Chapitre 2

Entropie relative

Une limite importante pour la notion d'entropie que nous avons introduite est qu'elle ne s'applique qu'à des suites de variables aléatoires i.i.d., or les applications de ce concept comme la compression mènent naturellement à des suites de variables fortement corrélées. Par exemple, dans un mot en français, on a beaucoup plus de chance de trouver un u si la lettre précédente est un q que si c'est un autre u. On aimerait faire apparaître cette structure supplémentaire dans les concepts utilisés.

Si l'on prend deux variables aléatoires X et Y, l'entropie de la variable aléatoire jointe s'exprime par

$$\begin{aligned} H(X,Y) &= -\sum_x \sum_y P(x,y) \log P(x,y) \\ &= -\sum_x \sum_y P(x,y) \log P(x) - \sum_x \sum_y P(x,y) \log P(y|x) \\ &= H(X) + H(Y|X) \\ &= H(Y) + H(X|Y). \end{aligned}$$

Définition 2.0.1. *Ce calcul motive la définition de l'entropie relative de Y par rapport à X,*

$$H(Y|X) = -\sum_x \sum_y P(x,y) \log P(y|x).$$

On remarque que X et Y sont indépendantes si et seulement si,

$$H(Y|X) = H(Y) \text{ et } H(X,Y) = H(X) + H(Y).$$

Définition 2.0.2. *Définissons à présent l'information mutuelle moyenne*

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= I(Y;X) \end{aligned}$$

Si l'on regarde plus précisément l'expression de l'information mutuelle, on obtient

$$\begin{aligned} I(X;Y) &= -\sum_x \sum_y P(x,y) \log P(x) + \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(y)} \\ &= \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\ &= D_{KL}\left(P(x,y) \| P(x)P(y)\right). \end{aligned}$$

Cela correspond à une "distance" entre les deux lois de probabilité $P(x,y)$ et $P(x)P(y)$ que l'on appelle distance de Kullback-Leibler.

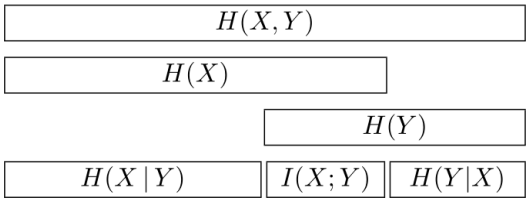
Définition 2.0.3. *Elle s'exprime de manière générale pour deux distributions de probabilité P et Q*

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

L'inégalité de Gibbs affirme que $D_{KL}(P,Q) \geq 0$ avec égalité si et seulement si $P(x) = Q(x)$ pour tout x. En particulier on a $I(X;Y) \geq 0$ ce qui implique la propriété suivante.

Proposition 2.0.4. *$H(X|Y) \leq H(X)$ avec égalité si et seulement si X et Y sont indépendants.*

On représente une synthèse de ces propriétés sur le diagramme ci-dessous.



Ces définitions nous mènent à la règle de chaînage fort utile pour les séquences de variables aléatoires.

Proposition 2.0.5. *Soit $(X_1, \ldots, X_n) \sim P(x_1, \ldots, x_n)$, alors*

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1} \ldots X_1).$$

Démonstration. On écrit $P(x_1,\dots,x_n) = P(x_1)P(x_2|x_1)\dots P(x_n|x_{n-1}\dots x_1)$, soit :

$$\begin{aligned}
H(X_1,\dots,X_n) &= -\sum_{x_1,\dots,x_n} P(x_1,\dots,x_n)\log P(x_1\dots x_n)\\
&= -\sum_{i=1}^{n}\sum_{x_1,\dots,x_i}\sum_{x_{i+1},\dots,x_n} P(x_1,\dots,x_n)\log P(x_i|x_{i-1}\dots x_1)\\
&= -\sum_{i=1}^{n}\sum_{x_1,\dots,x_i} P(x_1,\dots,x_i)\log P(x_i|x_{i-1}\dots x_1)\\
&= \sum_{i=1}^{n} H(X_i|X_{i-1}\dots X_1).
\end{aligned}$$

□

Corollaire 2.0.6. *$H(X_1,\dots,X_n) \leq \sum_{i=1}^{n} H(X_i)$ avec égalité si et seulement si les X_i sont indépendants.*

On définit l'information mutuelle conditionnelle de façon similaire.

Définition 2.0.7. *L'information mutuelle conditionnelle des variables aléatoires X et Y étant donné Z est :*

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(Y|Z) - H(Y|X,Z) = I(Y;X|Z).$$

Proposition 2.0.8. *Alors de même on a une loi de chaînage,*

$$I(X_1,\dots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1,\dots,X_{i-1}).$$

Démonstration.

$$\begin{aligned}
I(X_1,\dots,X_n;Y) &= H(X_1,\dots,X_n) - H(X_1,\dots,X_n|Y)\\
&= \sum_{i=1}^{n} H(X_i|X_1,\dots,X_{i-1}) - \sum_{i=1}^{n} H(X_i|X_1,\dots,X_{i-1},Y)\\
&= \sum_{i=1}^{n} I(X_i;Y|X_1,\dots,X_{i-1}).
\end{aligned}$$

□

Une question centrale dans cette théorie est de trouver les distributions qui maximisent l'entropie. Dans le cas des distributions discrètes et finies c'est la distribution uniforme.

Proposition 2.0.9. *$H(X) \leq \log|X|$ où $|X|$ est le nombre d'éléments de $\mathcal{A}$ pour lesquels la probabilité est strictement positive avec égalité si et seulement si la loi de probabilité est uniforme sur ces éléments de probabilité non nulle.*

Démonstration. Soit $U(x) = \frac{1}{|X|}$ pour tout x dans le support de la loi P notée S. On note alors que

$$D(P\|U) = \sum_{x\in S} P(x)\log\frac{P(x)}{U(x)} = \log|X| - H(X).$$

□

2.1 Lien entre théorie de l'information et PMU

Considérons un jeux de paris sur M chevaux concurrents. Si le k-ème cheval gagne, on gagne $o(k)$ fois sa mise. C'est-à-dire, si on a misé 1 euro, on obtient $o(k)$ euros si le cheval k gagne et rien sinon (la mise de 1 euro est perdue).

On suppose que le joueur distribue toute sa fortune sur les chevaux. Soit $b(k)$ la fraction pariée sur le cheval k. On a donc $b(i) \geq 0$ et $\sum_{k=1}^{M} b(k) = 1$.

Après n courses, la fortune du joueur est (en supposant que initialement $S_0 = 1$) :

$$S_n = \prod_{i=1}^{n} S(X_i) \quad \text{avec} \quad S(X) = b(X) \cdot o(X) \quad \text{et } X_i \text{ est le cheval vainqueur au tour } i.$$

Théoreme 2.1.1. *Si les X_i sont i.i.d. selon des probabilités $p_1, \ldots, p_M$, alors la fortune du joueur utilisant la stratégie b "croit" exponentiellement. C'est à dire que, presque surement, on a*

$$\lim_{n\to\infty} \frac{\log S_n}{n} = W(b,p) = \sum_{k=1}^{M} p_k \log(b(k) \cdot o(k))$$

.

Démonstration. Il suffit d'appliquer la loi (forte) des grands nombres et de noter que $W(b,p) = E[\log S(X)]$. □

Attention, rien ne dit que $W(b,p) \geq 0$!

Théoreme 2.1.2 (Critère de Kelly)**.** *La stratégie optimale est de prendre $b^* = p$ et dans ce cas $W(b^*,p) = \sum_{k=1}^{M} p_k \log o(k) - H(p)$.*

Démonstration. Il suffit d'écrire :

$$\begin{aligned} W(b,p) &= \sum_{k=1}^{M} p_k (\log o(k) + \log b(k)) \\ &= \sum_{k=1}^{M} p_k \log o(k) + \sum_{k=1}^{M} p_k \left(\log p_k + \log \frac{b(k)}{p_k} \right) \\ &= \sum_{k=1}^{M} p_k \log o(k) - H(p) - D(p\|b). \end{aligned}$$

Dans le cas équitable où $\sum_{k=1}^{M} \frac{1}{o(k)} = 1$, $r_k = \frac{1}{o(k)}$ est une distribution de probabilité qui correspond à l'estimation de la probabilité p_k par le bookmaker. On a alors :

$$\begin{aligned} W(b,p) &= \sum_{k=1}^{M} p_k \log (p_k \cdot o(k)) - D(p\|b) \\ &= D(p\|r) - D(p\|b). \end{aligned}$$

□

Si notre estimation b de p est meilleure que celle du bookmaker, dans le sens de la distance de Kullback-Leibler, alors $W(b,p) \geq 0$.

Dans le cas spécial où $o(i) = m$ pour tout i, on a $W(b^*,p) = \log m - H(p)$. Les courses à faible entropie sont les plus profitables.

2.2 Inégalité de traitement des données

On dit que les variables aléatoires X, Y, Z forment une chaîne de Markov dans cet ordre si la distribution de Z conditionnée par X, Y ne dépend que de Y et pas de X. Autrement dit, si leurs probabilités conjointes satisfont :

$$P(z|x,y) = P(z|y)$$

ce qui est équivalent à

$$P(x,y,z) = P(x)P(y|x)P(z|y).$$

Pour tout $x \in X, y \in Y$ et $z \in Z$. On note $X \to Y \to Z$.

Quelques conséquences simples sont les suivantes :

— $X \to Y \to Z$ si et seulement si X et Z sont conditionnellement indépendants étant donné Y. La markovité implique une indépendance conditionnelle car

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

— $X \to Y \to Z$ implique que $Z \to Y \to X$. Ainsi, la condition est parfois écrite $X \leftrightarrow Y \leftrightarrow Z$.

— Si $Z = f(Y)$, alors $X \to Y \to Z$. En effet, $P(z|y) = P(z|x,y) = \delta_{z=f(y)}$ qui vaut 1 si la condition est vérifiée et 0 sinon.

Nous pouvons maintenant prouver un théorème important et utile démontrant qu'aucun traitement de Y, déterministe ou aléatoire, ne peut augmenter l'information que Y contient sur X. Aucune manipulation astucieuse des données ne peut améliorer les inférences pouvant être tirées des données.

Théoreme 2.2.1 (Inégalité de traitement des données)**.** *Si $X \to Y \to Z$, alors*

$$I(X;Y) \geq I(X;Z).$$

En particulier, si $Z = g(Y)$, nous avons $I(X;Y) \geq I(X;g(Y))$.

Démonstration. Par la règle de chainage, nous pouvons étendre l'information mutuelle de deux manières différentes :

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y)$$

Étant donné que X et Z sont conditionnellement indépendants étant donné Y, nous avons $I(X;Z|Y) = 0$. Puisque $I(X;Y|Z) \geq 0$, nous avons $I(X;Y) \geq I(X;Z)$. □

Ainsi, les fonctions des données Y ne peuvent pas augmenter l'information sur X.

On a aussi montré dans la preuve que $I(X;Y|Z) \leq I(X;Y)$.

Ainsi, la dépendance entre X et Y est diminuée (ou reste inchangée) par l'observation d'une variable aléatoire "en aval" Z. Notez qu'il est également possible que $I(X;Y|Z) > I(X;Y)$ lorsque X, Y et Z ne forment pas une chaîne de Markov. Par exemple, laissons X et Y être des variables aléatoires binaires équitables indépendantes, et laissons $Z = X + Y$. Alors $I(X;Y) = 0$, mais $I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) = P(Z=1)H(X|Z=1) = \frac{1}{2}$ bit.

2.3 Statistiques exhaustives

Cette section est une illustration du pouvoir de l'inégalité de traitement des données pour clarifier une idée importante en statistiques. Supposons que nous ayons une famille de fonctions de masse de probabilité $\{f_\theta(x)\}$ indexée par θ et soit X un échantillon d'une distribution dans cette famille, c'est à dire des mesures d'une variable aléatoire satisfaisant cette loi. Soit $T(X)$ une statistique (fonction de l'échantillon) comme la moyenne de l'échantillon ou la variance de l'échantillon. Alors $\theta \to T(X) \to X$, et par l'inégalité de traitement des données, nous avons

$$I(\theta; T(X)) \geq I(\theta; X)$$

pour toute distribution sur θ. Si l'égalité est vérifiée, aucune information n'est perdue et on dit qu'une telle statistique *exhaustive* pour θ si elle contient toute l'information dans X sur θ.

Voici quelques exemples de statistiques exhaustives :

1. Soient $X_1, X_2, \ldots, X_n, X_i \in \{0,1\}$, une séquence indépendante et identiquement distribuée (i.i.d.) de lancers de pièce d'une pièce avec un paramètre inconnu $\theta = \Pr(X_i = 1)$. Étant donné n, le nombre de 1 est une statistique exhaustive pour θ. Ici, $T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^n X_i$. En fait, on peut montrer qu'étant donné T, toutes les séquences ayant autant de 1 sont également probables et indépendantes du paramètre θ. Plus précisément,

$$\Pr\left(\sum_{i=1}^n X_i = k\right) = \binom{n}{k}^{-1} \text{ si } \sum_{i=1}^n x_i = k, \quad 0 \text{ sinon.}$$

   Ainsi, $\theta \to \sum X_i \to (X_1, X_2, \ldots, X_n)$ forme une chaîne de Markov, et T est une statistique exhaustive pour θ.

   Les deux exemples suivants impliquent des densités de probabilité au lieu de fonctions de masse de probabilité, mais la théorie s'applique toujours.

2. Si X est distribué normalement avec une moyenne θ et une variance de 1, c'est-à-dire si

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}} = \mathcal{N}(\theta, 1)$$

   et $X_1, X_2, \ldots, X_n$ sont tirés indépendamment selon cette distribution, une statistique exhaustive pour θ est la moyenne de l'échantillon $X_n = \frac{1}{n}\sum_{i=1}^n X_i$.

   On peut vérifier que la distribution conditionnelle de $X_1, X_2, \ldots, X_n$, conditionnée par X_n ne dépend pas de θ.

3. Si $f_\theta = \text{Uniform}(\theta, \theta+1)$, une statistique exhaustive pour θ est

$$T(X_1, X_2, \ldots, X_n) = (\max\{X_1, X_2, \ldots, X_n\}, \min\{X_1, X_2, \ldots, X_n\}).$$

La preuve de ceci est légèrement plus compliquée, mais encore une fois, on peut montrer que la distribution des données est indépendante du paramètre étant donné la statistique T.

Définition 2.3.1. *Une statistique $T(X)$ est une statistique exhaustive minimale par rapport à $\{f_\theta(x)\}$ si elle est une fonction de toute autre statistique exhaustive U. C'est à dire que pour toute statistique exhaustive U, on a $\theta \to T(X) \to U(X) \to X$.*

Ainsi, une statistique exhaustive minimale compresse au maximum l'information sur θ dans l'échantillon. D'autres statistiques exhaustives peuvent contenir des informations supplémentaires non pertinentes. Par exemple, pour une distribution normale avec une moyenne θ, le couple de fonctions donnant la moyenne de tous les échantillons impairs et la moyenne de tous les échantillons pairs est une statistique exhaustive mais pas une statistique exhaustive minimale. Elles n'existent pas toujours mais leur existence est assurée sous des hypothèse faibles. Dans les exemples précédents, les statistiques exhaustives étaient également minimales.