Théorie de l'information et codage **2010/2011**

Cours 4 — 8 mars 2011

Enseignant: Marc Lelarge *Scribe: Louis Jachiet*

Pour information

- Page web du cours
  `http://www.di.ens.fr/~lelarge/info11.html`

Propriétés de l'entropie et de l'information mutuelle

4.1 Rappel

$(X, Y) \sim p(x, y) = \mathbb{P}(X = x, Y = y)$

Définition 4.1.1 *entropie (jointe) :* $H(X, Y) = -\sum_{x\in X}\sum_{y\in Y} p(x, y) \log p(x, y)$ *(logarithme en base 2)*

entropie conditionnelle :

$$\begin{aligned} H(Y|X) &= \sum_{x\in X} p(x)H(Y|X = x) \\ &= -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x) \log p(y|x) \\ &= -\sum_{x\in X}\sum_{y\in Y} p(x, y) \log p(y|x). \end{aligned}$$

Théorème 4.1.1 ***règle de la chaîne***

$$H(X, Y) = H(X) + H(Y|X)$$

Démonstration. On utilise $p(x, y) = p(x)p(y|x)$ de telle sorte que :

$$\begin{aligned} H(X, Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x, y) \log p(x, y) \\ &= -\sum_{x\in X}\sum_{y\in Y} p(x, y) \log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x, y) \log p(y|x) \\ &= H(X) + H(Y|X). \end{aligned}$$

□

Corollaire 4.1.1

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$

Définition 4.1.2 *L'information mutuelle est définie par :*

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= D(p(x,y)||p(x)p(y)), \end{aligned}$$

avec $D(p||q) = \sum_{x\in X} p(x) log \frac{p(x)}{q(x)}$ *la distance de Kullbak-Leibler.*

4.2 Règles de la chaîne

Théorème 4.2.1 ***(pour l'entropie)*** *Soit* $(X_1, ..., X_n) \sim p(x_1, ..., x_n)$*, alors*

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

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

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

□

Définition 4.2.1 *l'information mutuelle conditionnelle des v.a X et Y étant donné Z est :* $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$.

Théorème 4.2.2 ***(pour l'information)***

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

Démonstration.

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

□

4.3 Inégalités de convexité

Théorème 4.3.1 *$D(p||q) \geq 0$ avec égalité ssi $\forall x p(x) = q(x)$.*

Démonstration. Vue en TD. □

Corollaire 4.3.1 *$I(X;Y) \geq 0$ avec égalité ssi X et Y sont indépendantes*

Démonstration. Il suffit d'observer que : $I(X;Y) = D(p(x,y)||p(x)p(y))$ □

Corollaire 4.3.2
- *$D(p(y|x)||q(y|x)) \geq 0$ avec égalité ssi $p(y|x) = q(y|x), \forall y, x$ tq $p(x) > 0$.*
- *$I(X;Y|Z) \geq 0$ avec égalité ssi X et Y sont indépendantes conditionnellement à Z.*

Théorème 4.3.2 *$H(X) \leq log|\mathcal{X}|$ où $|\mathcal{X}|$ = nombre d'éléments dans le support de X, c'est à dire le nombre de x tels que $p(x) > 0$ avec égalité ssi $X \sim Unif(\mathcal{X})$.*

Démonstration. Soit $u(x) = \frac{1}{|\mathcal{X}|}$ la distribution uniforme sur $\mathcal{X}$. On note alors que

$$D(p||u) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{u(x)} = log|\mathcal{X}| - H(X)$$

□

Théorème 4.3.3 *$H(X|Y) \leq H(X)$ avec égalité ssi X et Y sont indépendantes.*

Démonstration. $0 \leq I(X;Y) = H(X) - H(X|Y)$. □

Exemple 4.3.1:

Attention l'entropie conditionnelle diminue en moyenne mais on peut avoir $H(X|Y = Y) > H(X)$ pour des y particuliers, comme le montre l'exemple suivant :

$X \setminus Y$	1	2
1	0	$\frac{3}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

On a $H(X) = H(1/8) \approx 0.544$ bit. $H(X|Y = 1) = 0$ bit et $H(X|Y = 2) = 1$ bit. L'incertitude sur X augmente si $Y = 2$ est observée et elle diminue (fortement !) si $Y = 1$ est observée. En moyenne, on a :

$$H(X|Y) = 3/4H(X|Y = 1) + 1/4H(X|Y = 2) = 1/4 \leq H(X).$$

Théorème 4.3.4 *$H(X_1, ..., X_n) \leq \sum_{i=1}^{n} H(X_i)$ avec égalité ssi les X_i sont indépendantes.*

Démonstration.

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

avec égalité ssi X_i est indépendante de $X_{i-1}, \ldots X_1$, c'est à dire si les X_i sont indépendantes entre elles. □

4.3.1 Inégalité logsum et applications

Théorème 4.3.5 *pour tout, $a_1...a_n b_1...b_n \geq 0$, on a :*

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \sum_{i=1}^{n} a_i \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

avec égalité ssi $\frac{a_i}{b_i} = cst.$

Notations Dans le théorème précédent, on utilise les conventions : $0 \times \log 0 = 0$, $a \times \log \frac{a}{0} = \infty$ pour tout $a > 0$ et $0 \times \log \frac{0}{0} = 0$.

Démonstration. On suppose tout d'abord que pour tout i, $a_i > 0$ et $b_i > 0$. Soit $f(t) = t \log t$. On a $f'(t) = \frac{\log t}{+} \frac{1}{\log(e)}$ et $f''(t) = \frac{1}{t \log e} > 0$ pour tout $t > 0$. Donc f est strictement convexe

et $\sum \alpha_i f(t_i) \leq f(\sum \alpha_i t_i)$, pour $\alpha_i \geq 0$, $\sum \alpha_i = 1$, $t_i > 0$. Il suffit alors de prendre

$$\alpha_i = \frac{b_i}{\sum b_j} \text{ et, } t_i = \frac{a_i}{b_i}.$$

Si $a_j = 0$ et $b_j = 0$ alors avec les conventions choisies, le j-ème terme du membre de gauche est nul et on peut supprimer a_j et b_j à gauche et à droite.

Si $a_j > 0$ et $b_j = 0$, le terme de gauche vaut ∞ donc l'inégalité est toujours valable.

Si $a_j = 0$ et $b_j > 0$, on peut supprimer le j-ème terme dans le mebre de gauche et en supprimant b_j dans le terme de droite, on obtient bien un majorant :

$$\sum_{i \neq j} a_i \log \frac{a_i}{b_i} \geq \sum_{i \neq j} a_i \log \frac{\sum_{i \neq j} a_i}{\sum_{i \neq j} b_i} \geq \sum_{i=1}^{n} a_i \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}.$$

□

Lemme 4.3.1 *$D(p||q)$ est convexe en la paires (p, q), c'est à dire, pour tout $\lambda \in \{0, 1\}$*

$$D(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$$

Démonstration. On applique l'inégalité logsum à x fixé :

$$(\lambda p_1(x) + (1-\lambda)p_2(x)) log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \leq \lambda p_1(x) log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x)|x) log \frac{(1-\lambda)p_2(x)}{(1\lambda)q_2(x)}.$$

□

Théorème 4.3.6 *$H(X)$ est une fonction concave de $p(x)$.*

Démonstration. Il suffit de noter que : $H(X) = log|\mathcal{X}| - D(p||u)$, où u les distribution uniforme. □

Théorème 4.3.7 *Soit $(X, Y) \sim p(x, y)$. L'information mutuelle $I(X, Y)$ est une fonction concave de $p(x)$, à $p(y|x)$ fixée, et convexe en $p(y|x)$, à $p(x)$ fixée.*

Démonstration. $I(X; Y) = H(Y) - \sum_x p(x)H(Y|X = x)$

Si $p(y|x)$ est fixée alors $p(y)$ est une fonction linéaire de $p(x)$. Donc $H(Y)$ qui est concave en $p(y)$ est concave en $p(x)$. Le second terme est linéaire en $p(x)$ donc la différence est concave.

Maintenant $p(x)$ est fixée. A $p_1(y|x)$ et $p_2(y|x)$, on associe :

$$\begin{aligned} p_1(x, y) &= p_1(y|x)p(x) \text{ et, } p_1(y) \\ p_2(x, y) &= p_2(y|x)p(x) \text{ et, } p_2(y). \end{aligned}$$

On définit alors $p_\lambda(y|x) = \lambda p_1(y|x) + (1-\lambda)p_2(y|x)$ ainsi que $p_\lambda(x, y) = p(x)p_\lambda(y|x) = \lambda p_1(x, y) + (1-\lambda)p_2(x, y)$ et $p_\lambda(y)$. Finalement, soit $q_\lambda(x, y) = p(x)p_\lambda(y)$ de telle sorte que $q_\lambda(x, y) = \lambda q_1(x, y) + (1-\lambda)q_2(x, y)$ et

$$I(X; Y) = D(p_\lambda(x, y)||q_\lambda(x, y)),$$

comme $D(p||q)$ est convexe en (p, q), $I(X, Y)$ est convexe en $p(y|x)$ à $p(x)$ fixée. □

4.4 Data Processing inequality

Définition 4.4.1 *(X, Y, Z) est une chaîne de Markov si $p(z|x, y) = p(z|y)$. On a donc $p(x, y, z) = p(x)p(y|x)p(z|y)$. On notera parfois $X \to Y \to Z$.*

Remarque 4.4.1 $p(x, z|y) = \dfrac{p(x, y, z)}{p(y)} = \dfrac{p(x, y)}{p(y)} \times p(z|y) = p(x|y)p(z|y)$ *donc $X \to Y \to Z$ ssi X et Z sont conditionnellement indépendantes étant donné Y donc $Z \to Y \to X$.*

Théorème 4.4.1 *Si $X \to Y \to Z$ alors* $\max\{I(X;Y), I(X;Z)\} \geq I(X;Z)$.

Démonstration.

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y). \end{aligned}$$

Comme X et Z sont conditonnellement indépendantes, on a $I(X;Z|Y) = 0$. Comme $I(X;Y|Z) \geq 0$, on a $I(X;Y) \geq I(X;Z)$. L'autre inégalité provient du fait que $Z \to Y \to X$ et de la symétrie de l'information mutuelle. □

Corollaire 4.4.1 *Si $X \to Y \to Z$ alors $I(X;Y|Z) \leq I(X;Y)$.*

Démonstration. En reprenant la décomposition de $I(X;Y,Z)$ ci-dessus, il suffit de noter que $I(X;Z) \geq 0$. □

4.5 Inégalité de Fano

On veut estimer $X \sim p(x)$ (p(x)>0 pour tout $x \in \mathcal{X}$) en observant Y avec $p(y|x)$. Soit $\hat{X}$ un estimateur de X, c'est à dire tel que $X \to Y \to \hat{X}$.

Théorème 4.5.1 *Pour tout $\hat{X}$ tel que $X \to Y \to \hat{X}$, soit $P_e = P(X \neq \hat{X})$ alors on a*

$$H(P_e) + P_e log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y),$$

où $H(p) = -p \log p - (1-p)\log(1-p)$.

Remarque 4.5.1 *$P_e = 0 \to H(X|Y) = 0 \to Y$ est une fonction de X (cf. exo)*

Démonstration. On définit :

$$E = \begin{cases} 1 & \text{si } \hat{X} \neq X \\ 0 & \text{sinon} \end{cases}$$

On a :

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) \\ &= H(E|\hat{X}) + H(X|E, \hat{X}). \end{aligned}$$

Comme $H(E|X, \hat{X}) = 0$ et $H(E|\hat{X}) \leq H(E) = H(P_e)$, et de plus,

$$\begin{aligned} H(X|E, \hat{X}) &= P(E = 0)H(X|\hat{X}, E = 0) + P(E = 1)H(X|\hat{X}, E = 1) \\ &\leq (1 - P_e) \times 0 + P_e \log(|\mathcal{X}| - 1), \end{aligned}$$

au final, on obtient :

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}).$$

Par l'inégalité 'data processing', on a : $I(X; \hat{X}) \leq I(X; Y)$ et donc $H(X|\hat{X}) \geq H(X|Y)$. □

4.6 Lien entre théorie de l'information et PMU

Objectif : définir une stratégie 'optimale' pour jouer aux courses.

Problème : M chevaux concurrents, p_i proba le $i$ème cheval gagne. Si le $i$ème cheval gagne, on gagne $o(i)$ fois sa mise. C'est à dire, si on a misé 1 euro, on obtient $o(i)$ euro si i gagne et rien sinon (la mise de 1 euro est perdue).

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

Après n courses, la fortune du joueur est (en supposant qu'initialement $S_0 = 1$), $S_n = \prod_{i=1}^{n} S(X_i)$ avec $S(X) = b(X)o(X)$ et X est le cheval vainqueur.

Théorème 4.6.1 *Si les X_i sont i.i.d. $\sim p(x)$ alors la fortune du joueur utilisant la stratégie b "croit" exponentiellement : $lim\frac{1}{n}log_2 S_n = W(b, p) = \sum_{k=1}^{m} p_k \log_2(b(k)o(k))$ où la convergence est en probabilité.*

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

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

Théorème 4.6.2 ***(Critère de Kelly)*** *La stratégie optimale est de prendre $b^* = p$ et dans ce cas $W(b^*, p) = \sum p_i \log o(i) - H(p)$ (optimum doubling rate)*

Démonstration. Il suffit d'écire :

$$W(b, p) = \sum p_i log o(i) - H(p) - D(p||b).$$

□

Dans le cas équitable : $\sum_i \frac{1}{o(i)} = 1$ alors $r_i = \frac{1}{o(i)}$ est une distribution de probabilité qui correspond à l'estimée de la probabilité p_i par le bookmaker. On a alors

$$W(b,p) = D(p||r) - D(p||b)$$

Si notre estimée b de p est meilleure que celle du bookmaker 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.