Théorie de l'information

Devoir domestique n°1

Exercice 1: Choix d'une distribution dans un modèle statistique paramétrique

Voici une distribution observée Q et trois distributions issues de lois Binomiales $P(\lambda)$.

Distribution / valeurs	0	1	2	3	4
Q	0,28	0,43	0,21	0,07	0,01
$B(4, p)$; $p=0.2$	0,4096	0,4096	0,1536	0,0256	0,0016
$B(4, p)$; $p=0.25$	0,3164063	0,421875	0,2109375	0,046875	0,00390625
$B(4, p)$; $p=0.3$	0,2401	0,4116	0,2646	0,0756	0,0081
$B(4, p)$; $p=0.35$	0,1785063	0,384475	0,3105375	0,111475	0,01500625

1. Quelle est celle qui approche le mieux Q au sens de Kullback-Leibler?

   Indication: calculez $K(Q, P_\lambda)$ *dans chaque cas.*

2. Même question au sens du χ^2 .

   Indication: $d_Q^{\chi^2}(Q, P_\lambda)=\sum_j \frac{(Q(j)-P_\lambda(j))^2}{Q(j)}$

Exercice 2: Segmentation

On considère le tableau ci-dessous, répartissant la population active occupée selon l'âge (A), le sexe (S) et la catégorie socioprofessionnelle (C) (source: INSEE, enquête emploi 2016):

1. Quelle variable auriez-vous envie de modéliser avec (= conditionner par) les deux autres?

   Indication: Après l'avoir justifié, calculez les informations mutuelles suivantes: $I(C,(A \times S))$; $I(S,(A \times C))$; $I(A,(C \times S))$. *Quelle est la plus grande? Conclusion?*

2. Calculer l'arbre de segmentation binaire le moins mauvais possible utilisant ces variables. Le résultat semble-t-il avoir un rapport avec celui de la question précédente?

   Indication: Calculez les informations mutuelles de chaque couple de variables. Pour chaque variable, faire la somme de ses informations mutuelles avec chacune des autres. Commencer l'arbre de segmentation par la variable ayant la somme la plus grande. Dans chaque classe de recette obtenue dans cette première partition, le choix de la seconde variable de segmentation a-t-il une importance?

Catégorie socioprofessionnelle des actifs occupés selon le sexe et l'âge

Âge	**De 15 à 29 ans**	***De 30 à 49 ans***	**De 30 à 39 ans**	**De 40 à 49 ans**	**De 50 à 59 ans**	**60 ans ou plus**
	Effectifs (en milliers)	Effectifs (en milliers)	Effectifs (en milliers)	Effectifs (en milliers)	Effectifs (en milliers)	Effectifs (en milliers)
SEXE : Femmes						
Agriculteurs	27,8	189,7	70,0	119,6	187,1	76,9
Artisans, com	117,4	914,0	357,9	556,1	525,8	184,8
Cadres et pro	564,9	2 638,5	1 209,0	1 429,5	1 161,6	360,0
Professions i	1 353,7	3 735,7	1 840,7	1 895,0	1 507,8	256,2
Employés	1 570,9	3 486,4	1 605,6	1 880,9	1 819,6	397,0
Ouvriers	1 271,6	2 648,6	1 285,9	1 362,7	1 300,4	180,5
SEXE : Hommes						
Agriculteurs	24,2	146,0	56,2	89,8	138,0	43,4
Artisans, com	79,2	645,6	258,4	387,2	378,7	128,7
Cadres et pro	315,7	1 538,5	685,3	853,2	719,0	240,1
Professions i	613,3	1 750,4	834,7	915,7	755,9	123,7
Employés	476,0	865,5	449,5	416,0	329,8	58,3
Ouvriers	1 085,8	2 133,9	1 068,4	1 065,5	987,9	130,1

Exercice 3: Recodage avec nombre de classes fixé.

Une petite enquête a fourni la distribution jointe suivante (exprimée en %) pour les deux variables:

- X = réponse à la question: "Quelle est la part de vos revenus que vous donnez à des associations d'utilité publique?"
- Y = réponse à la question: "Avez-vous voté lors des dernières élections?"

X / Y	0%	Entre 0 et 0.5%	Entre 0.5 et 1%	Entre 1 et 3%	Plus de 3%
Oui	2	6	8	5	1
Non	27	10	8	5	0

1 - Quel sont les meilleurs recodages en deux, puis en trois classes, de X ?

Indication: on ne peut agréger que des classes contiguës. Il y a donc 4 recodages binaires possibles de X. On calculera l'entropie de chacune des distributions binaires de X obtenues, et on choisira le recodage d'entropie maximale .

2 - Quel est le meilleur recodage de X en deux classes pour la prédiction de Y ?

Indication: on calculera l'information mutuelle de Y avec chacune des distributions binaires de X obtenues précédemment, et on choisira le recodage donnant l'information mutuelle maximale (pourquoi?)