



Sessão nº5

Redundância e Introdução à Codificação de Fonte



Redundância num conjunto binário



- Considere um conjunto binário onde a probabilidade de um dos textos em claro é p . Quando é que existe redundância?
 - Porque temos de usar pelo menos um símbolo para representar cada um dos textos então sempre que as suas probabilidades determinarem uma **entropia diferente de 1** existe redundância na representação destes textos

Nota: a entropia é uma medida da quantidade de informação média necessária para descrever os textos



Redundância num conjunto quaternário



- Dado o conjunto de 4 textos em claro onde $p(a)=0,5$, $p(b)=0,25$, $p(c)=0,125$ e $p(d)=0,125$ (logo $H(P)=1,75$) quando é que existe redundância?
 - Usando 2 dígitos binários na representação dos textos existe redundância porque a quantidade de informação necessária em média para os descrever é 1,75 bits.



Redundância numa linguagem



- Seja L uma linguagem natural. Define-se **redundância** de L como sendo

$$R_L = 1 - (H(L) / (N \times \log_2 |P|))$$

- $H(L)$ representa a entropia do conjunto dos textos de L ou a quantidade de informação média contida em cada letra dum texto “com significado”.
- N é o número médio de símbolos usados para representar os textos de L .
- $|P|$ é a dimensão do alfabeto usado na representação dos textos de L .



Redundância do Inglês



- Seja $|\mathbf{P}|=26$ a dimensão do alfabeto; então um texto pode ser representado com 1 símbolo de \mathbf{P} (letra), logo a quantidade de informação média associada a cada letra é $1 \times \log_2 26 \approx 4,7$ [bit] (numa sequência aleatória com letras equiprováveis).
- Considerando o valor empírico $H(L) \approx 1,25$ [bit/letra] vem

$$\begin{aligned}R_L &= 1 - (1,25 / (1 \times \log_2 26)) \\ &\approx 1 - (1,25/4,7) \approx \mathbf{0,75}\end{aligned}$$



Entropia duma linguagem



- Seja L uma linguagem natural. Define-se **entropia** de L como sendo

$$H(L) = \lim_{n \rightarrow \infty} H(\mathbf{P}^n) / n \text{ [bit/letra]}$$

- \mathbf{P}^n representa uma v.a. que tem como distribuição de probabilidades a distribuição que se observa para os n -gramas do texto



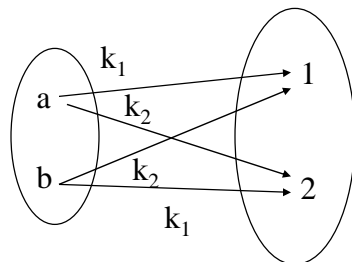
Entropia do Português e do Inglês (estimativas)



Ficheiro	H(X)	H(X, Y)/2	H(X, Y, Z)/3
pt.z26	3,980	3,678	3,282
trab.z26	4,081	3,780	3,416
tese22.z26	4,018	3,724	3,391
média (PT)	4,026	3,727	3,363
devices.z26	4,221	3,918	3,562
history3.z26	4,246	3,966	3,614
appnote.z26	4,176	3,825	3,382
bertrand.z26	4,159	3,835	3,510
média (UK)	4,201	3,886	3,517
ref. Stinson	4,190	3,900	-
rand.z26	4,700	4,700	4,696



Análise do sistema *One Time Pad* de Vernam



Dados:

$$p(a) = 1/4, \quad p(b) = 3/4$$

$$P(k_1) = 1/2, \quad p(k_2) = 1/2$$

Verifica-se que

$$H(P) \approx 0,81 \text{ [bit/texto]}$$

$$H(K) = 1 \text{ [bit/chave]}$$

$$H(C) = 1 \text{ [bit/cripto]}$$

$$H(K|C) = H(P) + H(K) - H(C) ?$$

$$H(K|C) \approx 1 + 0,81 - 1 \approx 0,81 \text{ é menor !!!}$$

• Mas se maximizarmos $H(P)$ temos que $H(K|C)=H(K)$



Conceitos básicos da Teoria da Informação

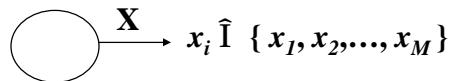


- Medida de quantidade de informação (entropia).
- Capacidade de informação dum canal.
- Codificação:

- **codificação de fonte**
 - [cifra]
 - codificação de canal



Modelo de fonte discreta sem memória

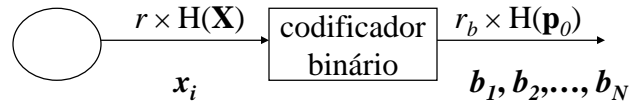


- x_i representa um símbolo da fonte
- $p(x_i)$ é a probabilidade dum símbolo tal que $\sum_{x_i \in X} p(x_i) = 1$
- fonte estacionária logo as probabilidades mantêm-se ao longo do tempo
- símbolos independentes (não existe memória)
- símbolos produzidos ao ritmo médio de r símbolos/s
- define-se **velocidade de informação** da fonte por

$$R = r \times H(X) \quad \text{onde } H(X) \text{ é a entropia da fonte}$$



Modelo genérico da codificação de fonte



- r_b representa o ritmo médio de dígitos binários produzidos pelo codificador
- $H(p_0)$ representa a entropia da fonte binária onde a probabilidade do símbolo $\mathbf{0}$ é p_0
- **importante:** no processo de codificação não pode haver “ganho” nem “perca” de informação
- Define-se **comprimento médio N** do código gerado como

$$N = \sum_{x_i \in X} p(x_i) \times n(x_i)$$

sendo $n(x_i)$ o nº de dígitos binários usados para representar o símbolo da fonte x_i



Teorema da codificação de fonte



- Shannon estabeleceu⁽¹⁾ que

$$H(\mathbf{X}) \leq N \leq H(\mathbf{X}) + \varepsilon \quad \text{com } \varepsilon \geq 0$$

– quando $N = H(\mathbf{X})$ o código diz-se **ótimo**

⁽¹⁾ *A mathematical theory of communication*, 1948



Classes de códigos



- Códigos **singulares**
– ex.: **1, 1, 1, 1**
- Códigos **não singulares** que **não têm decodificação única**
– ex.: **0, 010, 01, 10**
- Códigos **não singulares** de **decodificação única**
– ex. “comma code”: **0, 01, 011, 0111, ...**
 - Códigos **prefixo** ou **instantâneo**
– ex. código de Huffman: **0, 10, 110, 111**



Características dos códigos prefixo



- Os comprimentos das palavras do código satisfazem a **desigualdade de Kraft**

$$\sum_i 2^{-n(x_i)} \leq 1$$

- condição necessária para que o código seja prefixo
- atendendo a esta restrição demonstra-se que o **comprimento médio de qualquer código prefixo** satisfaz

$$H(X) \leq N \leq H(X) + 1$$



Comprimento médio dos códigos prefixo (demonstração)



- Por definição $N = \sum_{x_i \in X} p(x_i) \times n(x_i)$
- Os códigos prefixo satisfazem $\sum_i 2^{-n(x_i)} \leq 1$
- Os valores de $n(x_i)$ que minimizam N e que verificam a desigualdade de Kraft obtêm-se por

$$n(x_i)^* = -\log_2 p(x_i)$$

usando $n(x_i) = n(x_i)^*$ vem

$$N = \sum_{x_i \in X} p(x_i) \times n(x_i)^* = \sum_{x_i \in X} p(x_i) \times -\log_2 p(x_i) = H(\mathbf{X})$$

usando $n(x_i) \neq n(x_i)^*$

$$N = \sum_{x_i \in X} p(x_i) \times \lceil n(x_i)^* \rceil \leq \sum_{x_i \in X} p(x_i) \times (-\log_2 p(x_i) + 1) = H(\mathbf{X}) + 1$$

logo $H(\mathbf{X}) \leq N \leq H(\mathbf{X}) + 1$ c.q.d.



Código de dimensão fixa - binário natural



x_i	$p(x_i)$	$-\log_2 p(x_i)$	$C(x_i)$
a	1/2	1	00
b	1/4	2	01
c	1/8	3	10
d	1/8	3	11

- Entropia da fonte

$$H(\mathbf{X}) = \sum_{x_i \in X} p(x_i) \times -\log_2 p(x_i) = 1,75 \text{ [bit/símbolo]}$$

- Comprimento médio do código

$$N = \sum_{x_i \in X} p(x_i) \times n(x_i) = 2 \text{ dígitos}$$

– note-se que $N = H(\mathbf{X}) + 0,25$ sendo $\epsilon = 0,25$

- Redundância na representação dos símbolos da fonte

$$R = 1 - (1,75 / (2 \times \log_2 2)) = 0,125$$



Código de dimensão variável - “comma code”



x_i	$p(x_i)$	$-\log_2 p(x_i)$	$C(x_i)$
a	1/2	1	0
b	1/4	2	01
c	1/8	3	011
d	1/8	3	0111

- Entropia da fonte

$$\mathbf{H}(\mathbf{X}) = \sum_{x_i \in X} p(x_i) \times -\log_2 p(x_i) = 1,75 \text{ [bit/símbolo]}$$

- Comprimento médio do código

$$\mathbf{N} = \sum_{x_i \in X} p(x_i) \times n(x_i) = 1,875 \text{ dígitos}$$

– note-se que $\mathbf{N} = \mathbf{H}(\mathbf{X}) + 0,125$ sendo $\epsilon = 0,125$

- Redundância na representação dos símbolos da fonte

$$\mathbf{R} = 1 - (1,75 / (1,875 \times \log_2 2)) = 0,067$$

