Information Theory and Coding - Prof. Emere Telatar

Jean-Baptiste Cordonnier, Sebastien Speierer, Thomas Batschelet

October 5, 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*

$$\mathcal{C} \times \mathcal{D} : \mathcal{U} \times \mathcal{V} \to \{0, 1\}^*$$

as

$$(\mathcal{C} \times \mathcal{D})(u, v) \to \mathcal{C}(u)\mathcal{D}(v)$$

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

$$\mathcal{C}^* : \mathcal{U}^* \to \{0, 1\}^*$$

as

$$\mathcal{C}^*(u_1, u_n) = \mathcal{C}(u_1)...\mathcal{C}(u_n)$$

Kraft-sum **Definition:** The Kraftsum of a code C is $KS(C) = \sum_u 2^{-|C(u)|}$

- if C is prefix free then $KS(C) \leq 1$
- if C is non singular, then $KS(C) \leq 1 + \min_u |C(u)|$
- $KS(C^n) = KS(C)^n$

Theorem: for any U and associated $p(u)$ there exists a prefix free code C s.t.

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

Theorem: if $KS(C) \leq 1$ then there exists a prefix free code C' such that $|C(u)| = |C'(u)|$ for all u
Corollar: if C is uniquely decodable, then there exists C' that is prefix free with the same word lengths

Entropy **Definition:** the entropy of a random variable U is

$$H(U) = \sum_{u \in U} p(u) \log \frac{1}{p(u)} = E_U \left[\log \frac{1}{p(u)} \right]$$

Theorem: if C is uniquely decodable then $E[|C(U)|] \geq H(U)$

Properties of optimal prefix free codes

1. $p(u) < p(v) \to |u| \geq |v|$
2. The two longest codewords have the same length
3. The 2 least probable letters are assigned codewords that differ in the last bit

1.0.1 Hoffman algorithm

- Combine the 2 least likely symbols
- Sum their probability and assign it a new fictive symbol
- Repeat

2 Entropy and mutual information

Definition 2.1 (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 2.1.

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

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

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

Theorem 2.2.

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

with equality iff Us are independants

Corollary 2.2.1. *if $U_1, \dots, U_n$ are i.i.d. then $H(U_1 \dots U_n) = nH(U_1)$*

Definition 2.3 (Conditional entropy)**.**

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

Theorem 2.3.

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

Theorem 2.4.

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

Definition 2.4 (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

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

Definition 2.5 (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 2.5.

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

Notation 1.

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

Theorem 2.6.

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

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

Proof.

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

□

3 Data processing

Theorem 3.1. *$U - V - W$ is a MC $\iff I(U;W|V) = 0$*

Corollary 3.1.1. *$I(U;V) \geq I(U;W)$ 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 3.2. *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 3.3. *$I(U;V) = 0 \iff U \perp V$*

Definition 3.1 (Entropy rate of a stochastic process)**.** $\lim_{n\to\infty}\frac{1}{n}H(U^n)$ *if the limit exists.*

Theorem 3.4. *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. TODO: I didn't write the proof, Thomas can you write it ? □

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.