Information Theory and Coding - Prof. Emere Telatar

Jean-Baptiste Cordonnier, Sebastien Speierer, Thomas Batschelet

October 8, 2017

1 Data compression

Given an alphabet $\mathcal{U}$ (e.g. $\mathcal{U} = \{a, ..., z, A, ..., Z, ...\}$), we want to assign binary sequences to elements of $\mathcal{U}$, i.e.

$$e : \mathcal{U} \to 0, 1^* = \{\emptyset, 0, 1, 00, 01, ...\}$$

For $\mathcal{X}$ a set

$$\mathcal{X}^n \equiv \{(x_0...x_n), x_i \in \mathcal{X}\}$$
$$\mathcal{X}^* \equiv \bigcup_{n \geq 0} \mathcal{X}^n$$

Definition 1.1. *A code $\mathcal{C}$ is called singular if*

$$\exists (u, v) \in \mathcal{U}^2, u \neq v \quad s.t. \quad C(u) = C(v)$$

Non singular code is defined as opposite

Definition 1.2. *A code $\mathcal{C}$ is called uniquily decodable if*

$$\forall u_1, ..., u_n, v_1, ..., v_n \in \mathcal{U}^* \quad s.t. \quad u_1, ..., u_n \neq v_1, ..., v_n$$

we have

$$\mathcal{C}(u_1)\mathcal{C}(u_n) \neq \mathcal{C}(v_1)\mathcal{C}(v_n)$$

i.e, $\mathcal{C}^*$ is non-singular

Definition 1.3. *Suppose $\mathcal{C} : \mathcal{U} \to \{0, 1\}^*$ and $\mathcal{D} : \mathcal{V} \to \{0, 1\}^*$ we can define*

$$\mathcal{C} \times \mathcal{D} : \mathcal{U} \times \mathcal{V} \to \{0, 1\}^* \quad as \quad (\mathcal{C} \times \mathcal{D})(u, v) \to \mathcal{C}(u)\mathcal{D}(v)$$

Definition 1.4. *Given $\mathcal{C} : \mathcal{U} \to \{0, 1\}^*$, define*

$$\mathcal{C}^* : \mathcal{U}^* \to \{0, 1\}^* \quad as \quad \mathcal{C}^*(u_1, u_n) = \mathcal{C}(u_1)...\mathcal{C}(u_n)$$

Definition 1.5. *A code $\mathcal{U} \to \{0, 1\}^*$ is **prefix free** is for no $u \neq v$ $\mathcal{C}(u)$ is a prefix of $\mathcal{C}(v)$.*

Theorem 1.1. *If $\mathcal{C}$ is prefix free then $\mathcal{C}$ is uniquely decodable.*

Definition 1.6. ***Kraft sum**: Given $\mathcal{C} : \mathcal{U} \to \{0, 1\}^*$*

$$kraftsum(\mathcal{C}) = \sum 2^{length(\mathcal{C}(u))}$$

Lemma 1.2. *if $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ and $\mathcal{D} : \mathcal{V} \to \{0,1\}^*$ then $kraftsum(\mathcal{C} \times \mathcal{D}) = kraftsum(\mathcal{C}) j \times kraftsum(\mathcal{D})$*

Proof.

$$\begin{aligned} kraftsum(\mathcal{C} \times \mathcal{D}) &= \sum_{u,v} 2^{-(length(\mathcal{C}) * length(\mathcal{D}))} \\ &= \sum_{u} 2^{-length(\mathcal{C})} \sum_{v} 2^{-length(\mathcal{D})} \end{aligned}$$

□

Corollary 1.2.1. $kraftsum(\mathcal{C}^n) = (kraftsum(\mathcal{C}))^n$

Proposition 1.1. *if $\mathcal{C}$ is non-singular, then*

$$kraftsum(\mathcal{C}) \leq 1 + \max_{n} length(\mathcal{C}(u))$$

Theorem 1.3. *if $\mathcal{C}$ is uniquely decodable, then $kraftsum(\mathcal{C}) \leq 1$*

Proof. $\mathcal{C}$ is uniquely decodable $\equiv \mathcal{C}^*$ is non singular

$$\begin{aligned} &\Rightarrow kraftsum(\mathcal{C}^n) \leq 1 + \max_{u_1,\ldots,u_n} length(\mathcal{C}^n) \\ &\Rightarrow kraftsum(\mathcal{C})^n \leq 1 + nL, \quad L = \max length(\mathcal{C}(n)) \end{aligned}$$

A growing exp cannot be bounded by a linear function

$$\Rightarrow kraftsum(\mathcal{C}) \leq 1$$

□

Theorem 1.4. *Suppose $\mathcal{C} : \mathcal{U} \to \mathcal{N}$ is such that $\sum_u i^{\mathcal{C}(u)} \leq 1$, then, there exist a prefix-free code $\mathcal{C} : \mathcal{U} \to \{0,1\}$ s.t. $\forall length(\mathcal{C}(u)) = \mathcal{C}(u)$*

Proof. Let $\mathcal{U} = \{u_1, ..., u_n\}$ and $\mathcal{C}(u_1) \leq \mathcal{C}(u_2) \leq ... \leq \mathcal{C}(u_k) = \mathcal{C}_{max}$. Consider the complete binary tree up to depth $\mathcal{C}_{max}$ initially all nodes are available to be used as codewords. For $i = 1, 2, ..., n$, place $\mathcal{C}(u_i)$ at an available node at level $\mathcal{C}(u_i)$ remove all descendant of $\mathcal{C}(u_i)$ from the available list.

Corollary 1.4.1. *Suppose $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ is u.d., then there exist an $\mathcal{C}' : \mathcal{U} \to \{0,1\}^*$ which is prefix-free and $length(\mathcal{C}'(n)) = length(\mathcal{C}(n))$*

□

Example 1.
$\mathcal{U} = \{a, b, c, d\}$, $\mathcal{C} : \{0, 01, 011, 111\}$ and $\mathcal{C}' : \{0, 10, 110, 111\}$
In this case, decoding $\mathcal{C}$ may require delay, while decoding $\mathcal{C}'$ is instanteneous.

2 Alphabet with statistics

Suppose we have an alphabet $\mathcal{U}$, and suppose we have a random variable $\mathcal{U}$ taking values in $\mathcal{U}$. We denote by $p(u) = Pr(U = u), u \in \mathcal{U}$ with $p(u) \geq 0$ and $\sum_u p(u) = 1$.

Suppose we have a code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$. We then have $\mathcal{C}(u)$ a random binary string and $length(\mathcal{C}(u))$ a random integer.

Example 1.
$\mathcal{U} = \{a, b, c, d\}$
$p : \{0.5, 0.25, 0.125, 0.125\}$
$\mathcal{C} : \{0, 01, 110, 111\}$
then we have

$$length(\mathcal{C}(u)) = \begin{cases} 1, & p = 0.5 \\ 2, & p = 0.25 \\ 3, & p = 0.125 + 0.125 + 0.25 \end{cases}$$

We can measure how efficient $\mathcal{C}$ represents $\mathcal{U}$ by considering

$$E[length(\mathcal{C}(u))] = \sum_u p(u)\mathcal{C}(u) \quad \text{with} \quad \mathcal{C}(u) = length(\mathcal{C}(u))$$

Theorem 2.1. *if $\mathcal{C}$ is u.d., then*

$$E[length(\mathcal{C}(u))] \geq \sum_u p(u) \log(\frac{1}{p(u)})$$

Proof. let $\mathcal{C}(u) = length(\mathcal{C}(u))$, we know $\sum_u 2^{-\mathcal{C}(u)} \leq 1$ because $\mathcal{C}$ is u.d.

$$\begin{aligned} E[length(\mathcal{C}(u))] &= \sum_u p(u)\mathcal{C}(u) = \sum_u p(u) \log_2(\frac{1}{q(u)}) \\ &\equiv \sum_u p(u) \log(\frac{q(u)}{p(u)}) \leq 0 \\ &\equiv \sum_u p(u) \ln(\frac{q(u)}{p(u)}) \leq 0 \\ &\leq \sum_u p(u)[\frac{q(u)}{p(u)} - 1] = \underbrace{\sum_u q(u)}_{\leq 1} - \underbrace{\sum_u p(u)}_{=1} \leq 0 \end{aligned}$$

□

Theorem 2.2. *For any $\mathcal{U}$, there exists a prefix-free code $\mathcal{C}$ s.t.*

$$E[length(\mathcal{C}(u))] < 1 + \sum_{u \in \mathcal{U}} p(u) \log(\frac{1}{p(u)})$$

Proof. Given $\mathcal{U}$, let

$$\begin{aligned} &\mathcal{C}(u) = [\log(\frac{1}{p(u)})] < 1 + \log(\frac{1}{p(u)}) \\ &\Rightarrow \sum_u 2^{-\mathcal{C}(u)} \leq \sum_u p(u) = 1 \\ &\Rightarrow \sum_u p(u)\mathcal{C}(u) < \sum_u p(u) \log(\frac{1}{p(u)}) + \underbrace{1}_{\sum p(u)} \end{aligned}$$

□

Theorem 2.3. *The entropy of a RV $U \in \mathcal{U}$ is*

$$H(U) = \sum_{u \in \mathcal{U}} p(u) \log(\frac{1}{p(u)})$$

with $p(u) = Pr(U = u)$

Note that $H(U)$ is a fonction of the distribution $\mathcal{C}_u(.)$ of the RV U, it isn't a function of U.

$$H(U) = E[f(U)] \quad \text{where} \quad \log(\frac{1}{p(u)})$$

How to design optimal codes (in the sense of minimizing $E[length(\mathcal{C}(u))]$)?
Formally, given a random variable U, find $\mathcal{C}(u) \to \mathcal{N}$ s.t.

$$\sum_{u \in U} 2^{\mathcal{C}(u)} \leq 1 \quad \text{that minimizes} \quad \sum_{u \in U} p(u)\mathcal{C}(u)$$

Properties of optimal prefix-free codes

- if $p(u) < p(v)$ then $\mathcal{C}(u) \geq \mathcal{C}(v)$
- The two longest codewords have the same length
- There is an optimal code such that the two least probable letters are assigned codewords that differ in the last bit.

Observe that if $\mathcal{C}(u_1), ..., \mathcal{C}(u_{k-1}), \mathcal{C}(u_k)$ is a prefix-free collection of the property that

$$\begin{array}{ll} \mathcal{C}(u_{k-1}) & = \alpha 0 \\ \mathcal{C}(u_k) & = \alpha 1 \end{array} \quad \text{with} \quad \alpha \in \{0,1\}^*$$

then $\{\mathcal{C}(u_1), ..., \mathcal{C}(u_{k-2}, \alpha\}$ is also a prefix-free collection.
Also

$$\begin{aligned} \sum_{u \in \mathcal{U}} p(u) length(\mathcal{C}(u)) &= p(u_1) length(\mathcal{C}(u_1)) + ... + p(u_{k-2}) length(\mathcal{C}(u_{k-2})) + [p(u_{k-1}) + p(u_k)](length(\alpha) + 1) \\ &= (p(u_{k-1}) + p(u_k)) + \sum_{v \in \mathcal{V}} p(v) length(\mathcal{C}'(v)) \end{aligned}$$

So we have shown that with

$$E[length(\mathcal{C}(U)] = p(u_{k-1}) + p(u_k) + E[length(\mathcal{C}'(v))]$$

if $\mathcal{C}$ is optimal for U, then $\mathcal{C}'$ is optimal for V

3 Entropy and mutual information

Definition 3.1 (Joint entropy). *Suppose U, V are Random Variables with $p(u,v) = P(U = u, V = v)$, the joint entropy is*

$$H(UV) = \sum_{u,v} p(u,v) \log \frac{1}{p(u,v)}$$

Theorem 3.1.

$$H(UV) \leq H(U) + H(V)$$

with equality iff U and V are independants.

Proof. We want to show that

$$\sum_{u,v} p(u,v) \log \frac{1}{p(u,v)} \leq \sum_{u} p(u) \log \frac{1}{p(u)} + \sum_{v} p(v) \log \frac{1}{p(v)} \iff \sum_{u,v} p(u,v) \log \frac{p(u)p(v)}{p(u,v)} \leq 0$$

We use $\ln z \leq z - 1 \ \forall z$ (with equality iff $z = 1$):

$$\sum_{u,v} p(u,v) \log \frac{p(u)p(v)}{p(u,v)} \leq \sum_{u,v} p(u,v) \left[\frac{p(u)p(v)}{p(u,v)} - 1\right] = \sum_{u,v} p(u)p(v) - \sum_{u,v} p(u,v) = 1 - 1 = 0$$

□

Same definitions of entropy holds for n symbols.

Definition 3.2 (Joint Entropy)**.** *Suppose $U_1, U_2, \ldots, U_n$ are RVs and we are given $p(u_1 \ldots u_n)$, the joint entropy is*

$$H(U_1, \ldots, U_n) = \sum_{u_1 \ldots u_n} p(u_1 \ldots u_n) \log \frac{1}{p(u_1 \ldots u_n)}$$

Theorem 3.2.

$$H(U_1, \ldots, U_n) \leq \sum_{i=1}^{n} H(U_i)$$

with equality iff Us are independants

Corollary 3.2.1. *if $U_1, \ldots, U_n$ are i.i.d. then $H(U_1 \ldots U_n) = nH(U_1)$*

Definition 3.3 (Conditional entropy)**.**

$$H(U|V) = \sum_{u,v} p(u,v) \log \frac{1}{p(u|v)}$$

Theorem 3.3.

$$H(UV) = H(U) + H(V|U) = H(V) + H(U|V)$$

Theorem 3.4.

$$H(U) + H(V) \geq H(U,V) = H(V) + H(U|V)$$

Definition 3.4 (Mutual information)**.**

$$\begin{aligned} I(U;V) = I(V;U) &= H(U) - H(U|V) \\ &= H(V) - H(V|U) \\ &= H(U) + H(V) - H(UV) \end{aligned}$$

We can apply the chain rule on the entropy as follow

$$H(U_1, U_2, \ldots U_n) = H(U_1) + H(U_2|U_1) + \cdots + H(U_n|U_1, U_2 \ldots U_{n-1})$$

Definition 3.5 (Conditional mutual information)**.**

$$\begin{aligned} I(U;V|W) &= H(U|W) - H(U|VW) \\ &= H(V|W) - H(V|UW) \\ &= \mathbb{E}_{u,v,w}\left[\log \frac{p(uv|w)}{p(u|w)p(v|w)}\right] \end{aligned}$$

Theorem 3.5.

$$I(V; U_1 \ldots U_n) = I(V;U_1) + I(V;U_2|U_1) + \cdots + I(V;U_n|U_1 \ldots U_{n-1})$$

Notation 1.

$$U^n \triangleq (U_1, U_2, \dots U_n)$$

Theorem 3.6.

$$I(U;V|W) \geq 0$$

equality iff conditioned on w, u and v are independant, that is iff $U - V - W$ is a Markov chain.

Proof.

$$\begin{aligned}
I(U;V|W) &= \frac{1}{\ln 2} \sum_{u,v,w} p(u,v,w) \ln \frac{p(u|w)p(v|w)}{p(uv|w)} \\
&\geq \frac{1}{\ln 2} \sum_{u,v,w} p(u,v,w) \left[\frac{p(u|w)p(v|w)}{p(uv|w)} - 1 \right] \\
&= \frac{1}{\ln 2} \sum_{u,v,w} (p(w)p(u|w)p(v|w) - p(uvw)) \\
&= \frac{1}{\ln 2}(1-1) \\
&= 0
\end{aligned}$$

□

4 Data processing

Theorem 4.1. *$U - V - W$ is a MC $\iff$ $I(U;W|V) = 0$*

Corollary 4.1.1. *$I(U;V) \geq I(U;W)$ and by symetry of MC $I(W;V) \geq I(U;W)$*

Proof.

$$I(U;VW) = I(U;V) + I(U;W|V) = I(U;V)$$

and

$$I(U;VW) = I(U;W) + I(U;V|W) \geq I(U;W)$$

□

Theorem 4.2. *Given U a RV taking values in $\mathcal{U}$ then $0 \leq H(U) \leq \log|\mathcal{U}|$. $H(U) = 0$ iff U is constant, $H(U) = \log|\mathcal{U}|$ iff U is $p(u) = 1/|\mathcal{U}|$ for all u.*

Proof. For the lower bound,

$$H(U) = \sum_u \underbrace{p(u)}_{\geq 0} \underbrace{\log \frac{1}{p(u)}}_{\geq 0} \geq 0$$

For the upper bound,

$$\begin{aligned}
H(U) - \log|\mathcal{U}| &= \sum_u p(u) \log \frac{1}{p(u)} - \sum_u p(u) \log|\mathcal{U}| \\
&= \frac{1}{\ln 2} \sum_u p(u) \ln \frac{1}{|\mathcal{U}| p(u)} \\
&\leq \frac{1}{\ln 2} \sum_u p(u) \left(\frac{1}{|\mathcal{U}| p(u)} - 1 \right) \\
&= \frac{1}{\ln 2} \left[\sum_u \frac{1}{|\mathcal{U}|} - \sum_u p(u) \right] \\
&= 0
\end{aligned}$$

□

Theorem 4.3. $I(U;V) = 0 \iff U \perp V$

Definition 4.1 (Entropy rate of a stochastic process)**.** $\lim_{n\to\infty} \frac{1}{n} H(U^n)$ *if the limit exists.*

Theorem 4.4. *For stationary stochastic process U^n, the sequences*

$$a_n = \frac{1}{n} H(U^n) \text{ and } b_n = H(U_n|U^{n-1})$$

are positive and non increasing. Then $a = \lim_{n\to\infty} a_n$ and $b = \lim_{n\to\infty} b_n$ exists and $a = b$.

Proof.

$$\begin{aligned} b_{n+1} &= H(U_{n+1}|U_1, U_2, \ldots, U_n) \\ &\leq H(U_{n+1}|U_2, \ldots, U_n) \\ &= H(U_n|U_1, U_2, \ldots, U_{n-1}) \\ &= b_n \text{ , because } U_1 \ldots U_n \sim U_2 \ldots U_{n+1} \text{ (Stationarity).} \end{aligned}$$

Hence, it is non-increasing.

For the $\{a_n\}$, observe that

$$\begin{aligned} a_n = \frac{1}{n} H(U^n) &= \frac{1}{n}\Big[H(U_1) + H(U_2|U_1) + H(U_3|U^2) + \cdots + H(U_n|U^{n-1})\Big] \\ &= \frac{1}{n}\Big[b_1 + b_2 + \cdots + b_n\Big] \end{aligned}$$

and by the "Lemma", whenever $b_n \to b$, $a_n \to b$ ☐

Lemma 4.5 (Cesaro)**.** *Suppose $b_n \to b$,*

then,

$$a_n = \frac{1}{n}\Big[b_1 + b_2 + \cdots + b_n\Big] \text{ also converges and to 1.}$$

Proof. Since $b_n \to b$, $\Big(\equiv \forall \epsilon > 0 \text{ , } \exists\, n(\epsilon) \text{ s.t } \forall n > n(\epsilon)\ |b_n - b| < \epsilon\Big)$

$\exists B$ s.t. $|b_n| < B$ for all n.

Take $n > n_1(\epsilon) \triangleq \ldots$ then

$$|a_n - b| \leq \frac{|b_1 - b| + |b_2 - b| + |b_3 - b| + \cdots + |b_n - b|}{n}$$

$$\text{so } |a_n - b| \leq \frac{1}{n}\Bigg[\sum_{i=1}^{n_0(\epsilon)} \underbrace{|b_i - b|}_{2B} + \sum_{i=n_0(\epsilon)+1}^{n} \underbrace{|b_i - b|}_{\leq \epsilon}\Bigg] \leq \frac{n_0(\epsilon) 2B}{n} + \epsilon < 2\epsilon$$

$$\text{for } n > n_1(\epsilon) \triangleq \max, \{n_0(\epsilon) \frac{1}{\epsilon} n_0(\epsilon) 2B\}$$

☐

Appendices

A Markov chains

$U_1 - U_2 - \cdots - U_n$ forms a Markov chain if the joint probability distribution of the RVs is

$$p(a, b, c, d) = p(a)p(b|a)p(c|b)p(d|c)$$

which is equivalent to $(U_1, \ldots, U_{k-1})$ are independant of $(U_{k+1}, \ldots, U_n)$ when conditionned on U_k for any k.

Theorem A.1. *The reverse of a MC is a MC*

B Stochastic processes

A stochastic process is a collection $U_1, U_2 \ldots U_n$ of RVs each taking values in $\mathcal{U}$. It is described by its joint probability

$$p(u^n) = P(U_1 \ldots U_n = u_1 \ldots u_n) = P(U^n = u^n)$$

Definition B.1 (Stationary stochastic process)**.** *A process $U_1, U_2, \ldots$ is called stationary if for every n and k and $u_1 \ldots u_n$, we have*

$$p(u^n) = p(U_1 \ldots U_n = u_1 \ldots u_n) = p(U_{1+k} \ldots U_{n+k} = u_1 \ldots u_n)$$

In other words, the process is time shift invariant.