ACS : Information Theory

ABDELMALEK RHAYOUTE

November 15, 2022

Remark: I will not solve the easy questions

Exercise 2

Bayes' formula : $\begin{cases} P_{X|Y}(x \mid y) = \frac{P_{X,Y}(x,y)}{P_Y(y)} \\ P_{Y|X}(y \mid x) = \frac{P_{X,Y}(x,y)}{P_X(x)} \\ P_Y(y) = \sum_{x' \in \mathcal{X}} P_{X,Y}(x', y) \end{cases} \implies P_{X|Y}(x \mid y) = \frac{P_X(x) P_{Y|X}(y|x)}{\sum_{x'} P_X(x') P_{Y|X}(y|x')}$

Exercise 3

Given $P_{X,Y}$ of the $BSC(p)$ channel

Assume that $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, E\}$ and that $P_{X,Y}$ is given in the table

Y,X	0	1
0	$\frac{1-e}{2}$	0
1	0	$\frac{1-e}{2}$
E	$\frac{e}{2}$	$\frac{e}{2}$

$\implies$

$Y\|X$	0	1
0	$1-e$	0
1	0	$1-e$
E	e	e

e is called an erasure probability.

This channel model can also be written in a additive form,

$$Y = X \oplus W$$

where X is a Bern(0.5) random variable, W is a Bern(e) random variable on $\{E, 1\}$, X and W are independent, and $\oplus$ is the binary XOR operation.

Exercise 5

The KL divergence verifies a certain set of properties :

1. Asymmetry : $D_{KL}(P_X \| Q_X) \neq D_{KL}(Q_X \| P_X)$.
2. Null element : $D_{KL}(P_X \| P_X) = 0$.
3. Positivity: $D_{KL}(P_X \| Q_X) \geq 0$, for all laws $(P_X, Q_X)! = 0$

Gibbs' inequality:

Suppose that

$$P = \{p_1, \ldots, p_n\}$$

is a discrete **probability distribution**. Then for any other probability distribution

$$Q = \{q_1, \ldots, q_n\}$$

the following inequality between positive quantities (since P and Q are between zero and one)

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

with equality if and only if

$$p_i = q_i$$

for all i. Put in words, the **information entropy** of a distribution P is less than or equal to its **cross entropy** with any other distribution Q.

The difference between the two quantities is the **Kullback–Leibler divergence** or relative entropy

$$D_{\mathrm{KL}}(P\|Q) \equiv \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geq 0.$$

Proof :

First method : Let I denote the set of all i for which p_i is non-zero. Then, since $\ln x \leq x-1$ for all $x>0$, with equality if and only if $x=1$, we have:

$$-\sum_{i\in I} p_i \ln \frac{q_i}{p_i} \geq -\sum_{i\in I} p_i \left(\frac{q_i}{p_i}-1\right)$$

= - $\sum_{i\in I} q_i + \sum_{i\in I} p_i = -\sum_{i\in I} q_i + 1 \geq 0$

The last inequality is a consequence of the p_i and q_i being part of a probability distribution. Specifically, the sum of all non-zero values is 1. Some non-zero q_i, however, may have been excluded since the choice of indices is conditioned upon the p_i being non-zero. Therefore the sum of the q_i may be less than 1.

So far, over the index set I, we have:

$$-\sum_{i\in I} p_i \ln \frac{q_i}{p_i} \geq 0$$

or equivalently

$$-\sum_{i\in I} p_i \ln q_i \geq -\sum_{i\in I} p_i \ln p_i$$

Both sums can be extended to all $i=1,\ldots,n$, i.e. including $p_i=0$, by recalling that the expression $p\ln p$ tends to 0 as p tends to 0, and $(-\ln q)$ tends to ∞ as q tends to 0. We arrive at

$$-\sum_{i=1}^{n} p_i \ln q_i \geq -\sum_{i=1}^{n} p_i \ln p_i$$

Second method : Below we give a proof based on Jensen's inequality:

Because log is a concave function, we have that:

$$\sum_i p_i \log \frac{q_i}{p_i} \leq \log \sum_i p_i \frac{q_i}{p_i} = \log \sum_i q_i \leq 0$$

The first inequality is due to Jensen's inequality, and the last is due to the same reason given in the above proof.

Third method :

Log sum inequality

==Statement==

Let $a_1,\ldots,a_n$ and $b_1,\ldots,b_n$ be nonnegative numbers. Denote the sum of all a_is by a and the sum of all b_is by b. The log sum inequality states that

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

==Proof==

Notice that after setting $f(x) = x\log x$ we have

$$\begin{aligned}\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} &= \sum_{i=1}^{n} b_i f\left(\frac{a_i}{b_i}\right) = b \sum_{i=1}^{n} \frac{b_i}{b} f\left(\frac{a_i}{b_i}\right) \\ &\geq b f\left(\sum_{i=1}^{n} \frac{b_i}{b}\frac{a_i}{b_i}\right) = b f\left(\frac{1}{b}\sum_{i=1}^{n} a_i\right) = b f\left(\frac{a}{b}\right) \\ &= a \log \frac{a}{b},\end{aligned}$$

where the inequality follows from **Jensen's inequality** since $\frac{b_i}{b} \geq 0$, $\sum_{i=1}^{n} \frac{b_i}{b} = 1$, and f is convex.

—Let $P = (p_i)_{i\in N}$ and $Q = (q_i)_{i\in N}$ be pmfs. In the log sum inequality, substitute $n = \infty$, $a_i = p_i$ and $b_i = q_i$ to get

$$D_{\mathrm{KL}}(P\|Q) \equiv \sum_i p_i \log_2 \frac{p_i}{q_i} \geq 1 \log \frac{1}{1} = 0$$

Fourth method : As particular case of Bregman divergence
Exercise 6

Here are a few special cases of entropy:
1. The entropy of a constant random variable, $X = c$ with probability 1 , is given by

$$H(X) = 0.$$

2. The entropy of a random variable X uniformly distributed over $\mathcal{X}$, c.à.d, $P_X(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$, is given by

$$H(X) = \log_2(|\mathcal{X}|).$$

3. Let X be a binary Bernoulli random variable Bern(p) where $p = P_X(1)$. The entropy of X is given by

$$H(X) = H_2(p) - p\log_2(p) - (1-p)\log_2(1-p) = H_2(p).$$

$\mathrm{H}_2(p)$ is commonly known as the binary entropy function. It is maximal when $p = 0.5$ (equals 1 bit), and is minimal when $p = 0$ or $p = 1$, (equals 0 bits). It is symmetric around $p = 0.5$.

Exercise 7

Discrete entropy $H(X)$ satisfies the following properties :
1. Minimum : $H(X) \geq 0$ for all discrete distributions P_X, and the minimum is achieved for a degenerate distribution, i.e., X is a constant.
2. Maximum: $H(X) \leq \log_2(|\mathcal{X}|)$ for all discrete distributions P_X, and the maximum is achieved for the uniform distribution over $\mathcal{X}$.
3. Data Processing Inequality (DPI) : the entropy of a function f of X, is no greater than the entropy of X, i.e.,

$$H(f(X)) \leq H(X)$$

with equality iif f is a one-to-one function.

Corollary of Gibbs' inequality

Information entropy of P is bounded by:

$$H(p_1, \ldots, p_n) \leq \log n.$$

The proof is trivial – simply set $q_i = 1/n$ for all "i".
Exercise 9
Conditional entropy $H(X \mid Y)$ satisfies the following properties
1. Asymmetry : $H(X \mid Y) \neq H(Y \mid X)$
2. Minimum: $H(X \mid Y) \geq 0$ for all $P_{X,Y}$, and the minimum is achieved when $P_{X|Y}$ is degenrate, i.e., X is a function of Y.
3. Upper bound: conditional entropy is always smaller than the individual entropy

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

with equality iif X and Y are independent : $P_{X,Y}=P_X P_Y$.

4. Maximum: $H(X \mid Y) \leq \log_2(|\mathcal{X}|)$ for all $P_{X,Y}$ and the maximum is achieved when X and Y are independent, and X is uniform, i.e.,

$$\begin{cases} Independence : \forall(x,y) \quad P_{X,Y}(x,y)=P_X(x)P_Y(y) \\ P_{X\mid Y}(x \mid y)=\frac{P_{X,Y}(x,y)}{P_Y(y)}, P_{Y\mid X}(y \mid x)=\frac{P_{X,Y}(x,y)}{P_X(x)} \\ H(X \mid Y)-\displaystyle\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\log_2\left(P_{X\mid Y}(x \mid y)\right) \\ \qquad =\displaystyle\sum_{y\in\mathcal{Y}} P_Y(y)H(X \mid Y=y) \\ H(X \mid Y=y)-\sum_{x\in\mathcal{X}} P_{X\mid Y}(x \mid y)\log_2\left(P_{X\mid Y}(x \mid y)\right) \end{cases} \Longrightarrow \text{H(X} \mid Y)=H(X)$$

Propertie 2 : When X is a function of Y
Hint :

$$\log_2\left(P_{X\mid Y}(x \mid y)\right.=0$$

Exercise 10

Compute the conditional entropy $H(Y \mid X)$ where $Y=X\oplus W$, where X follows a Bern(1/2), independent from W which follows a Bern(p) and $\oplus$ is the binary XOR operation. (Hint : $P_{X,Y}$ and $P_{Y\mid X}$ were given previously in example 2)

Propertie : The individual entropies $H(X)$ and $H(Y)$, the joint entropy $H(X,Y)$ and the conditional entropies $H(X \mid Y)$ et $H(Y \mid X)$ can be related as follows

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

Exercise 13

The differential entropy of a real-valued Gaussian random variable $X_G \sim \mathcal{N}\left(\mu,\sigma^2\right)$ is given by

$$h\left(X_G\right)=\frac{1}{2}\log_2\left(2\pi e\sigma^2\right)$$

Exercise 14

Mutual information is positive.
Hint : proof is similar to the proof of positivity of the KL divergence.
Proof :

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

If we use this defintion and the exercise 5, we deduct simply many versions of the proof.

Exercise 18

Shannon formula for the AWGN :

$$\log_2(1+SNR)$$

Let $X\sim\mathcal{N}(0,P)$ and $W\sim\mathcal{N}\left(0,\sigma^2\right)$ be two independent Gaussian random variables. Assume that we observe Y given by

$$Y=X+W.$$

This model describes the so called Additive White Gaussian Noise (AWGN) channel with input signal X, additive noise W, and output signal Y.

The mutual information between X and Y is given by :

$$I(X;Y)=\frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right).$$

Proof We have :

$$\begin{aligned} I(X;Y) &= h(X)-h(Y \mid X) \\ &= h(X)-h(Y \mid X) \\ &= h(X)+h(Y)-h(X;Y) \end{aligned}$$

Because X $\perp\!\!\!\perp$W then

$$h(Y) = \frac{1}{2} log(2\pi \exp(\sigma^2 + P))$$

Or

$$h(Y|X) = h(W) = \frac{1}{2} log(2\pi \exp \sigma^2)$$

Finally :

$$I(X;Y) = h(X) - h(Y \mid X) = \frac{1}{2} log(2\pi \exp(\sigma^2 + P)) - \frac{1}{2} log(2\pi \exp \sigma^2) = \frac{1}{2} \log_2 \left(1 + \frac{P}{\sigma^2}\right)$$