Théorie de l'information

D. Kokel

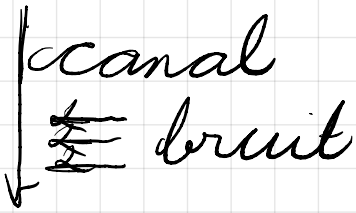
06/01/2022

Théorie des codes

Probabilité conditionnelle

Motivation

X = emetteur de symboles, messages



Y = message reçu

où X et Y sont des espaces de proba.

Si un mot x est transmis (avec proba $p(x)$), quel est la proba que y soit reçu? On la note :

$$p(\ |x) : Y \longrightarrow [0,1]$$

où $p(y|x)$ est la proba que y soit reçu, étant donné que x soit transmis.

On définit une <u>proba jointe</u> par

$$p : X \times Y \longrightarrow [0,1]$$
$$(x,y) \longmapsto p(x)\,p(y|x) =: p(x,y)$$

Nous avons également une <u>proba sur Y</u>, induit par $p = p_X : X \to [0,1]$ et $p(\ |x)$:

$$p_Y : Y \longrightarrow [0,1]$$
$$y \longmapsto p_Y(y) = \sum_{x \in X} p(x)\,p(y|x)$$

Exercice Montrer que la proba jointe $p : X \times Y \longrightarrow [0,1]$ et la proba p_Y sont des probabilités, étant données p_X et des probas conditionelles $p(\ |x) : Y \longrightarrow [0,1]$, pour $x \in X$, $p(x) \neq 0$.

On peut également définir

$$p(\ |y) : X \longrightarrow [0,1]$$

et

$$p(\ |x) : y \longrightarrow [0,1]$$

en fonction d'une proba jointe, pour tout $x \in X$ et $y \in Y$ avec $p_X(x) \neq 0$ et $p_Y(y) \neq 0$.

Remarque. Etant donnée

$$p : X \times Y \longrightarrow [0,1]$$

on en déduit

$$p_X : X \longrightarrow [0,1] \text{ par } p_X(x) = \sum_{y \in Y} p(x,y)$$

et

$$p_Y : Y \longrightarrow [0,1] \text{ par } p_Y(y) = \sum_{x \in X} p(x,y)$$

Inversement,

$$p_X \text{ et } p(\ |x) \Longleftrightarrow p(x,y) = p_X(x)\, p(y|x)$$

$$\forall x \in X,\ p_X(x) \neq 0$$

Entropie conditionelle

Nous avons vu que

$$0 \leqslant H(X) \leqslant \log_2 |X|,$$

avec l'interprétation de $H(X)$ comme le contenu de l'info dans X. L'objectif est de définir le contenu de l'info dans Y, après avoir pris connaissance de (x dans) X. C'est l'entropie conditionnelle ($H(Y|x)$ ou) $H(Y|X)$.

Défn Pour $x \in X$ on définit

$$H(Y|x) = \sum_{y \in Y} p(y|x) \log_2 p(y|x)^{-1}$$

et l'entropie conditionnelle

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|x) = \mathbb{E}(H(Y|x)),$$

où on voit

$$H(Y|x) : X \longrightarrow \mathbb{R},$$

une variable aléatoire.

Défn L'information mutuelle est

$$I(X,Y) = \sum_{x,y} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right).$$

$p(y|x)$

En utilisant les identités :

$$p(x,y) = p(x)\,p(y|x) = p(y)\,p(x|y),$$

on en déduit :

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

Preuve

$$I(X,Y) = \sum_{x,y} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

$$= \sum_{x,y} p(x)\,p(y|x) \log_2 \left(\frac{p(y|x)}{p(y)} \right)$$

car

$$\begin{cases} = \sum_y \left(\sum_x p(x)\,p(y|x) \log_2 \left(p(y)^{-1} \right) \right) \\ - \sum_x p(x) \left(\sum p(y|x) \log_2 \left(p(y|x)^{-1} \right) \right) \end{cases}$$

$$\log_2 \left(\frac{p(y|x)}{p(y)} \right) = -\log_2 \left(p(y|x)^{-1} \right) + \log \left(p(y)^{-1} \right)$$

$$= \sum_y \left(\sum_x p(x)\,p(y|x) \right) \log_2 \left(p(y)^{-1} \right)$$

$$- \sum_x p(x)\,H(y|x)$$

$$= \sum_y p(y) \log_2 \left(p(y)^{-1} \right) - H(y|x)$$

$$= H(y) - H(y|x). \qquad \square$$

Par symétrie on $= H(x) - H(x|y).$

Rappel : $I(X,Y) = H(X) - H(X|Y)$
(Lemme) $= H(Y) - H(Y|X)$.

Remarque

$$0 \leqslant H(Y|X) \leqslant H(Y).$$

Si $H(Y|X) = 0$, on a $I(X,Y) = H(Y)$, et on dit que Y est **dépendent** de X.

Si $H(Y|X) = H(Y)$, donc $I(X,Y) = 0$, on dit que Y est **indépendent** de X.

* **Exercice** Dans le cas $I(X,Y) = 0$, montrer que $p(x,y) = p(x)\,p(y)$, pour tout x et y.

Inversement, si $p(x,y) = p(x)\,p(y)$, alors

$$I(X,Y) = \sum_{x,y} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)\,p(y)} \right)$$

$$= \sum_{x,y} p(x,y) \log_2 1 = 0.$$

Conclusion.

X et Y sont indépendents ($I(X,Y) = 0$) ssi $p(x,y) = p(x)\,p(y)$.

* Besoin du lemme de Gibbs (pas fait).

Capacité du canal de transmission

Pour un canal (avec bruit) ayant probabilité conditionelle $p(\ |x): Y \longrightarrow [0,1]$, on définit la capacité du canal

$$C = \max_{p: X \to [0,1]} I(X,Y).$$

$\leq \log_2 |X|$.

$\leq \max_{p: X \to [0,1]} H(X)$

Remarques. On peut supposer que le canal est *sans mémoire*, i.e. que la fonction $p(\ |x)$ ne change pas en fonction du symbole précédent. Sinon la définition la capacité est plus compliquée.

Ici on peut supposer que $X = Y$, en tant qu'ensembles, mais leurs fonctions de probabilité peuvent être différentes à cause de bruit.

On dit que le canal est *symétrique* si $p(y|x) = \varepsilon$ pour tout $y \neq x$. Alors $p(x|x) = 1 - (n-1)\varepsilon$ si $|y| = n$.

Théorème (Shannon, Canal bruyant)
Soit donné un canal avec bruit (i.e. $p(\ |x)$) et un $\delta > 0$. Pour tout $R < C$* (la capacité), il existe un codage qui permet la transmission de l'information avec moins de δ proba d'erreur.

Idée. Un canal avec bruit a une capacité intrinsèque et on doit chercher un code adapté à ce canal pour transmission fiable (avec rendement maximum C*).

* Si $X = \{0,1\}$ on doit normaliser la capacité par $1/\log_2 |X|$.

Canal binaire symétrique sans mémoire
On va étudier un canal **binaire** ($X = \{0,1\}$) **symétrique sans mémoire**.
On suppose que X émet des symboles successifs, et que la proba $p(\ |x)$ est

stationnaire, i.e. elle ne change pas avec le temps.

$p(y\|x)$		0	1	$= x$
y	0	$1-\varepsilon$	ε	
	1	ε	$1-\varepsilon$	

Le canal est sans bruit si $\varepsilon = 0$.

Défn. On pose

$$H(\varepsilon) = \varepsilon \log_2 \varepsilon^{-1} + (1-\varepsilon)\log_2(1-\varepsilon)^{-1}.$$

On note que $H(\varepsilon)$ est l'entropie d'un ensemble $\mathcal{X} = \{0,1\}$ avec $p(0) = \varepsilon$, $p(1) = 1-\varepsilon$.

Par symétrie on a

$$H(\varepsilon) = H(1-\varepsilon).$$

La capacité du canal bin, sym, sans mém, avec proba d'erreur ε est déterminée par

$$H(y|x) = \varepsilon \log_2 \varepsilon^{-1} + (1-\varepsilon)\log_2(1-\varepsilon)^{-1}$$
$$= H(\varepsilon),$$

alors

$$H(y|\mathcal{X}) = \sum_{x \in \mathcal{X}} p(x) H(\varepsilon) = H(\varepsilon),$$

et par conséquent, $H(Y|X)$ est indépendent de la propa sur X. Alors on obtient le maximum de

$$I(X,Y) = H(X) - H(X|Y)$$
$$= 1 - H(X|Y) = 1 - H(\varepsilon)$$

pour la proba uniforme sur X.

<u>Attention</u>. Il faut établir que $H(X|Y)$ est bien $H(\varepsilon)$ pour ce maximum ou montrer que $H(Y) \leqslant 1$, avec max quand $H(X) = 1$.

$$= H(Y) - H(Y|X) = H(Y) - H(\varepsilon)$$

<u>Preuve</u>

$$H(Y) = p_Y(0) \log_2(p_Y(0)^{-1}) + p_Y(1) \log_2(p_Y(1)^{-1})$$

où

$$p_Y(0) = p(0,0) + p(1,0)$$
$$= p_X(0)\, p(0|0) + p_X(1)\, p(1|0)$$
$$= p_X(0)(1-\varepsilon) + p_X(1)\,\varepsilon$$
$$= p_X(0) - \varepsilon(p_X(0) - p_X(1))$$
$$= 1/2 \text{ si } p_X(0) = 1/2.$$

Donc si la proba sur X est uniforme, la proba sur Y est aussi uniforme. □

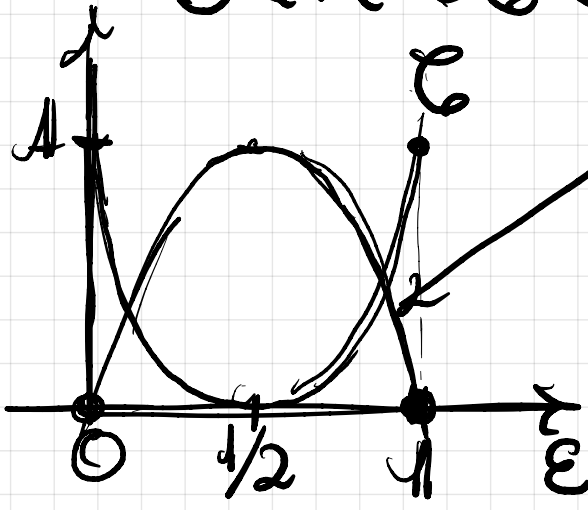
Alors la capacité du canal est $C = 1 - H(\varepsilon)$.

Si $\varepsilon = 1/2$, la capacité est zéro, i.e. on ne peut pas passer de l'information — l'info mutuelle entre X = mots trans et Y = mots reçus est 0.

Si $\varepsilon = 0$, la capacité est 1.

Conclusion. Pour une capacité $C = 1 - H(\varepsilon)$, $0 < C < 1$, on peut choisir un codage convenable avec rendement

$$0 < R < C < 1.$$



$$H(\varepsilon) = \varepsilon \log_2(\varepsilon^{-1}) + (1-\varepsilon)\log_2(1-\varepsilon)^{-1}$$

Attention. Un tel codage va regrouper des bits pour

transmettre de l'information fiablement.

On va maintenant étudier les constructions explicites pour realiser des bons codes.

Distance de Hamming et distance minimiem

On va étudier des codes en blocs, $C \subseteq A^n$, pour un alphabet A. On va souvant prendre $A = \{0,1\}$.

Défn La distance de Hamming est une fonction

$$d : A^n \times A^n \longrightarrow \mathbb{N}$$

définie par

$$d(x,y) = \left| \{ i : 1 \leq i \leq n, x_i \neq y_i \} \right|$$

où $x = x_1 \cdots x_n$ et $y = y_1 \cdots y_n$

La distance minimum d'un codage $C : X \longrightarrow A^n$ est

$$d(C) = \min \{ d(x,y) : x \neq y \in C(X) \}$$

si C est sans pertes, sinon 0.

Pour $C = C(X) \subseteq A$, on définit

$$d(C) = \min\{d(x,y) : x \neq y \in C\}.$$

Défn La boule de rayon t autour de $x \in A^n$ est

$$B(x,t) = \{y \in A^n : d(x,y) \leq t\}.$$

La boule de rayon t autour de $C \subseteq A^n$ est

$$B(C,t) = \bigcup_{x \in C} B(x,t)$$

On va s'intéresser aux codes linéaires, où $A = \mathbb{F}_p$ est un corps fini (par exemple $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ pour p premier), surtout pour $p=2$, on a

$\mathbb{F}_2 = \{0,1\}$ (ensemble)
avec addition et mult :

+	0	1
0	0	1
1	1	0

·	0	1
0	0	0
1	0	1

$+ =$ XOR $\quad\quad$ $\cdot =$ ET (AND)

$\mathbb{F}_p = \{0, 1, \dots, p-1\}$.

$\oplus$ = la reste de la somme après division par p

$$x \oplus y = \begin{cases} x+y < p \\ x+y-p \text{ s'il dépasse } p-1. \end{cases}$$

$\odot$ = la reste après division par p du produit.

Ex. $p = 7$ $\quad -1 = 6$

$\odot$	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	(1)	2	3	4	5	6
2	0	2	4	6	(1)	3	5
3	0	3	6	2	5	(1)	4
4	0	4	(1)	5	2	6	3
5	0	5	3	(1)	6	4	2
6	0	6	5	4	3	2	(1)

Inverses:

On dit que $b = \bar{a}^1$ dans $\mathbb{F}_p$ si $ab = 1$.

Par exemple:
$2 \cdot 4 = 1$, et
$3 \cdot 5 = 1$,
alors
$4 = 2^{-1}$, etc.

On identifie $a \in \mathbb{F}_p$ avec n'importe quel élément, où l'ensemble

$$a + p\mathbb{Z} = \{a + kp : k \in \mathbb{Z}\}.$$

Un <u>code linéaire</u> est un code $C \subseteq A^n = \mathbb{F}_p^n$, où C est un sous-espace vectoriel.

<u>Remarque</u> $\mathbb{F}_p^n$ est un espace vectoriel:

suffisants
- $(0,\dots,0) \in \mathbb{F}_p^n$ est le vecteur 0.
- $(x_1,\dots,x_n)+(y_1,\dots,y_n) = (x_1+y_1,\dots,x_n+y_n)$
- $a(x_1,\dots,x_n) = (ax_1,\dots,ay_n)$ où $a \in \mathbb{F}_p$, $(x_1,\dots,x_n) \in \mathbb{F}_p^n$.

(Comme $a \in \mathbb{N}$, on peut définir

$a(x_1,\dots,x_n) = \underbrace{(x_1,\dots,x_n)+\dots+(x_1,\dots,x_n)}_{a \text{ fois}}$

Pour $p=2$, on a juste $a=0$ ou $a=1$:

$0\cdot(x_1,\dots,x_n) = (0,\dots,0)$, et

$1\cdot(x_1,\dots,x_n) = (x_1,\dots,x_n)$.

La distance de Hamming est linéaire (sur $\mathbb{F}_p^n$), alors

$$d(x,y) = d(x+z, y+z),$$

où $x = (x_1, \dots, x_n)$,

$y = (y_1, \dots, y_n)$, et

$z = (z_1, \dots, z_n)$.

Alors, si on met $z = -y$, on voit que

$$d(x,y) = d(x-y, 0).$$

Donc si $\mathcal{C}$ est linéaire

$$d(\mathcal{C}) = \min\{d(x,0) : x \neq 0, x \in \mathcal{C}\}$$
$$= \min\{d(x,y) : x \neq y, x,y \in \mathcal{C}\}.$$

car $x - y \in \mathcal{C}$ et $d(x,y) = d(x-y, 0)$

Il suffit de tester $|\mathcal{C}| - 1$ distances pour déterminer $d(\mathcal{C})$, au lieu de $|\mathcal{C}|^2 - |\mathcal{C}|$.