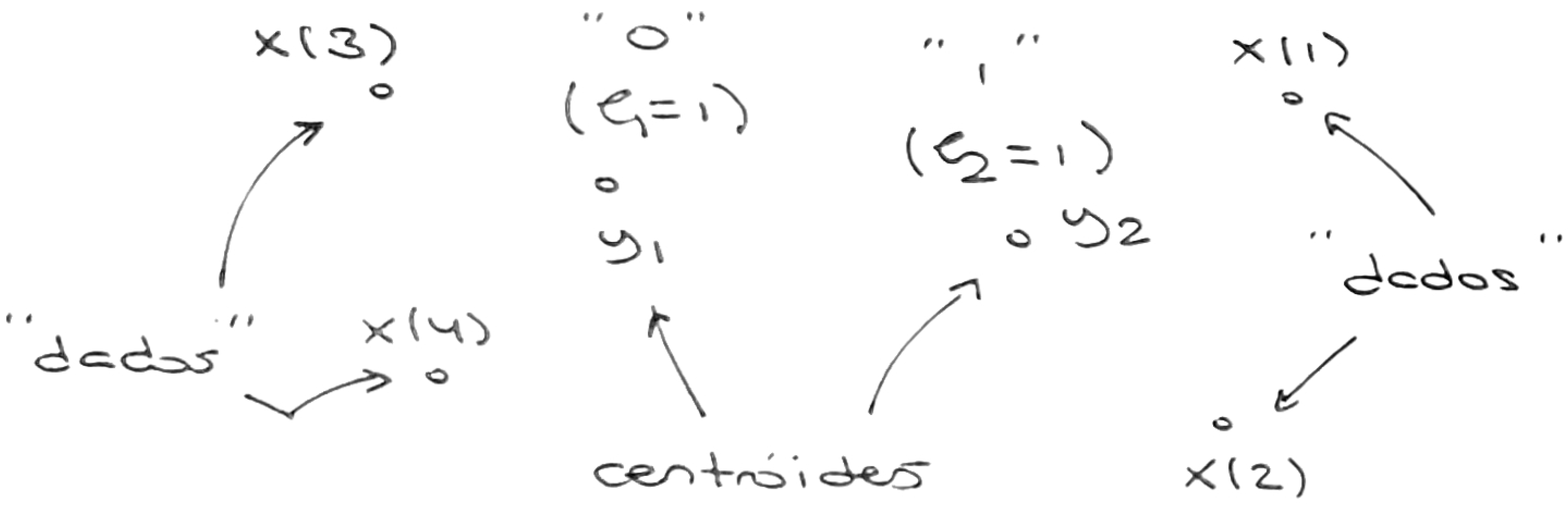
UFRJ — COPPE — PEE — CPE723 — Otimização Natural

Simulated Annealing — Aula 06 — Deterministic Annealing

Voltando à Aula 03, Exemplo 4:



$$D = \frac{1}{N}\sum_{n=1}^{N} d(x(n), y_{k(n)}) = D(Y)$$

$$R = \frac{1}{N}\sum_{n=1}^{N} \ell_{k(n)} = \frac{1+1+\ldots+1}{N} = 1$$

"data rate"
(taxa de dados)

(1 bit/vetor)

ℓ_i $(i=1,2)$ ("length"): comprimento da "palavra binária" transmitida.

Projeto: $Y^* = \underset{Y}{\operatorname{argmin}}\, D(Y)$

Minimização de $D(Y) = \frac{1}{N} \sum d(x(n), y_{k(n)})$ por $\boxed{\text{método gradiente}}$, ou seja,

solução via GLA (Generalized Lloyd Algorithm*) (ou LBG). Duas etapas.

Etapa 1 (Condição da Partição — "calcular partição, dado Y fixo")

$$D = \frac{1}{4}\left(\|x(1) - y_{\bigcirc}\|^2 + \|x(2) - y_{\bigcirc}\|^2 + \|x(3) - y_{\bigcirc}\|^2 + \|x(4) - y_{\bigcirc}\|^2 \right)$$

$$k(1) = \arg\min_i \|x(1) - y_i\|$$

$$\forall n : k(n) = \arg\min_i \|x(n) - y_i\|$$

(*) A. Gersho e R.M. Gray. Vector Quantization and Signal Compression, Ed. Kluwer, 1992.

Y. Linde, A. Buzo e R.M. Gray. An algorithm for vector quantizer design. IEEE Trans. Commun., vol. 28, pp. 84-95, 1980.

Etapa 2 (Condição do Centróide — "calcular Y, dada uma partição fixa")

$$D = \frac{1}{4}\left(\|x(1) - y_2\|^2 + \|x(2) - y_2\|^2 + \|x(3) - y_1\|^2 + \|x(4) - y_1\|^2 \right)$$

$\frac{dD}{dy_k} = ?$ ⟶ derivada de função escalar em relação a um vetor (slide seguinte)

$$\frac{d}{dy_2}\left(\|x(1) - y_2\|^2\right) = \frac{d}{dy_2}\left(x^T(1)x(1)\right) - \frac{d}{dy_2}\left(2x^T(1)y_2\right) + \frac{d}{dy_2}\left(y_2^T y_2\right) = -2x(1) + 2y_2$$

$$\frac{dD}{dy_2} = \frac{1}{4}\left(-2x(1) - 2y_2 - 2x(2) - 2y_2\right) = 0$$ (otimizando...)

$x(1) \in$ cluster 1

$x(2) \in$ cluster $\cancel{2}$ 1

$$y_2 = \frac{x(1) + x(2)}{N_1}$$

N_1: nº de vetores $x(n) \in$ cluster 1

Em geral: $\boxed{y_k = \frac{1}{N_k} \sum x(n) \text{ "} \in y_k \text{"}}$

Derivada da função escalar em relação a um vetor:

$$\frac{d(y^T y)}{dy} = \begin{bmatrix} \dfrac{d(y_1^2 + \dots + y_m^2)}{dy_1} \\ \vdots \\ \dfrac{d(y_1^2 + \dots + y_m^2)}{dy_m} \end{bmatrix} = \begin{bmatrix} 2y_1 \\ \vdots \\ 2y_m \end{bmatrix} = 2y$$

(vetor)

$$\frac{d(x^T y)}{dy} = \begin{bmatrix} \dfrac{d(x_1 y_1 + \dots + x_m y_m)}{dy_1} \\ \vdots \\ \dfrac{d(x_1 y_1 + \dots + x_m y_m)}{dy_m} \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = x$$

(vetor)

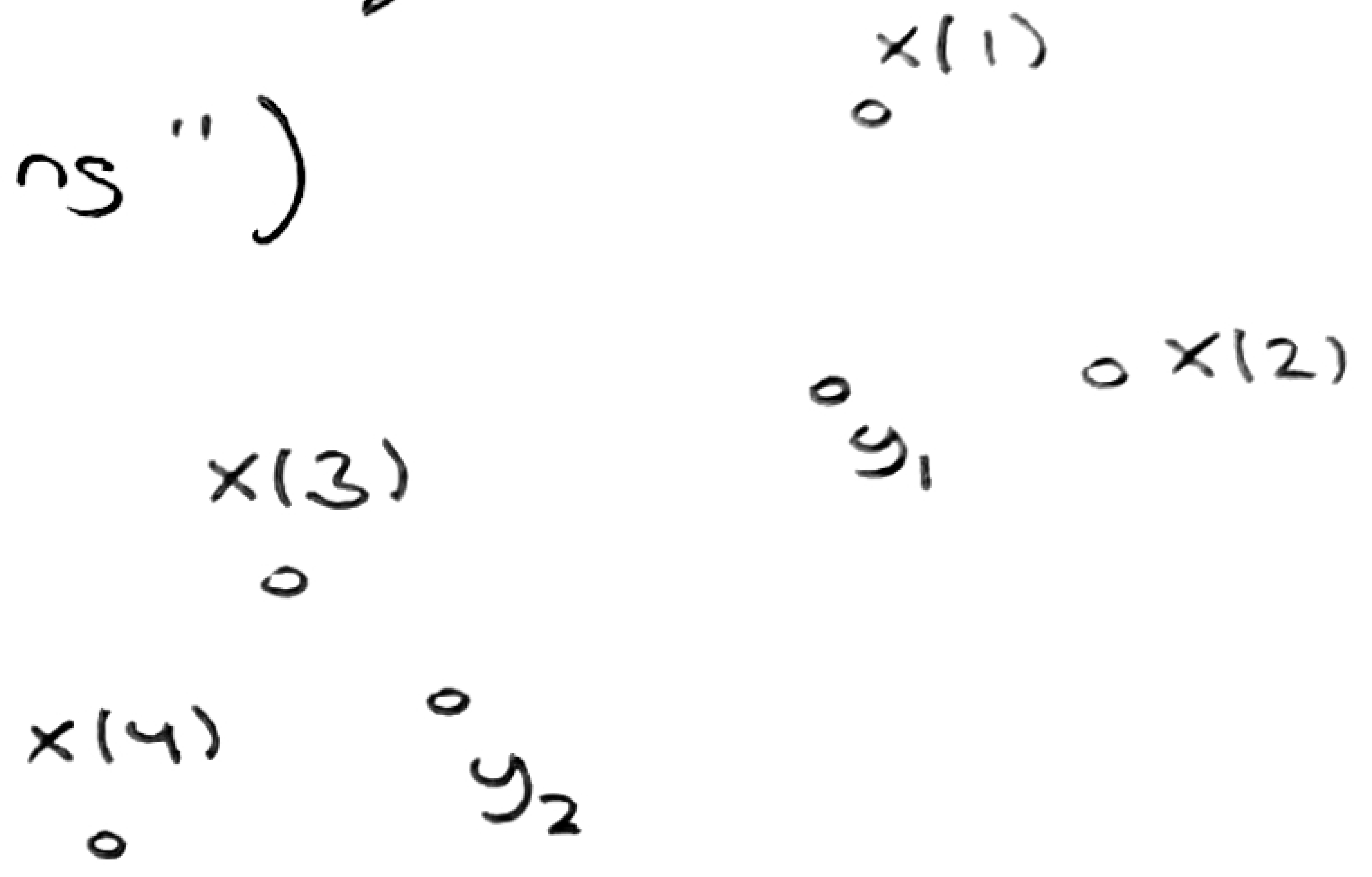
— Deterministic Annealing (DA) (Rose, 1993 e Rose, 1998)

- Minimização de funções não-convexas (habilidade de evitar mínimos locais)
- Aplicabilidade a diferentes tipos de problemas
- Número mínimo de restrições (máxima entropia)
- Analogia com termodinâmica estatística
- Ler: Rose, 1993 (Seções I e II)

  Rose, 1998 (até o final da seção II.A)

— Deterministic Annealing (DA) (Rose, 1993 e Rose, 1998)

Vamos estudar DA através de um exemplo ("soft clustering")



Matriz $p_{Y|X}$:

$(p_{Y|X}(y|x))$

	x(1)	x(2)	x(3)	x(4)
y_1	0.7	0.8	0.4	0.3
y_2	0.3	0.2	0.6	0.7

$$D = \sum_x p_x \sum_y p_{Y|X} d_{XY} = D(p_{Y|X}) = \frac{1}{4}\Big(0.7\,\|x(1)-y_1\|^2 + 0.3\,\|x(1)-y_2\|^2$$

$$+ 0.8\,\|x(2)-y_1\|^2 + 0.2\,\|x(2)-y_2\|^2$$

$$+ 0.4\,\|x(3)-y_1\|^2 + 0.6\,\|x(3)-y_2\|^2$$

$$+ 0.3\,\|x(4)-y_1\|^2 + 0.7\,\|x(4)-y_2\|^2\Big)$$

$$\left(D = \sum_x \sum_y p_{XY} d_{XY}\right)$$

Observação: uma forma alternativa de codificar o estado ("Y") para otimização em problemas de "clustering": ao invés de usar o dicionário Y como estado, podemos usar a $\boxed{\text{partição}}$ ("hard" ~~x~~ ou "soft") como estado:

$$
\begin{array}{ccc}
 & & \begin{array}{cccc} x(1) & x(2) & \cdots & x(1800) \end{array} \leftarrow \text{dados} \\
P_{Y|X} = & \begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_8 \end{array} & \left[\begin{array}{cccc} 1 & 0 & & 1 \\ 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 0 \end{array}\right] \\
 & & \begin{array}{cccc} \downarrow & \downarrow & & \downarrow \\ \Sigma=1 & \Sigma=1 & & \Sigma=1 \end{array}
\end{array}
$$

(hard)

$k=8$ centróides

ou seja, "estado = [1 2 7 ... 5 8 1]"

agora o estado fica discreto; e uma perturbação possível seria sortear uma posição do estado e perturbá-la para outro inteiro de 1 a 8. O resto do SA permanece igual. Y passa a ser consequência do estado.

Entropia de uma variável aleatória X:

$$H(x) = -\sum_{k=1}^{K} p_k \log p_k \quad (\log_2 \longrightarrow \text{em bits} \; ; \; \ln \longrightarrow \text{em nats})$$

Variável aleatória uniforme:

x	p(X=x)	código
0	0.25	00
1	0.25	01
2	0.25	11
3	0.25	10

$$H = -0.25 \log_2 0.25 - 0.25 \log_2 0.25$$
$$- 0.25 \log_2 0.25 - 0.25 \log_2 0.25$$

$H = 2$ bits $\quad (\log_2 0.25 = -2)$

Outra distribuição de probabilidades:

x	p(X=x)	código
0	0.5	0
1	0.25	10
2	0.125	110
3	0.125	111

(-2)

(-3)

$$H = -0.5 \log_2 0.5 - 0.25 \log_2 0.25$$
$$- 0.125 \log_2 0.125 - 0.125 \log_2 0.125$$

$H = 0.5 + 0.5 + 0.375 + 0.375 = 1.75$ bits

Entropia conjunta de variáveis aleatórias X e Y :

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

$$\left(H(X,Y) = -\sum_x \sum_y p_{xy} \log p_{xy} \right) \quad (\text{forma "abreviada"})$$

$$H(X,Y) = -\sum_x \sum_y p(y|x)\,p(x) \underbrace{\log p(y|x)\,p(x)} \rightarrow \left(\log p(x) + \log p(y|x) \right)$$

$$H(X,Y) = -\sum_x p(x) \log p(x) \underbrace{\sum_y p(y|x)}_{1} - \sum_x p(x) \sum_y p(y|x) \log p(y|x)$$

$$H(X,Y) = H(X) + \boxed{H(Y|X)}$$

$$\left(H(X,Y) = -\sum_x p_x \log p_x \overbrace{\sum_y p_{y|x}}^{1} - \overbrace{\sum_x p_x \sum_y p_{y|x} \log p_{y|x}}^{H(Y|X)} \right)$$

Maximização de $H(Y|X)$:

(partição com "máxima entropia")

$$\begin{array}{c|cccc} & x(1) & x(2) & x(3) & x(4) \\ \hline y_1 & 0.5 & 0.5 & 0.5 & 0.5 \\ y_2 & 0.5 & 0.5 & 0.5 & 0.5 \end{array}$$

($p_{Y|X}$)

Maximização de $H(Y|X)$ $\underline{\underline{e}}$ minimização de D:

$$\boxed{J = D - TH}$$

$$J = \sum_x p_x \sum_y p_{y|x} d_{xy} + T \sum_x p_x \sum_y p_{y|x} \log p_{y|x} +$$

parcela para garantir que $\sum_y p_{y|x} = 1$, caso contrário $p_{y|x} = 0$ leva a $J \longrightarrow -\infty$ (solução inválida).