Théorie de l'information

Devoir domestique n°1

TD 2, exercice 1: Choix d'une distribution dans un modèle statistique paramétrique

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

Distribution / valeurs	0	1	2	3	≥ 4
Q	0.31	0.32	0.19	0.08	0.1
$P_\lambda ; \lambda = 1$	.368	.368	.184	.061	.019
$P_\lambda ; \lambda = 1.25$	.287	.358	.224	.093	.038
$P_\lambda ; \lambda = 1.5$	.223	.335	.251	.126	.066

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

TD 2, exercice 2: Segmentation

On considère le tableau ci-dessous, répartissant la population active occupée selon le sous-emploi (SE), le sexe (S) et le diplôme (D) (source: INSEE, enquête emploi 2016):

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

   Indication: Calculez les informations mutuelles suivantes: I(SE,(S × D)); I(S,(SE × D)); I(D,(SE × 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?

	INDICATEURS	Population en sous-emploi (en milliers)	Population non en sous-emploi (en milliers)
Diplôme le plus élevé obtenu	*SEXE*		
Ensemble	***Ensemble***	1,723.6	24,793.3
	Femmes	1,208.5	11,647.9
	Hommes	515.0	13,403.9
Non déclaré	***Ensemble***	7.0	73.5
	Femmes	4.3	32.5
	Hommes	2.7	40.8
Diplôme supérieur	***Ensemble***	238.1	6,027.7
	Femmes	164.8	3,066.6
	Hommes	73.3	2,980.9
BTS, DUT ou autre diplôme de niveau bac + 2	***Ensemble***	186.5	4,150.7
	Femmes	130.1	2,235.4
	Hommes	56.4	1,888.4
Baccalauréat ou brevet professionnel ou autre diplôme de ce niveau	***Ensemble***	430.0	4,945.0
	Femmes	310.1	2,317.9
	Hommes	119.9	2,605.1
CAP, BEP ou autre diplôme de ce niveau	***Ensemble***	445.6	6,012.4
	Femmes	317.8	2,398.4
	Hommes	127.8	3,631.0
Brevet des collèges	***Ensemble***	107.4	1,171.2
	Femmes	73.3	542.7
	Hommes	34.1	621.7
Aucun diplôme ou CEP	***Ensemble***	309.0	2,578.9
	Femmes	208.2	1,053.6
	Hommes	100.8	1,525.0

TD 2, exercice 6: 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: "utilisez-vous une application smartphone pour réserver un billet de train?"
- Y = réponse à la question: "Avez-vous changé de smartphone cette dernière année?"

Y \ X	jamais	rarement	moyennement souvent	souvent	toujours
Oui	0	7	4	11	15
Non	8	33	10	11	1

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?)