MAPSI – Examen final – 60pts

Durée : 2 heures

Seuls documents autorisés : Calculatrice, antisèche recto-verso
– Barème indicatif –

Exercice 1 (10pts) – **Le verdict des urnes**

Soit deux urnes mystérieuses remplies de boules de couleur dans laquelle nous avons effectué 100 tirages avec remise pour le résultat suivant :

Couleur	R	B	V	J
Nb boules	35	10	32	23

Urne A

Couleur	R	B	V	J	N
Nb boules	28	9	25	33	5

Urne B

Q 1.1 (3pts) Supposons que l'on tire successivement (toujours avec remise) une boule Jaune et une Rouge dans une même urne ; est-il plus probable que ce tirage ait été effectué dans l'urne A ou dans la B sachant qu'un expert nous indique qu'il y a 55% de chance que le tirage vienne de A ? La conclusion aurait-elle été la même sans l'avis de l'expert ? Donner le détail des probabilités à calculer pour répondre.

Q 1.2 (3pts) Le tirage d'une boule Rouge dans l'urne A s'aparente à un processus de Bernoulli. L'expert nous indique que la probabilité de victoire historique est de $p = 32\%$ (variance $p(1-p)$). Peut-on affirmer -avec une confiance de 95%- que le taux de boules Rouges dans l'urne A a augmenté dans la période de notre tirage ?

Q 1.3 (1.5pt) Nous sommes passés précedemment par un célèbre théorème pour vérifier une hypothèse... Mais nous aurions en fait pu directement calculer la probabilité de tirer 35 boules rouges en 100 tirages sous l'hypothèse $p = 32\%$. Poser les calculs (sans faire l'application numérique) et expliquer comment tirer une conclusion.

Q 1.4 (2.5pts) Peut-on affirmer que les deux urnes sont en fait probablement identiques ? Toujours avec une confiance de 95%. Pour des raisons numériques, nous prendrons l'urne B comme distribution théorique et nous testerons si A lui correspond.

Exercice 2 (10pts) – **ML et MAP**

Soit X une variable aléatoire, avec une fonction de densité suivante :

$$f_X(x) = \begin{cases} \frac{1}{Z}\frac{2}{\theta^2}(\theta - x) & \text{pour } x \in [0,1] \\ 0 & \text{sinon} \end{cases}$$

où Z est une constante de normalisation pour que f_X soit bien homogène à une fonction de densité. Deux échantillons sont extraits : $x_1 = 0.5$, $x_2 = 0$.

Q 2.1 (2pts) Donner l'expression de Z pour que f_X soit une fonction de densité (sans la calculer).

Q 2.2 (4pts) Estimation par ML Estimez le paramètre θ par maximum de vraisemblance.

Q 2.3 (3pts) Estimation par MAP Estimez le paramètre par la méthode de maximum a posteriori (MAP) en sachant que θ est issue d'une variable aléatoire Θ suivant une loi avec densité :

$$\pi(\theta) = \begin{cases} \frac{1}{8}\theta & \text{pour } 0 < \theta \leq 16 \\ 0 & \text{sinon} \end{cases}$$

Q 2.4 (1pt) A priori, quelle est la valeur la plus vraisemblable pour θ ? De ce fait, la valeur optimale obtenue en MAP est-elle cohérente avec la valeur obtenue en MV ?

Exercice 3 (22pts) – Analyse des logs d'un système de transport en commun

Nous disposons de données de log des usagers issues d'un système de transport en commun (type RATP). Les données sont collectées par jour et par station et agrégées par quart d'heure. Tous ces comptages sont stockés dans une matrice 3D :

$$X \in \mathbb{N}^{S \times J \times H}, \qquad x_{s,j,h} \text{ donne le nombre de validations à la station s, le jour j pour le quart d'heure h}$$

Q 3.1 (3.5pts) On s'intéresse à la variable aléatoire $X_{s,h}$ donnant le nombre de validations pour une station s et un créneau horaire h. On a choisi un modèle de Poisson : $p(X_{s,h} = k) = \frac{1}{k!}e^{-\lambda}\lambda^k$. Par rapport aux données à disposition, indiquer l'hypothèse nécessaire pour formuler simplement le calcul de la vraisemblance, poser le problème d'optimisation et le résoudre pour identifier les paramètres optimaux de loi.
La page wikipedia de la loi de Poisson indique que l'espérance de la loi est λ ; cela est-il cohérent avec votre calcul ?

Q 3.2 (1pt) Combien de paramètres faut-il optimiser ? Nous utiliserons des paramètres avec des indices explicites dans la suite de l'exercice.

Q 3.3 Détection et caractérisation des anomalies au niveau du réseau.

[Les questions suivantes contiennent essentiellement de la modélisation et du protocole : choix de lois de probabilités, décompte des paramètres, ...]

Q 3.3.1 (1pt) Une *situation* à l'instant h le jour j correspond donc à un vecteur donnant les comptages d'usagers par station : $\mathbf{x}_{h,j} \in \mathbb{R}^S$. En faisant l'hypothèse (forte) d'une indépendance des stations donner le calcul de la vraisemblance d'une *situation*.

Q 3.3.2 (1.5pt) On définit arbitrairement 3 catégories de vraisemblances correspondant respectivement à des situations normales (N), dégradées (D) et bloquées (B). Chaque journée correspond donc à un enchaînement de *situations*. Quel modèle utiliser pour modéliser une journée séquentiellement ? Détailler les paramètres nécessaire à la mise en place du modèle. Rappeler comment sont appris les paramètres d'un tel modèle en pratique.

Q 3.3.3 (1.5pt) Un point critique dans les réseaux de transport concerne la prédiction de retour à la normale quand la situation est bloquée. Nous voulons trouver une loi de probabilité dont l'espérance correspondra au temps moyen nécessaire pour un retour à la normale : que proposez-vous (en détail) ?

Q 3.4 On s'intéresse maintenant au même problème, mais sous un angle assez différent. Un agent de l'autorité de transport a effectué un étiquetage des *situations* à la main : pour chaque pas de temps, il a donc associé une lettre N, D ou B pour l'ensemble du réseau. Par ailleurs, vous disposez toujours de votre calcul de la vraisemblance qui, à partir des logs observés, vous donne un réel associé à ce même pas de temps. Nous souhaitons développer un modèle statistique modélisant ces données et cet étiquetage, avec l'idée de pouvoir plus efficacement étiqueter les données de comptage du futur.

Q 3.4.1 (1.5pt) Quel modèle permet de modéliser séquentiellement ces couples (vraisemblance des observations, étiquetage) ? Détailler les paramètres de ce modèle et leur nombre.

Q 3.4.2 (1pt) A partir des données et des étiquetages manuels, comment apprendre les paramètres optimaux du modèle ?

Q 3.4.3 (1pt) Détailler le protocole et les algorithmes à utiliser pour prédire l'état du réseau sur une journée future j où seules la matrice des comptages $X_j \in \mathbb{R}^{S \times H}$ est disponible.

Q 3.4.4 (1pt) Le modèle développé dans cette question (Q3.4) vous semble-t-il plus ou moins intéressant/performant que celui de la question (Q3.3) ? Pourquoi (en une phrase) ?

Q 3.5 Implémentation. Pour les questions de code, on propose de revenir à la première partie de l'énoncé (Q3.1 & Q3.3). Les seules informations disponibles en entrée sont des données de comptage :

```
import numpy as np
import math
X # np.array compose des donnees de validations
print(X[s,j,h]) # nb validations: station s, jour j, horaire h
S, J, H = X.shape # recuperation des dimensions
# fonctions utiles (et nouvelles):
# fact = np.vectorize(math.factorial) # factorielle sur un vecteur
# use: fact(np.array([2, 3, 4])
```

Q 3.5.1 (1pt) Donner les **signatures des méthodes** permettant : (1) d'optimiser les lois de Poisson, (2) calculer la log-vraisemblance d'une situation, (3) calculer la log-vraisemblance d'une journée.
Note : même si les calculs sont triviaux, ils seront obligatoirement inclus dans des méthodes.
Note 2 : le but de l'exercice est de NE PAS utiliser de variable globale. Vos signatures incluront toutes les informations nécessaire pour le calcul.
Note 3 : vous pouvez répondre à la question en même temps.
Note 4 : donner en commentaires les dimensions des structures de données les plus importantes (entrées/sorties des méthodes).

Q 3.5.2 (4pts) Donner les codes associées à ces méthodes.

Q 3.5.3 (2pts) Donner la méthode permettant de passer de `X` à une séquence d'états à partir de deux seuils de log-vraisemblance fournis.
Note : chaque journée correspondra à une séquence.

Q 3.5.4 (2pts) Donner le code permettant d'évaluer quantitativement les résultats par rapport à la vérité terrain fournie par l'agent de la Q3.4. Vous calculerez le taux de bonne classification et la matrice de confusion.

Exercice 4 (18pts) – HMM

Soit un HMM dont l'état est représenté par les variables V_t qui peuvent prendre k valeurs, et l'observation est représentée par la variable O_t qui peut prendre m valeurs différentes.

Q 4.1 (2pts) Donner le nombre de paramètres de ce modèle.

Q 4.2 (4pts) Les propositions sont-elles vraies (justifier vos réponses) :

- Connaître la valeur de O_1 n'a aucune influence sur la distribution de O_3.
- Connaître la valeur de V_1 n'a aucune influence sur la distribution de O_3.
- Connaître la valeur de O_1 n'a aucune influence sur la distribution de O_3 si on connaît déjà la valeur de O_2.
- Connaître la valeur de O_1 n'a aucune influence sur la distribution de O_3 si on connaît déjà la valeur de V_2.

Q 4.3 Avec les données suivante :

1. $k = 2$: 2 états A et B,
2. $m = 2$: 2 observations 0 et 1
3. $P(V_0 = A) = 0.99$
4. $\forall t > 0, P(V_t = x | V_{t-1} = x) = 0.99$
5. $\forall t \geq 0, P(O_t = 0 | V_t = A) = 0.8$ et $P(O_t = 1 | V_t = B) = 0.9$

Q 4.3.1 (2pts) Compléter l'ensemble des paramètres du HMM. À votre avis (c'est-à-dire sans calcul), quelle est la séquence d'observation (O_0, O_1, O_2) la plus probable ?

Q 4.3.2 (2pts) Calculer la probabilité de la séquence d'observations $(0, 1, 0)$

Q 4.3.3 (2pts) Calculer la séquence d'états la plus probable, d'après cette même séquence d'observations $(0, 1, 0)$.

Q 4.4 Chaîne de Markov des états

Q 4.4.1 (2pts) En ne considérant que la chaîne de Markov des états (V_t), montrer que cette chaîne est ergodique et calculer la distribution π^* vers laquelle converge $P(V_t)$.

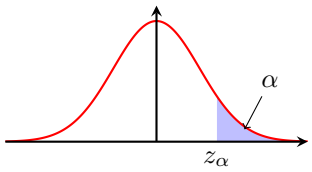
Q 4.4.2 (2pts) En déduire, pour chaque observation possible (0 ou 1) de la variable O_t, l'état le plus probable (la valeur la plus probable de V_t) en régime permanent.

Q 4.4.3 (2pts) Comparer la séquence d'états la plus probable étant donnée la séquence d'observation $(0, 1, 0)$ et la séquence des états les plus probables en régime permanent pour chaque observation de la séquence. Comment expliquez-vous ce résultat ?

Extrait de la table de la loi normale

Dans le tableau ci-contre

$P(Z > z_\alpha) = \alpha$ avec $Z \sim \mathcal{N}(0,1)$



z_α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2297	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0859	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0466	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233

Extraut de la table du χ^2

La table ci dessous donne la valeur de seuil $c_{r,\alpha}$ telle que $P(Z \geq c_{r,\alpha}) = \alpha$ avec $Z \sim \chi^2_{(r)}$ une variable aléatoire suivant un χ^2 à r degrés de libertés.

$n \setminus \alpha$	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
1	0,0000393	0,000157	0,000982	0,00393	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21	10,6
3	0,0717	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,3	12,8
4	0,207	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3	14,9
5	0,412	0,554	0,831	1,15	1,61	9,24	11,1	12,8	15,1	16,7
6	0,676	0,872	1,24	1,64	2,20	10,6	12,6	14,4	16,8	18,5
7	0,989	1,24	1,69	2,17	2,83	12,0	14,1	16,0	18,5	20,3
8	1,34	1,65	2,18	2,73	3,49	13,4	15,5	17,5	20,1	22,0
9	1,73	2,09	2,70	3,33	4,17	14,7	16,9	19,0	21,7	23,6
10	2,16	2,56	3,25	3,94	4,87	16,0	18,3	20,5	23,2	25,2
11	2,60	3,05	3,82	4,57	5,58	17,3	19,7	21,9	24,7	26,8
12	3,07	3,57	4,40	5,23	6,30	18,5	21,0	23,3	26,2	28,3
13	3,57	4,11	5,01	5,89	7,04	19,8	22,4	24,7	27,7	29,8
14	4,07	4,66	5,63	6,57	7,79	21,1	23,7	26,1	29,1	31,3
15	4,60	5,23	6,26	7,26	8,55	22,3	25,0	27,5	30,6	32,8