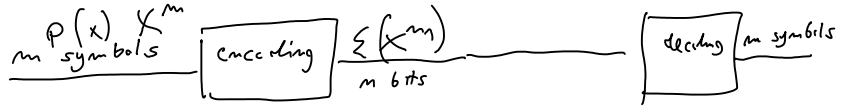
2 - Coding (compresion)

4 października 2010
13:04

2.1 Motivation



Imagine we use a 4-letter alphabet $\{a,b,c,d\}$
and probabilities for respective symbols in the message are:

	a	b	c	d
$P(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

How to encode symbols in order to obtain largest $\frac{m}{n}$

(i) stupid $a=00$ $b=01$ $c=10$ $d=11$ $n=2\cdot m$ $\frac{m}{n}=\frac{1}{2}$

(ii) better $a=0$ $b=10$ $c=110$ $d=111$

$n = m\left(\frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{4}\cdot 3\right) = m\cdot 1{,}75$ $\quad \frac{m}{n} = \frac{1}{1{,}75}$

Intuitions

- for long sequences of symbols we will expect roughly $p(x)\cdot m$ symbols of x (typical sequences) other sequences will almost never happen so we only need to encode typical sequences

- przy kodowaniu (ii) mamy rzadziej wystąpienie a i 1
przy kodowaniu (i) częściej 0 niż 1 - nie optymalne wykorzystanie kanału

2.2 Law of large numbers (empirical mean approach the mean for large number of repts)
(The most important law of statistics)

X - random variable $x \in X$

$X_1, \dots X_N$ - N independent realization of X

$\overline{X_N} = \frac{\sum_i X_i}{N}$ estimator of the mean $\quad \left\{ \text{we want to prove that } \overline{X_N} \xrightarrow{N\to\infty} E(X) = \sum p_i x_i \right.$

Theorem:

Assume X_i indep. random variables with $|E(X)| < \infty$, $|E(X^2)| < \infty$

Then for any ε $\quad p\left(|\overline{X_N} - E(X)| > \varepsilon\right) \xrightarrow{N\to\infty} 0$

Proof:

$P(x_1, \dots x_N) = \prod_i p(x_i)$

$E(\overline{X_N}) = E(X)$ $\qquad E\left(\left|\overline{X_N} - E(X)\right|^2\right) = E\left(\overline{X_N}^2\right) - E(X)^2$

$E(X_N^2) = \frac{N E(X^2)}{N^2} + \frac{1}{N^2}\sum_{i\neq j} E(X)^2 = \frac{E(X^2)}{N} + \frac{N-1}{N} E(X)^2$

$E\left(|\overline{X_N} - E(X)|^2\right) = \frac{E(X^2)}{N} - \frac{E(X)^2}{N} = \frac{E\left((X - E(X))^2\right)}{N} \xrightarrow{N\to\infty} 0$

Nie wprost: Gdyby $\exists_{\varepsilon>0}$ t. że $p\left(|\overline{X_N} - E(X)| > \varepsilon\right) \xrightarrow{N\to\infty} a \quad a > 0$

to znaczyłoby że $E\left(|\overline{X_N} - E(X)|^2\right) \geq a\cdot\varepsilon^2$ sprzeczność. ∎

jeśli $\alpha \varepsilon^2$ możemy wybrać nie $\frac{1}{N}$ to ale wtedy

$\varepsilon^2 \leq \frac{1}{N}$ $\quad \varepsilon N^{\frac{1}{2}} \xrightarrow{N\to\infty} 0$

Uwaga: Dla jakiego N prawdopodobieństwo że $|x_N - E(X)| > \varepsilon$ będzie wciąż nie α

$\frac{E(|x - E(x)|^2)}{N} \geq \alpha \varepsilon^2$ $\quad N < \frac{\mathrm{Var}X}{\alpha \varepsilon^2}$

Prawdopodobieństwo że $|x_n - E(x)| > \varepsilon$ jest $\alpha < \frac{\mathrm{Var}X}{N\varepsilon^2}$

Fakt With probability $> 1 - \frac{\mathrm{Var}X}{N\varepsilon^2}$ a sequence is $|x_n - E(x)| < \varepsilon$

2.3 Typical sequences

$x \in X$ $\quad \underbrace{X_1, \ldots, X_n}_{x^n}$ n independent realization

$P(x^n) = p(x_1) \cdot \ldots \cdot p(x_n)$

$\log p(x^n) = \sum_{x \in X} n_x \log p(x)$

n_x ↑ number of occurences of x

By law of large numbers:

$\frac{1}{n} \log p(x^n) = \sum_{i=1}^{n} \frac{1}{n} \log p(x_i) \xrightarrow{n\to\infty} E(\log p(X)) =$

$= \sum_x p(x) \log p(x)$

Definition:

Shannon entropy $H(X) = -\sum_x p(x) \log p(x)$

$\left(\begin{matrix} x=0 & p \\ x=1 & 1-p \end{matrix} \right.$ $\quad p = \frac{1}{2}$ $H(x) = 1$, $p=1$ $p=0$ $H(x) = 0$ $\left.\right)$

Definition (the most important concept of information theory)

A sequence $x^n = \{x_1, \ldots, x_n\}$ is called ε-typical iff

$$\left| -\frac{1}{n} \log p(x^n) - H(x) \right| < \varepsilon$$

$T_\varepsilon^n = \{x^n : x^n \text{ - } \varepsilon\text{-typical}\}$

↑ set of ε-typical sequences x^n

Facts $\forall_{\delta > 0}$, for sufficiently large n:

(i) $\quad pr[x^n \in T_\varepsilon^n] > 1 - \delta$ $\quad$ asymptotically all sequences are typical

(ii) $\quad (1-\delta) 2^{n(H(x) - \varepsilon)} \leq |T_\varepsilon^n| \leq 2^{n(H(x) + \varepsilon)}$

Proof:

(i) by law of large numbers

$pr[x^n \in T_\varepsilon^n] > 1 - \underbrace{\frac{\mathrm{Var} \log p(X)}{n\varepsilon^2}}_{\delta}$

(ii) $x^n \in T_\varepsilon^n \quad 2^{-n(H(x)+\varepsilon)} \leq p(x^n) \leq 2^{-n(H(x)-\varepsilon)}$

$$\sum_{x^n \in T_\varepsilon^n} p(x^n) \le 1$$

$$2^{-n(H(x)+\varepsilon)} \sum_{x^n \in T_\varepsilon^n} \le 1 \Rightarrow |T_\varepsilon^n| \le 2^{n(H(x)+\varepsilon)}$$

$$p(x^n) \le 2^{-n(H(x)-\varepsilon)}$$

$$|T_\varepsilon^n| > (1-\delta)\, 2^{n(H(x)-\varepsilon)}$$

$$\sum_{x^n \in T_\varepsilon^n} p(x^n) > 1-\delta$$

All typical sequences are almost equally probable (Asymptotic equipartition property). $\simeq 2^{-n(H(x)\pm\varepsilon)}$

All proofs may assume that we have only typical sequences since a-typical sequences are negligible for large n

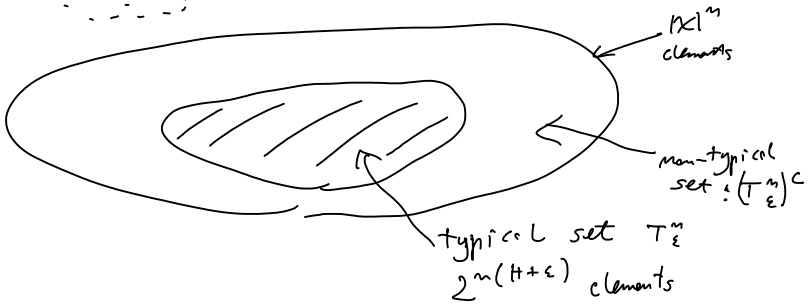
2.4 Shannon coding theorem (data compression)

X^n i.i.d distributed $p(X)$

For large n
The average length of code words per symbol

$$H(x)-\varepsilon \le \frac{1}{n} E(|C(x^n)|) \overset{(i)}{\le} H(x)+\varepsilon$$

Proof:
(i)



We label typical sequences using $n(H+\varepsilon)+1$ bit sequences + one bit 0 for a prefix

The atypical sequences can be encoded using $< n \log|X|+1$ bits (acctually less but it does not matter...) + one bit 1 for a prefix

$|C(x^n)|$ length of a codeword encoding x^n

On average expected length of codeword will be:

$$E(|C(x^n)|) = \sum_{x^n \in T_\varepsilon^n} p(x^n)\,|C(x^n)| + \sum_{x^n \in (T_\varepsilon^n)^c} p(x^n)\,|C(x^n)|$$

$$\le \sum_{x^n \in T_\varepsilon^n} p(x^n)\left[n(H+\varepsilon)+2\right] + \sum_{x^n \in T_\varepsilon^n} p(x^n)\cdot(n\log|X|+2)$$

$$\le n(H+\varepsilon)+2 + \delta(n\log|X|+2) = n(H+\varepsilon')$$

$$\varepsilon' = \varepsilon + \frac{2}{n} + \delta\left(\log|X| + \frac{2}{n}\right)$$ — arbitrarily small

So $E\left(\frac{1}{n}|C(x^n)|\right) \leq H(X) + \varepsilon$

(ii) (to prove strictly we need Kraft inequality ...)

But intuitively it is clear we need at least $2^{n(H(x)-\varepsilon)}$ codewords so $n(H(X)-\varepsilon)$ bits since messages are uniformly distributed.

More precisely if someone took a smaller set $|S| \leq 2^{n(H(X)-\varepsilon-\varepsilon')}$

$$Pr(x^n \in S) \leq Pr(x^n \in S \wedge x^n \in T_\varepsilon^n) + Pr(x^n \in S \wedge x^n \in T_\varepsilon^{n\,c})$$
$$\leq 2^{n(H(x)-\varepsilon-\varepsilon')} \cdot 2^{-n(H(x)-\varepsilon)} + \delta = 2^{-n\varepsilon'} + \delta$$

so it will go to 0 for $n \to \infty$.

2.5 Properties of Entropies ...

Shann entropy

$$H(X) = -\sum_x p(x) \log p(x)$$ (measure of uncertainty)

Joint entropy:

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y)$$

Conditional entropy

$$H(X|Y=y) = -\sum_x p(x|y) \log p(x|y)$$

$$H(X|Y) = -\sum_x p(y)\, p(x|y) \log p(x|y)$$

Lemma: $p(x)$, $q(x)$ – prob. distribution

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \geq 0$$ (relative entropy)

$D(p\|q) = 0$ iff $p = q$ $\quad D(p\|q) \neq D(q\|p)$

Proof:

– $\log x$ – concave function $\quad \sum_i p_i \log x_i \leq \log \sum_i p_i x_i$

$$-\sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_x p(x) \log \frac{q(x)}{p(x)} \leq 0$$ ▨

Properties:

(i) $0 \leq H(X) \leq \log |\mathcal{X}|$

(ii) $H(X,Y) \leq H(X) + H(Y)$ { equality for independent variables

(iii) $H(X,Y) = H(X|Y) + H(Y)$ { chain rule

(iv) $H(X|Y) \leq H(X)$

(v) $H(X,Y|Z) = H(X|YZ) + H(Y|Z)$ { chain rule for cond probs

Proofs:

(i) take $q(x) = \frac{1}{|\mathcal{X}|}$, $p(x)$

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_x p(x) \log p(x) + \sum_x p(x) \log |\mathcal{X}| \geq 0$$

$H(X) \leq \log |\mathcal{X}|$

(ii) $D(p(x,y) \| p_x(x) \cdot p_y(y)) =$ $p_x(x)$, $p_y(y)$ - marginal distributions

$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p_x(x) \cdot p_y(y)} = -H(X,Y) + H(X) + H(Y) \geq 0$

(iii) $H(X|Y) = -\sum_{x,y} p(x,y) \log p(x,y) = -\sum p(x,y) \log p(x|y) \cdot p(y) =$

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

(iv) from (ii) and (iii)

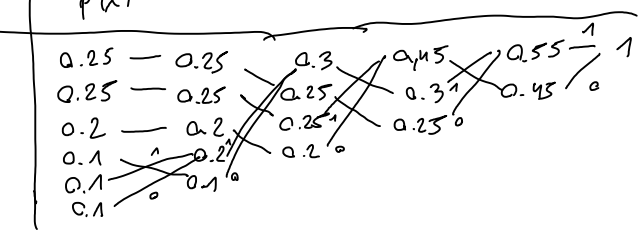
2.6 Huffman coding

$x_i \quad p(x)$

- we take two x with smallest probabilities assign 0 and 1 and group them into one new symbol

Example:

codeword	X	p(x)
10	1	0.25
01	2	0.25
00	3	0.2
110	4	0.1
1111	5	0.1
1110	6	0.1



on average: $\frac{1}{2} \cdot 2 + \frac{1}{5} \cdot 2 + 0.1 \cdot 3 + 0.2 \cdot 4 = 1.4 + 0.3 + 0.8 = 2.5$ bits

$H(X) = 2.461$