Divergence de Kullback- Liebler ou entropie relative

Soit deux loi $\mathcal{P}$ et $\mathcal{Q}$ définies sur un même support $\mathcal{A}$, la divergence de Kullback-Liebler est donnée par

$$\mathcal{D}(\mathcal{P}||\mathcal{Q}) = \sum_{x \in \mathcal{A}} p(x) log_2 \left(\frac{p(x)}{q(x)} \right)$$

avec les conventions
$0 log_2 \left(\frac{0}{0}\right) = 0, \; 0 log_2 \left(\frac{0}{q}\right) = 0 \; et \; p log_2 \left(\frac{p}{0}\right) = \infty$

Divergence de Kullback-Liebler

Soit deux loi $\mathcal{P}$ et $\mathcal{Q}$ définies sur un même support $\mathcal{A}$,

1. $\mathcal{D}(\mathcal{P}||\mathcal{Q}) \geq 0$
2. $\mathcal{D}(\mathcal{P}||\mathcal{Q}) = 0$ si et seulement si $\mathcal{P}$ et $\mathcal{Q}$ sont identiques

la démonstration repose sur la concavité de la fonction $\log_2$ (faite en td)

Divergence de Kullback et le PMU

Objectif : définir une stratégie optimale pour jouer aux courses

contexte :

n courses indépendantes, M chevaux concurrents, $C_1, C2, \cdots C_M$.
Soit X_k la variable qui donne le numéro du cheval gagnant de la k-ème course .
On suppose les X_k indépendantes et suivant la même loi p .
Notons $p_i = P(X_k = i)$

Soit $o(i)$ le coefficient multiplicateur de la mise si le le cheval C_i gagne.
Dans le cas $\sum_{i=1}^{M} \frac{1}{o(i)} = 1$, r définie par $r(i) = \frac{1}{o(i)}$ est la distribution de probabilité qui correspond à l'estimée de la probabilité p par le bookmaker

Divergence de Kullback et le PMU

Le joueur joue tout son pécule S sur les chevaux.
On note s_i la somme jouée sur le cheval C_i et $b(i) = \frac{s_i}{s}$.

Après la première course le pécule du joueur est $S_1 = b(X_1)o(X_1)$,
au terme de la $n$ème course est $S_n = \prod_{k=1}^{n} S(X_i)$
alors la fortune du joueur utilisant la stratégie b vérifie

$$lim_{n \to \infty} \frac{1}{n} Log_2(S_n) = W(b,p) \quad avec \quad W(b,p) = \sum_{i=1}^{M} p_i log_2(b(i)o(i))$$

$$W(b,p) = \sum_{i=1}^{M} p_i log_2(o(i)) - H(p) - \mathcal{D}(p||b)$$

La stratégie optimale (critère de Kelly) est de prendre $b = p$

$$W(b,p) = \mathcal{D}(p||r) - \mathcal{D}(p||b)$$

Divergence de Kullback-Liebler par rapport à une loi uniforme

Soit $\mathcal{P}$ une loi définie sur $\mathcal{A}$ de cardinal fini, et $\mathcal{U}$ la loi uniforme définie sur ce même support

$$\boxed{\mathcal{D}(p_X||\mathcal{U}) = H(\mathcal{U}) - H(X)}$$

$$\boxed{H(X) = H(\mathcal{U}) - \mathcal{D}(p_X||\mathcal{U})}$$

Divergence de Kullback-Liebler

Soit $\mathcal{P}$ une loi définie sur un support $\mathcal{A}$, on définit sur $\mathcal{A} \times \mathcal{A}$

1. p_0 definie par

$$p_0(x, y) = \begin{cases} p(x) & si \quad y = x \\ 0 & si \quad y \neq x \end{cases}$$

2. $p \cdot p$ définie par

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

$$\boxed{H(X) = \mathcal{D}(p_0 || p \cdot p)}$$

L'entropie peut-être interprétée comme la divergence entre la probabilité qui correspond à une corrélation parfaite (p_0) et celle correspondant à l'indépendance (p.p)

Divergence de Kullback et Raffinement

Définition : Soit $\mathcal{Q}$ et $\mathcal{R}$ deux partitions d'un même ensemble $\mathcal{A}$.
On dit que $\mathcal{R}$ est un raffinement de $\mathcal{Q}$ si tout élément de $\mathcal{Q}$ peut s'écrire comme réunion d'élément de $\mathcal{R}$
Propriété : soit P et M deux lois de probabilités définies sur le même support '$\mathcal{A}$ et soit $\mathcal{Q}$ et $\mathcal{R}$ deux partitions finies , avec $\mathcal{R}$ raffinement de $\mathcal{Q}$.
notons : $P_{\mathcal{Q}}$ la loi de probabilités des éléments de $\mathcal{Q}$ induite par la probabilité P :

$$(\forall Q \in \mathcal{Q}) P_{\mathcal{Q}}(Q) = \sum_{q \in Q} P(q)$$

Alors

$$\mathcal{D}(P_{\mathcal{Q}} || M_{\mathcal{Q}}) \leq \mathcal{D}\left(P_{\mathcal{R}} || M_{\mathcal{R}}\right)$$
$$H(P_{\mathcal{Q}}) \leq H\left(P_{\mathcal{R}}\right)$$

Entropie conjointe et Entropies marginales

Entropie conjointe :

$$H(X,Y) = -\sum_{x\in\mathcal{A}_X}\sum_{y\in\mathcal{A}_Y} p(x,y)log_2(p(x,y))$$

Entropies marginales

$$\begin{array}{rcl} H(X) & = & -\sum_{x\in\mathcal{A}_X} p(x)log_2(p(x)) \\ H(Y) & = & -\sum_{y\in\mathcal{A}_y} p(y)log_2(p(y)) \end{array}$$

première relation :

$$\boxed{\begin{array}{rcl} H(X,Y) & \geq & H(X) \\ & \geq & H(y) \end{array}}$$

Relation entre Entropie conjointe et Entropies marginales

Entropie conjointe et entropies marginales lorsque X et Y sont indépendantes

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

Entropie conjointe et entropies marginales

$$\boxed{H(X,Y) \leq H(X) + H(Y)}$$

$$H(X,Y) = H(X) + H(Y) - \underbrace{\sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x,y) log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right)}_{>0 \ (cf \ ingalité \ de \ Gibbs)}$$

Information mutuelle

L'information mutuelle $I(X, Y)$ est la divergence de Kullback entre la loi conjointe et le produit de ces marginales.

$$\boxed{I(X,Y) = \sum_{x \in \mathcal{A}_X} \sum_{y \in \mathcal{A}_Y} p(x,y) log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)}$$

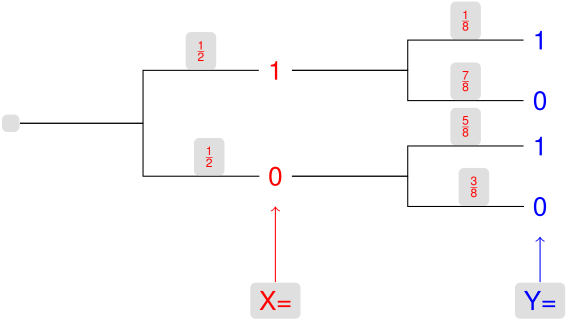
propriété : l'information mutuelle est symétrique $I(X, Y) = I(Y, X)$
cas particuliers :

1. $I(X, X) = H(X)$
2. pour un couple (X, Y) est un couple indépendant : $I(X, Y) = 0$

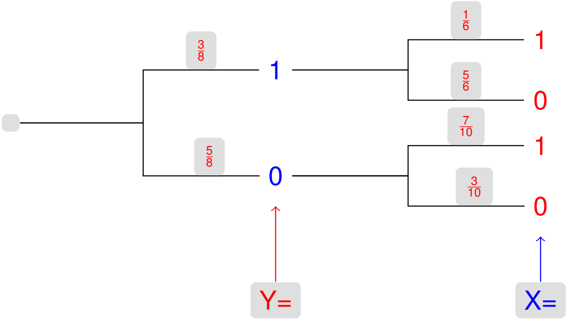
Entropie, entropies marginales, information mutuelle

$$\boxed{H(X, Y) = H(X) + H(Y) - I(X, Y)}$$

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



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



$$
\begin{aligned}
H(Y) &= -\frac{3}{8} log_2\left(\frac{3}{8}\right) - \frac{5}{8} log_2\left(\frac{5}{8}\right)\\
&= 3\left(\frac{3}{8} + \frac{5}{8}\right) - \frac{3log_2(3) + 5log_2(5)}{8}\\
&= 3 - \frac{3log_2(3) + 5log_2(5)}{8}\\
&= 0.954434
\end{aligned}
$$

$$
\begin{aligned}
I(X,Y) &= 1.954434 - 1.748999\\
&= 0.205435
\end{aligned}
$$

Entropie conditionnelle

$$H(Y|X=x) = -\sum_{y\in\mathcal{A}_Y} p(y|x) log_2 (p(y|x))$$

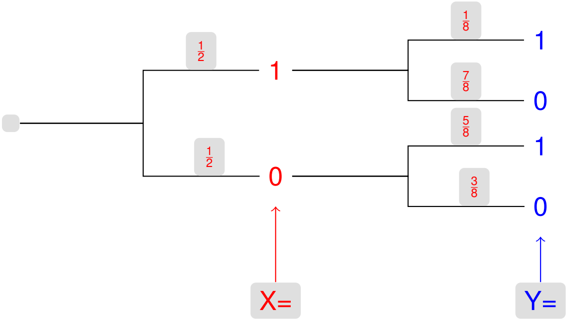
$$H(Y|X=x) = -\sum_{x\in\mathcal{A}_X} p(x) H(Y|X=x)$$

$$\boxed{H(Y|X) = -\sum_{x\in\mathcal{A}_X}\sum_{y\in\mathcal{A}_Y} p(x,y) log_2 (p(y|x))}$$

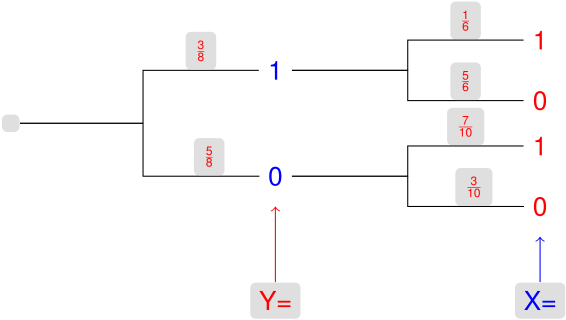
Cas particulier

1. pour le couple (X,X) on a $H(X,X)=0$
   $(\forall (x_i,x_j)\in\mathcal{A}_X^2) p(x_i|x_j) = \delta_{ij}$, et donc $(\forall x_j \in \mathcal{A}_X) H(X|x_j) = 0$
2. pour un couple (X,Y) indépendant $H(Y|X) = H(Y)$:
   $(\forall (x_i,y_j)\in\mathcal{A}_X\times\mathcal{A}_Y) p(y_j|x_i) = p(y_j)$, et donc
   $(\forall x_i \in \mathcal{A}_X) H(Y|x_i) = H(Y)$

Exemple $H(Y|X)$



Exemple $H(X|Y)$



Entropie conditionnelle : règle du chaînage

$$\boxed{\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ H(X,Y) &= H(Y) + H(X|Y) \end{aligned}}$$

$$\begin{aligned} H(X,Y,Z) &= H(X,Y) + H(z|(X,Y)) \\ &= H(Z|(X,Y)) + H(Y|X) + H(X) \end{aligned}$$

généralisation

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1)$$

majoration de l'entropie conditionnelle

$$\boxed{\begin{array}{lcr} H(Y|X) & \leq & H(Y)) \\ H(X|Y) & \leq & H(X) \end{array}}$$

avec égalité ssi X et Y indépendants

Information mutuelle et entropie conditionnelle

attention l'entropie conditionnelle diminue en moyenne mais pour un y particulier on peut avoir $H(X|Y=y) > H(X)$

Exemple

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

$H(X) = H\left(\mathcal{B}\left(\frac{1}{8}\right)\right) = 0.544 bits$
$H(X|Y=1) = 0$ et $H(X|Y=2) = 1 bit$
mais en moyenne

$$\begin{aligned} H(X|Y) &= \tfrac{3}{4}H(X|Y=1) + \tfrac{1}{4}H(X|Y=2) \\ &= \frac{1}{4} \\ &\leq H(X) \end{aligned}$$

Information mutuelle et entropie conditionnelle

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

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

Entropie conditionnelle nulle

1. dans quel cas $I(X,Y)=0$?
2. Comment interpréter $H(Y|X)=0$ ie $I(X,Y)=H(Y)$
   1. $H(Y|X) = -\sum_{x\in\mathcal{A}_X}\sum_{y\in\mathcal{A}_Y} p(x,y)log_2\left(\frac{p(x,y)}{p(x)}\right)$
   2. $H(Y|X)=0$ ssi $(\forall x\in\mathcal{A}_X)(\forall y\in\mathcal{A}_Y)\,(p(x,y)>0 \Longrightarrow p(x,y)=p(x))$
   3. $(\forall x\in\mathcal{A}_X)\,(p(x)>0 \Longrightarrow (\exists ! y\in\mathcal{A}_Y))p(x,y)=p(x))$
   4. Y est une fonction déterministe de X

prévisions météorologiques

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

	temps : pluie	temps : soleil
temps prévu : pluie	$\frac{1}{12}$	$\frac{1}{6}$
temps prévu : soleil	$\frac{1}{12}$	$\frac{2}{3}$

1. quelle est la probabilité que le système donne une prévision incorrecte ?
2. Calculer l'information mutuelle entre le temps prévu et le temps effectif
3. Comparer à un système de prévisions qui prédit systématiquement le soleil

test de dépistage

Un test de dépistage pour une maladie est censé discriminer entre le situations $X \in \{C, \overline{C}\}$ en donnant un resultat $Y \in \{T^+, T^-\}$

	$Y = T^+$	$Y = T^-$
$X = C$	0.07	0.01
$X = \overline{C}$	0.03	0.89

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

1. Calculer $H(X)$, $I(X, Y)$, $H(X|Y)$, r
2. Que signifie $r = 0$?, $r = 1$?