Théorie de l'information

Cours 5 - Entropies conjointe et conditionnelles, information mutuelle

Laurent Oudre
laurent.oudre@univ-paris13.fr

Université Paris 13, Institut Galilée
Master Ingénierie et Innovations en Images et Réseaux - 1ère année
2017-2018

Sommaire

Sommaire

Passage à deux variables aléatoires

- On a vu que l'entropie mesurait l'information moyenne apportée par une source (ou de façon équivalente par une variable aléatoire)
- Nous allons dans ce cours estimer l'information moyenne apportée par un système de deux variables aléatoires
- Cela va nous servir à étudier deux sources, mais surtout à étudier les liens entre l'entrée et la sortie d'un canal de communication.

Entropie conjointe

Entropie conjointe

Étant données deux variables aléatoires X et Y à valeurs respectivement dans $\mathcal{X}$ et $\mathcal{Y}$, on définit **l'entropie conjointe** des variables aléatoires X et Y et on note $H(X,Y)$ la quantité :

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(X=x,Y=y)\log_2\left(P_{XY}(X=x,Y=y)\right)$$

Propriétés

- $H(X,Y)$ correspond à l'information apportée par le système (X,Y)
- On a les propriétés :
  1. $\boxed{H(X,Y) = H(Y,X)}$
  2. $\boxed{H(X,Y) \geq 0}$
  3. $\boxed{H(X,Y) \leq H(X) + H(Y)}$
     avec égalité si et seulement si X et Y sont indépendantes
  4. $\boxed{H(X,Y) \geq H(X)}$
- 1 et 2 sont triviales : dans la suite, démonstration de 3 et 4

Démonstration propriété 3 : $H(X,Y) \leq H(X) + H(Y)$

- Notons :
  - $p_{XY}(x,y) = P_{XY}(X=x,Y=y)$
  - $q_{XY}(x,y) = P_X(X=x)P_Y(Y=y)$
- Ce sont bien toutes les deux des lois de probabilités car :

$$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} q_{XY}(x,y) = \sum_{x\in\mathcal{X}} P_X(X=x)\sum_{y\in\mathcal{Y}} P_Y(Y=y) = 1$$

- Nous allons comparer $p_{XY}(x,y)$ et $q_{XY}(x,y)$ grâce à l'entropie relative.

Démonstration propriété 3 : $H(X,Y) \leq H(X) + H(Y)$

On a d'après la propriété démontrée au cours précédent :

$$\begin{aligned}
D_{KL}(p||q) \geq 0 \quad &\iff \quad \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\log_2\left(\frac{p_{XY}(x,y)}{q_{XY}(x,y)}\right) \geq 0 \\
&\iff \quad \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(X=x,Y=y)\log_2\left(\frac{P_{XY}(X=x,Y=y)}{P_X(X=x)P_Y(Y=y)}\right) \geq 0 \\
&\iff \quad \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(X=x,Y=y)\log_2\left(P_{XY}(X=x,Y=y)\right) \\
&\qquad\quad -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(X=x,Y=y)\log_2\left(P_X(X=x)P_Y(Y=y)\right) \geq 0 \\
&\iff \quad -H(X,Y) - \sum_{x\in\mathcal{X}}\log_2\left(P_X(X=x)\right)\sum_{y\in\mathcal{Y}} P_{XY}(X=x,Y=y) \\
&\qquad\quad -\sum_{y\in\mathcal{Y}}\log_2\left(P_Y(Y=y)\right)\sum_{x\in\mathcal{X}} P_{XY}(X=x,Y=y) \geq 0 \\
&\iff \quad H(X,Y) \leq H(X) + H(Y)
\end{aligned}$$

Interprétation propriété 3 : $H(X, Y) \leq H(X) + H(Y)$

- On a une égalité ssi $p_{XY}(x, y) = q_{XY}(x, y)$, donc ssi

$$P_{XY}(X = x, Y = y) = P_X(X = x)P_Y(Y = y) \iff X \text{ et } Y \text{ sont indépendantes}$$

- Cette propriété nous dit que, si l'on considère deux sources X et Y ensemble, elles ne peuvent pas apporter plus d'information que la somme des apports d'information de chacune

Interprétation propriété 3 : $H(X, Y) \leq H(X) + H(Y)$

Prévisions météo

X source qui nous dit s'il fera *soleil* ou *pluie*
Y source qui nous dit s'il fera *chaud* ou *froid*

X, Y	$Y =$ chaud	$Y =$ froid
$X =$ soleil	0.3	0.2
$X =$ pluie	0.1	0.4

$H(X, Y) = 1.8464$ bits, $X(X) = 1$ bit, $H(Y) = 0.9710$ bits

$$H(X, Y) < H(X) + H(Y)$$

Individuellement les sources nous donnent beaucoup d'information, mais ensemble on a finalement moins d'information que la somme des deux informations car il existe un lien entre elles.

Démonstration propriété 4 : $H(X, Y) \geq H(X)$

$$\begin{aligned}
H(X, Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(X = x, Y = y)\log_2\left(P_{XY}(X = x, Y = y)\right)\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{Y|X=x}(Y = y|X = x)P_X(X = x)\log_2\left(P_{Y|X=x}(Y = y|X = x)P_X(X = x)\right)\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{Y|X=x}(Y = y|X = x)P_X(X = x)\log_2\left(P_{Y|X=x}(Y = y|X = x)\right)\\
&\quad -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{Y|X=x}(Y = y|X = x)P_X(X = x)\log_2\left(P_X(X = x)\right)
\end{aligned}$$

Or $P_{Y|X=x}(Y = y|X = x) \leq 1$ donc le premier terme est ≥ 0

$$\begin{aligned}
H(X, Y) &\geq -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{Y|X=x}(Y = y|X = x)P_X(X = x)\log_2\left(P_X(X = x)\right)\\
&\geq -\sum_{x\in\mathcal{X}} P_X(X = x)\log_2\left(P_X(X = x)\right)\sum_{y\in\mathcal{Y}} P_{Y|X=x}(Y = y|X = x)\\
&\geq -\sum_{x\in\mathcal{X}} P_X(X = x)\log_2\left(P_X(X = x)\right) = H(X)
\end{aligned}$$

Interprétation propriété 4 : $H(X, Y) \geq H(X)$

- On pourrait démontrer de la même manière que $H(X, Y) \geq H(Y)$
- Cette propriété nous dit que deux sources donnent forcément plus d'information qu'une seule source, ce qui est conforme à notre intuition
- On a une égalité si les deux sources contiennent exactement la même information

Sommaire

Liens entre variables aléatoires

- On a vu que ce n'était pas parce qu'on considérait deux sources qu'on obtenait beaucoup plus d'information
- On a en effet :

$$H(X) \leq H(X,Y) \leq H(X) + H(Y)$$

- Qu'est-ce qui explique les variations de $H(X,Y)$ entre ces deux bornes ?
- Les liens éventuels entre les variables : il arrive parfois que connaissant Y, la source X donne moins d'information, tout simplement parce que Y nous a déjà transmis une partie de l'information contenue dans X
- Pour étudier ceci, on va introduire la notion d'entropie conditionnelle

Entropie conditionnelle

Entropie conditionnelle

Étant données deux variables aléatoires X et Y à valeurs respectivement dans $\mathcal{X}$ et $\mathcal{Y}$, on définit **l'entropie conditionnelle** de Y sachant X et on note $H(Y|X)$ la quantité :

$$H(Y|X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(X=x, Y=y)\log_2\left(P_{Y|X=x}(Y=y|X=x)\right)$$

Propriétés

- $H(Y|X)$ correspond à l'incertitude restante sur Y lorsqu'on connait X.
- Si on appelle X et Y respectivement l'entrée et la sortie d'un canal de communication, $H(X|Y)$ représente l'incertitude restant sur le symbole émis une fois qu'on l'a reçu
- On a les propriétés suivantes :
  1. $\boxed{H(Y|X) \geq 0}$
  2. $\boxed{H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)}$
  3. $\boxed{H(X) \geq H(X|Y)}$
  4. $\boxed{H(X|X) = 0}$
- 1 est triviale : dans la suite, démonstration des autres

Démonstration propriété 2 : $H(X,Y) = H(X) + H(Y|X)$

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\log_2\left(p_{XY}(x,y)\right) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\log_2\left(p_{Y|X}(y|x)p_X(x)\right) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\log_2\left(p_{Y|X}(y|x)\right) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\log_2\left(p_X(x)\right) \\
&= H(Y|X) - \sum_{x\in\mathcal{X}}\log_2\left(p_X(x)\right)\sum_{y\in\mathcal{Y}} p_{XY}(x,y) \\
&= H(Y|X) + H(X)
\end{aligned}
$$

Interprétation propriété 2 : $H(X,Y) = H(X) + H(Y|X)$

- On démontrerait de la même façon que $H(X,Y) = H(X|Y) + H(Y)$
- Considérons deux sources X et Y : cette propriété nous dit que l'information fournie par les deux sources est la somme de l'information fournie par l'une des sources et de l'incertitude restante sur l'autre source

Interprétation propriété 2 : $H(X,Y) = H(X) + H(Y|X)$

Prévisions météo

X source qui nous dit s'il fera *soleil* ou *pluie*
Y source qui nous dit s'il fera *chaud* ou *froid*

X, Y	$Y =$ chaud	$Y =$ froid
$X =$ soleil	0.3	0.2
$X =$ pluie	0.1	0.4

$Y\|X$	$Y =$ chaud	$Y =$ froid
$X =$ soleil	0.6	0.4
$X =$ pluie	0.2	0.8

$H(X,Y) = 1.8464$ bits, $X(X) = 1$ bit, $H(Y|X) = 0.8464$ bits et $H(Y) = 0.9710$ bits

$$\underbrace{H(X,Y)}_{\text{information fournie par les deux sources}} = \underbrace{H(X)}_{\text{information fournie par } X} + \underbrace{H(Y|X)}_{\text{incertitude restante sur } Y}$$

On vérifie ici que si l'on connaît déjà X, Y donne moins d'information que si on avait aucune information a priori

Démonstration propriété 3 : $H(X|Y) \leq H(X)$

- On a démontré tout à l'heure que $H(X,Y) \leq H(X) + H(Y)$
- Comme on a également $H(X,Y) = H(X|Y) + H(Y)$, on a naturellement

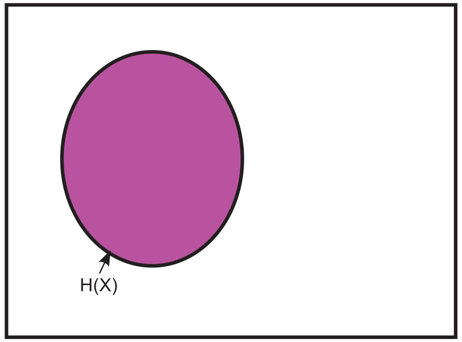
$$H(X|Y) \leq H(X)$$

- Cela signifie qu'une connaissance a priori ne peut que diminuer l'incertitude que l'on a, ce qui est conforme à notre intuition

Démonstration propriété 4 : $H(X|X) = 0$

- On a $p_{X|X}(x|x) = \frac{p_{XX}(x,x)}{p_X(x)} = 1$
- Il arrive naturellement que $H(X|X) = 0$
- Ceci est logique, car si on connaît déjà X, il ne reste plus aucune incertitude sur X !

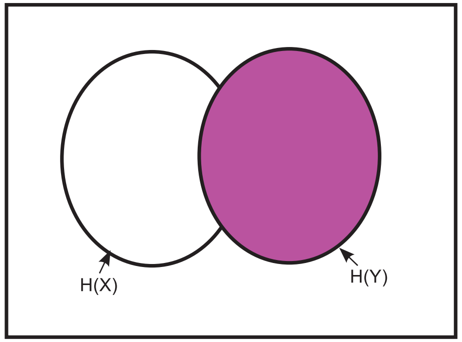
Bilan



$H(X)$: Entropie de X

- Information moyenne fournie par X
- Incertitude sur la variable aléatoire X

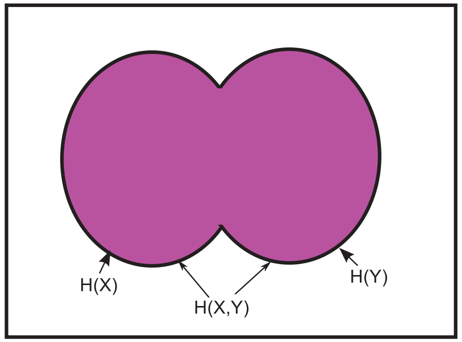
Bilan



$H(Y)$: Entropie de Y

- Information moyenne fournie par Y
- Incertitude sur la variable aléatoire Y

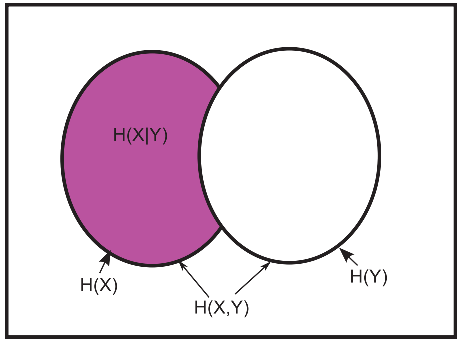
Bilan



$H(X, Y)$: Entropie conjointe de X et Y

- Information moyenne fournie par le système (X, Y)
- Attention, non égale à priori à la somme de $H(X)$ et $H(Y)$

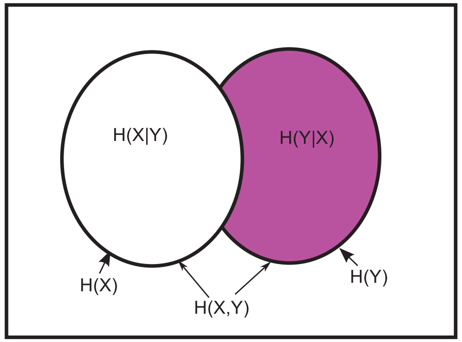
Bilan



$H(X|Y)$: Entropie conditionnelle de X sachant Y

- Incertitude restante sur X si l'on connaît Y
- Information manquante sur X malgré la connaissance de Y

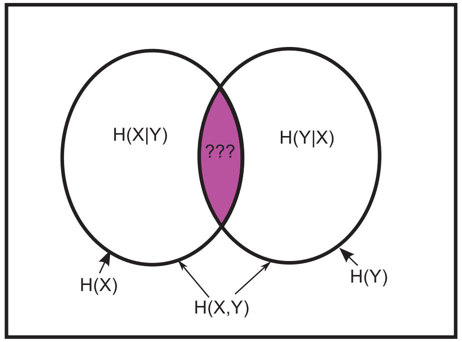
Bilan



$H(Y|X)$: Entropie conditionnelle de Y sachant X

- Incertitude restante sur Y si l'on connaît X
- Information manquante sur Y malgré la connaissance de X

Bilan



Il reste une quantité à considérer :

- Elle représente l'information contenue à la fois dans X et dans Y
- On l'appellera **information mutuelle** (ou transinformation)

Sommaire

Information mutuelle

Information mutuelle

Étant données deux variables aléatoires X et Y à valeurs respectivement dans $\mathcal{X}$ et $\mathcal{Y}$, on définit **l'information mutuelle** de X et Y (ou la transinformation) et on note $I(X;Y)$ la quantité :

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(X=x,Y=y)\log_2\left(\frac{P_{XY}(X=x,Y=y)}{P_X(X=x)P_Y(Y=y)}\right)$$

Interprétation

- On peut remarquer que : $I(X;Y) = D_{KL}(p_{XY}(x,y)||p_X(x)p_Y(y))$
- $I(X;Y)$ est donc un indicateur d'indépendance de variables aléatoires.
- Plus $p_{XY}(x,y)$ *ressemble* à $p_X(x)p_Y(y)$, plus $I(X,Y)$ sera faible et moins les variables seront corrélées.

Information mutuelle

On a les propriétés suivantes :

1. $\boxed{I(X;Y) = I(Y;X)}$
2. $\boxed{I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)}$
   Interprétation : l'information mutuelle correspond à la diminution du degré d'incertitude sur X due à Y (ou l'inverse)
3. $\boxed{I(X;Y) = H(X) + H(Y) - H(X,Y)}$
4. $\boxed{I(X;X) = H(X)}$
5. $\boxed{I(X;Y) \geq 0}$
   Égalité si et seulement si X et Y sont indépendantes

Ces résultats seront démontrés en exercice

Information mutuelle

Prévisions météo

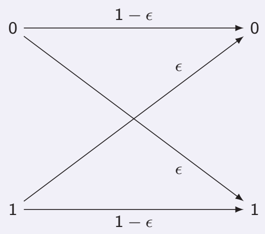
X source qui nous dit s'il fera *soleil* ou *pluie*
Y source qui nous dit s'il fera *chaud* ou *froid*

X,Y	$Y=$ chaud	$Y=$ froid
$X=$ soleil	0.3	0.2
$X=$ pluie	0.1	0.4

$H(X,Y) = 1.8464$ bits, $X(X) = 1$ bit, $H(Y) = 0.9710$ bits donc $I(X;Y) = 0.1246$ bits
Exprime le fait qu'il y a une dépendance entre X et Y et mesure l'information commune aux deux sources.

Exemple d'un canal

Canal binaire symétrique



On suppose que les symboles 0 et 1 sont équiprobables à l'émission

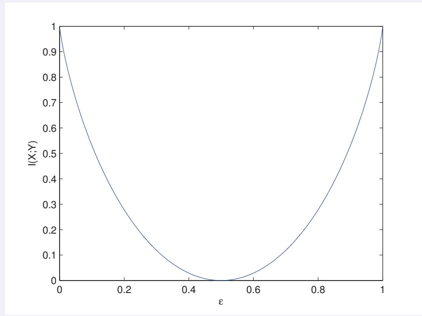
- $H(X) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$ bit
- $p_Y(0) = p_{Y|X}(0|0)p_X(0) + p_{Y|X}(0|1)p_X(1)$ $= \frac{1-\epsilon}{2} + \frac{\epsilon}{2} = \frac{1}{2}$
- $p_Y(1) = \frac{1}{2}$
- $H(Y) = 1$ bit

$$H(Y|X) = -\frac{1-\epsilon}{2}\log_2(1-\epsilon) - \frac{\epsilon}{2}\log_2(\epsilon) - \frac{1-\epsilon}{2}\log_2(1-\epsilon) - \frac{\epsilon}{2}\log_2(\epsilon)$$
$$I(X;Y) = 1 + (1-\epsilon)\log_2(1-\epsilon) + \epsilon\log_2(\epsilon)$$

Exemple d'un canal

Canal binaire symétrique (suite)

$$I(X;Y) = 1 + (1-\epsilon)\log_2(1-\epsilon) + \epsilon\log_2(\epsilon)$$



- Si $\epsilon = \frac{1}{2}$, il y a tellement d'erreurs qu'il n'y a plus de liens entre X et Y
- Si $\epsilon = 0$, $X = Y$ donc X et Y contiennent la même information : leur information mutuelle est donc maximale
- Si $\epsilon = 1$, X est exactement l'inverse de Y donc ils contiennent d'une certaine façon la même information
- Plus $I(X;Y)$ est faible, plus on aura du mal à retrouver la valeur de X à partir de Y