VITAMINES A

Théorie de l'information

Le but de la science est de rechercher l'explication la plus simple de faits complexes. Nous sommes susceptibles de commettre l'erreur de penser que les faits sont simples, parce que la simplicité est l'objet de notre quête. La devise de tout physicien devrait être : « Cherche la simplicité et méfie-t'en ».

ALFRED NORTH WHITEHEAD

L'ENTROPIE et l'information mutuelle moyenne sont les deux grandeurs fondamentales de la théorie de l'information. Nous definrons tout d'abord formellement ces deux quantités à partir des distributions de probabilité et nous verrons ensuite comment on les utilise dans un cas concret.

A.1 Préliminaires

Je présume que le lecteur est au moins un familier des notions de base utilisées en probabilité. Néanmoins, à la demande de certains lecteurs, je vais quand même donner un certain nombre de définitions formelles.

A.1.1 Espace de probabilité

- Soit Ω l'ensemble des épreuves ω, içi supposé dénombrable (et souvent fini).
- Soit T la tribu formée des sous ensembles de Ω (appelés événements probabilisables). Elle est souvent égale à l'ensemble $\mathcal{P}(\Omega)$ des parties de Ω. Le couple (Ω, T) est appelé : espace mesurable.
- Alors la mesure de probabilité $P(E)$, ou loi de probabilité (ou encore loi de tirage au sort), définie pour tout événement E de T, lors d'un tirage au sort sur T, est une application de (Ω, T) dans $[0, 1] \subset \boldsymbol{R}$ possédant toute les propriétés habituelles :
  1. positive : $P(E) \geq 0$,
  2. normée : $P(\Omega) = 1$ avec $P(\emptyset) = 0$,
  3. complémentarité : $P(\text{non } E) = P(E^c) = 1 - P(E)$,

4. σ-additive : pour tout ensemble dénombrable $\{E_i\}$ d'événements 2 à 2 disjoints, la mesure de l'union est la somme des mesures : $P(\cup_i E_i) = \sum_i P(E_i)$ si $E_i \cap E_j = \emptyset \quad \forall i, j$. Cette égalité est aussi connue sous le nom d'*axiome d'additivité*

La donnée (Ω, T, P) d'un ensemble Ω d'épreuves, de la tribu T et d'une loi de probabilité P est dite *un espace de probabilité*. [1]

A.1.2 Application mesurable

Soit (Ω, T) et (Ω', T') deux espaces mesurables. Par définition, une application $X(\cdot)$ de (Ω, T) dans (Ω', T') est mesurable si, pour tout élément A' de la tribu T', l'ensemble $X^{-1}(A')$ défini par :

$$X^{-1}(A') := \{\omega \in \Omega \quad : \quad X(\omega) \in A'\} \tag{A.1}$$

appartient à la tribu T.

En d'autres termes, l'ensemble Ω' est muni de la tribu T' qui assure que tout se passe bien, c'est-à-dire que les événements que l'on peut définir à partir de (Ω', T') correspondent en amont à des événements de (Ω, T).

A.1.3 Variable aléatoire

De façon intuitive, une variable aléatoire (en abrégé v.a.) réelle est un nombre réel dépendant du hasard. Cette variable, que nous appellerons X, prend donc ses valeurs dans $\boldsymbol{R}$, cette valeur dépendant d'une expérience aléatoire.

De façon plus formelle, une variable aléatoire X est une *application mesurable* de l'espace (Ω, T) dans l'espace (Ω', T'). Notez que X n'est pas une variable formelle telle que celle qu'on peut rencontrer en Algèbre en manipulant des polynomes par exemple.

Le cas classique, qui est cité dans la plupart des cours sur les probabilités et les variables aléatoires, est celui où (Ω', T') correspond à $(\boldsymbol{R}, \mathcal{B})$ c'est-à-dire à l'ensemble $\boldsymbol{R}$ des nombres réels muni de la tribu de Borel[2] $\mathcal{B}$. Un événement typique de $\mathcal{B}$ est alors un ouvert $]a, b[$, il lui correspond, dans Ω, l'ensemble $\mathcal{E}$ des ω tels que $a < X(\omega) < b$. Cet ensemble $\mathcal{E}$ doit être un événement probabilisable, c'est-à-dire qu'il doit appartenir à T. En simplifiant, on peut désigner cet événement par : "$a < X < b$".

De façon plus pragmatique on dira que X est une variable aléatoire prenant des valeurs x sur un alphabet $\mathcal{A}_X = \{a_1, a_2, \ldots, a_I\}$, avec l'ensemble des probabilités $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$, où $P(x = a_i) = p_i \geq 0$ et $\sum_{a_i \in \mathcal{A}_X} P(X = a_i) = 1$.

Le nombre d'éléments de cet alphabet sera noté $\mid \mathcal{A} \mid$.

A.1.4 Quelques autres définitions et propriétés des probabilités

Un ensemble joint XY, est un ensemble où chaque événement est une paire ordonnée x, y avec $x \in \mathcal{A}_X = \{a_1, a_2, \ldots, a_I\}$ et $y \in \mathcal{A}_Y = \{b_1, b_2, \ldots, b_J\}$.

[1] **Rappels** : soit Ω un ensemble quelconque et F une famille des parties de Ω ($F \subset \mathcal{P}(\Omega)$). Alors, F est une algèbre de Boole ou **clan** si :

- $F \neq \emptyset$
- $\forall\, P \in F, \quad P^c \in F$
- $\forall\, P_i, P_j \in F, \quad P_i \cup P_j \in F$

Si la dernière propriété reste vraie pour tout ensemble dénombrable de P_i, alors F est appelé σ-algèbre de Boole ou **tribu**, c'est-à-dire que l'on a : $\forall P_i (i = 1, 2, \ldots, \infty) \quad P_i \in F, \quad$ et $\quad \cup_{i=1}^{\infty} P_i \in F$

[2] C'est la tribu engendrée par la classe des ouverts. Les éléments de cette tribu sont dits boréliens.

Bien sûr, les deux variables X et Y ne sont pas forcément indépendantes. Dans la notation on pourra supprimer la virgule et écrire xy en lieu et place de x, y.

La probabilité marginale se calcule à partir de la probabilité jointe :

$$P(x = a_i) = \sum_{y \in \mathcal{A}_Y} P(X = a_i, y) \tag{A.2}$$

ou encore, en utilisant une notation abrégée :

$$P(x) = \sum_{y \in \mathcal{A}_Y} P(x, y) \tag{A.3}$$

La probabilité conditionnelle se définit par :

$$P(x = a_i \mid y = b_j) = \frac{P(x = a_i, y = b_j)}{P(y = b_j)} \qquad \text{si} \quad P(y = b_j) \neq 0 \tag{A.4}$$

On préfère souvent écrire la probabilité jointe en fonction des probabilités conditionnelles si bien que l'on obtient (sous couvert de certaines hypothèses $\mathcal{H}$) la règle du produit :

Règle du produit :

$$P(x, y \mid \mathcal{H}) = P(x \mid y, \mathcal{H}) P(y \mid \mathcal{H}) \tag{A.5}$$

Règle de la somme obtenue en écrivant la définition de la probabilité marginale :

$$\begin{aligned} P(x \mid \mathcal{H}) &= \sum_y P(x, y \mid \mathcal{H}) && \text{(A.6)} \\ &= \sum_y P(x \mid y, \mathcal{H}) P(y \mid \mathcal{H}) && \text{(A.7)} \end{aligned}$$

Théorème de Bayes obtenu grâce à la règle du produit :

$$\begin{aligned} P(y \mid x, \mathcal{H}) &= \frac{P(x \mid y, \mathcal{H}) P(y \mid \mathcal{H})}{P(x \mid \mathcal{H})} && \text{(A.8)} \\ &= \frac{P(x \mid y, \mathcal{H}) P(y \mid \mathcal{H})}{\sum_{y'} P(x \mid y', \mathcal{H}) P(y' \mid \mathcal{H})} && \text{(A.9)} \end{aligned}$$

Indépendance : Deux variables aléatoires X et Y sont dites indépendantes si et seulement si :

$$P(x, y) = P(x) P(y) \tag{A.10}$$

A.1.5 Signification de la probabilité

Les probabilités sont utilisées lorsqu'il s'agit de caractériser les fréquences d'occurence de certains événements lors d'expériences où l'incertitude entre en jeu. Désolé, mais il n'y a pas de définition non circulaire de "fréquence".

De manière plus générale, les probabilités sont utilisées pour décrire le degré de croyance qu'il faut accorder à des propositions qui font intervenir des variables aléatoires. Si ces croyances satisfont un certain nombre de règles de cohérence appelées axiomes de Cox, alors on peut leur affecter des probabilités. Les probabilités permettent de décrire des hypothèses ainsi que les déductions qui peuvent être faites à partir de ces hypothèses, si bien que deux personnes ayant les mêmes

hypothèses et munies de données identiques arriveront aux mêmes conclusions. Cette façon de voir les probabilités est le point de vue Bayésien. Beaucoup pensent que c'est une interprétation subjective des probabilités étant donné qu'elles dépendent ainsi d'un certain nombre d'hypothèses. D'un autre côté, nous soutiendrons qu'il est impossible de faire des inférences à partir de données sans avoir d'*a priori* sur celles-ci. Plus loin nous discuterons du théorème NFL (No Free Lunch theorem) qui *démontre* que l'apprentissage ne peut se faire que sous contrainte d'hypothèses. Néanmoins il faut que le lecteur soit prévenu que cette approche Bayésienne des probabilités n'est pas universellement reconnue, le monde des probabilités ayant été dominé depuis toujours par l'idée que celles-ci ne peuvent servir qu'à décrire que des variables aléatoires.

A.1.6 Information associée à un événement

Soit E un événement ($E \subset T$), la quantité d'information associée à la réalisation de l'événement E, ou encore information propre de E est définie et mesurée en *bit* par la quantité réelle ou infinie :

$$h(E) = -\log_2(P(E)) \tag{A.11}$$

Lorsque $P(E) = 0$, on pose $h(E) = \infty$.

Si on utilise un log à base e on a affaire au *neper* ou le *nat*, en base 10, l'unité sera le *hartley*. Dans la suite du texte, sauf cas explicite, la base du logarithme ne sera pas spécifiée.

$h(E)$ mesure "l'information" obtenue lorsque l'événement E s'est réalisé.

Lorsque 2 événements E et F sont indépendants on a : $P(E \cap F) = P(E) \cdot P(F)$, ce qui conduit immédiatement à :

$$h(E \cap F) = h(E) + h(F) \tag{A.12}$$

Dans le cas général on a : $P(E \cap F) = P(E) \cdot P(F \mid E)$, ce qui donne :

$$h(E \cap F) = h(E) + h(F \mid E) \tag{A.13}$$

A.1.7 Information mutuelle associée à deux événements

On appelle information mutuelle de deux événements E et F la quantité :

$$i(E; F) = h(E) + h(F) - h(E \cap F) \tag{A.14}$$

$i(E; F)$ est nulle si E et F sont deux événements indépendants.

D'après A.11 on a :

$$\begin{aligned} i(E; F) &= \log(P(E \cap F)) - \log(P(E) \cdot P(F)) \\ &= \log\left[\frac{P(F \mid E)}{P(F)}\right] \end{aligned} \tag{A.15}$$

ou encore l'expression symétrique obtenue en changeant en E et F.

On dit que $i(E; F)$ mesure l'information que la réalisation d'un événément apporte sur l'autre. Cette grandeur peut être négative.

A.2 Entropie et autres grandeurs liées à la quantité d'information

Définition de l'entropie :

Soit X une v.a. à valeurs dans un ensemble $\mathcal{A}_X$ fini ou dénombrable, c'est-à-dire, une application d'un espace probabilisé (Ω, T, P) dans $\mathcal{A}_X$. On appelle entropie[3] de la v.a. X la quantité :

$$H(X) = -\sum_{x \in \mathcal{A}_X} P(x) \log(P(x)) \tag{A.16}$$

où $P(x)$ est la probabilité de l'événement "$X = x$".

Soit alors $x_1, \ldots, x_n$ les n éléments de $\mathcal{A}_X$ et p_i la probabilité de l'événement "$X = x_i$". On aura donc :

$$H(X) = -\sum_{i=1}^{n} p_i \log(p_i) = H(p_1, \ldots, p_n) \tag{A.17}$$

$H(X)$ ne dépend que de la loi de X, c'est-à-dire de la distribution de probabilité $p = \{p_1, \ldots, p_n\}$ et non des valeurs (x_i) effectivement prises par la v.a. X.

Il ne s'agit donc pas d'une fonction de la v.a. X. La notation $H(X)$ doit se lire "entropie de la distibution de probabilité de la v.a. X".

L'entropie $H(X)$ est la quantité d'information qu'apporte, en moyenne, une réalisation de X. Ou encore : c'est une mesure de l'incertitude (l'aléa) attachée au phénomène aléatoire modélisé par X.

Concrètement, $H(X)$ représente le nombre minimal de bits permettant de coder, en moyenne, une réalisation de X.

Propriétés de l'entropie :

- $H(X) \geq 0$, avec égalité ssi $p_i = 1$ pour un i. La v.a. X est presque sûrement égale à une quantité certaine. La réalisation d'un phénomène entièrement prévisible n'apporte aucune information.
- $H(X) = 0$ pour $P(x) = 0$ car on sait que $\lim_{u \to 0^+} [u \log(1/u)] = 0$
- $H(X) \leq \log(\mid X \mid)$ avec égalité ssi $p_i = 1/ \mid X \mid \quad \forall i$, ($\mid X \mid$ étant le nombre d'éléments de $\mathcal{A}_X$). L'entropie $H(p_1, \ldots, p_n)$ est donc maximale pour la distribution uniforme (voir exercice 4 pour la démonstration).

Entropie d'un couple de v.a. : L'entropie du couple de v.a. (X, Y) où X et Y sont deux v.a. définies sur le même espace (Ω, T, P), à valeurs dans $\mathcal{A}_X$ et $\mathcal{A}_Y$ respectivement, est appellée *information conjointe* et vaut selon la définition A.16 :

$$H(X, Y) = -\sum_{x,y \in \mathcal{A}_X \otimes \mathcal{A}_Y} P(x, y) \log(P(x, y)) \tag{A.18}$$

Cette entropie est additive pour des v.a. indépendantes :

$$H(X, Y) = H(X) + H(Y) \quad \text{ssi} \quad P(x, y) = P(x)P(y) \tag{A.19}$$

(voir exercice 5 pour la démonstration).

L'information conditionnelle de X sachant que $y = b_k$ se définit de la même façon :

$$H(X \mid y = b_k) = -\sum_{x \in \mathcal{A}_X} P(x \mid y = b_k) \log(P(x \mid y = b_k)) \tag{A.20}$$

[3] C'est l'allemand Rudolf Clausius qui a parlé pour la première fois d'entropie *Verwandlungsinhalt* traduit en anglais par *transformation content* et qui a donné en français entropie (qui en grec signifie "tourner dans")

L'information conditionnelle de X connaissant Y sera la moyenne sur tous les y de l'entropie conditionnele de X sachant un y :

$$H(X \mid Y) = -\sum_{y \in \mathcal{A}_Y} P(y) \left[\sum_{x \in \mathcal{A}_X} P(x \mid y) \log(P(x \mid y)) \right] \quad \text{(A.21)}$$

$$= -\sum_{xy \in \mathcal{A}_X \otimes \mathcal{A}_Y} P(x,y) \log(P(x \mid y)) \quad \text{(A.22)}$$

C'est aussi l'incertitude moyenne que l'on a sur x quand y est connu.

L'entropie jointe, l'entropie marginale et l'entropie conditionnelle sont liées par les relations suivantes :

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y) \quad \text{(A.23)}$$

(voir exercice 8 pour la démonstration). En clair : le contenu en information de XY est l'information apportée par X augmentée de l'information apportée par Y sachant X.

Information mutuelle moyenne L'information mutuelle moyenne entre 2 v.a. X et Y est définie par :

$$I(X;Y) = + \sum_{x,y \in \mathcal{A}_X \otimes \mathcal{A}_Y} P(x,y) \log(\frac{P(x,y)}{P(x)P(y)}) \quad \text{(A.24)}$$

$I(X;Y) \geq 0$. Elle est nulle si les deux v.a. sont indépendantes.

La distance entropique entre X et Y peut être définie comme la différence entre leur entropie jointe et leur information mutuelle.

$$D_H(X,Y) = H(X,Y) - I(X;Y) \quad \text{(A.25)}$$

Cette quantité satisfait les axiomes d'une distance :

- $D_H(X,Y) \geq 0$
- $D_H(X,X) = 0$
- $D_H(X,Y) = D_H(Y,X)$
- $D_H(X,Z) \leq D_H(X,Y) + D_H(Y,Z)$

L'information mutuelle entre X et Y sachant que $z = c_k$ est la moyenne selon z de l'information mutuelle entre les v.a. X et Y dans l'ensemble joint $P(x,y \mid z = c_k)$:

$$I(X;Y \mid z = c_k) = H(X \mid z = c_k) - H(X \mid Y, z = c_k) \quad \text{(A.26)}$$

L'information mutuelle entre X et Y sachant Z est la moyenne selon z de la quantité précédente :

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z) \quad \text{(A.27)}$$

Relations diverse avec information mutuelle et entropie

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad \text{(A.28)}$$

$$I(X;Y) = H(Y) - H(Y \mid X) \quad \text{(A.29)}$$

$$H(X \mid Y) \leq H(X) \quad \text{(A.30)}$$

$$H(X,Y) \leq H(X) + H(Y) \leq 2H(X,Y) \quad \text{(A.31)}$$

La figure A.1 montre les relations qui existent entre l'entropie jointe, l'entropie marginale, l'entropie conditionnelle et l'information mutuelle.

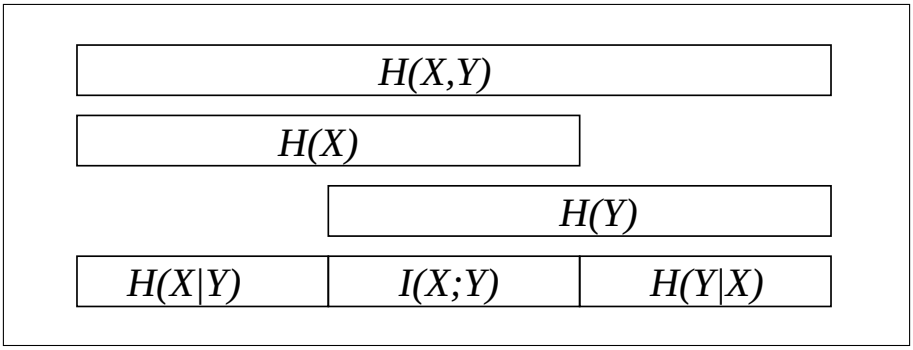

FIG. A.1 – Illustration des relations existant entre l'entropie jointe, l'entropie marginale, l'entropie conditionnelle et l'information mutuelle.

A.2.1 Autres définitions utiles

Divergences en information. Dès que l'on parle de quantité d'information, c'est l'entropie et la définition donnée plus haut qui vient à l'esprit. Ce n'est cependant pas la seule mesure de la quantité d'information. Mathématiquement, n'importe quelle membre de la famille suivante, appellée *divergences en information* ou encore *δ-déviations* :

$$D_\delta(p,q) = \frac{1}{\delta(1-\delta)}(1 - \int p^\delta q^{1-\delta}), \quad \delta \in [0,1] \tag{A.32}$$

est généralement une bonne mesure de la quantité d'information permettant de discriminer entre deux distributions de probabilité p et q.

Parmi celles-ci, les plus importantes sont la 1-déviation ou divergence de Kullback-Leibler ou encore entropie relative que j'explicite au paragraphe suivant et la 1/2-déviation (distance de Hellinger) :

$$D_{1/2}(p,q) = 2 \int (\sqrt{p} - \sqrt{q})^2 \tag{A.33}$$

Entropie de Kullback-Leibler. C'est une grandeur utilisée en reconnaissance d'images et en réseaux de neurones. Elle est définie par :

$$D_1(p,q) = D_{KL}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \tag{A.34}$$

pour deux distributions de probabilités $P(x)$ et $Q(x)$ définies sur le même alphabet $\mathcal{A}$. Une propriété de la divergence de KL est l'additivité pour un ensemble de distribution indépendantes. Par exemple, si l'espace des états est fini et si Q est une distribution uniforme ($q_i = 1/n$), alors nous retrouvons l'entropie à une constante près :

$$D_{KL}(P \parallel Q) = \log n + \sum P \log Q. \tag{A.35}$$

L'entropie relative vérifie l'inégalité de Gibbs : $D_{KL}(P \parallel Q) \geq 0$, avec égalité si $P = Q$ (voir exercice 9). Ce n'est pas une vraie distance même si on emploie souvent ce terme.

L'exercice 7 utilise ce théorème.

Fonction convexe. Une fonction $f(x)$ est convexe sur un intervalle $[a,b]$, si $\forall x_1, x_2 \in [a,b]$ et $0 \leq \lambda \leq 1$ on a (voir la figure A.2) :

$$f(\lambda x_2 + (1-\lambda)x_1) \leq \lambda f(x_2) + (1-\lambda) f(x_1) \tag{A.36}$$

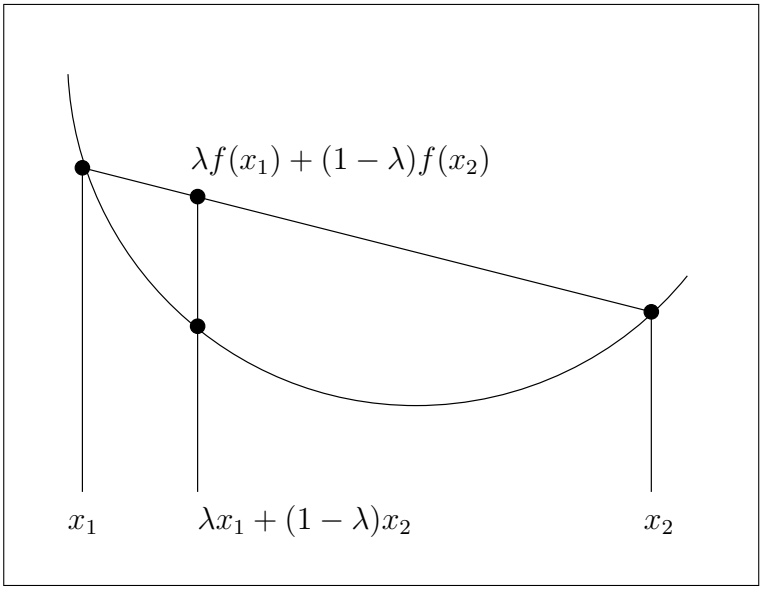

FIG. A.2 – Définition d'une fonction convexe

Par exemple les fonction $x^2, e^x, \log(1/x)$ et $x \log x$ sont strictement convexes sur leurs domaines de définition.

Inégalité de Jensen. Si on note l'espérance par $E[\cdot]$, si f est une fonction convexe et si x est une variable aléatoire, alors on a :

$$E[f(x)] \geq f(E[x]) \tag{A.37}$$

Si on a l'égalité et que f est strictement convexe alors la variable aléatoire est une constante (avec la probabilité 1).

A.3 Exercices

Exercice 1 : Soit $p_a = 0.1$, $p_b = 0.2$, $p_c = 0.7$, $f(a) = 10$, $f(b) = 5$, $f(c) = 10/7$.
Quelle est l'espérance de $f(x)$?
Quelle est l'espérance de $1/P(x)$?
Pour des données quelconques, quelle est l'espérance de $1/P(x)$?

Exercice 2 : Soit $p_a = 0.1$, $p_b = 0.2$, $p_c = 0.7$, $f(a) = 0$, $f(b) = 1$, $f(c) = 0$.
Quelle est l'espérance de $f(x)$?

Exercice 3 : Soit $p_a = 0.1$, $p_b = 0.2$, $p_c = 0.7$. Quelle est la probabilité pour que $P(x) \in [0.15, 0.5]$?
Que vaut :

$$P\left(\left|\log \frac{P(x)}{0.2}\right| > 0.05\right) ?$$

Exercice 4 : Prouver que $H(X) \leq \log(\mid X \mid)$ avec égalité ssi $p_i = 1/ \mid X \mid \quad \forall i$, ($\mid X \mid$ étant le nombre d'éléments de $\mathcal{A}_X$).
Prouver que l'entropie $H(X)$ est donc maximale pour la distribution uniforme.

Exercice 5 : Prouver que l'entropie jointe est additive pour des v.a. indépendantes :

$$H(X,Y) = H(X) + H(Y) \quad \text{ssi} \quad P(x,y) = P(x)P(y)$$

Exercice 6 : Soit U, V et W trois variables aléatoires indépendantes d'entropie H_u, H_v, H_w. Soit alors les variables $X = (U, V)$, $Y = (V, W)$.
Calculer $H(X, Y)$, $H(X \mid Y)$ et $I(X; Y)$.

Exercice 7 : Prouver que $H(X \mid Y) \leq H(X)$

Exercice 8 : Prouver que $H(X, Y) = H(X) + H(Y \mid X)$

Exercice 9 : Prouver que l'entropie relative vérifie l'inégalité de Gibbs : $D_{KL}(P \parallel Q) \geq 0$, avec égalité si $P = Q$

Exercice 10 : Soit un ensemble XY dont la distribution jointe est la suivante :

$P(x,y)$		x			
		1	2	3	4
y	1	1/8	1/16	1/32	1/32
	2	1/16	1/8	1/32	1/32
	3	1/16	1/16	1/16	1/16
	4	1/4	0	0	0

a/ Calculer l'entropie jointe $H(X, Y)$.
b/ Calculer les entropies marginales $H(X)$ et $H(Y)$.
c/ Pour chaque valeur de y quelle est l'entropie conditionelle $H(X \mid y)$?
d/ Que vaut l'entropie conditionnelle $H(X \mid Y)$?
e/ Et celle de Y sachant X ?
f/ Calculer l'information mutuelle de X et Y.

Exercice 11 : Théorème sur le traitement des données.

Considérons un ensemble MDR où M est le monde, D les données et R le résultat d'un traitement sur les données. Ces trois variables forment une chaîne de Markov :

$$m \rightarrow d \rightarrow r$$

si bien que la probabilité $P(m, d, r)$ peut s'écrire :

$$P(m, d, r) = P(m)P(d \mid m)P(r \mid d)$$

Montrer que l'information que R contient sur le monde M, $I(M; R)$ est plus petite ou égale à l'information que D contient sur le monde, $I(M; D)$. C'est un théorème qui dit qu'il faut faire attention à la définition de ce qu'est « l'information » mais aussi qu'il faut faire attention au traitement des données.

A.4 Corrigé des exercices

Solution de l'exercice 1 : $E[f(x)] = \sum p(x)f(x) = 0.1 \times 10 + 0.2 \times 5 + 0.7 \times 10/7 = 3$
$E[1/P(x)] = \sum_{x \in \mathcal{A}_X} P(x)/P(x) = \sum_{x \in \mathcal{A}_X} 1 = \mid \mathcal{A}_X \mid$

■

Solution de l'exercice 2 : $E[f(x)] = \sum p(x)f(x) = 0.1 \times 0 + 0.2 \times 1 + 0.7 \times 0 = 0.2$

■

Solution de l'exercice 3 :

$$\begin{aligned} P\left(P(x) \in [0.15, 0.5]\right) &= 0.2 \\ P\left(\left|\log \frac{P(x)}{0.2}\right| > 0.05\right) &= p_a + p_c = 0.8 \end{aligned}$$

■

Solution de l'exercice 4 :

$$\begin{aligned} H(x) &= \sum_i p_i \log 1/p_i \\ \frac{\partial H}{\partial p_i} &= \log 1/p_i - 1 \end{aligned}$$

Pour tenir compte de la contrainte $\sum_i p_i = 1$ nous utiliserons un multiplicateur de Lagrange et chercherons à maximiser $G(X)$:

$$\begin{aligned} G(x) &= H(X) + \lambda(\sum p_i - 1) \\ \text{soit :} \quad \frac{\partial G}{\partial p_i} &= \log 1/p_i - 1 + \lambda \end{aligned}$$

dont l'extremum est tel que :

$$\begin{aligned} \log 1/p_i - 1 + \lambda &= 0 \\ \text{soit :} \quad \log p_i &= 1 - \lambda = \text{constante} \end{aligned}$$

Les p_i sont donc tous égaux, la distribution est uniforme.

Pour montrer que cet extremum est bien un maximum, il faut étudier le signe de la dérivée seconde :

$$\frac{\partial^2 G}{\partial p_i \partial p_j} = -\frac{1}{p_i}\delta_{ij} \qquad < 0$$

∎

Solution de l'exercice 5 : Si les variables aléatoires sont indépendantes on a :

$$\begin{aligned} P(x,y) &= P(x)P(y) \\ H(X,Y) &= \sum_{x,y \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) \log \frac{1}{P(x,y)} \quad \text{par définition} \\ &= \sum_{x,y \in \mathcal{A}_X \mathcal{A}_Y} P(x)P(y) \log \frac{1}{P(x)P(y)} \\ &= \sum_{y \in \mathcal{A}_Y} P(y) \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)} + \sum_{x \in \mathcal{A}_X} P(x) \sum_{y \in \mathcal{A}_Y} P(y) \log \frac{1}{P(y)} \\ &= \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)} + \sum_{y \in \mathcal{A}_Y} P(y) \log \frac{1}{P(y)} \\ &= H(X) + H(Y) \end{aligned}$$

∎

Solution de l'exercice 6 :

$$\begin{aligned} H(X,Y) &= H(U,V,V,W) = H(U,V,W) = H_u + H_v + H_w \\ H(X \mid Y) &= H(U,V \mid V,W) = H(U \mid W) = H_u \\ I(X;Y) &= H(X) - H(X \mid Y) = H_u + H_v - H_u = H_v \end{aligned}$$

∎

Solution de l'exercice 7 : On fait apparaître $H(X)$ à partir de la définition de $H(X \mid Y)$:

$$\begin{aligned} H(X \mid Y) &= \sum_y P(y) \sum_x P(x \mid y) \log \frac{1}{P(x \mid y)} \\ &= \sum_{xy} P(x,y) \log \frac{1}{P(x \mid y)} \end{aligned}$$

que l'on peut écrire en utilisant le théorème de Bayes :

$$\begin{aligned} H(X \mid Y) &= \sum_{xy} P(x)P(y \mid x) \log \frac{P(y}{P(x)P(y \mid x)} \\ &= \sum_{xy} P(x)P(y \mid x) \log \frac{1}{P(x)} + \sum_x P(x) \sum_y P(y \mid x) \log \frac{P(y)}{P(y \mid x)} \\ &= H(X) + \sum_x P(x) \sum_y P(y \mid x) \log \frac{P(y)}{P(y \mid x)} \end{aligned}$$

d'après l'inégalité de Gibbs (avec içi $P(x) \equiv P(y)$ et $Q(x) \equiv P(y \mid x)$) le deuxième terme est ≥ 0 avec égalité si $P(y) = P(y \mid x)$, c'est-à-dire si les deux distributions sont indépendantes.

∎

Solution de l'exercice 8 :

$$\begin{aligned} H(X,Y) &= \sum_{xy} P(x,y) \log \frac{1}{P(x,y)} \\ &= \sum_{xy} P(x)P(y \mid x) \log \frac{1}{P(x)P(y \mid x)} \\ &= \sum_x P(x) \log \frac{1}{P(x)} + \sum_x P(x) \sum_y P(y \mid x) \log \frac{1}{P(y \mid x)} \\ &= H(X) + H(Y \mid X) \end{aligned}$$

∎

Solution de l'exercice 9 :

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

On va utiliser l'inégalité de Jensen qui dit que :

$$E[f(u)] \geq f(E[u]) \quad \text{si } f \text{ est convexe}$$

Posons alors $u = Q/P$ et $f(u) = \log 1/u$. Il vient d'une part :

$$E[f(u)] = \sum_x P(x) \log \frac{1}{u} = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

qui est l'entropie relative. D'autre part nous avons :

$$\begin{aligned} E[f(u)] &\geq f(E[u]) \\ &\geq f(E[Q/P]) \\ &\geq f(\sum_x P(x)Q(x)/P(x)) \\ &\geq \log \frac{1}{\sum_x Q(x)} \\ &\geq 0 \end{aligned}$$

On a égalité ssi $u = P/Q$ est une constante, c'est-à-dire si $P(x) = Q(x)$

∎

Solution de l'exercice 10 :

a/ par définition l'entropie jointe est :

$$\begin{aligned} H(X,Y) &= \sum_{x,y} P(x,y) \log \frac{1}{P(x,y)} \\ &= 27/8 \quad \text{bits} \end{aligned}$$

b/ par définition les entropies marginales $H(X)$ et $H(Y)$ sont données par :

$$\begin{aligned} H(X) &= \sum_x P(x) \log \frac{1}{P(x)} \\ H(Y) &= \sum_y P(y) \log \frac{1}{P(y)} \end{aligned}$$

A partir des probabilités jointes nous trouvons les marginales :

$$\begin{aligned} P(X) &= \sum_x P(x,y) \\ P(Y) &= \sum_y P(x,y) \end{aligned}$$

Nous obtenons alors le tableau suivant :

$P(x,y)$		x				$P(y)$
		1	2	3	4	
y	1	1/8	1/16	1/32	1/32	1/4
	2	1/16	1/8	1/32	1/32	1/4
	3	1/16	1/16	1/16	1/16	1/4
	4	1/4	0	0	0	1/4
P(x)		1/2	1/4	1/8	1/8	

on obtient ainsi :

$$\begin{aligned} H(X) &= 1/2 \log 2 + 1/4 \log 4 + 1/8 \log 8 + 1/8 \log 8 \\ &= 7/4 \quad \text{bits} \\ H(Y) &= 4 \times 1/4 \log 4 \\ &= 2 \quad \text{bits} \end{aligned}$$

c/ par définition l'entropie conditionnelle de X sachant que $y = b_k$ est :

$$\begin{aligned} H(X \mid y = b_k) &= -\sum_{x \in \mathcal{A}_X} P(x \mid y = b_k) \log(P(x \mid y = b_k)) \\ \rightarrow \quad H(X \mid y = 1) &= P(x = 1 \mid y = 1) \log(P(x = 1 \mid y = 1)) \\ &+ P(x = 2 \mid y = 1) \log(P(x = 2 \mid y = 1)) \\ &+ P(x = 3 \mid y = 1) \log(P(x = 3 \mid y = 1)) \\ &+ P(x = 4 \mid y = 1) \log(P(x = 4 \mid y = 1)) \end{aligned}$$

Nous savons d'autre part que :

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

si bien que nous pouvons en déduire le tableau suivant :

$P(x \mid y)$		x				$H(X \mid y = b_k)$
		1	2	3	4	
y	1	1/2	1/4	1/8	1/8	7/4
	2	1/4	1/2	1/8	1/8	7/4
	3	1/4	1/4	1/4	1/4	2
	4	1	0	0	0	0

on obtient ainsi :

$$\begin{aligned} H(X \mid y = 1) &= 1/2 \log 2 + 1/4 \log 4 + 1/8 \log 8 + 1/8 \log 8 \\ &= 7/4 \quad \text{bits} \\ H(X \mid y = 2) &= 1/4 \log 4 + 1/2 \log 2 + 2/8 \log 8 \\ &= 7/4 \quad \text{bits} \\ H(X \mid y = 3) &= 4/4 \log 4 \\ &= 2 \quad \text{bits} \\ H(X \mid y = 4) &= 1 \log 1 + 0 \\ &= 0 \quad \text{bits} \end{aligned}$$

d/ par définition de l'information conditionnelle de X connaissant Y comme étant la moyenne sur tous les y de l'entropie conditionnele de X sachant un y :

$$\begin{aligned} H(X \mid Y) &= -\sum_{y \in \mathcal{A}_Y} P(y) \left[\sum_{x \in \mathcal{A}_X} P(x \mid y) \log(P(x \mid y)) \right] \\ &= \sum_y P(y) H(X \mid y = b_k) \\ &= 1/4 \times 7/4 + 1/4 \times 7/4 + 1/4 \times 2 \\ &= 11/8 \quad \text{bits} \end{aligned}$$

e/ L'entropie conditionnelle $H(Y \mid X)$ étant donnée par :

$$H(Y \mid X) = \sum_x P(x) H(Y \mid x = a_k)$$

et sachant que :

$$H(Y \mid x = a_k) = -\sum_y P(y, x = a_k) \log P(y \mid x = a_k)$$

et que :

$$P(y \mid x) = \frac{P(x, y)}{P(x)}$$

nous obtenons le tableau suivant :

$P(y \mid x)$		x			
		1	2	3	4
y	1	1/4	1/4	1/4	1/4
	2	1/8	1/2	1/4	1/4
	3	1/8	1/4	1/2	1/2
	4	1/2	0	0	0
$H(Y \mid x = a_k)$		7/4	3/2	3/2	3/2

on obtient ainsi :

$$\begin{aligned}
H(Y \mid x = 1) &= 1/4 \log 4 + 1/8 \log 8 + 1/8 \log 8 + 1/2 \log 2 = 7/4 \quad \text{bits} \\
H(Y \mid x = 2) &= 1/4 \log 4 + 1/2 \log 2 + 1/4 \log 4 = 3/2 \quad \text{bits} \\
H(Y \mid x = 1) &= 1/4 \log 4 + 1/4 \log 4 + 1/2 \log 2 = 3/2 \quad \text{bits} \\
H(Y \mid x = 1) &= 1/4 \log 4 + 1/4 \log 4 + 1/2 \log 2 = 3/2 \quad \text{bits}
\end{aligned}$$

si bien que l'on trouve :

$$H(X \mid Y) = 1/2 \times 7/4 + 1/4 \times 3/2 + 1/8 \times 3/2 + 1/8 \times 3/2 = 13/8 \quad \text{bits}$$

f/ L'information mutuelle peut se calculer de différentes manières, par exemple en utilisant :

$$\begin{aligned}
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
&= 7/4 + 2 - 27/8 \\
&= 3/8 \quad \text{bits}
\end{aligned}$$

■

Solution de l'exercice 11 : ■

Bibliographie