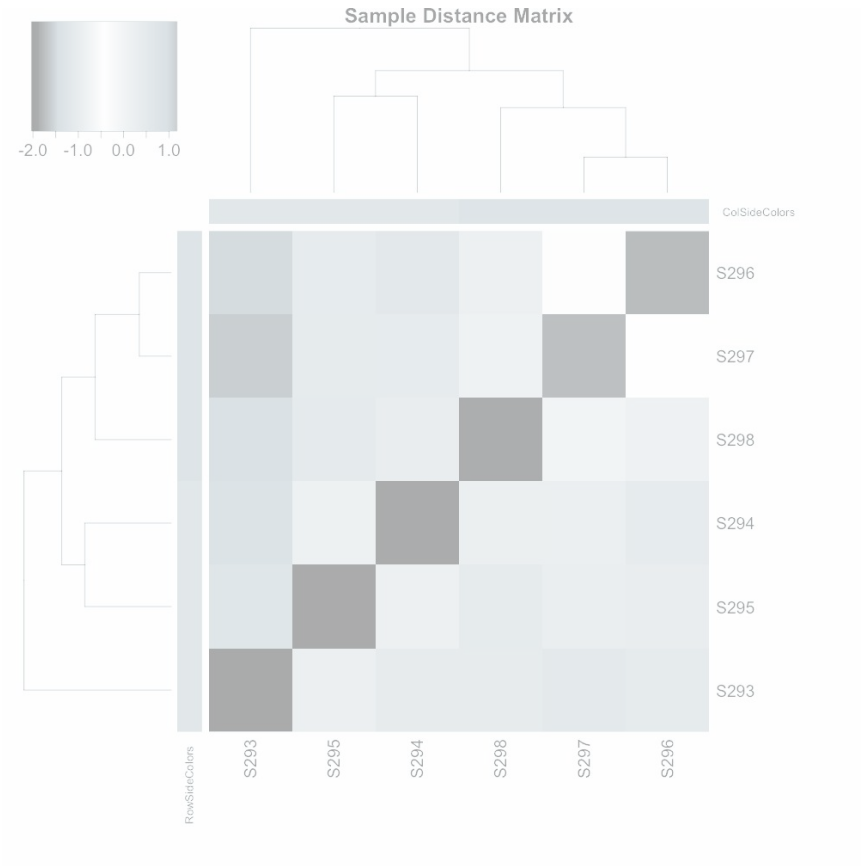


Aula 12

Medidas de Proximidade

Rafael Geraldeli Rossi

Conteúdo

- Em muitas aplicações de mineração de dados, como agrupamento de dados, análise de *outliers*, e classificações baseada em vizinhos mais próximos, é necessário definir quais objetos são mais *"semelhantes"*, ou encontram-se mais *"próximos"* no espaço *n*-dimensional, e quais objetos são *"diferentes"* dos outros

- Mas o que é próximo (similar ou dissimilar)??

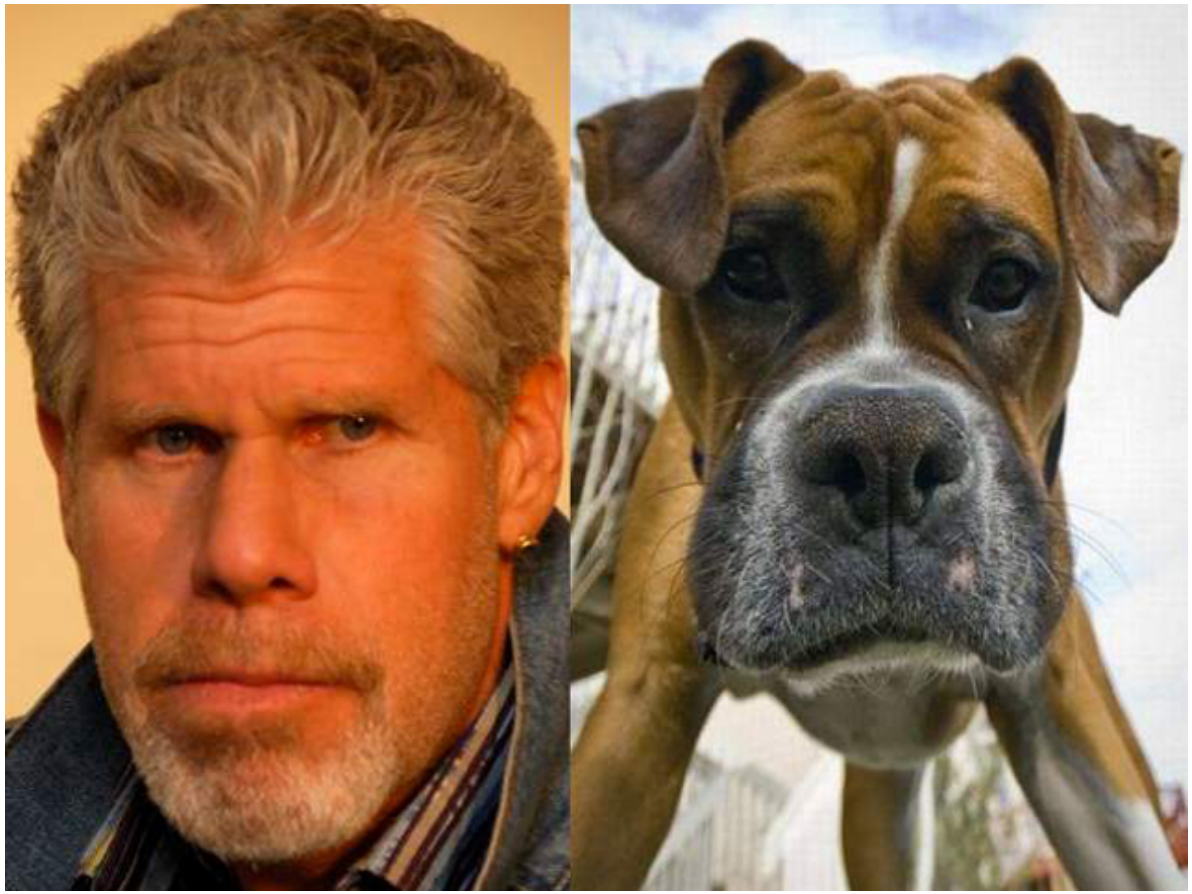
Essas imagens são similares?

Essas imagens são similares?

Essas imagens são similares?

Essas imagens são similares?

Essas imagens são similares?



Essas imagens são similares?

Essas imagens são similares?

Essas imagens são similares?

- **Medidas de proximidade**: **medidas de similaridade** e **medidas de dissimilaridade**
  - Uma medida de **similaridade** retorna o valor 0 se dois objetos são *"diferentes"*
  - Tipicamente o valor 1 indica uma similaridade completa, isto é, os objetos são idênticos
  - Uma medida de **dissimilaridade** retorna o valor 0 se os objetos são idênticos
  - Quanto maior o valor de dissimilaridade, mais *"diferentes"* são os objetos

- **Notações**
  - $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_n\}$: conjunto de exemplos
  - $A = \{a_1, a_2, ...a_m\}$: conjuto de atributos
  - $\mathbf{x}_i$: vetor de características do i-ésimo exemplo
  - $x_{i,j}$: valor do j-ésimo atributo do exemplo i

- Um atributo nominal pode ter dois ou mais estados
- A dissimilaridade entre dois objetos i e j, tratada por $d(i, j)$, pode ser computada baseada na taxa de discordância

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \frac{|A| - count(x_{i,k} == j_{j,k})}{|A|} \quad (1)$$

em que $count(x_{i,k} == x_{j,k})$ é o número de valores de atributos iguais entre os exemplos $\mathbf{x}_i$ e $\mathbf{x}_j$

- De maneira semelhante, a similaridade pode ser computada como

$$sim(\mathbf{x}_i, \mathbf{x}_j) = 1 - dist(\mathbf{x}_i, \mathbf{x}_j) = \frac{count(x_{i,k} == j_{j,k})}{|A|} \quad (2)$$

- Para atributos binários simétricos, cada estado é igualmente importante

Tabela: Exemplo de um conjunto de dados com atributos binários simétricos

ID Cliente	Fumante	Estudante	Casado	Trabalhador
1	sim	não	sim	sim
2	não	sim	não	não
3	sim	não	não	não

- Para atributos binários assimétricos, os dois estados não são considerados igualmente importantes
  - Os valores 1 (positivos, *true*, sim, etc.) são considerados mais significativos do que os que tem valores 0.

Tabela: Exemplo de um conjunto de dados com atributos binários simétricos

ID Estudante	**D1**	**D2**	**D3**	**D4**	**D5**	...	**D98**	**D99**	**D100**
1	1	0	0	0	1	...	0	0	0
2	0	0	0	0	0	...	0	0	1
3	0	0	0	0	0	...	0	0	0

- Para ambos os tipos de atributos binários, a seguinte tabela é utilizada para calcular a (dis)similaridade

Tabela: Tabela de contingência para atributos binários

		Objeto j	
		1	**0**
Objeto i	**1**	f_{11}	f_{10}
	0	f_{01}	f_{00}

- Similaridade entre objetos i e j que contém atributos binários simétricos (Coeficiente de Casamento Simples (CCS))

$$ccs(\mathbf{x}_i, \mathbf{x}_j) = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} \tag{3}$$

- Similaridade entre objetos i e j que contém atributos binários assimétricos (Coeficiente de Jaccard)

$$J(\mathbf{x}_i, \mathbf{x}_j) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \tag{4}$$

Exemplo

$i = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$

$j = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$

$f_{01} = 2,\ f_{10} = 1,\ f_{00} = 7,$ e $f_{11} = 0$

$$CCS = \frac{0 + 7}{2 + 1 + 0 + 7} = 0,7$$

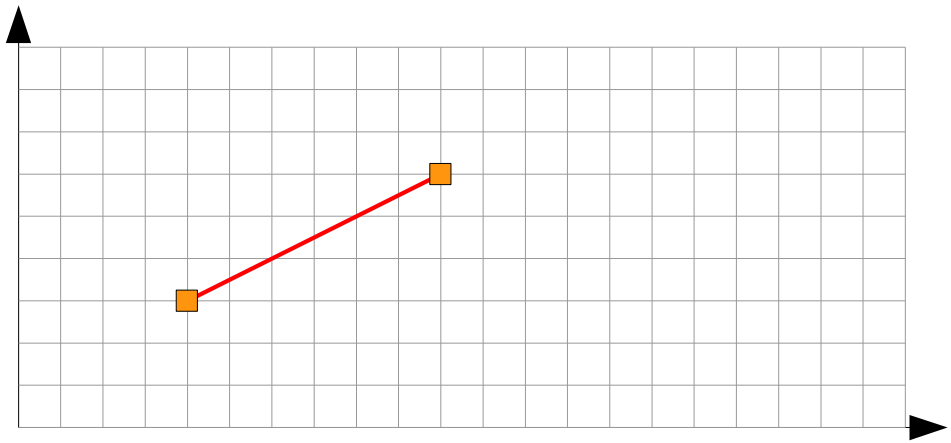
$$J = \frac{0}{2 + 1 + 0} = 0$$

- Atributos nominais podem ser codificados usando atributos binários assimétricos
  - Criar um atributo binário para cada um dos M estados
  - O valor do atributo representando o estado é definido em 1 e os demais valores são definidos em 0
  - Com isso, pode-se utilizar medidas de similaridade apresentadas na Seção 2 (Atributos Binários)

Tabela: Conversão de um atributo nominal em atributos binários assimétricos

Valor nominal para cor	**cor$_1$**	**cor$_2$**	**cor$_3$**	**cor$_4$**	**cor$_5$**
verde	1	0	0	0	0
vermelho	0	1	0	0	0
azul	0	0	1	0	0
cinza	0	0	0	1	0
preto	0	0	0	0	1

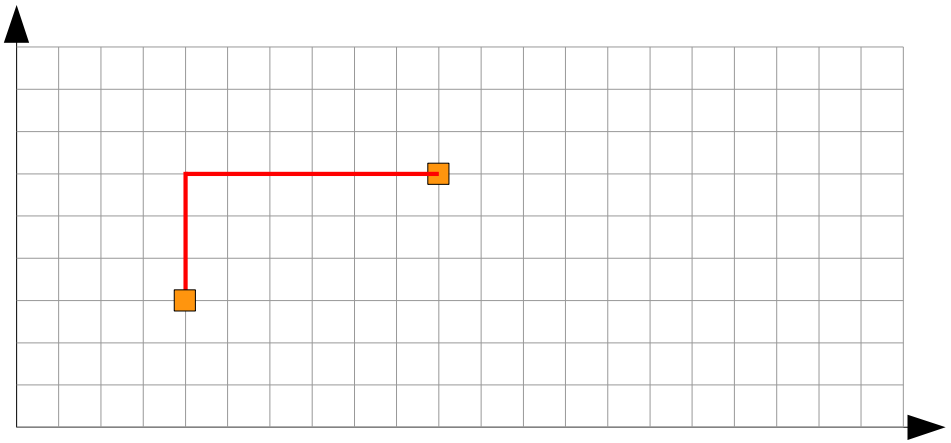
- A medida de distância mais utilizada para dados numéricos é a **distância Euclidiana**
  - Linha reta entre dois objetos



  - Sejam dois objetos descritos por m atributos numéricos $\mathbf{x}_i = (x_{i,1}, x_{i,2}, x_{i,3}, \ldots, x_{i,m})$ e $\mathbf{x}_j = (x_{j,1}, x_{j,2}, x_{j,3}, \ldots, x_{j,m})$
  - A distância Euclidiana entre dois objetos $\mathbf{x}_i$ e $\mathbf{x}_j$ é definida como

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{i,1} - x_{j,1})^2 + (x_{i,2} - x_{j,2})^2 + \cdots + (x_{i,m} - x_{j,m})^2} \quad (5)$$

- Distância Euclidiana
  - Exemplo
    $\mathbf{x}_i = (0, 2)$
    $\mathbf{x}_j = (5, 1)$
    $dist(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(0-5)^2 + (2-1)^2} = \sqrt{25+1} = 5,1$
- Outra medida bastante conhecida é a **distância Manhattan**
  - Também conhecida como distância *City Block* (distância em blocos entre quaisquer dois pontos em uma cidade)



- Distância Manhattan
  - A distância Manhattan entre dois objetos i e j é definida como

$$dist(\mathbf{x}_i, \mathbf{x}_j) = |x_{i,1} - x_{j,1}| + |x_{i,2} - x_{j,2}| + \cdots + |x_{i,m} - x_{j,m}| \quad (6)$$

  - Exemplo
    $i = (0, 2)$
    $j = (5, 1)$
    $dist(\mathbf{x}_i, \mathbf{x}_j) = |0 - 5| + |2 - 1| = 6$
- A **distância Minkowski** (L_{norm}) é uma generalização das distâncias Euclidiana e Manhattan

$$dist(\mathbf{x}_i, \mathbf{x}_j) = (|x_{i,1} - x_{j,1}|^h + |x_{i,2} - x_{j,2}|^h + \cdots + |x_{i,m} - x_{j,m}|^h)^{\frac{1}{h}} \quad (7)$$

na qual h é um número real tal que $h \geq 1$

- A **distância Supremum** (L_{max}, L_{infty})
  - Generalização da distância Minkowski para $h \to \infty$

$$dist(\mathbf{x}_i, \mathbf{x}_j) = (\lim_{h \to \infty} (\sum_{f=1}^{m} |x_{i,f} - x_{j,f}|^h)^{\frac{1}{h}} \quad (8)$$

  - É a diferença máxima entre quaisquer atributos dos objetos
  - Exemplo
    $\mathbf{x}_i = (0, 2)$
    $\mathbf{x}_j = (5, 1)$
    $dist(\mathbf{x}_i, \mathbf{x}_j) = argmax(|0 - 5|, |2 - 1|) = 5$

- Pode ser atribuído um peso para ser dada maior/menor importância a um atributo no cálculo da distância

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{w_1(x_{i,1} - x_{j,1})^2 + w_2(x_{i,2} - x_{j,2})^2 + \cdots + w_,(x_{i,m} - x_{j,m})^2} \quad (9)$$

- Os valores de um atributo ordinal possuem uma ordem significativa entre eles
- A magnitude entre valore sucessivos é desconhecida
- Seja z o número de possíveis estados que um atributo f ordinal pode ter
- Ao ordenar os valores do atributo f, tem-se valores que varia de 1 a R_f, sendo R_f o vamor valor do ranking
- **Exemplo**: $\{\texttt{baixo, médio, alto}\} \rightarrow \{\texttt{1, 2, 3}\}$

- A computação da dissimilaridade com respeito a f envolve os seguintes passos:
  1. Substitua cada $x_{i,f}$ por sua posição correspondente no *ranking* $(r_{i,f})$
  2. Uma vez que cada atributo pode ter um número diferente de estados, é necessário mapear o intervalo de valores de cada atributo em $[0, 1]$ para que cada atributo tenha um peso igual

$$z_{i,f} = \frac{r_{i,f} - 1}{R_f - 1} \tag{10}$$

  3. Utilizar uma função de dissimilaridade para atributos numéricos utilizando z_{if} para representar o valor o atributo f para o objeto i

Exemplo

Tabela: Conjunto de dados com atributos ordinais

Objeto	Idade	Altura	Escolaridade
1	Adolescente	Alto	Fundamental
2	Adulto	Médio	Médio
3	Idoso	Pequeno	Superior
4	Adulto	Alto	Nenhuma
5	Adolescente	Pequeno	Médio

Tabela: Conjunto de dados com atributos ordinais transformados em atributos numéricos

Objeto	Idade	Altura	Escolaridade
1	0,00	1,00	0,33
2	0,50	0,50	0,66
3	1,00	0,00	1,00
4	0,50	1,00	0,00
5	0,00	0,00	0,66

Tabela: Matriz de dissimilaridades (distância Euclidiana)

	1	**2**	**3**	**4**	**5**
1	–				
2	0,78	–			
3	1,56	0,78	–		
4	0,60	0,60	1,50	–	
5	1,05	0,70	1,10	1,30	–

- Em muitos conjuntos de dados reais, objetos são descritos por atributos de diferentes tipos
- Uma abordagem é agrupar cada tipo de atributo e executar algoritmos de mineração de dados para cada tipo
  - É viável se as análises derivam em resultados compatíveis
  - Em aplicações reais, é pouco provável que este tipo de análise gere resultados compatíveis

- Uma abordagem preferível é processar todos os tipos de atributos juntos
  - Colocar os atributos em uma escala comum ($[0, 1]$)
  - Função de distância

$$dist(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{f=1}^{m} \delta_{\mathbf{x}_i,\mathbf{x}_j}^{(f)} |x_{i,f} - x_{j,f}|}{\sum_{f=1}^{m} \delta_{\mathbf{x}_i,\mathbf{x}_j}^{(f)}} \tag{11}$$

na qual $\delta_{\mathbf{x}_i,\mathbf{x}_j}^{(f)} = 0$ se
1) $x_{i,f}$ ou $x_{j,f}$ têm valores ausentes, ou
2) se $x_{i,f} = x_{j,f} = 0$ e o atributo f é assimétrico binário (caso contrário $\delta_{\mathbf{x}_i,\mathbf{x}_j}^{(f)} = 1$)

Exemplo

Tabela: Conjunto de dados composto por atributos por diferentes tipos

Objeto	Fumante	Escolaridade	Idade
1	1	Fundamental	18
2	0	Médio	23
3	0	Nenhuma	32
4	1	Nenhuma	27
5	1	Superior	36

Tabela: Matriz de distâncias

	1	2	3	4	5
1	–				
2	1,60	–			
3	2,10	1,16	–		
4	0,83	1,88	1,27	–	
5	1,67	2,16	2,22	1,5	–

- Um documento pode ser representado por centenas de atributos, cada qual armazena a frequência de um termo particular (ex: palavra simples e frases) em um documento.
- Cada documento é representado por um vetor termo $\times$ frequência

Tabela: Exemplo de uma matriz documento $\times$ termo (atributo $\times$ valor)

	Site	Google	Rede	...	Campeonato	Brasileiro	Futebol	...	Juros	Banco
Doc 1	1	6	0	...	0	0	0	...	0	0
Doc 2	0	7	8	...	0	0	0	...	0	0
Doc 3	3	0	10	...	0	0	0	...	0	0
Doc 4	0	0	0	...	4	4	2	...	0	0
Doc 5	0	0	0	...	3	0	0	...	0	0
Doc 6	0	0	0	...	3	2	1	...	0	0
Doc 7	0	0	0	...	0	0	0	...	9	5
Doc 8	0	0	0	...	0	0	0	...	7	0
Doc 9	0	0	0	...	0	0	0	...	4	5

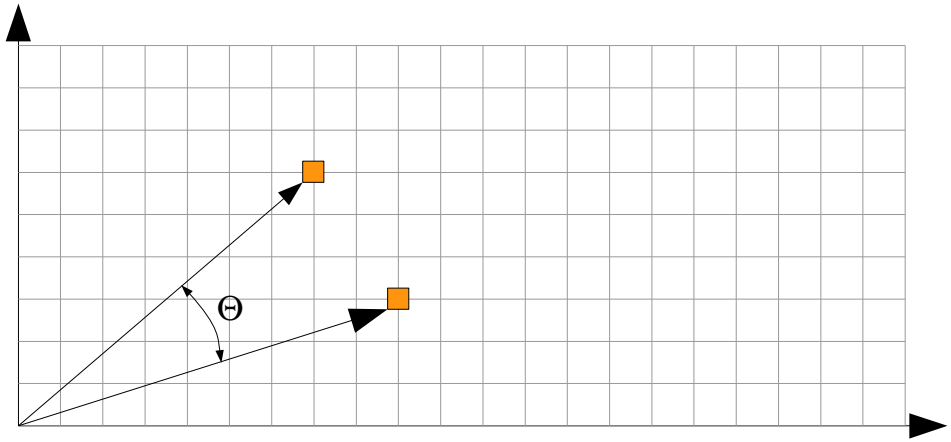
- As medidas para dados numéricos apresentadas na Seção 3 não funcionam bem para dados numéricos **esparsos**
- Os vetores termo $\times$ frequência são tipicamente longos e esparsos (isto é, possuem muitos valores 0)
- Dois vetores termo $\times$ frequência podem ter muitos valores 0 em comum, mas este fato não os fazem documentos similares
- **Similaridade Cosseno**: foca apenas nas palavras que os documentos têm em comum e na frequência de tais palavras

$$cos(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{||\mathbf{x}_i||\ ||\mathbf{x}_j||} \tag{12}$$

na qual $||\mathbf{x}_i||$ (norma euclideana) é definida como

$$\sqrt{x_{i,1}^2 + x_{i,2}^2 + \cdots + x_{i,m}^2}$$

- *Cosseno* $= 0$ significa que os dois vetores possuem um ângulo de 90°(ortogonal), ou seja não contém nenhum atributo em comum
- Quanto valor próximo o valor $cos(\mathbf{x}_i, \mathbf{x}_j)$ de 1, menor o ângulo entre os dois vetores, e mais similar são os objetos



Exemplo

$x = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$

$y = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$

$x.y = 3 * 1 + 2 * 0 + 0 * 0 + 5 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 2 * 1 + 0 * 0 + 0 * 2 = 5$

$||x|| = \sqrt{3 * 3 + 2 * 2 + 0 * 0 + 5 * 5 + 0 * 0 + 0 * 0 + 2 * 2 + 0 * 0 + 0 * 0} = 6,48$

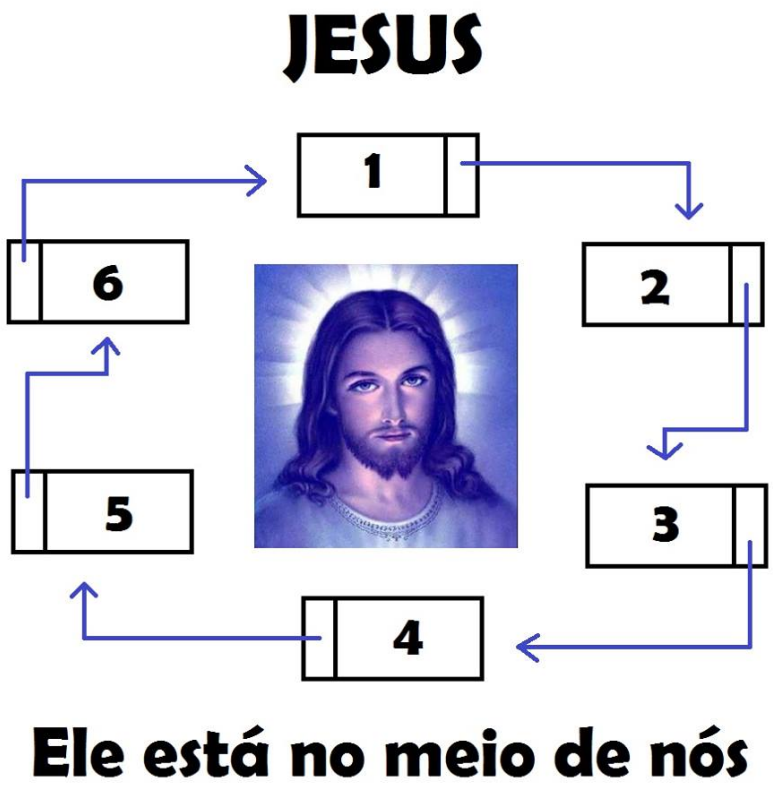
$||y|| = \sqrt{1 * 1 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 1 * 1 + 0 * 0 + 2 * 2} = 2,24$

$cos(x, y) = 0,31$

Material Complementar

- Aula 7 – Medidas de Distância
  `http://www.facom.ufu.br/~elaine/disc/MFCD/Aula7-MedidasDistancia.pdf`

Imagem do Dia



Inteligência Artificial
http://lives.ufms.br/moodle/

Rafael Geraldeli Rossi
rafael.g.rossi@ufms.br

Slides baseados em [Han et al., 2011] e [Tan et al., 2005]

Referências Bibliográficas I


Han, J., Kamber, M., and Pei, J. (2011).
Data Mining: Concepts and Techniques.
The Morgan Kaufmann Series in Data Management Systems.
Elsevier.


Tan, P.-N., Steinbach, M., and Kumar, V. (2005).
Introduction to Data Mining.
Addison-Wesley.