Théorie de l'Information

Philippe Duchon

ENSEIRB

2007-08

"Théorie de l'information"...

- Définir proprement la *quantité d'information* apportée par la connaissance d'un fait
- Idée (très) informelle : c'est le *nombre de questions à réponse oui/non que cette connaissance nous évite d'avoir à poser*

Ce dont on va parler...

- **Notions mathématiques :** entropie, information mutuelle
- **Codage et décodage :** codes, déchiffrabilité, efficacité d'un code
- **Transmission de l'information :** comment transmettre correctement une information en présence d'erreurs de transmission
- **Compression des données :** essentiellement, compression sans pertes ("conservative")
- Le langage mathématique est celui de la **théorie des probabilités :** variables aléatoires et processus discrets (élémentaires)

Ce dont on ne va *pas* parler. . .

- **Cryptographie :** cours en deuxième année
- On s'intéresse à transmettre de l'information, pas à la cacher ou à la protéger
- Les erreurs sont le fait d'imperfection du système de transmission, pas d'interférences malveillantes d'un adversaire

Historiquement

- **Rudolf Clausius** (1822-1888) : définit l'entropie comme une mesure du *désordre* d'un système
- **Ludwig Boltzmann** (1844-1906) : l'entropie d'un état macroscopique est proportionnelle au *logarithme* du nombre d'états microscopiques correspondants
- **Ronald Fisher** (1890-1962) : utilise le mot *information* dans un contexte mathématique
- **Harry Nyquist** (1889-1976) (bruit, fréquence d'échantillonnage), **Ralph Hartley** (1888-1970)
- **Claude Shannon** (1916-2001), premiers théorèmes sur l'information en théorie de la communication

Quantité d'information

- $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé, $A \in \mathcal{A}$ un événement (de probabilité non nulle)
- On définit la **quantité d'information** apportée par la réalisation de A (**self-information**), la quantité

$$I(A) = -\log \mathbb{P}(A).$$

- On utilise quasi systématiquement le *logarithme à base* 2, l'unité d'information est alors le **bit**

Événements indépendants

- Si (et seulement si) les événements A et B sont **indépendants**,

$$I(A \cap B) = I(A) + I(B).$$

- "La quantité d'information apportée par la *réalisation conjointe* de deux événements indépendants, est égale à la *somme* de celles apportées séparément par chacun d'eux."

Retour sur la définition

La propriété d'additivité en cas d'indépendance implique l'usage de la fonction logarithme :

Proposition

Si on souhaite définir la quantité d'information par une fonction positive de la probabilité : $I(A) = f(\mathbb{P}(A))$, *avec* $f : (0,1) \to \mathbb{R}^+$, *avec*

- f *continue*
- $I(A \cap B) = I(A) + I(B)$ *dès que* A *et* B *sont indépendants*

alors I *est forcément de la forme*

$$I(A) = -C \log(\mathbb{P}(A))$$

avec $C > 0$ *une constante.*

Système simple

Définition (Système simple)

*On appelle **système simple**, soit une variable aléatoire discrète X, soit la donnée d'une partition $\Omega = A_1 + A_2 + \cdots + A_n$ d'un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$.*

Les deux définitions sont fonctionnellement équivalentes (on ne s'intéresse pas aux valeurs de la variable aléatoire).

Notation : $X = \{A_i\}_{1 \leq i \leq n}$, $p_i = \mathbb{P}(A_i)$

Entropie : un exemple

L'enseignant choisit aléatoirement un élève parmi les présents, et on considère la variable aléatoire X qui désigne le *rang* où se trouve assis l'élève choisi.

- Que vaut l'information "l'élève choisi est au premier rang" ?
- Que vaut l'information "l'élève choisi est au rang x" ? (ça *dépend* de x, sauf si tous les rangs comptent le même nombre d'élèves)
- Que vaut **en moyenne** l'information du rang auquel se trouve l'élève choisi ? **(notion d'entropie)**

Entropie d'un système simple

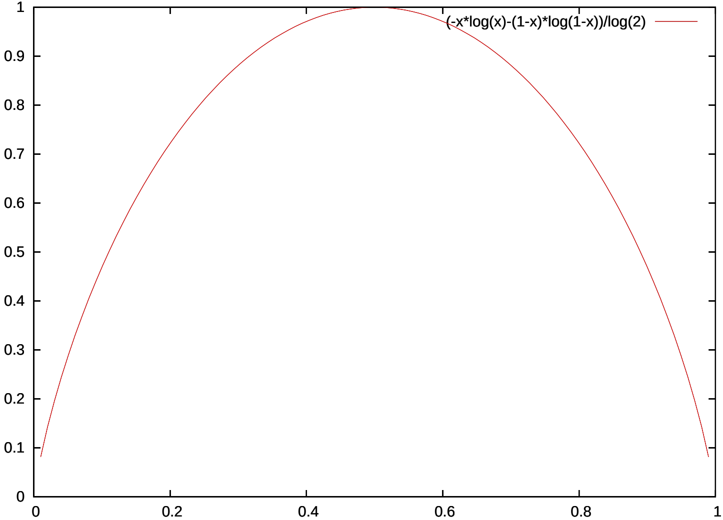
Définition (Entropie d'un système simple)

On appelle ***entropie du système*** X*, l'espérance de la quantité d'information apportée par la réalisation d'un événement du système.*

$$H(X) = \mathbb{E}(I(X)) = \sum_{A \in X} \mathbb{P}(A) I(A) = -\sum_i p_i \log(p_i).$$

(convention : pour $x = 0$, $x \log(x) = 0$)

Entropie d'un système de deux événements



Entropie des lettres du français

- On considère un alphabet de 29 symboles : 26 lettres, l'espace, le point et "les autres signes de ponctuation".
- Si on accorde la même probabilité 1/29 à chaque symbole : $H(U_{29}) = \log_2(29) \simeq 4.848$ bits.
- Si on prend des probabilités plus proches de la fréquence empirique des lettres dans la langue française :

E, espace	0.14	I, N, T	0.08
A, R, S	0.06	D, L, O, U	0.04
C, M, P	0.02	ponctuation	0.018
point	0.014	12 autres lettres	0.004

on obtient $H(X) \simeq 4.05$ bits, soit un peu plus de 80% de l'entropie du système uniforme (l'entropie est peu sensible à des variations autour de l'équiprobabilité)

Lemme de Gibbs

Le "lemme de Gibbs" permet, entre autres, de prouver que l'entropie est maximale pour une répartition uniforme des probabilités.

Lemme (Gibbs)

Soient $\mathbf{p} = (p_1, \ldots, p_n)$ *et* $\mathbf{q} = (q_1, \ldots, q_n)$ *deux vecteurs de probabilités* ($p_i \geq 0$, $q_i \geq 0$, $\sum_i p_i = \sum_i q_i = 1$).
Alors

$$-\sum_{i=1}^{n} p_i \log(p_i) \leq -\sum_{i=1}^{n} p_i \log(q_i),$$

avec égalité si et seulement si $p_i = q_i$ *pour tout* i.

Entropie maximale d'un système à n symboles

Proposition

Si X est un système simple à n symboles, alors

$$H(X) \leq \log_2(n),$$

avec égalité si et seulement si $p_i = 1/n$ pour tout $i, 1 \leq i \leq n$.

Système composé

Définition

*On appelle **système composé**, soit la donnée de deux (ou plus) variables aléatoires discrètes (X, Y) sur un même espace probabilisé, soit un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ muni de deux (ou plus) partitions :*

$$\Omega = A_1 + \cdots + A_m = B_1 + \cdots + B_n.$$

Comme pour un système simple, les deux définitions sont fonctionnellement équivalentes, dans la mesure où on ne s'intéresse pas aux valeurs des variables aléatoires.

On utilise les notations suivantes, moyennement cohérentes mais consacrées par l'usage :

$$p_i = \mathbb{P}(A_i) \quad q_j = \mathbb{P}(B_j) \quad p_{ij} = \mathbb{P}(A_i \cap B_j).$$

Exemples de systèmes composés

Les systèmes composés prennent leur sens lorsque les variables aléatoires qui les composent ne sont pas indépendantes.
Quelques exemples de situations pouvant être modélisées de la sorte :

- deux variables relatives à l'état d'un même système, l'une pouvant être "facile à mesurer" alors que l'autre l'est moint ;
- deux symboles émis consécutivement par une même source ;
- l'une des variables peut représenter le symbole réellement émis par une source, tandis que l'autre représente le symbole reçu après transmission sur un canal bruité.

Entropie totale, information mutuelle, entropie mutuelle

Soit $(X.Y)$ un système composé. On définit les quantités suivantes :

Définition (Entropie totale)

*L'**entropie totale** du système est celle définie par la partition formée de toutes les intersections de parties des deux partitions :*

$$H(X,Y) = \mathbb{E}(I(XY)) = -\sum_{i,j} p_{ij} \log(p_{ij})$$

Ici, XY doit être compris comme désignant la partition de Ω formée de toutes les intersections d'un événement de la première partition avec un événement de la seconde :

$$XY = \{A_i \cap B_j : 1 \leq i \leq m, 1 \leq j \leq n\}$$

Entropie totale, information mutuelle, entropie mutuelle

Soit $(X.Y)$ un système composé. On définit les quantités suivantes :

Définition (Information mutuelle)

*L'**information mutuelle** entre deux événements A_i et B_j, la "part commune"*

$$I(A_i, B_j) = I(A_i) + I(B_j) - I(A_i \cap B_j) = \log\left(\frac{p_{ij}}{p_i q_j}\right).$$

*L'**information mutuelle** entre les deux variables X et Y, l'espérance de l'information mutuelle entre événements :*

$$I(X, Y) = \sum_{A \in X} \sum_{B \in Y} \mathbb{P}(A \cap B) I(A, B) = \sum_{i,j} p_{ij} \log\left(\frac{p_{ij}}{p_i q_j}\right).$$

Liens entre les différentes quantités

On a une première relation de décomposition (ce n'est rien d'autre que la *linéarite de l'espérance* appliquée à la définition de l'information mutuelle entre événements) :

Première décomposition de l'information mutuelle

$$H(X,Y) = H(X) + H(Y) - I(X,Y)$$

Signe de l'information mutuelle

Remarque

Il est important de noter que $I(A, B)$ peut être de signe quelconque :

- positif, si A et B "se ressemblent" (cas limite, $A = B$)
- négatif, si A et B sont "presque incompatibles" (cas limite, A et B disjoints, l'information mutuelle vaut $-\infty$)
- nul, si A et B sont indépendants.

En revanche, on verra que $I(X, Y)$ est toujours positif ou nul.

Entropie mutuelle de deux variables aléatoires

Théorème

Soit (X, Y) un système composé.

- $I(X, Y) \geq 0$, avec égalité si et seulement si X et Y sont deux variables aléatoires indépendantes ;
- de manière équivalente, $H(X, Y) \leq H(X) + H(Y)$, avec égalité si et seulement si X et Y sont deux variables aléatoires indépendantes.

Entropie mutuelle de deux variables aléatoires

Théorème

Soit (X, Y) un système composé.

- $I(X, Y) \geq 0$, avec égalité si et seulement si X et Y sont deux variables aléatoires indépendantes ;
- de manière équivalente, $H(X, Y) \leq H(X) + H(Y)$, avec égalité si et seulement si X et Y sont deux variables aléatoires indépendantes.

Si vous n'avez pas eu la preuve du théorème au tableau, la réclamer

Entropie conditionnelle

- On part des probabilités conditionnelles :

$$\mathbb{P}(B_j|A_i) = \frac{\mathbb{P}(A_i \cap B_j)}{\mathbb{P}(A_i)} = \frac{p_{ij}}{p_i}$$

Entropie conditionnelle

- On part des probabilités conditionnelles :

$$\mathbb{P}(B_j|A_i) = \frac{\mathbb{P}(A_i \cap B_j)}{\mathbb{P}(A_i)} = \frac{p_{ij}}{p_i}$$

- On peut réécrire la définition de l'information mutuelle :

$$I(A_i, B_j) = \log\left(\frac{p_{ij}}{p_i q_j}\right) = -\log(q_j) + \log\left(\frac{p_{ij}}{p_i}\right)$$

Entropie conditionnelle

- On part des probabilités conditionnelles :

$$\mathbb{P}(B_j|A_i) = \frac{\mathbb{P}(A_i \cap B_j)}{\mathbb{P}(A_i)} = \frac{p_{ij}}{p_i}$$

- On peut réécrire la définition de l'information mutuelle :

$$I(A_i, B_j) = \log\left(\frac{p_{ij}}{p_i q_j}\right) = -\log(q_j) + \log\left(\frac{p_{ij}}{p_i}\right)$$

- On définit l'information conditionnelle
$I(B_j|A_i) = -\log(p_{ij}/p_i)$

Entropie conditionnelle

- On part des probabilités conditionnelles :

$$\mathbb{P}(B_j|A_i) = \frac{\mathbb{P}(A_i \cap B_j)}{\mathbb{P}(A_i)} = \frac{p_{ij}}{p_i}$$

- On peut réécrire la définition de l'information mutuelle :

$$I(A_i, B_j) = \log\left(\frac{p_{ij}}{p_i q_j}\right) = -\log(q_j) + \log\left(\frac{p_{ij}}{p_i}\right)$$

- On définit l'information conditionnelle $I(B_j|A_i) = -\log(p_{ij}/p_i)$
- On a ainsi la décomposition (pour l'information) :

$$I(A_i, B_j) = I(B_j) - I(B_j|A_i)$$

Entropie conditionnelle

- On part des probabilités conditionnelles :

$$\mathbb{P}(B_j|A_i) = \frac{\mathbb{P}(A_i \cap B_j)}{\mathbb{P}(A_i)} = \frac{p_{ij}}{p_i}$$

- On peut réécrire la définition de l'information mutuelle :

$$I(A_i, B_j) = \log\left(\frac{p_{ij}}{p_i q_j}\right) = -\log(q_j) + \log\left(\frac{p_{ij}}{p_i}\right)$$

- On définit l'information conditionnelle
$I(B_j|A_i) = -\log(p_{ij}/p_i)$
- On a ainsi la décomposition (pour l'information) :

$$I(A_i, B_j) = I(B_j) - I(B_j|A_i)$$

- Ou encore :

$$I(B_j) = I(A_i, B_j) + I(B_j|A_i)$$

Interprétations des informations mutuelle/conditionnelle

- L'information $I(B_j)$ s'interprète comme la quantité d'information obtenue si l'on apprend que B_j se réalise.

Interprétations des informations mutuelle/conditionnelle

- L'information $I(B_j)$ s'interprète comme la quantité d'information obtenue si l'on apprend que B_j se réalise.
- L'information mutuelle $I(A_i, B_j)$ s'interpète comme la quantité d'information **sur** B_j apportée par la (connaissance de la) réalisation de A_i (ou, symétriquement, l'information sur A_i apportée par la réalisation de B_j)

Interprétations des informations mutuelle/conditionnelle

- L'information $I(B_j)$ s'interprète comme la quantité d'information obtenue si l'on apprend que B_j se réalise.
- L'information mutuelle $I(A_i, B_j)$ s'interpète comme la quantité d'information **sur** B_j apportée par la (connaissance de la) réalisation de A_i (ou, symétriquement, l'information sur A_i apportée par la réalisation de B_j)
- L'information conditionnelle $I(B_j|A_i)$ s'interprète comme la quantité d'information "résiduelle" obtenue lorsque l'on apprend que B_j se réalise, alors qu'on savait déjà que A_i se réalise (et donc, la probabilité perçue de B_j était la probabilité conditionnelle)

Entrope conditionnelle

Définition : entropie conditionnelle

On appelle **entropie conditionnelle** de Y sachant X, l'espérance mathématique (sur toutes les réalisations possibles de X et Y) de l'information conditionnelle de la réalisation de Y sachant celle de X :

$$
\begin{aligned}
H(Y|X) &= \mathbb{E}(I(Y|X)) = \sum_{A\in X}\sum_{B\in Y}\mathbb{P}(A\cap B)I(B|A) \\
&= -\sum_{i,j} p_{ij}\log\left(\frac{p_{ij}}{p_i}\right)
\end{aligned}
$$

D'autres décompositions

- $H(X, Y) = H(X) + H(Y|X)$ (l'information complète sur (X, Y) est la somme de l'information sur X seule, et de l'information résiduelle sur Y lorsque X est connue)

D'autres décompositions

- $H(X, Y) = H(X) + H(Y|X)$ (l'information complète sur (X, Y) est la somme de l'information sur X seule, et de l'information résiduelle sur Y lorsque X est connue)
- $H(X, Y) = H(Y) + H(X|Y)$

D'autres décompositions

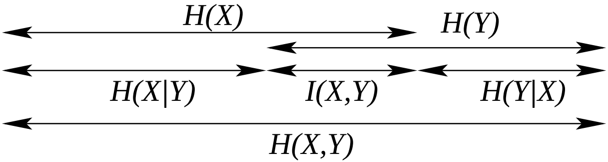
- $H(X, Y) = H(X) + H(Y|X)$ (l'information complète sur (X, Y) est la somme de l'information sur X seule, et de l'information résiduelle sur Y lorsque X est connue)
- $H(X, Y) = H(Y) + H(X|Y)$
- $H(X) = I(X, Y) + H(X|Y)$ (l'information sur X seule est la somme de l'information commune à X et Y, et de l'information résiduelle sur X sachant Y)

D'autres décompositions

- $H(X, Y) = H(X) + H(Y|X)$ (l'information complète sur (X, Y) est la somme de l'information sur X seule, et de l'information résiduelle sur Y lorsque X est connue)
- $H(X, Y) = H(Y) + H(X|Y)$
- $H(X) = I(X, Y) + H(X|Y)$ (l'information sur X seule est la somme de l'information commune à X et Y, et de l'information résiduelle sur X sachant Y)
- $H(Y) = I(X, Y) + H(Y|X)$

D'autres décompositions

- $H(X, Y) = H(X) + H(Y|X)$ (l'information complète sur (X, Y) est la somme de l'information sur X seule, et de l'information résiduelle sur Y lorsque X est connue)
- $H(X, Y) = H(Y) + H(X|Y)$
- $H(X) = I(X, Y) + H(X|Y)$ (l'information sur X seule est la somme de l'information commune à X et Y, et de l'information résiduelle sur X sachant Y)
- $H(Y) = I(X, Y) + H(Y|X)$
- Mnémotechniquement :



Remarque : entropie conditionnelle nulle

- On sait à quoi correspond le cas $I(X, Y) = 0$: X et Y indépendants ;

Remarque : entropie conditionnelle nulle

- On sait à quoi correspond le cas $I(X, Y) = 0$: X et Y indépendants ;
- Peut-on interpréter le cas "complémentaire" $H(Y|X) = 0$, *i.e.* $I(X, Y) = H(Y)$?

Remarque : entropie conditionnelle nulle

- On sait à quoi correspond le cas $I(X,Y)=0$: X et Y indépendants ;
- Peut-on interpréter le cas "complémentaire" $H(Y|X)=0$, *i.e.* $I(X,Y)=H(Y)$?
- $H(Y|X) = -\sum_{i,j} p_{ij} \log(p_{ij}/p_i)$ est **nul** si et seulement si les (i,j) tels que $p_{ij}>0$ donnent tous $p_{ij}=p_i$ (sinon un des logarithmes sera strictement négatif)

Remarque : entropie conditionnelle nulle

- On sait à quoi correspond le cas $I(X,Y)=0$: X et Y indépendants ;
- Peut-on interpréter le cas "complémentaire" $H(Y|X)=0$, *i.e.* $I(X,Y)=H(Y)$?
- $H(Y|X)=-\sum_{i,j} p_{ij}\log(p_{ij}/p_i)$ est **nul** si et seulement si les (i,j) tels que $p_{ij}>0$ donnent tous $p_{ij}=p_i$ (sinon un des logarithmes sera strictement négatif)
- c'est-à-dire si, pour tout i (tel que $p_i>0$), il existe un j (forcément **unique**) tel que $p_{ij}=p_i$

Remarque : entropie conditionnelle nulle

- On sait à quoi correspond le cas $I(X, Y) = 0$: X et Y indépendants ;
- Peut-on interpréter le cas "complémentaire" $H(Y|X) = 0$, *i.e.* $I(X, Y) = H(Y)$?
- $H(Y|X) = -\sum_{i,j} p_{ij} \log(p_{ij}/p_i)$ est **nul** si et seulement si les (i, j) tels que $p_{ij} > 0$ donnent tous $p_{ij} = p_i$ (sinon un des logarithmes sera strictement négatif)
- c'est-à-dire si, pour tout i (tel que $p_i > 0$), il existe un j (forcément **unique**) tel que $p_{ij} = p_i$
- c'est-à-dire si Y s'écrit comme une fonction déterministe de X (la connaissance de X donne à coup sûr la valeur de Y).

Exercice : entropie et décision

Une naissance a eu lieu un jour de la semaine, sur lequel on n'a aucune information. On a le choix de recevoir la réponse à une question :

- Q_1 : "la naissance a-t-elle eu lieu (a) lundi, mardi ou mercredi, ou (b) jeudi, vendredi, samedi ou dimanche ?"
- Q_2 : "la naissance a-t-elle eu lieu (a) samedi, ou (b) dimanche, ou (c) un autre jour ?"
- Q_3 : "la naissance a-t-elle eu lieu (a) lundi ou mardi, ou (b) mercredi ou jeudi, ou (c) vendredi, samedi ou dimanche ?"

Décrire la situation par trois systèmes composés, et proposer un critère de choix de la question à poser.

Exercice : prévisions météorologiques (janvier 2005)

On teste un système de prévisions météorologiques pour lequel on a obtenu, sur un an, les fréquences de résultats suivantes :

	Temps : Pluie	Temps : Soleil
Prévu : Pluie	$1/12$	$1/6$
Prévu : Soleil	$1/12$	$2/3$

- Quelle est la probabilité que le système donne une prévision incorrecte ?
- Calculer l'information mutuelle entre le temps prévu et le temps effectif.
- Comparer à un système de "prévisions" qui prédit systématiquement du soleil. Commenter.

Exercice : test de dépistage

On considère une situation dans laquelle un test de dépistage pour une maladie est censé discriminer entre les situations $X \in \{V, nV\}$, en donnant un résultat $Y \in \{T^+, T^-\}$. Les probabilités respectives des différents cas sont les suivantes :

	T^+	T^-
V	0.07	0.01
nV	0.03	0.89

On définit l'*efficacité* du test comme étant $r = \frac{I(X,Y)}{H(X)}$.

- Quelles valeurs peut, *a priori*, prendre r ?
- Calculer $H(X)$, $I(X, Y)$ et r.
- À quoi correspondraient des valeurs $r = 0$? $r = 1$?