)].$$

Donc la dérivée de la log-vraisemblance est égale à :

$$\frac{\partial \log L(\mathbf{x}, \mu)}{\partial \mu} = \sum_{i=1}^{6} \left[-x_i + \frac{x_i - 1}{\mu} \right].$$

En remplaçant les x_i par leurs valeurs, on obtient :

$$\frac{\partial \log L(\mathbf{x}, \mu)}{\mu} = -20 + \frac{14}{\mu}.$$

Cette expression est égale à 0 lorsque $-20\mu + 14 = 0$, autrement dit $\mu = 14/20 = 0,7$.

Q 1.2 (2pt) Un expert de la banque affirme qu'*a priori* la distribution des paramètres μ dans la région où est situé le distributeur suit une loi Beta de paramètres $\alpha = 7$ et $\beta = 21$. On rappelle que

la loi Beta est définie de la manière suivante :

$$\text{Beta}(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

En utilisant cet apriori, calculez par *maximum a posteriori* la valeur du paramètre μ.

Pour le *maximum a posteriori*, on calcule $L(\mathbf{x},\mu)\times\pi(\mu)$, où π est la loi Beta indiquée ci-dessus. En passant au log, on obtient donc :

$$\log(L(\mathbf{x},\mu)\times\pi(\mu)) = \sum_{i=1}^{6}[-\mu x_i + (x_i-1)\log\mu + (x_i-1)\log x_i - \log(x_i!)] + \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) + (\alpha-1)\log\mu + (\beta-1)\log(1-\mu).$$

Par conséquent, en remplaçant les x_i par leurs valeurs et en dérivant, on obtient :

$$\begin{aligned}\frac{\partial\log(L(\mathbf{x},\mu)\times\pi(\mu))}{\mu} &= -20 + \frac{14}{\mu} + \frac{\alpha-1}{\mu} - \frac{\beta-1}{1-\mu} = -20 + \frac{14}{\mu} + \frac{6}{\mu} - \frac{20}{1-\mu} = -20 + \frac{20}{\mu} - \frac{20}{1-\mu} \\ &= \frac{20\times(-\mu(1-\mu)+(1-\mu)-\mu)}{\mu(1-\mu)} = \frac{20}{\mu(1-\mu)}(\mu^2-3\mu+1).\end{aligned}$$

La dérivée s'annule lorsque $\mu^2-3\mu+1=0$, autrement dit lorsque $\mu=(3\pm\sqrt{5})/2$. Mais comme μ doit appartenir à l'intervalle $[0,1]$, la seule solution possible est $\mu=(3-\sqrt{5})/2\approx 0,382$.

Exercice 2 – Boursicotage (3pt)

Votre banquier vous propose un placement. Avant de l'accepter, vous avez observé le rendement de ce dernier sur chaque semaine pendant environ un an (exactement 50 semaines). Pour 1000 € investis, vous avez relevé les gains suivants :

gain	5 €	10 €	20 €	50 €
nombre de semaines	10	15	20	5

Autrement dit, pendant 10 semaines, un placement de 1000 € aurait rapporté chaque semaine 5 €, pendant 15 semaines il aurait rapporté chaque semaine 10 €, etc.

Q 2.1 (1pt) Calculez la moyenne et la variance de cet échantillon.

La moyenne $\overline{x}$ de l'échantillon est égale à

$$\overline{x} = \frac{10\times5+15\times10+20\times20+5\times50}{10+15+20+5} = \frac{850}{50} = 17.$$

La variance S^2 est égale à :

$$S^2 = \frac{1}{50}\left[10\times(5-17)^2+15\times(10-17)^2+20\times(20-17)^2+5\times(50-17)^2\right] = \frac{7800}{50} = 156.$$

Q 2.2 (2pt) Le banquier vous annonce que des études menées par la banque montrent que le gain moyen de ce placement est de 20 € par semaine, avec un écart-type de 10 €. Selon un test d'hypothèse de niveau de confiance 95%, pouvez-vous confirmer que ce gain par semaine n'a pas baissé ?

Appelons X la variable aléatoire « gain par semaine » et $\overline{X}$ le gain moyen par semaine sur l'ensemble des échantillons possibles. Dans notre test d'hypothèses, nous avons l'hypothèse nulle $H_0 : \mu = 20$ et la contre-hypothèse $H_1 : \mu < 20$. On a $n = 50$ observations. Donc, d'après le théorème central limite, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, où $\sigma = 10$. Le test revient donc à chercher le seuil c tel que :

$$\begin{aligned}
\alpha &= P(\text{rejeter } H_0 | H_0 \text{ est vraie}) = 5\% \\
&= P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{c-\mu}{\sigma/\sqrt{n}} \middle| \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \mu = 20, \sigma = 10\right) \\
&= P\left(\frac{\overline{X}-20}{10/\sqrt{50}} < \frac{c-20}{10/\sqrt{50}} \middle| \frac{\overline{X}-20}{10/\sqrt{50}} \sim \mathcal{N}(0,1)\right)
\end{aligned}$$

D'après la table de la loi normale, on a $\frac{c-20}{10/\sqrt{50}} \approx -1,645$. Donc $c \approx 20 - 1,645 \times 10/\sqrt{50} \approx 17,674$. Comme, d'après la question précédente, $\overline{x} = 17 < c$, on doit rejeter le fait que le gain par semaine n'a pas baissé.

Exercice 3 – Indépendance (3pt)

Un site de vente sur internet a recensé le volume de ventes d'un produit en fonction de l'âge de l'acheteur. Soit X la variable « âge de l'acheteur », dont le domaine est {jeune,vieux}, et Y la variable « volume de ventes », dont le domaine est {petit,moyen,élevé}. Sur les 100 dernières ventes, le site a enregistré la répartition suivante :

$X \backslash Y$	petit	moyen	élevé
jeune	7	22	11
vieux	13	28	19

À un niveau de confiance de 95%, peut-on considérer que les variables X et Y sont indépendantes ? Vous justifierez votre réponse.

On commence par rajouter les marginales :

$X \backslash Y$	petit	moyen	élevé	total
jeune	7	22	11	40
vieux	13	28	19	60
total	20	50	30	100

On en déduit donc ce que le tableau de contingence que l'on devrait obtenir théoriquement si X et Y sont bien indépendants est :

$X \backslash Y$	petit	moyen	élevé	total
jeune	8	20	12	40
vieux	12	30	18	60
total	20	50	30	100

Par conséquent, la statistique d^2 est égale à :

$$\begin{aligned} d^2 &= \frac{(7-8)^2}{8} + \frac{(22-20)^2}{20} + \frac{(11-12)^2}{12} + \frac{(13-12)^2}{12} + \frac{(28-30)^2}{30} + \frac{(19-18)^2}{18} \\ &= \frac{1}{8} + \frac{1}{5} + \frac{(1}{12} + \frac{1}{12} + \frac{2}{15} + \frac{1}{18} \approx 0,6805. \end{aligned}$$

Dans ce test d'indépendance, on a $(3-1) \times (2-1) = 2$ degrés de liberté. D'après la table de la loi du χ^2 à 2 degrés de liberté, pour $\alpha = 5\%$, $d^2_\alpha = 5,99$. Comme $d^2 < d^2_\alpha$, on en déduit que l'on peut considérer X et Y comme indépendantes.

Exercice 4 – Mixtures et processeurs (6pt)

Sur un ordinateur muni d'un processeur à deux cœurs, on a noté les temps (en millisecondes) d'inaction des cœurs. On sait que, pour chaque cœur, la distribution de probabilité de ces temps est une distribution exponentielle (on rappelle que la distribution exponentielle de paramètre λ a pour fonction de densité $f(x;\lambda) = \lambda e^{-\lambda x}$). La distribution des temps d'inaction (variable X) sur le processeur est donc une mixture de distributions exponentielles :

$$p(x) = \pi_1 \lambda_1 e^{-\lambda_1 x} + (1-\pi_1)\lambda_2 e^{-\lambda_2 x}.$$

Un logiciel a relevé les temps d'inaction suivants, sans indiquer à quels cœurs ils correspondent :

3	5	2

Q 4.1 (1pt) Soit $\Theta = \{\pi_1; \lambda_1; \lambda_2\}$. On introduit la variable de classe Y spécifiant à quelle loi exponentielle on appartient ($Y = 1$ si le temps correspond au 1er cœur, $Y = 2$ sinon). Cette variable est non observée. Soit x_i une valeur observée dans le tableau ci-dessus. Donnez l'expression mathématique de $p(x_i|Y = y, \Theta)$.

$$p(x_i|y,\Theta) = \begin{cases} \lambda_1 e^{-\lambda_1 x_i} & \text{si } y = 1 \\ \lambda_2 e^{-\lambda_2 x_i} & \text{si } y = 2 \end{cases}$$

Q 4.2 (1pt) Indiquez l'expression de la fonction $Q_i(y)$ de l'algorithme EM en fonction de la $i$ème observation x_i et des paramètres de Θ.

$$Q_i(y) = p(y|x_i,\Theta) = \frac{p(x_i|y,\Theta)p(y|\Theta)}{p(x_i|\Theta)} \propto p(x_i|y,\Theta)p(y|\Theta) = \begin{cases} \pi_1 \lambda_1 e^{-\lambda_1 x_i} & \text{si } y = 1 \\ (1-\pi_1)\lambda_2 e^{-\lambda_2 x_i} & \text{si } y = 2 \end{cases}$$

Q 4.3 (1pt) Donnez l'expression mathématique de la fonction que l'on maximise dans l'étape M de l'algorithme EM en fonction des paramètres $\Theta = \{\pi_1; \lambda_1; \lambda_2\}$, des fonctions Q_i et des observations x_i.

À l'étape M, on doit calculer $\text{Argmax}_\Theta \log L(\mathbf{x}, \Theta)^1$, défini par :

$$\begin{aligned}\log L(\mathbf{x}, \Theta)^1 = & Q_1(y_1)\log[\pi_1\lambda_1 e^{-\lambda_1 x_1}] + Q_1(y_2)\log[(1-\pi_1)\lambda_2 e^{-\lambda_2 x_1}] + \\ & Q_2(y_1)\log[\pi_1\lambda_1 e^{-\lambda_1 x_2}] + Q_2(y_2)\log[(1-\pi_1)\lambda_2 e^{-\lambda_2 x_2}] + \\ & Q_3(y_1)\log[\pi_1\lambda_1 e^{-\lambda_1 x_3}] + Q_3(y_2)\log[(1-\pi_1)\lambda_2 e^{-\lambda_2 x_3}] + \text{constante}\end{aligned}$$

où la constante est égale à $-\sum_{i=1}^{3}\sum_{j=1}^{2} Q_i(y_j)\log Q_i(y_j)$. Cette expression étant constante par rapport aux paramètres de Θ, on n'en tient pas compte dans les calculs de EM.

Q 4.4 (3pt) En partant de $\Theta^0 = \{\pi_1 = 0,4; \lambda_1 = 2; \lambda_2 = 3\}$, appliquez une étape de l'algorithme EM.

À l'étape E, on calcule la distribution $Q_i(y)$ pour chaque observation de la base de données. D'après la question **??**, on a donc :

$$\begin{aligned}Q_1(y) &\propto \begin{bmatrix} 0,4\times 2\times e^{-2\times 3} \\ 0,6\times 3\times e^{-3\times 3}\end{bmatrix} = \begin{bmatrix} 0,001983 \\ 0,000222\end{bmatrix} \propto \begin{bmatrix} 0,9 \\ 0,1\end{bmatrix} \\ Q_2(y) &\propto \begin{bmatrix} 0,4\times 2\times e^{-2\times 5} \\ 0,6\times 3\times e^{-3\times 5}\end{bmatrix} = \begin{bmatrix} 0,00003632 \\ 0,00000055\end{bmatrix} \propto \begin{bmatrix} 0,985 \\ 0,015\end{bmatrix} \\ Q_3(y) &\propto \begin{bmatrix} 0,4\times 2\times e^{-2\times 2} \\ 0,6\times 3\times e^{-3\times 2}\end{bmatrix} = \begin{bmatrix} 0,014652 \\ 0,004462\end{bmatrix} \propto \begin{bmatrix} 0,767 \\ 0,233\end{bmatrix}\end{aligned}$$

À l'étape M, on doit calculer $\text{Argmax}_\Theta \log L(\mathbf{x}, \Theta)^1$. D'après la question **??**, on a :

$$\begin{aligned}\log L(\mathbf{x}, \Theta)^1 = & 0,9\log[\pi_1\lambda_1 e^{-3\lambda_1}] + 0,1\log[(1-\pi_1)\lambda_2 e^{-3\lambda_2}] + \\ & 0,985\log[\pi_1\lambda_1 e^{-5\lambda_1}] + 0,015\log[(1-\pi_1)\lambda_2 e^{-5\lambda_2}] + \\ & 0,767\log[\pi_1\lambda_1 e^{-2\lambda_1}] + 0,233\log[(1-\pi_1)\lambda_2 e^{-2\lambda_2}]\end{aligned}$$

En dérivant l'expression ci-dessus par rapport à π_1, on obtient :

$$\frac{\partial \log L(\mathbf{x}, \Theta)^1}{\partial \pi_1} = (0,9 + 0,985 + 0,767)\frac{1}{\pi_1} - (0,1 + 0,015 + 0,233)\frac{1}{1-\pi_1} = \frac{2.652}{\pi_1} - \frac{0,348}{1-\pi_1}.$$

Cette dérivée s'annule pour $\pi_1 = 0,884$.
En dérivant l'expression par rapport à λ_1, on obtient :

$$\begin{aligned}\frac{\partial \log L(\mathbf{x}, \Theta)^1}{\partial \lambda_1} &= 0,9 \times \left(\tfrac{1}{\lambda_1} - 3\right) + 0,985 \times \left(\tfrac{1}{\lambda_1} - 5\right) + 0,767 \times \left(\tfrac{1}{\lambda_1} - 2\right) \\ &= \tfrac{2,652}{\lambda_1} - 9,159\end{aligned}$$

Cette dérivée s'annule pour $\lambda_1 = 2,652/9,159 = 0,28955$.
En dérivant l'expression ci-dessus par rapport à λ_2, on obtient :

$$\begin{aligned}\frac{\partial \log L(\mathbf{x}, \Theta)^1}{\partial \lambda_2} &= 0,1 \times \left(\tfrac{1}{\lambda_2} - 3\right) + 0,015 \times \left(\tfrac{1}{\lambda_2} - 5\right) + 0,233 \times \left(\tfrac{1}{\lambda_2} - 2\right) \\ &= \tfrac{0,348}{\lambda_2} - 0,841\end{aligned}$$

Cette dérivée s'annule pour $\lambda_2 = 0,348/0,841 = 0,4138$.

Exercice 5 – Indépendance en probabilité (4pt)

Soit 5 variables aléatoires A, B, C, D, E dont la distribution de probabilité jointe se décompose selon l'expression suivante :

$$P(A, B, C, D, E) = P(A) \times P(B|A) \times P(C) \times P(D|B, C) \times P(E|D).$$

Q 5.1 (1pt) Exprimez la distribution $P(B, C)$ en fonction de $P(A, B, C, D, E)$.

Comme nous l'avons vu en cours, $P(B, C) = \sum_{A,D,E} P(A, B, C, D, E)$. Autrement dit, pour supprimer des variables d'une probabilité jointe, il faut effectuer des sommations sur ces variables.

Q 5.2 (0,5pt) Rappelez sous quelle condition sur $P(B, C)$ les deux variables B et C sont indépendantes.

B et C sont indépendantes si $P(B, C) = P(B) \times P(C)$.

Q 5.3 (2,5pt) Déterminez si B et C sont des variables aléatoires indépendantes. Vous justifierez mathématiquement votre réponse.

B et C sont indépendantes si $P(B, C) = P(B) \times P(C)$. Comme nous l'avons vu précédemment, $P(B, C) = \sum_{A,D,E} P(A, B, C, D, E)$. On va donc sommer sur A, D et E, la loi jointe :

$$\begin{aligned} P(A, B, C, D) &= \sum_E [P(A) \times P(B|A) \times P(C) \times P(D|B, C) \times P(E|D)] \\ &= P(A) \times P(B|A) \times P(C) \times P(D|B, C) \times \sum_E P(E|D). \end{aligned}$$

Or $\sum_E P(E|D)$ est un vecteur de taille la dimension de D constitué uniquement de 1. On peut donc le supprimer de l'équation (puisque tous les produits sont tensoriels (terme à terme)). On obtient donc :

$$P(A, B, C, D) = P(A) \times P(B|A) \times P(C) \times P(D|B, C).$$

De même, on peut supprimer A et D :

$$\begin{aligned} P(B, C) &= \sum_A \sum_D P(A) \times P(B|A) \times P(C) \times P(D|B, C) \\ &= \sum_A [P(A) \times P(B|A)] \times P(C) \times \sum_D P(D|B, C). \end{aligned}$$

Comme précédemment, $\sum_D P(D|B, C) = 1$. De plus, $P(A) \times P(B|A) = P(A, B)$. Donc $\sum_A [P(A) \times P(B|A)] = \sum_A P(A, B) = P(B)$. On en déduit donc que :

$$P(B, C) = P(B) \times P(C).$$

Donc B et C sont bien des variables indépendantes.