cadeias de Markov

- $p_{t+1}(x_{t+1}) = \sum_{x_1,\ldots,x_t,x_{t+2},\ldots,x_N} p_N(x_1, x_2, \cdots x_N)$
- $p_{t+1}(x_{t+1}) = \sum_{x_t \in \mathscr{X}} p_t(x_t) w(x_t \to x_{t+1})$
- notação vetorial e matricial
  - $\vec{p}_t =$ vetor coluna com elementos $p_t(a_1), p_t(a_2), \ldots, p_t(a_M)$ com $\mathscr{X} = \{a_1, \ldots, a_M\}$
  - $\Pi_{y,x} = w(x \to y)$ com $x \in \mathscr{X}$ e $y \in \mathscr{X}$ é a matriz de Markov
  - $\vec{p}_{t+1} = \Pi \vec{p}_t = \Pi^t \vec{p}_1$.
  - distribuição estacionária, $\vec{p}^* = \lim_{t\to\infty} \vec{p}_t$
  - $\vec{p}^* = \Pi \vec{p}^*$, e $\vec{p}^*$é um vetor próprio à direita de Π

com valor próprio 1

Exemplo:

ver exercicio 11

Duas caixas e N partículas

Em cada passo uma partícula é escolhida ao acaso e mudada de caixa

O número de partículas numa caixa, 0<=n<= N, é uma cadeia de Markov

$n \longrightarrow n-1$ ou $n \longrightarrow n+1$

Constroi a matriz Π

correlação e informação mútua

definições

- Entropia conjunta.
  $S_{X,Y} = -\sum_{x,y\in\mathscr{X}} p_{X,Y}(x,y)\log_2 p_{X,Y}(x,y)$

  $p_{xy}(x,y) = p_x(x)\, p_{y|x}(y|x)$
- Entropia condicional
  $S_{Y|X} = -\sum_{x,y\in\mathscr{X}} p_X(x)\, p_{Y|X}(y|x)\, \log_2 p_{Y|X}(y|x)$
- Regra da cadeia
  $S_{X,Y} = S_X + S_{Y|X}$
- Informação mútua, $I_{X,Y}$ mede a redução na incerteza em X pelo conhecimento de Y e é simétrico em X e Y. Temos
  $I_{X,Y} = \sum_{x,y\in\mathscr{X}} p_{X,Y}(x,y)\log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}$, $I_{X,Y} \geq 0$ e $I_{X,Y} = 0$ se e só se X e Y são independentes.
  $I_{X,Y} = S_Y - S_{Y|X} = S_X - S_{X|Y}$

Entropia Conjunta

Exercicio 14

Formula de Bayes $p_{xy}(x,y) = p_x(x)\, p_{y|x}(y|x)$

$$S_{x,y} = -\sum_{x,y} p_{x,y}(x,y) \log p_{x,y}(x,y)$$

$$S_{x,y} = -\sum_{x,y} p_x(x)\, p_{y|x}(y|x) \log\left[p_x(x)\, p_{y|x}(y|x)\right]$$

$$S_{y,x} = \underbrace{-\sum_{x,y} p_x(x)\, p_{y|x}(y|x) \log\left(p_{y|x}(y|x)\right)}_{S_{y|x}}$$

$$\sum_y p_{y|x}(y|x) = 1$$

$$-\sum_{x,y} p_x(x)\, p_{y|x}(y|x) \log p_x(x)$$

Então

$$-\sum_{x,y} p_X(x)\, p_{Y|X}(y|x) \log p_X(x) = S_X$$

Portanto

$$S_{Y,X} = S_X + S_{Y|X}$$

ou

$$S_{Y,X} = S_Y + S_{X|Y}$$

Informação mútua

definição

$$I_{x,y} = \sum_{x,y} p_{x,y}(x,y) \log\left[\frac{p_{x,y}(x,y)}{p_x(x)\, p_y(y)}\right]$$

$$= -S_{y,x} + S_x + S_y$$

$$I_{x,y} = -\left(S_{y|x} + S_x\right) + S_x + S_y = -S_{y|x} + S_y$$

$$I_{x,y} = -\left(S_{x|y} + S_y\right) + S_x + S_y = -S_{x|y} + S_x$$

Informação mútua condicional

Exercicio 15

definições

$$\boxed{I_{X,Y|Z} = \sum_{z,x,y} p_Z(z)\, p_{X,Y|Z}(x,y|z) \log \frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)\; p_{Y|Z}(y|z)}}$$

$$p_{X,Y,Z}(x,y,z) = p_Z(z)\, p_{X,Y|Z}(x,y|z)$$

$$p_{Y|XZ}(y|x,z) = p(y|z, y|x)$$

definições

$$\boxed{I_{X,YZ} = \sum_{z,x,y} p_{X,YZ}(x,y\,z) \log \frac{p_{X,YZ}(x,y\,z)}{p_X(x)\; p_{YZ}(yz)}}$$

$$p_{X,YZ}(x,y\,z) = p_{XZ}(x,z)\, p_{Y|XZ}(y|x,z) = p_{XZ}(x,z)\, p_{Y|Z,Y|X}(y|z, y|x)$$

$$p_{Y|Z,Y|X}(y|z, y|x) = p_{Y|Z,X}(y|z, x) / p_{X|Z}(x|z)$$

$$p_{X,YZ}(x,yz) = p_{XZ}(x,z)\,\frac{p_{Y|Z,X}(y|z,x)}{p_{X|Z}(x|z)}$$

$$p_{YZ}(yz) = p_{Y|Z}(y|z)\,p_Z(z)$$

$$I_{X,YZ} = \sum_{z,x,y} p_{X,YZ}(x,yz)\log\frac{p_{X,YZ}(x,yz)}{p_X(x)\,p_{YZ}(yz)}$$

$$I_{X,YZ} = \sum_{z,x,y} p_{X,YZ}(x,yz)\log\left[\frac{p_{XZ}(x,z)}{p_X(x)}\,\frac{p_{Y|Z,X}(y|z,x)}{p_{X|Z}(x|z)\,p_{Y|Z}(y|z)\,p_Z(z)}\right]$$

$$\boxed{I_{X,YZ} = I_{X,Z} + I_{X,Y|Z}}$$

Regra da cadeia para Informação Mútua

taxa de entropia

$X_{t-1} \to X_t \to X_{t+1}$

- para uma cadeia de Markov em que $X \to Y \to Z$ estão ordenados temporalmente $I_{X,Z} \leq I_{X,Y}$

Exercicio 15

  - demonstração: $I_{X,(YZ)} = I_{X,Z} + I_{X,Y|Z} = I_{X,Y} + \underbrace{I_{X,Z|Y}}_{=0}$ como $I_{X,Z|Y} = 0$ devido a que X e Z são indpendentes dado Y e $I_{X,Y|Z} \geq 0$ temos $I_{X,Z} \leq I_{X,Y}$

$I_{X,Z} = I_{X,Y} - \underbrace{I_{X,Y|Z}}_{\geq 0} \leq I_{X,Y}$

- dada a sequência, $\underline{X_N} = \{X_t\}$, $t = 1, ..., N$ define-se a taxa de entropia como $h_{\underline{X}} = \lim_{N\to\infty} S_{\underline{X_N}}/N$
- para uma cadeia de Markov em que existe uma distribuição de probabilidade estacionária $p^*(x) = \lim_{N\to\infty} p_N^*(x)$ mostra-se que:

$$h_{\underline{X}} = -\sum_{x,y\in\mathscr{X}} p^*(x)w(x\to y)\log w(x\to y)$$

compressão de dados

- definição do problema
- dada uma sequência de variáveis aleatórias , $\underline{X} = \{X_1, X_2, \cdots, X_N\}$com $X_i \in \mathscr{X}$ pretende-se armazenar informação de uma realização $\underline{x} = \{x_1, x_2, \cdots, x_N\}$ da forma mais compacta possível.
- codificação
  - define-se um mapeamento w entre cada possível sequência em $\mathscr{X}^N$ numa string binária de tamanho arbitrário $\{0,1\}^*$, $\underline{x} \longmapsto w(\underline{x})$
- compressão de uma sequência
  - uma sequência grande (stream) é partida em blocos de tamanho N
  - cada bloco é codificado com w
  - as palavras de código são concatenadas para formar uma sequência codificada mais compacta

compressão de dados

Example 1.14 Take $N = 1$, and consider a random variable X which takes values in $\mathcal{X} = \{1, 2, \ldots, 8\}$ with probabilities $p(1) = 1/2$, $p(2) = 1/4$, $p(3) = 1/8$, $p(4) = 1/16$, $p(5) = 1/32$, $p(6) = 1/64$, $p(7) = 1/128$, and $p(8) = 1/128$. Consider the two codes w_1 and w_2 defined by the table below:

x A B C . . . .

x	$p(x)$	$w_1(x)$	$w_2(x)$
1	1/2	000	0
2	1/4	001	10
3	1/8	010	110
4	1/16	011	1110
5	1/32	100	11110
6	1/64	101	111110
7	1/128	110	1111110
8	1/128	111	11111110

(1.31)

These two codes are instantaneous. For instance, looking at the code w_2, the encoded string 10001101110010 can be parsed in only one way, since each symbol 0 ends a codeword. It thus corresponds to the sequence $x_1 = 2, x_2 = 1, x_3 = 1, x_4 = 3, x_5 = 4, x_6 = 1, x_7 = 2$. The average length of code w_1 is $L(w_1) = 3$, and the average length of code w_2 is $L(w_2) = 247/128$. Notice that w_2 achieves a shorter average length because it assigns the shortest codeword (namely 0) to the most probable symbol (i.e. $x = 1$).

000 e 0000 não poderiam ser simultâneamente palavras de código porque 000 é a string inicial de 0000

. . . $\boxed{11}$ . . .   . A . . . .

110 ?

texto codificado   texto descodificado

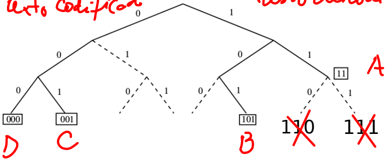


Figure: códigos de descodificação única. Nenhuma palavra de código é antecessora de outra palavra de código.

. . . 101 . . . B

teorema da compressão ótima

$p_x(x)$ = probabilidade do caracter x no texto (fonte)
$S_x = -\sum_x p_x(x) \log p_x(x)$

- comprimento médio do código
  $L(w) = \sum_{\underline{x} \in \mathscr{X}^{\mathscr{N}}} p(\underline{x}) l_w(\underline{x})$

  $l_w(\underline{x})$ = comprimento do código de $\underline{x}$

- Teorema da compressão ótima

  Shannon

  - Seja L_N^* o comprimento ótimo de um código para a sequência $\underline{X} = \{X_1, X_2, \cdots, X_N\}$com entropia $S_{\underline{X}}$
    (1) para qualquer $N \geq 1$ temos $S_{\underline{X}} \leq L_N^* \leq S_{\underline{X}} + 1$
    (2) se a fonte tiver uma taxa de entropia finita, $h = \lim_{N \to \infty} S_{\underline{X}}/N$
    então $\lim_{N \to \infty} L_N^*/N = h$

teorema da compressão ótima

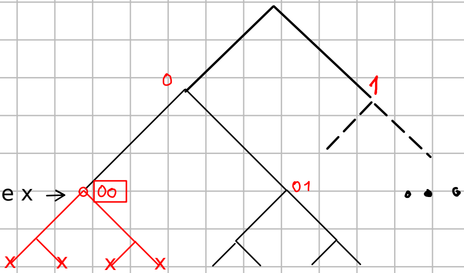
- Consideremos um código binário instantâneo. Seja L_M o comprimento da palavra de código maior. Para cada palavra de código de comprimento $l_w(\underline{x})$ o número de descendentes de comprimento L_M não usados é $2^{L_M - l_w(\underline{x})}$. O número total de códigos não usados no nível L_M é $\sum_{\underline{x}\in\mathscr{X}^{\mathscr{N}}} 2^{L_M - l_w(\underline{x})}$. Este número tem que ser inferior ao número total possível com L_M bits, 2^{L_M}, ou seja $\sum_{\underline{x}\in\mathscr{X}^{\mathscr{N}}} 2^{-l_w(\underline{x})} \leq 1$.

.

Dado $\underline{x}$

$l_w(\underline{x}) = 2$

codigo de x $\rightarrow$



$L_M = 4$

Há 4 códigos no nível $L_M = 4$

que não podem ser usados
por se ter usado $\boxed{00}$

$$2^{L_M - l_w(\underline{x})} = 4$$

teorema da compressão ótima

- Para dados com distribuição $p_{\underline{X}}(\underline{x})$ o comprimento médio é $L = \sum_{\underline{x}\in\mathscr{X}^{\mathscr{N}}} p_{\underline{X}}(\underline{x}) l_w(\underline{x})$. Pode minimizar-se L com a restrição $\sum_{\underline{x}\in\mathscr{X}^{\mathscr{N}}} 2^{-l_w(\underline{x})} = a \leq 1$. Ou seja:

$$\frac{\partial}{\partial l_w(\underline{x})}\left[\sum_{\underline{x}\in\mathscr{X}^{\mathscr{N}}} p_{\underline{X}}(\underline{x}) l_w(\underline{x}) + \beta\left(\sum_{\underline{x}\in\mathscr{X}^{\mathscr{N}}} 2^{-l_w(\underline{x})} - a\right)\right] = 0$$

multiplicador de Lagrange
minimização com restrição

$$\frac{\partial}{\partial l_w(\underline{x})} \sum_{\underline{x}} p_{\underline{X}}(\underline{x})\, l_w(\underline{x}) = p_{\underline{X}}(\underline{x})$$

$$\frac{\partial}{\partial l_w(\underline{x})} 2^{-l_w(\underline{x})} = \frac{\partial}{\partial l_w(\underline{x})} e^{-\ln 2\, l_w(\underline{x})} =$$

$$= -\ln 2\; e^{-\ln 2\, l_w(\underline{x})}$$

$$= -\ln 2\; 2^{-l_w(\underline{x})}$$

$$p_{\underline{x}}(\underline{x}) - \beta \ln 2\; 2^{-l_w(\underline{x})} = 0$$

teorema da compressão ótima

- Obtém-se $p_{\underline{X}}(\underline{x}) - \beta \ln 2\, 2^{-l_w(\underline{x})} = 0$. Como $\sum_{\underline{x} \in \mathscr{X}^{\mathscr{N}}} p_{\underline{X}}(\underline{x}) = 1$ temos $\beta \ln 2 = \frac{1}{a}$ e $l_w(\underline{x}) = -\log_2 \frac{p_{\underline{X}}(x)}{a}$. Como $a \leq 1$ o menor $l_w^*(\underline{x})$ consegue-se com $a = 1$ e $l_w^*(\underline{x}) = -\log_2 p_{\underline{X}}(\underline{x})$. Então $L^* = -\sum_{\underline{x} \in \mathscr{X}^{\mathscr{N}}} p_{\underline{X}}(\underline{x}) \log_2 p_{\underline{X}}(\underline{x}) = S_{\underline{X}}$. Como L^* é inteiro temos $S_{\underline{X}} \leq L^* \leq S_{\underline{X}} + 1$. o comprimento médio do código ótimo é igual à entropia da fonte
- Dividindo por N e tomando o limite $N \to \infty$ temos $h \leq \lim_{N \to \infty} \frac{L^*}{N} \leq h$, com h a taxa de entropia, $h = \lim_{N \to \infty} \frac{S_{\underline{X}}}{N}$

A entropia da fonte de informação determina a compressão máxima possível

algoritmo de Huffman

Para encontrar um código instantâneo ótimo.

- Dado um vetor de probabilidade, $p_X(x)$ dos caracteres do alfabeto, $\mathscr{X}$
  - repetir até $\mathscr{X}$ ter 1 elemento
    - remover os dois caracteres de probabilidade mais baixa do alfabeto, c_{i_1} e c_{i_2} criando o alfabeto $\mathscr{X} = \mathscr{X} \setminus \{c_{i_1}, c_{i_2}\}$.
    - adicionar ao alfabeto um caracter c_p, $\mathscr{X} = \{\mathscr{X}, c_p\}$, atribuindo a probabilidade, $p_X(c_p) = p_X(c_{i_1}) + p_X(c_{i_2})$
    - os dois caracteres removidos são os filhos 1 e 2 de c_p
- seguir a árvore binária desde a raiz atribuindo ao Filho 1 bit 0 e Filho 2 bit 1. Os nodos sem filhos são as palavras de código do alfabeto original $\mathscr{X}$

c_p

c_{i_1} c_{i_2}

canais de comunicação

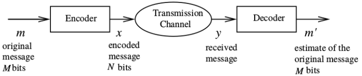


Fig. 1.3 Typical flowchart of a communication device.

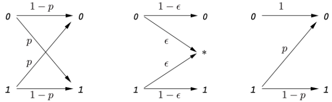


Fig. 1.4 Three communication channels. *Left*: the binary symmetric channel. An error in the transmission, in which the output bit is the opposite of the input one, occurs with probability p. *Middle*: the binary erasure channel. An error in the transmission, signaled by the output $*$, occurs with probability ϵ. *Right*: the Z channel. An error occurs with probability p whenever a 1 is transmitted.

Figure: canais binários de comunicação

- mensagem codificada $\underline{x} = (x_1, x_2, \cdots, x_N)$, with N bits
- mensagem recebida $\underline{y} = (y_1, y_2, \cdots, y_N)$, (Binary Channel)

Canais de comunicação

- canal sem memória $Q(\underline{y}|\underline{x}) = \prod_{i=1,\ldots,N} Q(y_i|x_i)$
- capacidade do canal, $C = \max_{p(x)} I_{X,Y}$ com

  p(x) = distribuição de probabilidade da fonte

  $$I_{X,Y} = \sum_{x\in\mathscr{X}} \sum_{y\in\mathscr{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \text{ e } p(x,y) = p(x)Q(y|x)$$
- C=0 se X e Y não estão correlacionados $p(x,y) = p(x)p(y)$ e se $p(x,y) = p(x)\delta_{y,x}$, $C = 1$

0<= C <=1 caracteriza a "qualidade" do canal de comunicação

Uma maior correlação entre a mensagem recebida
e a mensagem codificada enviada significa maior capacidade

Códigos de correção de erros

- mensagem original $\underline{m} = (m_1, m_2, \cdots, m_M)$, M bits (*Block*)
- mensagem codificada $\underline{x}(\underline{m}) = (x_1, x_2, \cdots, x_N)$, $N \geq M$
- mensagem recebida $\underline{y} = (y_1, y_2, \cdots, y_N)$
- mensagem descodificada $\underline{m}' = d(\underline{y})$

$0 \leq R \leq 1$

- Taxa do código: $R = \frac{M}{N}$ (preferencialmente a maior possível)
- Probabilidade de erro (no *Block* $\underline{m}$):
$$P_B(\underline{m}) = \sum_{\underline{y}} Q\left(\underline{y}|\underline{x}(\underline{m})\right) I(d(\underline{y}) \neq \underline{m})$$
- Prob. erro média: $P_B^{av} = \frac{1}{2^M} \sum_{\underline{m}} P_B(\underline{m})$