TP1-TID

Guillaume Bernard-Reymond et Lorenzo Gaggini

October 2023

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

Dans cet exercice, tous les résultats seront arrondis à 10^{-4} près afin de pouvoir comparer les résultats.

1. Pour calculer la divergence de Kullback-Leibler, nous avons utilisé la formule suivante :

$$K(Q, P_\lambda) = -\sum_{i=0}^{4} \ln\left(\frac{P_\lambda(i)}{Q(i)}\right) Q(i)$$

et ce pour chaque valeur de p.
Voici les résultats obtenus :

	$K(Q, P_\lambda)$
$B(4, p); p = 0.2$	$0,0688$
$B(4, p); p = 0.25$	$0,0105$
$B(4, p); p = 0.3$	$0,0100$
$B(4, p); p = 0.35$	$0,0554$

Au sens de Kullback-Leibler, il faudrait choisir la distribution $B(4, 0.25)$ afin d'approcher au mieux Q.

2. Pour approcher Q au sens du χ^2, c'est la formule donnée dans l'énoncé qui a été utilisée :

$$d_Q^{\chi^2}(Q, P_\lambda) = \sum_{i=0}^{4} \frac{(Q(i) - P_\lambda(i))^2}{Q(i)}$$

Voici les résultats obtenus :

	$d_Q^{\chi^2}(Q, P_\lambda)$
$B(4, p); p = 0.2$	$0,1113$
$B(4, p); p = 0.25$	$0,0162$
$B(4, p); p = 0.3$	$0,0215$
$B(4, p); p = 0.35$	$0,1168$

Cette fois-ci, il vaudrait mieux choisir la distribution $B(4; 0.3)$ pour approcher Q.

Exercice 2 : Segmentation

Dans cet exercice, il a d'abord fallu faire un tri des données en supprimant la colonne "De 30 à 49 ans" pour éviter d'avoir des redondances dans nos données. Une fois ce tri fait, nous avons diviser par l'effectif total pour obtenir la loi conjointe de nos trois variables : catégorie socio-professionnelle (C), l'âge (A) et le sexe (S).

Ensuite nous avons sommer certains éléments du tableau de la loi de (C, A, S) pour obtenir les lois de A, de S, de C mais aussi de $(A \times S$, de $A \times C$ et de $C \times S$.

Les résultats seront de nouveau arrondis à 10^{-4} près.

1. Il semble plutôt logique de regarder quelle catégorie socio-professionnelle occupe une personne selon son âge et non autre chose. Si certes, il y une dépendance entre le sexe et, la catégorie professionnelle et l'âge de l'individu, le sexe et l'âge sont des données bien intrinsèques de l'individu et qui ne sont pas emmenés à être modifiée selon la catégorie professionnelle. Vérifions tout ceci par le calcul :

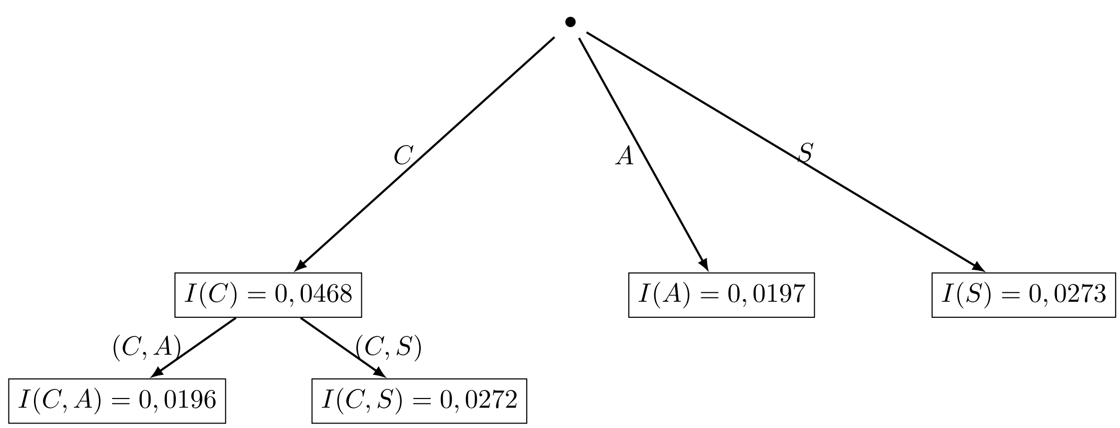
   - $I\left(C,(A \times S)\right) = -\sum_{c \in C} \sum_{(a,s) \in A \times S} P^{(C,A,S)}(c,a,s) \ln\left(\frac{P^{(C,A,S)}(c,a,s)}{P^{C}(c) \times P^{A \times S}(a,s)}\right) \approx 0,0402$
   - $I\left(S,(A \times C)\right) \approx 0,0207$
   - $I\left(A,(C \times S)\right) \approx 0,0207$

   On retrouve donc bien par le calcul notre intuition à savoir que l'information mutuelle des variables C et $A \times S$ est bien supérieure aux autres possibilités.

2. (a) **Calcul des informations mutuelles de chaque couple de variables :**
      - $I(A,C) \approx 0,0196$
      - $I(A,S) \approx 0(3 \times 10^{-5})$
      - $I(S,C) \approx 0,0272$

   (b) **Calcul de l'information pour chaque variable :**
      - $I(S) = I(S,C) + I(S,A) \approx 0,0273$
      - $I(C) = I(C,S) + I(C,A) \approx 0,0468$
      - $I(A) = I(A,S) + I(A,C) \approx 0,0197$

   (c) **Arbre de segmentation**



   Le choix de la seconde variable de segmentation, n'a pas d'importance, car on commençant par C, on a déjà conditionné l'arbre.

Exercice 3 : Recodage avec nombre de classes fixé :

Pour effectuer les différents codages, il nous a fallu dans un premier temps diviser par l'effectif total (72) pour déterminer la loi de (X, Y) puis de sommer certaines valeurs du tableau pour obtenir la loi de Y.

Le codage ne peut se faire qu'en classes contiguës. On notera dans toute la suite :

- $\{a\}$ entre 0% ;
- $\{b\}$ entre 0 et $0,5$% ;
- $\{c\}$ entre $0,5$ et 1% ;
- $\{d\}$ entre 1 et 3% ;
- $\{e\}$ plus de 3%.

Les résultats sont arrondis à 10^{-4} près.

1. (a) **Codage en deux classes :**
   - $\{a\};\{b,c,d,e\} : H(X) = -p_a \ln(p_a) - p_{(b,c,d,e)} \ln(p_{(b,c,d,e)}) \approx 0,6741$
   - $\{a,b\};\{c,d,e\} : H(X) \approx 0,6616$
   - $\{a,b,c\};\{d,e\} : H(X) \approx 0,4275$
   - $\{a,b,c,d\};\{e\} : H(X) \approx 0,0732$

   Pour le recodage de X en deux classes, on choisira le recodage $\{a\};\{b,c,d,e\}$.

   (b) **Codage en trois classes :**
   - $\{a\};\{b\};\{c,d,e\} : H(X) = -p_a \ln(p_a) - p_b \ln(p_b) - p_{(c,d,e)} \ln(p_{(c,d,e)}) \approx 1,0683$
   - $\{a,b\};\{c\};\{d,e\} : H(X) \approx 0,9150$
   - $\{a,b,c\};\{d\};\{e\} : H(X) \approx 0,6060$
   - $\{a\};\{b,c\};\{d,e\} : H(X) \approx 0,9412$
   - $\{a\};\{b,c,d\};\{e\} : H(X) \approx 0,8721$
   - $\{a,b\};\{c,d\};\{e\} : H(X) \approx 0,8529$

   Pour le recodage de X en trois classes, on choisira le recodage $\{a\};\{b\};\{c,d,e\}$.

2. **Recodage en deux classes de X pour la prédiction de Y**

   A finir, c'est du copier coller en dessous :

   - $\{a\};\{b,c,d,e\} : I(X,Y) = -p_a \ln(p_a) - p_{(b,c,d,e)} \ln(p_{(b,c,d,e)}) \approx 0,6741$
   - $\{a,b\};\{c,d,e\} : H(X) \approx 0,6616$
   - $\{a,b,c\};\{d,e\} : H(X) \approx 0,4275$
   - $\{a,b,c,d\};\{e\} : H(X) \approx 0,0732$