TP1 TID

SCAIA Matteo, MARIAC Damien

October 20, 2024

Contents

1 Exercice 1

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).

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, con	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, con	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

Figure 1: Tableau répartissant la population active occupée selon des catégories

1.1 Modélisation et variable

Tout d'abord de manière intuitive, nous avons envie de modéliser la variable socioprofessionnelle avec les deux autres. Cependant, nous devons le montrer de manière formelle. Grâce au code fourni dans la partie 4, nous calculons l'information mutuelle de chacune des variables.

Premièrement, nous calculons l'entropie de chacune de ces variables.
Pour la variable A, nous avons le tableau suivant (en fréquence).

	15-29 ans	30-39 ans	40-49 ans	50-59 ans	60+ ans
Age	0.1866	0.2419	0.2730	0.2441	0.0542

Table 1: Distribution par âge (A)

Nous pouvons calculer l'entropie de A.

$$H(A) = -\sum_{n=1}^{6} p_i \log_2(p_i) = 2,1833$$

De même manière, nous calculons l'entropie de C et S.

	Femme	Homme
Proportion	0,6589	0,3411

Table 2: Distribution par sexe (S)

	Agriculteur	Artisans	Cadres	Profession In	Employes	Ouvrier
Proportion	0,0207	0,0740	0,1875	0,2512	0,2240	0,2423

Table 3: Distribution par catégorie socioprofessionnelle (C)

Nous obtenons.

$$H(S) = 0,9258 \qquad\qquad H(C) = 2,3266$$

A présent, nous devons calculer les valeurs suivantes : $H(A,S)$, $H(A,C)$ et $H(S,C)$.

	15-29 ans	30-39 ans	40-49 ans	50-59 ans	60+
Femme	0,1220	0,1584	0,1803	0,1618	0,0362
Homme	0,0645	0,0834	0,0927	0,0823	0,1802

Table 4: Distribution jointe sexe (S) et âge (A)

	15-29 ans	30-39 ans	40-49 ans	50-59 ans	60+
Agriculteur	0,0013	0,0031	0,0052	0,0081	0,0030
Artisans	0,0049	0,0153	0,0234	0,0225	0,0080
Cadres	0,0219	0,0471	0,0568	0,0468	0,0149
Professions In	0,0489	0,0665	0,0699	0,0563	0,0094
Employes	0,0509	0,0511	0,0571	0,0535	0,0113
Ouvrier	0,0586	0,0586	0,0604	0,0569	0,0077

Table 5: Distribution jointe (C) et âge (A)

	Agriculteur	Artisans	Cadres	Profession IN	Employes	Ouvrier
Femmes	0.0119	0.0433	0,1175	0.1705	0.1810	0.1344
Homme	0.0087	0.0307	0,0700	0.0807	0.0430	0.1079

Table 6: Distribution jointe (C) et (S)

Nous obtenons les valeurs suivantes.

$$H(A,S) = 3,1092 \qquad H(A,C) = 4,4817 \qquad H(C,S) = 3,2242$$

De plus, nous obtenons pour $H(A,S,C)$ la valeur suivante.

$$H(A,C,S) = -\sum_{n=1}^{72} p_i \log_2(p_i) = 5,3778$$

Nous pouvons calculer les informations mutuelles.

$$I(C,(AS)) = H(C) + H(A,S) - H(A,S,C) = 0,0580$$

$$I(A,(SC)) = 0,029803$$

$$I(S,(CA)) = 0,029813$$

Cherchons le rapport entre l'information mutuelle et la variable conditionnée le plus élevé.

$$R_1 = \frac{I(C,(AS))}{H(A,C)} = 0,0187$$

$$R_2 = \frac{I(A,(SC)))}{H(S,C)} = 0,0092$$

$$R_3 = \frac{I(S,(CA)))}{H(A,C)} = 0,0066$$

Le rapport R_1 est le plus élevé. Donc c'est la variable C modélisé par les deux autres (A et S) qui nous donne le plus d'information.

1.2 Arbre de segmentation binaire

Grâce à la partie 1.1, nous pouvons calculer les informations mutuelles suivantes.

$$I(A,S) = H(A) + H(S) - H(A,S) = 5,1104 * 10^{-5}$$

$$I(A,C) = 0,02830$$

$$I(S,C) = 0,02831$$

Faisons, la somme de ses informations mutuelles avec chacune des deux autres.

$$\bar{I_A} = I(A,S) + I(A,C) = 0,02835$$

$$\bar{I_S} = 0,02836$$

$$\bar{I_C} = 0,0566$$

Nous avons donc $\bar{I_C}$ qui est le plus élevé. Ainsi, l'arbre de segmentation commencera avec la variable C.

2 Exercice 2

2.1 Recodages

2.1.1 Recodage en 2 variables

En agrégeant seulement les classes contigues, nous avons 4 possibilités de regroupement binaire de X.

$$Z1 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%, 1 - 3\%, > 3\%\}\}$$
$$Z1 = \{\{0\%, 0 - 0.5\%\}, \{0.5 - 1\%, 1 - 3\%, > 3\%\}\}$$
$$Z1 = \{\{0\%, 0 - 0.5\%, 0.5 - 1\%\}, \{1 - 3\%, > 3\%\}\}$$
$$Z1 = \{\{0\%, 0 - 0.5\%, 0.5 - 1\%, 1 - 3\%\}, \{> 3\%\}\}$$

L'entropie de chacun de ses recodages se calcule numeriquement et donne :

$$H(Z1) = -(29/72 * log2(29/72) + 43/72 * log2(43/72)) = 0.973$$
$$H(Z2) = -(29/72 * log2(29/72) + 43/72 * log2(43/72)) = 0.954$$
$$H(Z3) = -(29/72 * log2(29/72) + 43/72 * log2(43/72)) = 0.617$$
$$H(Z4) = -(29/72 * log2(29/72) + 43/72 * log2(43/72)) = 0.106$$

Le meilleur recodage est donc le premiers c'est à dire : $Z1 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%, 1 - 3\%, > 3\%\}\}$

2.1.2 Recodage en 3 variables

En procédant de la meme facon, on considere alors 6 cas :

$$Z1 = \{\{0\%\}, \{0 - 0.5\%\}, \{0.5 - 1\%, 1 - 3\%, > 3\%\}\}$$
$$Z2 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%\}, \{1 - 3\%, > 3\%\}\}$$
$$Z3 = \{\{0\%\}, \{0 - 0.5\%, 0.5 - 1\%, 1 - 3\%\}, \{> 3\%\}\}$$
$$Z4 = \{\{0\%, 0 - 0.5\%\}, \{0.5 - 1\%, 1 - 3\%\}, \{> 3\%\}\}$$
$$Z5 = \{\{0\%, 0 - 0.5\%\}, \{, 0.5 - 1\%\}, \{1 - 3\%, > 3\%\}\}$$
$$Z6 = \{\{0\%, 0 - 0.5\%, 0.5 - 1\%\}, \{1 - 3\%\}, \{> 3\%\}\}$$

L'entropie est calculé numériquement :

$$H(Z1) = 1.541$$

$$H(Z2) = 1.462$$
$$H(Z3) = 1.068$$
$$H(Z4) = 1.040$$
$$H(Z5) = 1.320$$
$$H(Z6) = 0.684$$

Et nous remarquons que le meilleure recodage en 3 variables est Z1.

2.2 Le meilleur recodage de X pour prédire Y

Il sagit ici de recoder X en réduisant l'incertitude sur Y. Nous devons donc trouver le Z_k qui maximise l'information mutuelle entre Z_k et Y.
C'est à dire, trouvons le recodage Z_k qui maximise :

$$I(Z_k; Y) = \sum_{z \in Z_k} \sum_{y \in Y} p(z, y) \log \left(\frac{p(z, y)}{p(z)p(y)} \right)$$

<u>Détaillons le calcul pour le premier recodage :</u>

$$p(Z1 = 0\%) = \frac{29}{72}, \quad p(Z1 \neq 0\%) = \frac{43}{72}$$
$$p(Y = \text{Oui}) = \frac{29}{72}, \quad p(Y = \text{Non}) = \frac{43}{72}$$

$$p(Z1 = 0\%, Y = \text{Oui}) = \frac{2}{72}, \quad p(Z1 = 0\%, Y = \text{Non}) = \frac{27}{72}$$
$$p(Z1 \neq 0\%, Y = \text{Oui}) = \frac{20}{72}, \quad p(Z1 \neq 0\%, Y = \text{Non}) = \frac{23}{72}$$

$$I(Z1; Y) = p(Z1 = 0\%, Y = \text{Oui}) \log_2 \left(\frac{p(Z1 = 0\%, Y = \text{Oui})}{p(Z1 = 0\%)p(Y = \text{Oui})} \right) + \dots$$
$$+ p(Z1 = 0\%, Y = \text{Non}) \log_2 \left(\frac{p(Z1 = 0\%, Y = \text{Non})}{p(Z1 = 0\%)p(Y = \text{Non})} \right) + \dots$$
$$+ p(Z1 \neq 0\%, Y = \text{Oui}) \log_2 \left(\frac{p(Z1 \neq 0\%, Y = \text{Oui})}{p(Z1 \neq 0\%)p(Y = \text{Oui})} \right) + \dots$$
$$+ p(Z1 \neq 0\%, Y = \text{Non}) \log_2 \left(\frac{p(Z1 \neq 0\%, Y = \text{Non})}{p(Z1 \neq 0\%)p(Y = \text{Non})} \right)$$
$$= 0.176$$

En faisant de meme pour chaque information mutuelle, on trouve :

$$I(Z1, Y) = 0.178$$

$$I(Z2, Y) = 0.091$$

$$I(Z3, Y) = 0.033$$

$$I(Z4, Y) = 0.024$$

L'information mutuelle la plus grande est le recodage Z1. Cela signifie que Z1 est le meilleur recodage qui permet de prédir Y.

3 Exercice 3

3.1 Description

On considere le tableau recensent des especes de champignons suivant 4 variables qualitatifs.

Table 7: Caractéristiques des espèces de champignons

Espèce	Comestible	Chapeau	Tige	Couleur
a	o	a	e	b
b	o	a	e	j
c	o	a	e	b
d	o	pl	f	j
e	o	pl	f	b
f	n	po	f	r
g	n	po	f	j
h	n	po	e	r
i	n	a	f	j
j	n	pl	f	j

On cherche à faire un arbre de discrimination qui permet de predire la comestibilité à partir des autres caractéristique, et qui soit le plus court possible

Pour ce faire, on calcul les informations mutuelles totales entre chaque variable

3.2 Calculs

3.2.1 Détails pour le premier calcul

On considere que la variable comestible est noté X_1 et la variable chapeau X_2. On calcul l'information mutuelle avec $I(X_1, X_2) = H(X_1)+H(X_2)-H(X_1, X_2)$.

Table 8: Croissement entre comestibilité (X1) et forme du chapeau (X2)

X1\X2	Po	a	Pl
O	0	3	2
N	3	1	1

$$I(X1, X2) = H(X1) + H(X2) - H(X1, X2) = 0.277$$

En procédant de même pour chaque cas, on a :

$I(X_1, X_3) = 0.086$; $I(X_1, X_4) = 0.357$; $I(X_2, X_3) = 0.257$; $I(X_2, X_4) = 0.370$; $I(X_3, X_4) = 0.093$;

3.3 Informations mutuelles totales

On calcul l'information mutuelles totales pour une variable X_k comme la somme des informations mutuelles avec les autres variables. C'est à dire :

$\bar{I_1} = I(X_1, X_2) + I(X_1, X_3) + I(X_1, X_4) = 0.720$
$\bar{I_2} = I(X_1, X_2) + I(X_2, X_3) + I(X_2, X_4) = 0.905$
$\bar{I_3} = I(X_1, X_3) + I(X_2, X_3) + I(X_3, X_4) = 0.437$
$\bar{I_4} = I(X_1, X_4) + I(X_2, X_4) + I(X_3, X_4) = 0.820$

Ainsi, X2 ("forme du chapeau") est la variable qui donne le plus d'information sur les autres.

Nous allons alors analyser les information mutuelles en fonction de leur forme

3.4 Information mutuelle par forme

3.4.1 Chapeau de forme pointu

On remarque qu'aucun des champigons pointu est commestible.(cf tableau)

3.4.2 Chapeau de forme arrondi

Si on considère que les chapeau de forme arrondi, on obtient les tableaux suivant:

Table 9: X1 et X3 avec chapeau arrondi

X1\X3	E	F
O	3	0
N	0	1

Table 10: X1 et X4 avec chapeau arrondi

X1\X4	b	j	r
O	2	1	0
N	0	1	0

Table 11: X3 et X4 avec chapeau arrondi

X3\X4	b	j	r
E	2	1	0
F	0	1	0

Les information mutuelles sont alors :
$I(X_1, X_3) = 0.562$; $I(X_1, X_4) = 0.216$; $I(X_3, X_4) = 0.216$;
On a alors :

3.4.3 Chapeau de forme plat

4 ANNEXE