Medidas de Similaridade

Prof. Sandro Jerônimo de Almeida

Fontes das Imagens: www.shutterstock.com ou próprio autor

PUC Minas Virtual

Agrupamento (*Clustering*)

Cluster

Coleção de objetos que são similares uns aos outros (de acordo com algum critério de similaridade pré-fixado e dissimilares a objetos pertencentes a outros clusters.

Análise de cluster (*clustering*)

Separa os objetos em grupos com base na similaridade, e em seguida atribuir rótulos a cada grupo.

Mas como medir a similaridade entre os dados?

Como medir a similaridade entre os dados?

Exemplo

Cliente	Salário	Idade	Gastos Roupas
1	3.000	22	100
2	1.000	18	50
3	9.000	48	600

João tem 24 anos, mensalmente recebe um salário de R$ 8.500 e não gasta com roupas.

- **João está mais "próximo" de qual cliente?**

João e sua similaridade

- A idade do cliente 1
- Gasta como o cliente 2
- O salário do cliente 3

Processo

1° Normalizar os dados

2° Medir a distância

Normalização dos Dados

Processo de Normalização

Colocar os dados entre 0 e 1

Para cada coluna
- O maior valor vira 1
- O menor valor vira 0
- Demais “proporcionais”

Podemos usar a equação:

$$L_n = \frac{L_a - L_{\min}}{L_{\max} - L_{\min}}$$

Dados Originais

Cliente	Salário	Idade	Gastos Roupas
1	3.000	22	100
2	1.000	18	50
3	9.000	48	600
João	8.500	24	0

Dados Normalizados (salário)

Menor valor da coluna → $L_{min} = 1.000$

Maior valor da coluna → $L_{max} = 9.000$

Salário Norm. Cliente 1 → $L_n = (3.000 - L_{min}) / (L_{max} - L_{min})$

$= (3.000 - 1.000) / (9.000 - 1.000)$

$= 2.000 / 8.000 = 0,25$

Normalização dos Dados

Dados Originais

Cliente	Salário	Idade	Gastos Roupas
1	3.000	22	100
2	1.000	18	50
3	9.000	48	600
4	8.500	24	0

$$L_n = \frac{L_a - L_{\min}}{L_{\max} - L_{\min}}$$

Dados Normalizados

Cliente	Salário	Idade	Gastos Roupas
1	0,25		
2	0	0	
3	1	1	1
4			0

Normalização dos Dados

Dados Originais

Cliente	Salário	Idade	Gastos Roupas
1	3.000	22	100
2	1.000	18	50
3	9.000	48	600
4	8.500	24	0

Dados Normalizados

Cliente	Salário	Idade	Gastos Roupas
1	0,25	0,13	0,17
2	0	0	0,08
3	1	1	1
4	0,94	0,20	0

$$L_n = \frac{L_a - L_{\min}}{L_{\max} - L_{\min}}$$

Como medir a similaridade entre os dados?

Exemplo – DADOS NORMALIZADOS

Cliente	Salário	Idade	Gastos Roupas
1	0,25	0,13	0,17
2	0	0	0,08
3	1	1	1
4	0,94	0,20	0

João tem 24 anos, mensalmente recebe um salário de R$ 8.500 e não gasta com roupas.

- **João está mais "próximo" de qual cliente?**

João e sua similaridade

- A idade do cliente 1
- Gasta como o cliente 2
- O salário do cliente 3

Processo

1° Normalizar os dados

2° Medir a distância

Distância Euclidiana

Dados Normalizados

Cliente	Salário	Idade	Gastos Roupas
1	0,25	0,13	0,17
2	0	0	0,08
3	1	1	1

Salário: 0,94
Idade: 0,20
Gastos roupas: 0

Distância Euclidiana

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}$$

Distância Euclidiana

Dados Normalizados

Cliente	Salário *x*	Idade *y*	Gastos Roupas *z*
1	0,25	0,13	0,17
2	0	0	0,08
3	1	1	1

João

Cliente Id: 4
Salário (x): 0,94
Idade (y): 0,20
Gastos roupas (z): 0

Distância Euclidiana

$$d(1,4) = \sqrt{|x_1 - x_4|^2 + |y_1 - y_4|^2 + |z_1 - z_4|^2} = 0{,}71$$

$$d(2,4) = \sqrt{|x_2 - x_4|^2 + |y_2 - y_4|^2 + |z_2 - z_4|^2} = 0{,}96$$

$$d(3,4) = \sqrt{|x_3 - x_4|^2 + |y_3 - y_4|^2 + |z_3 - z_4|^2} = 1{,}28$$

João se parece mais com o cliente 1 (menor distância)

Medindo outros tipos de atributos

Variáveis binárias

Atributos simétricos: sexo (M/F)

Atributo assimétricos: resultado de um exame (positivo/negativo)

Variáveis nominais

Generalização de variáveis binárias que podem ter mais de dois valores, ex: vermelho, amarelo, azul, verde

Variáveis ordinárias

Uma variável ordinária em que a ordem dos valores é importante.

Exemplo: atributo Medalha, cujos valores são: Bronze, Prata e Ouro

PUC Minas
Virtual