Réparti 1

Durée : 2H

Seuls documents autorisés : antisèche recto-verso, calculatrice
– Barême indicatif –

Exercice 1 (6 pts) – Modélisation d'auteurs

Nous nous intéressons à la modélisation de 5 auteurs célèbres. Des experts nous ont donné 100 mots par auteur qui les caractérisent. Au total, cela nous donne un dictionnaire de 350 mots d'intérêt sur lesquels nous allons travailler. Nous disposons aussi d'une base de 20 textes par auteur, cette base est pré-traitée de sorte à ne conserver que les 350 mots qui nous intéressent. Le texte le plus long est composé de 180 mots. Pour simplifier la modélisation, nous pourrons faire l'hypothèse que tous les textes font 180 mots (les textes plus courts étant complétés par des mots fantômes). Nous utiliserons les notations suivantes : le i^e texte de la base, $\mathbf{t}_i = \{w_1, \ldots, w_{|\mathbf{t}_i|}\}$, est un ensemble de mot w. Le vecteur $\mathbf{y} = \{y_1, \ldots, y_{100}\}, y_i \in \{1, \ldots, 5\}$ indique l'auteur de chaque texte $\mathbf{t}_i$. Comme dans l'exercice vu en TD, nous allons faire l'hypothèse (très naïve) d'une indépendance des mots dans tous les textes.

Q 1.1 (0.5 pt) Pour rappel, lorsque une variable Z suit une loi binomiale $\mathcal{B}(p, n)$, nous avons $\mathbb{P}(Z = k) = \binom{n}{k} p^k (1-p)^{n-k}$. Selon la loi binomiale, donner la probabilité d'observer k fois le mot w_j dans un document. Bien distinguer les paramètres que nous allons apprendre dans la suite de l'exercice et les constantes dont vous indiquerez les valeurs.

Q 1.2 (0.5 pt) Pour travailler plus simplement, nous allons mettre en forme les données textuelles dans une matrice X, où chaque texte correspondra à une ligne. Indiquer le contenu et les dimensions de cette matrice. Dans la suite, un texte i sera une ligne de X que nous noterons $\mathbf{x}_i$.

Q 1.3 (1 pt) Donner la probabilité d'observation d'un texte en fonction des x_{ij} (issus de X) et des paramètres des binomiales des différents mots.

Q 1.4 (2 pts) Méthodologie générale. Comment procéder pour construire un classifieur d'auteurs ? Indiquer rapidement les grandes étapes, en particulier : expliciter la log-vraisemblance, donner la formulation du problème d'optimisation à résoudre et donner le nombre total de paramètres à optimiser. Indiquer également rapidement comment procéder pour l'inférence sur de nouveaux textes.

Q 1.5 (2 pts) Résoudre le problème du maximum de vraisemblance et exprimer les paramètres optimaux en fonction des x_{ij} et des paramètres des binomiales. La solution est finalement assez intuitive : à quoi correspond-elle ?

Exercice 2 (4.5 pts) **– Auteurs... Anonymes**

[le cadre et les notations sont les mêmes que dans l'exercice précédent, mais les questions sont indépendantes]
Dans le cas où nous ne disposons plus des données de labélisation d'auteurs $\mathbf{y}$, nous voudrions tout de même réussir à regrouper ensemble les textes qui se ressemblent. L'idée serait d'initialiser 5 modèles et de procéder à une optimisation itérative avec un algorithme EM. Les paramètres p_j associé au kème modèle seront notés $p_{j,k}$.

Q 2.1 (1 pt) Proposer une initialisation rapide et grossière pour 5 modèles toujours basés sur des binomiales en tirant parti des informations des experts.

Q 2.2 (1 pt) Dans l'algorithme EM, la première étape consiste à estimer les probabilités des valeurs cachées sachant les valeurs observées (E). En faisant l'hypothèse que les 5 classes de textes sont équiprobables ($\frac{1}{5}$), donner la formule permettant de calculer $Q_i(k) = p(y_i = k|\mathbf{x}_i)$ pour un texte i.

Q 2.3 (1 pt) Donner la formule de la log-vraisemblance par rapport aux $Q_i(k)$, aux observations x_{ij} et aux paramètres des modèles $p_{j,k}$. En quoi consiste l'étape M de l'algorithme ? Expliquer le principe sans calculer.

Q 2.4 (1.5 pt) Résoudre l'étape M et exprimer les paramètres optimaux en fonction des x_{ij} et des $Q_i(k)$. Interpréter la solution obtenue.

Exercice 3 (4pts) **– indépendances et calcul**

Dans cet exercice, on étudie des distributions jointes de 3 variables aléatoires binaires A, B, C.

Q 3.1 (2pts) Dans la distribution du tableau ci-dessous, vérifier les indépendances par paires et l'indépendance mutuelle *des variables* A, B, C.

$P(A, B, C)$

		C	
B	A	0	1
0	0	0.25	0.00
	1	0.00	0.25
1	0	0.00	0.25
	1	0.25	0.00

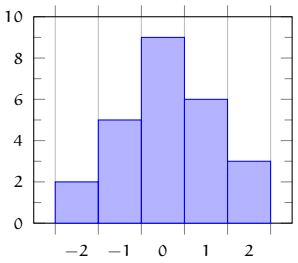
Q 3.2 (2pts) Dans la distribution du tableau ci-dessous, en considérant uniquement les évènements A = 1, B = 1, C = 1, vérifier la propriété d'indépendances 2 à 2 puis sur les 3 évènements et conclure sur l'indépendance mutuelle de ces 3 évènements.

$P(A, B, C)$

		C	
B	A	0	1
0	0	0.34	0.00
	1	0.10	0.06
1	0	0.16	0.10
	1	0.20	0.04

Exercice 4 (6pts) – Comptage, ajustement et vraisemblance

Soit un tableau de contingence d'un échantillon iid d'une variable X, distribuée suivant 5 classes $C_{(i)}, i \in \{-2,-1,0,1,2\}$.

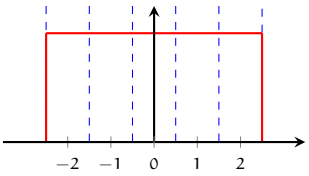


Classes	$C_{(-2)}$	$C_{(-1)}$	$C_{(0)}$	$C_{(1)}$	$C_{(2)}$	
Effectifs	2	5	9	6	3	25

Note : dans ce tableau, les effectifs ont été choisis faibles afin de simplifier les calculs. On ne tiendra donc pas compte des problèmes de taille limite pour l'utilisation des tests statistiques.

Q 4.1 (2pts) Modèle uniforme

Dans cette question, on étudie une distribution de X uniforme sur les 5 classes.

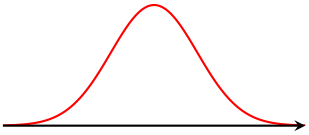


Q 4.1.1 Si la distribution est uniforme sur les 5 classes, quelle est la probabilité $P(X \in C_{(i)})$

Q 4.1.2 Peut-on estimer au seuil de 5% si cette hypothèse est acceptable ?

Q 4.2 (1pt) Modèle gaussien

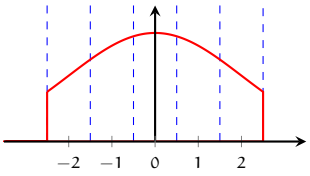
Dans cette question et la suivante, on étudie une distribution gaussienne pour la variable continue X. Pour se faire, on précise qu'une classe $C_{(i)}$ représente l'intervalle $[i-0.5, i+0.5]$.



En considérant que tout élément de la classe $C_{(i)}$ a pour valeur i, calculer le meilleur estimateur de la moyenne μ et de la variance σ^2.

Q 4.3 (3pts) Modèle gaussien tronqué

Dans la suite, on considère que X suit une loi normale de moyenne 0 et d'écart-type 2, *tronquée dans l'intervalle* $[-2.5, 2.5]$:



$$X \in [-2.5, 2.5] \sim \mathcal{N}_{[-2.5,2.5]}(\mu = 0, \sigma^2 = 4)$$

Q 4.3.1 Pour une variable $Y \sim \mathcal{N}(\mu = 0, \sigma^2 = 4)$ (*non tronquée*), calculer les probabilités $P(Y \in C_{(i)}), \forall i \in \{-2,-1,0,1,2\}$.

Petite aide : il n'y a que 3 valeurs à chercher dans la table de la loi normale centrée réduite pour répondre à cette question.

Q 4.3.2 En déduire les probabilité $P(X \in C_{(i)}), \forall i \in \{-2,-1,0,1,2\}$ pour la distribution gaussienne tronquée $\mathcal{N}_{[-2.5,2.5]}(\mu = 0, \sigma^2 = 4)$.

Q 4.3.3 Peut-on estimer au seuil de 5% que la distribution de cet échantillon est la distribution gaussienne tronquée $\mathcal{N}_{[-2.5,2.5]}(\mu = 0, \sigma^2 = 4)$?

Q 4.4 (1pt) Sélection de modèle

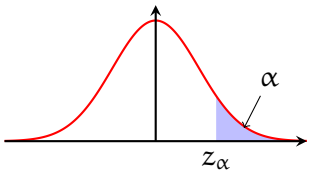
Q 4.4.1 D'après les résultats des questions précédentes, quel modèle vous semble le plus acceptable pour X entre la distribution uniforme et la distribution gaussienne tronquée $\mathcal{N}_{[-2.5,2.5]}(\mu = 0, \sigma^2 = 4)$?

Q 4.4.2 En utilisant un critère de vraisemblance de l'échantillon de taille 25, quel modèle vous semble le plus acceptable pour X entre la distribution uniforme et la distribution gaussienne tronquée $\mathcal{N}_{[-2.5,2.5]}(\mu = 0, \sigma^2 = 4)$?

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

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