Aprendizagem de Dados Simbólicos e/ou Numéricos

Francisco de A.T. de Carvalho

Transformação de Variáveis 1/15

- Homogeneização de Variáveis
  - Mudança de escala
    - Exemplo de mudança de escala
      Faixa etária c/ O = {Jovem, Adulto, Idoso}
    - Considerando-se a ordem
      Jovem < Adulto < idoso
      tem-se uma variável qualitativa ordinal
    - Não considerando-se essa ordem
      tem-se uma variável qualitativa nominal
  - Mudança de Codificação

Transformação de Variáveis 2/15

- Homogeneização de Variáveis
  - Mudança de Codificação
    - Exemplo de mudança de codificação
      Faixa etária c/ O = {Jovem, Adulto, Idoso}
    - {Jovem} $\cup$ {Idoso} = Não adulto
      O' = {Adulto, Não Adulto}

Transformação de Variáveis 3/15

- Transformação Quantitativo-Quantitativo: Normalização de Variáveis
  - Unidades diferentes
  - Dispersão heterogênea

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, j = 1,\ldots,n \qquad z_{ij} = \frac{x_{ij} - x_j^{\min}}{x_j^{\max} - x_j^{\min}}, j = 1,\ldots,n$$

Transformação de Variáveis 4/15

- Transformação Quantitativo-Quantitativo: Normalização de Variáveis

$$z_{ij} = \frac{x_{ij}}{\bar{x}_j}, j = 1, \ldots, n \qquad z_{ij} = \frac{x_{ij}}{x_j^{\max}}, j = 1, \ldots, n$$

$$z_{ij} = \frac{x_{ij}}{s_j}, j = 1, \ldots, n$$

Transformação de Variáveis 5/15

- Transformação Quantitativo-Qualitativo
  - O é um intervalo
  - Trata-se de dividir O
    - em intervalos contíguos e
    - associar a cada um deles uma categoria

Transformação de Variáveis 6/15

- Transformação Quantitativo-Qualitativo
  - Perda de Informação em dois níveis
    - Perde-se a distinção entre indivíduos que agora assumem a mesma categoria
    - Perde-se a amplitude da diferença entre indivíduos que agora assumem a mesma categoria

Transformação de Variáveis 7/15

- Transformação Quantitativo-Qualitativo
  - O intervalo O pode ser dividido de várias maneiras
    - O usuário escolhe os limites dos subintervalos
      - Exemplo: idade com O = [0, 150]

$0 < x_{kj} \leq 20 \Rightarrow z_{kj}$ = Jovem

$20 \leq x_{kj} < 60 \Rightarrow z_{kj}$ = Adulto

$60 < x_{kj} \leq 150 \Rightarrow z_{kj}$ = Idoso

O' = { Jovem, Adulto, Idoso}

Transformação de Variáveis 8/15

- Transformação Quantitativo-Qualitativo
  - O intervalo O pode ser dividido de várias maneiras
    - O usuário divide O em intervalos iguais
      - $I_1, \ldots, I_k$, k intervalos

$$\text{Amplitude} = \frac{x_j^{max} - x_j^{min}}{k}$$

Transformação de Variáveis 9/15

- Transformação Quantitativo-Qualitativo
  - O intervalo O pode ser dividido de várias maneiras
    - Divisão em Intervalos de Efetivos Iguais
      - A construção da função de repartição empírica F permite a obtenção dessa divisão

$$J_1 = F^{-1}\left([0,\frac{1}{k}]\right)\cdots J_i = F^{-1}\left(]\frac{i-1}{k},\frac{i}{k}]\right)\cdots J_k = F^{-1}\left(]\frac{k-1}{k},\frac{k}{k}]\right)$$

Transformação de Variáveis 10/15

- Transformação Quantitativo-Qualitativo
  - O intervalo O pode ser dividido de várias maneiras
    - Divisão em intervalos segundo os métodos de ligação hierárquicos unidimensionais
    - Divisão em intervalos segundo o método de hierárquico de Ward (Minimização da variância)

Transformação de Variáveis 11/15

- Transformação Qualitativo-Qualitativo
  - Ordinal -> Nominal
    - Por mudança de estrutura (escala) ou
    - por mudança de codificação

Transformação de Variáveis 12/15

- Transformação Qualitativo-Binário
  - Nominal -> Binário
    - Codificação disjuntiva completa
      - Ex: Cor dos Olhos: 1(verde); 2(azul); 3(marron)
        Idade: 1(0 a 20); 2(20 a 50); 3(>50a)

	Cor dos Olhos	Idade
k	1	2
l	2	1

Transformação de Variáveis 13/15

- Transformação Qualitativo-Binário
  - Nominal -> Binário
    - Codificação disjuntiva completa

		Cor dos Olhos			Idade	
	Verde	Azul	Marron	0-20	20-50	>50
k	1	0	0	0	1	0
l	0	1	0	1	0	0

Transformação de Variáveis 14/15

- Transformação Qualitativo-Binário
  - Ordinal -> Binário
    - Codificação Aditiva
      - Ex: Cor dos Olhos: 1(verde); 2(azul); 3(marron)
        Idade: 1(0 a 20); 2(20 a 50); 3(>50a)

	Cor dos Olhos	Idade
k	1	2
l	2	1

Transformação de Variáveis 15/15

- Transformação Qualitativo-Binário
  - Nominal -> Binário
    - Codificação disjuntiva completa

		Cor dos Olhos			Idade	
	Verde	Azul	Marron	0-20	20-50	>50
k	1	0	0	1	1	0
l	0	1	0	1	0	0

Funções de Proximidade 1/9

- Índice de Similaridade
  - É uma função

$$s : E \times E \rightarrow \Re_+$$
$$(k,l) \rightarrow s(k,l)$$

  - Tal que

- $s(k,l) = s(l,k) \quad \forall (k,l) \in E \times E$
- $s(k,k) = s(l,l) = s_{\max} > s(k,l) \quad \forall (k,l) \in E \times E \text{ com } k \neq l$

Funções de Proximidade 2/9

- Índice de Dissimilaridade
  - É uma função

$$d : E \times E \rightarrow \Re_+$$
$$(k,l) \rightarrow d(k,l)$$

  - Tal que
    - $d(k,l) = d(l,k) \quad \forall (k,l) \in E \times E$
    - $d(k,k) = 0 \quad \forall k \in E$

Funções de Proximidade 3/9

- Similaridade x Dissimilaridade

$$d(k,l) = s_{\max} - s(k,l)$$

$$d(k,l) = \frac{1}{1 + s(k,l)}$$

Funções de Proximidade 4/9

- Propriedades das Funções de Dissimilaridade

(1) $d(k,l) = 0 \Rightarrow k = l$

(2) $d(k,l) \leq d(k,m) + d(l,m), \forall (k,l,m) \in E \times E \times E$

(3) $d(k,l) \leq Max\{d(k,m), d(l,m)\}, \forall (k,l,m) \in E \times E \times E$

Funções de Proximidade 5/9

- Tipos de Funções de Dissimilaridade

Funções de dissimilaridade	Propriedades (1)	(2)	(3)
Índice de distância	X		
Distância	X	X	X
Ultramétrica	X	X	X

Funções de Proximidade 6/9

- Espaço Métrico
  - Se d é uma distancia definida em E, então
    - (E,d) é um espaço métrico
- Relação de Ordem
  - Uma medida de proximidade r definida em E
    - induz uma semi-ordem no conjunto $E \times E$

Funções de Proximidade 7/9

- Relação de Ordem
  - Para cada par (a,b), (c,d) $\in$ E $\times$ E $\preceq$ é definido por:

$$(a,b) \preceq_r (c,d) \Leftrightarrow \begin{cases} r(a,b) \leq r(c,d), \text{ se r é uma dissimilaridade} \\ r(a,b) \geq r(c,d), \text{ se r é uma similaridade} \end{cases}$$

Funções de Proximidade 8/9

- Relação de Ordem
  - Propriedades
    - Reflexiva

$$(a,b) \preceq (a,b), \forall a,b \in E$$

    - Transitiva

$$\forall (a,b),(c,d),(e,f) \in E \times E$$

$$(a,b) \preceq_r (c,d) \wedge (c,d) \preceq_r (e,f) \Rightarrow (a,b) \preceq_r (e,f)$$

Funções de Proximidade 9/9

- Relação de Ordem
  - Propriedades
    - Equivalência
      - Duas funções de proximidade r_1 e r_2 definidas sobre E são equivalentes se
        - as correspondentes semi-ordens são idênticas:

$$r_1 \approx r_2 \Leftrightarrow \preceq_{r_1} = \preceq_{r_2}$$

Funções de Dissimilaridade 1/12

- Dados quantitativos

$$\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})^T \qquad \mathbf{x}_l = (x_{l1}, \ldots, x_{lp})^T$$

- Distância Quadrática

$$d^2(k,l) = (\mathbf{x}_k - \mathbf{x}_l)^T \mathbf{M} (\mathbf{x}_k - \mathbf{x}_l)$$

Funções de Dissimilaridade 2/12

- Dados quantitativos
  - Distância Quadrática: casos especiais
    - Distância Euclidiana Usual:
      - M = I

$$d^2(k,l) = (\mathbf{x}_k - \mathbf{x}_l)^T \mathbf{I} (\mathbf{x}_k - \mathbf{x}_l) = \sum_{j=1}^{p} (x_{kj} - x_{lj})^2$$

Funções de Dissimilaridade 3/12

- Dados quantitativos
  - Distância Quadrática: casos especiais
    - Distância Euclidiana Ponderada pela Variância:
      - M = D^{-1}, D = diag(s_1^2, …,s_p^2)

$$d^2(k,l) = (\mathbf{x}_k - \mathbf{x}_l)^T \mathbf{D}^{-1} (\mathbf{x}_k - \mathbf{x}_l) = \sum_{j=1}^{p} \left(\frac{x_{kj} - x_{lj}}{s_j} \right)^2$$

Funções de Dissimilaridade 4/12

- Dados quantitativos
  - Distância Quadrática: casos especiais
    - Distancia de Mahalanobis:
      - M = V^{-1}, V = matriz de variâncias-covariâncias

$$d^2(k,l) = (\mathbf{x}_k - \mathbf{x}_l)^T \mathbf{V}^{-1} (\mathbf{x}_k - \mathbf{x}_l)$$

Funções de Dissimilaridade 5/12

- Dados quantitativos
  - Distancia de Minkowsky

$$d_\lambda(k,l) = \left[\sum_{j=1}^{p} \left| x_{kj} - x_{lj} \right|^\lambda \right]^{\frac{1}{\lambda}}, \quad \lambda = 1,2,\ldots$$

Funções de Dissimilaridade 6/12

- Dados quantitativos
  - Distancia de Minkowsky: casos especiais
    - Distancia de City-Block (Canberra, Métrica L_1):
      - $\lambda=1$

$$d_1(k,l) = \sum_{j=1}^{p} \left| x_{kj} - x_{lj} \right|$$

Funções de Dissimilaridade 7/12

- Dados quantitativos
  - Distancia de Minkowsky: casos especiais
    - Distancia Euclidiana usual:
      - $\lambda=2$

$$d_2(k,l) = \sqrt{\sum_{j=1}^{p} \left(x_{kj} - x_{lj}\right)^2}$$

Funções de Dissimilaridade 8/12

- Dados quantitativos
  - Distancia de Minkowsky: casos especiais
    - Distancia de Chebyshev:
      - $\lambda \rightarrow +\infty$

$$d_{\infty}(k,l) = \underset{j}{Max} \left| x_{kj} - x_{lj} \right|$$

Funções de Dissimilaridade 9/12

- Dados quantitativos
  - Distancia do Qui-quadrado

$$d^2(k,l) = \sum_{j=1}^{p} \frac{1}{x_{\bullet j}} \left(\frac{x_{kj}}{x_{k\bullet}} - \frac{x_{lj}}{x_{l\bullet}} \right)^2$$

  - onde

$$x_{\bullet j} = \sum_{i=1}^{n} x_{ij}, \quad x_{k\bullet} = \sum_{j=1}^{p} x_{kj}, \quad x_{l\bullet} = \sum_{j=1}^{p} x_{lj}$$

Funções de Dissimilaridade 10/12

- Dados quantitativos
  - Coeficiente de Canberra

$$d(k,l) = \frac{1}{p}\sum_{j=1}^{p}\frac{\left|x_{kj} - x_{lj}\right|}{x_{kj} + x_{lj}}$$

Funções de Dissimilaridade 11/12

- Dados quantitativos
  - Coeficiente de Bray-Curtis

$$d(k,l) = \frac{1}{p} \frac{\sum_{j=1}^{p} \left| x_{kj} - x_{lj} \right|}{\sum_{j=1}^{p} \left(x_{kj} + x_{lj} \right)}$$

Funções de Dissimilaridade 12/12

- Dados quantitativos
  - Coeficiente de Bhattacharyya

$$d(k,l) = \sum_{j=1}^{p} \sqrt{\sqrt{x_{kj}} - \sqrt{x_{lj}}}$$

Funções de Similaridade 1/3

- Dados Quantitativos

$$\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})^T \qquad \mathbf{x}_l = (x_{l1}, \ldots, x_{lp})^T$$

- Produto Vetorial

$$d(k,l) = \left| \sum_{j=1}^{p} x_{kj} x_{lj} \right|$$

Funções de Similaridade 2/3

- Dados Quantitativos
  - Similaridade Angular

$$d(k,l) = \frac{\sum_{j=1}^{p} x_{kj} x_{lj}}{\sqrt{\left[\sum_{j=1}^{p} (x_{kj})^2\right]\left[\sum_{j=1}^{p} (x_{lj})^2\right]}}$$

Funções de Similaridade 3/3

- Dados Quantitativos
  - Covariância

$$d(k,l) = \left| \sum_{j=1}^{p} (x_{kj} - \bar{x}_j)(x_{lj} - \bar{x}_j) \right|$$

  - Correlação

$$d(k,l) = \left| \frac{\sum_{j=1}^{p} (x_{kj} - \bar{x}_j)(x_{lj} - \bar{x}_j)}{\sqrt{\left[\sum_{j=1}^{p} (x_{kj} - \bar{x}_j)^2 \right] \left[\sum_{j=1}^{p} (x_{lj} - \bar{x}_j)^2 \right]}} \right|$$

Funções de Proximidade 1/24

- Dados Binários

$$\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})^T \qquad \mathbf{x}_l = (x_{l1}, \ldots, x_{lp})^T$$

- Número de vezes em que $x_{kj} = x_{lj} = 1$

$$a = (\mathbf{x}_k)^T \mathbf{x}_l$$

- Número de vezes em que $x_{kj} = 0$ e $x_{lj} = 1$:

$$b = (\mathbf{1} - \mathbf{x}_k)^T \mathbf{x}_l$$

Funções de Proximidade 2/24

- Dados Binários
  - Número de vezes em que $x_{kj} = 1$ e $x_{lj} = 0$:

$$c = (\mathbf{x}_k)^T (\mathbf{1} - \mathbf{x}_l)$$

  - Número de vezes em que $x_{kj} = x_{lj} = 0$:

$$d = (\mathbf{1} - \mathbf{x}_k)^T (\mathbf{1} - \mathbf{x}_l)$$

Funções de Proximidade 3/24

- Dados Binários

k \ l	1	0	Σ
1	a	b	a + b
0	c	d	c + d
Σ	a + c	b + d	P = a+b+c+d

Funções de Proximidade 4/24

- Dados Binários
  - Coeficientes de Concordância
    - Russel and Rao

$$s(k,l) = \frac{a}{a+b+c+d} = \frac{a}{p}$$

$$d(k,l) = 1 - \frac{a}{a+b+c+d} = \frac{b+c+d}{p}$$

Funções de Proximidade 5/24

- Dados Binários
  - Coeficientes de Concordância
    - Sokal e Michener

$$s(k,l) = \frac{a+d}{a+b+c+d} = \frac{a+d}{p}$$

$$d(k,l) = 1 - \frac{a+d}{a+b+c+d} = \frac{b+c}{p}$$

Funções de Proximidade 6/24

- Dados Binários
  - Coeficientes de Concordância
    - Roger e Tanimoto

$$s(k,l) = \frac{a+d}{a+d+2(b+c)}$$

$$d(k,l) = 1 - \frac{a+d}{a+d+2(b+c)} = \frac{2(b+c)}{a+d+2(b+c)}$$

Funções de Proximidade 7/24

- Dados Binários
  - Coeficientes de Concordância

$$s(k,l) = \frac{2(a+d)}{2(a+d)+b+c}$$

$$d(k,l) = 1 - \frac{2(a+d)}{2(a+d)+b+c} = \frac{b+c}{2(a+d)+b+c}$$

Funções de Proximidade 8/24

- Dados Binários
  - Coeficientes de Concordância
    - Jacquard

$$s(k,l) = \frac{a}{a+b+c}$$

$$d(k,l) = 1 - \frac{a}{a+b+c} = \frac{b+c}{a+b+c}$$

Funções de Proximidade 9/24

- Dados Binários
  - Coeficientes de Concordância
    - Dice

$$s(k,l) = \frac{2a}{2a+b+c}$$

$$d(k,l) = 1 - \frac{2a}{2a+b+c} = \frac{b+c}{2a+b+c}$$

Funções de Proximidade 10/24

- Dados Binários
  - Coeficientes de Concordância
    - Kulczynski

$$s(k,l) = \frac{a}{b+c} \qquad s(k,l) = \frac{a+d}{b+c)}$$

Funções de Proximidade 11/24

- Dados Binários
  - sabendo-se que

$$\sum_{j=1}^{p} x_{kj}\, x_{lj} = a \qquad \sum_{j=1}^{p} (x_{kj})^2 = a + b \qquad \sum_{j=1}^{p} (x_{lj})^2 = a + c$$

  - Produto vetorial

$$s(k,l) = \sum_{j=1}^{p} x_{kj}\, x_{lj} = a$$

Funções de Proximidade 12/24

- Dados Binários
  - Similaridade Angular

$$s(k,l) = \frac{\sum_{j=1}^{p} x_{kj} x_{lj}}{\left\{\left[\sum_{j=1}^{p} (x_{kj})^2\right]\left[\sum_{j=1}^{p} (x_{lj})^2\right]\right\}^{\frac{1}{2}}} = \left[\left(\frac{a}{a+b}\right)\left(\frac{a}{a+c}\right)\right]^{\frac{1}{2}}$$

Funções de Proximidade 13/24

- Dados Binários
  - Similaridade Angular (bis)

$$s(k,l) = \left[\left(\frac{a}{a+b} \right) \left(\frac{a}{a+c} \right) \left(\frac{d}{b+d} \right) \left(\frac{d}{c+d} \right) \right]^{\frac{1}{2}}$$

  - Covariância

$$s(k,l) = \left| \sum_{j=1}^{p} x_{kj} x_{lj} - \frac{1}{p} \left(\sum_{j=1}^{p} x_{kj} \right) \left(\sum_{j=1}^{p} x_{lj} \right) \right| = \left| \frac{ad - bc}{a+b+c+d} \right|$$

Funções de Proximidade 14/24

- Dados Binários
  - Correlação

$$s(k,l) = \frac{\sum_{j=1}^{p} x_{kj}\, x_{lj} - \frac{1}{p}\left(\sum_{j=1}^{p} x_{kj}\right)\left(\sum_{j=1}^{p} x_{lj}\right)}{\left[\sum_{j=1}^{p} (x_{kj})^2 - \frac{1}{p}\left(\sum_{j=1}^{p} x_{kj}\right)^2\right]^{\frac{1}{2}}\left[\sum_{j=1}^{p} (x_{lj})^2 - \frac{1}{p}\left(\sum_{j=1}^{p} x_{lj}\right)^2\right]^{\frac{1}{2}}}$$

$$= \frac{|ad - bc|}{[(a+b)(c+d)(a+c)(b+d)]^{\frac{1}{2}}}$$

Funções de Proximidade 15/24

- Dados Binários
  - Outros coeficientes
    - Kulczynski

$$s(k,l) = \frac{1}{2}\left[\frac{a}{a+b} + \frac{a}{a+c}\right]$$

$$s(k,l) = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right]$$

Funções de Proximidade 16/24

- Dados Binários
  - Outros coeficientes
    - Hamman

$$s(k,l) = \frac{\left|(a+d)-(b+c)\right|}{a+b+c+d}$$

    - Yule

$$s(k,l) = \frac{\left|ad-bc\right|}{ad+bc}$$

Funções de Proximidade 17/24

- Dados Qualitativos

Sejam

- n_a: número de vezes em que as variáveis concordam para k e l
- n_{b+c} : número de vezes em que as variáveis discordam para k e l

Funções de Proximidade 18/24

- Dados Qualitativos
  - Coeficientes de Concordância

$$s(k,l) = \frac{n_a}{n_a + n_{b+c}} \qquad d(k,l) = \frac{n_{b+c}}{n_a + n_{b+c}}$$

$$s(k,l) = \frac{2n_a}{2n_a + n_{b+c}} \qquad d(k,l) = \frac{n_{b+c}}{2n_a + n_{b+c}}$$

Funções de Proximidade 19/24

- Dados Qualitativos
  - Coeficientes de Concordância

$$s(k,l) = \frac{n_a}{n_a + 2n_{b+c}} \qquad d(k,l) = \frac{2n_{b+c}}{n_a + 2n_{b+c}}$$

$$s(k,l) = \frac{n_a}{n_{b+c}}$$

Funções de Proximidade 20/24

- Dados Qualitativos
  - Dados qualitativos nominais
    - Codificação Disjuntiva Completa
    - Funções para dados binários
  - Dados qualitativos ordinais
    - Codificação Aditiva
    - Funções para dados binários

Funções de Proximidade 21/24

- Dados Heterogêneos
  - Transformação de todas as variáveis em binárias
    - Funções para variáveis binárias
    - Inconveniente: perda de informação

Funções de Proximidade 22/24

- Dados Heterogêneos
  - Coeficiente de proximidade combinado

$$c(k,l) = w_b\, c_b(k,l) + w_n\, c_n(k,l) + w_o\, c_o(k,l) + w_q\, c_q(k,l)$$

  - Cuidados a serem observados
    - Coeficientes de proximidade de mesmo sentido (similaridade ou dissimilaridade)
    - Intervalos de variação dos coeficientes próximos
    - Pesos adequados (ex., número de variáveis de cada tipo)

Funções de Proximidade 23/24

- Dados Heterogêneos
  - Variáveis assumindo valores no intervalo [0,1]
    - Binárias: já estão no intervalo [0,1]
    - Nominais: Aplicar a codificação disjuntiva completa
    - Ordinais: Aplicar a codificação aditiva
    - Quantitativa: normalização

$$z_{kj} = \frac{x_{kj} - x_j^{\min}}{x_j^{\max} - x_j^{\min}}$$

    - Usar a distancia euclidiana

Funções de Proximidade 24/24

- Dados Heterogêneos
  - Coeficiente de Similaridade Geral (Gower)

$$s(k,l) = \frac{\sum_{j=1}^{p} I_j(k,l)\, s_i(k,l) w_i}{\sum_{j=1}^{p} I_j(k,l)}, 0 \le s_i \le 1$$

$$I_j(k,l) = \begin{cases} 1, \text{se k e l podem ser comparados segundo } y_j \\ 0, \text{senão} \end{cases}$$