UNIDADE 2 -Medidas de similaridade e de dissimilaridade

Marta Noronha

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
CURSO: CIÊNCIA DE DADOS
APRENDIZADO DE MÁQUINA II

SUMÁRIO

Definição de cluster

- Conjunto de dados formado por p objetos dado por $D = \{v_1, v_2, ..., v_i, ..., v_p\}$.
- Problema de mapear k clusters no conjunto de dados D onde cada objeto v_i (onde $1 \leq i \leq p$) deve ser mapeado em um cluster j (onde $1 \leq j \leq k$).
- $f : D \rightarrow \{1, 2, ..., k\}$.
- $C_j = \{v_i | f(t_i) = C_j\}$ e $v_i \in D$.
- Cada v_i é um vetor de dados contendo n atributos.
- Para uso no material de estudo, $v_1 = \{x_1, x_2, ..., x_i, ..., x_n\}$ e $v_2 = \{y_1, y_2, ..., y_i, ..., y_n\}$ onde cada x_i e y_i corresponde a um atributo.

Média, variância, desvio padrão e coeficiente de variação

Média:

- A média de um atributo x, dada por $\overline{x}$, é dada pela soma de todos os valores observados para este atributo em cada objeto do conjunto de dados.
- Ideal para representação de dados homogêneos.
- *Outliers* forçam um deslocamento da média.
- Se for medido sobre a população é indicado por μ. Sobre amostras, $\overline{x}$.

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Média, variância, desvio padrão e coeficiente de variação

Variância:

- Indica a variação nos valores observados de um atributo, ou seja, o espalhamento ou dispersão.
- Se for medido sobre a população é indicado por σ^2. Sobre amostras, s^2.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Média, variância, desvio padrão e coeficiente de variação

Desvio padrão:

- Mede a dispersão da distribuição dos dados em relação à média.
- Se for medido sobre a população é indicado por σ. Sobre amostras, s.

$$s = \sqrt{s^2}$$

Média, variância, desvio padrão e coeficiente de variação

Coeficiente de variação (CV):

- Mede a razão entre desvio padrão e a média.
- Não possui unidade de medida.
- Mede a variabilidade dos dados em relação à média.
- Quanto maior o CV, maior a variância.

$$CV = \frac{s}{\bar{x}} * 100$$

Exercício

Em uma empresa trabalham 8 pessoas. Destas, 2 possuem 20 anos, 3 possuem 24, 1 possui 32 e 2 possuem 38. Calcule:

- A média;
- A variância;
- O desvio padrão;
- O coeficiente de variação.

Exercício

	Idade	(xi-x)^2
	20	56,25
	20	56,25
	24	12,25
	24	12,25
	24	12,25
	32	20,25
	38	110,25
	38	110,25
média	27,5	
variância	48,75	
Desvio padrão	6,9821200219	
CV	0,2538952735	

Exercícios resolvidos

	Idade	(xi-x)^2
	20	56,25
	20	56,25
	24	12,25
	24	12,25
	24	12,25
	32	20,25
	38	110,25
	38	110,25
média	27,50	
variância	48,75	
Desvio padrão	6,98	
CV	25,39	

	Idade	(xi-x)^2
	20	13,14
	20	13,14
	24	0,14
	24	0,14
	24	0,14
	25	1,89
	26	5,64
	26	5,64
média	23,63	
variância	4,98	
Desvio padrão	2,23	
CV	9,45	

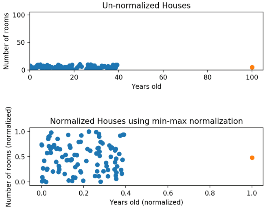
	Idade	(xi-x)^2
	20	315,06
	20	315,06
	24	189,06
	24	189,06
	24	189,06
	32	33,06
	38	0,06
	120	6.765,06
média	37,75	
variância	999,44	
Desvio padrão	31,61	
CV	83,75	

Normalização dos dados

Normalização Min-max

- Redimensiona o intervalo dos dados em $[0, 1]$.

$$MinMax = \frac{x - x_{min}}{x_{max} - x_{min}}$$



Fonte: Rudd et al. 2020

Normalização dos dados

Sem normalização

Idade	(xi-x)^2
20	56,25
20	56,25
24	12,25
24	12,25
24	12,25
32	20,25
38	110,25
38	110,25
média	27,50
variância	48,75
Desvio padrão	6,98
CV	25,39

Idade	(xi-x)^2
20	13,14
20	13,14
24	0,14
24	0,14
24	0,14
25	1,89
26	5,64
26	5,64
média	23,63
variância	4,98
Desvio padrão	2,23
CV	9,45

Idade	(xi-x)^2
20	315,06
20	315,06
24	189,06
24	189,06
24	189,06
32	33,06
38	0,06
120	6.765,06
média	37,75
variância	999,44
Desvio padrão	31,61
CV	83,75

Com normalização

Idade	Normalizado	(xi-x)^2
20	0,00	0,02
20	0,00	0,02
20	0,00	0,02
20	0,00	0,02
20	0,00	0,02
20	0,00	0,02
20	0,00	0,02
21	1,00	0,77
média	0,13	
variância	0,11	
desvio padrão	0,33	
CV	264,58	

Idade	Normalizado	(xi-x)^2
20	0,00	0,17
20	0,00	0,17
24	0,22	0,04
24	0,22	0,04
24	0,22	0,04
32	0,67	0,06
38	1,00	0,34
38	1,00	0,34
média	0,42	
variância	0,15	
desvio padrão	0,39	
CV	93,09	

Idade	Normalizado	(xi-x)^2
20	0,00	0,37
20	0,00	0,37
24	0,67	0,00
24	0,67	0,00
24	0,67	0,00
25	0,83	0,05
26	1,00	0,16
26	1,00	0,16
média	0,60	
variância	0,14	
desvio padrão	0,37	
CV	61,59	

Idade	Normalizado	(xi-x)^2
20	0,00	0,03
20	0,00	0,03
24	0,04	0,02
24	0,04	0,02
24	0,04	0,02
32	0,12	0,00
38	0,18	0,00
120	1,00	0,68
média	0,18	
variância	0,10	
desvio padrão	0,32	
CV	178,11	

Matriz de covariância

A matriz quadrada que contém as variâncias e covariâncias associadas aos atributos do conjunto de dados.
A diagonal da matriz contém os desvios das variáveis.
Os demais contêm as covariâncias entre cada par de atributos.

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x-\bar{x})*(y-\bar{y})}{n}$$

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}$$

Propriedades de medidas de distância

Considere três pontos de dados a,b e c:

- Distância mínima é igual a zero: A distância de qualquer ponto a ele mesmo é igual a zero.
- Positividade: Sendo $a \neq b$, $dist(a,b) > 0$.
- Simetria: Sendo $a \neq b$, $dist(a,b) = dist(b,a)$.
- Inequalidade triangular: A soma de dois lados quaisquer de um triângulo representado no espaço euclidiano é maior do que o terceiro lado. $dist(a,b) + dist(b,c) > dist(a,c)$

Medidas de (dis)similaridades para dados numéricos

- Em Shirkhorshidi, Aghabozorgi e Wah 2015 é apresentado um estudo de comparação de medidas de similaridades e dissimilaridades em dados contínuos.
- O estudo considera conjuntos de dados com baixa e alta dimensionalidade para analisar o comportamento das medidas em diferentes contextos de dimensionalidade e domínio.
- As medidas são núcleos do algoritmos e afetam diretamente no desempenho destes.
- Não existe uma medida que é capaz de ter o melhor desempenho em todos os tipos de conjuntos de dados, porém conhecê-las pode ajudar a selecionar uma que melhor se adapte ao problema.

Família de Minkowski

- Ideal para clusters isolados ou compactos, caso contrário pode ocorrer dominância pelos atributos de maior escala.
- Dominância por atributos de maior escala.
- Normalizar os dados soluciona o problema da dominância.
- Valor de m é um positivo real com $m \geq 1$.

$$d_{min} = \left(\sum_{i=1}^{n} |x_i - y_i|^m\right)^{\frac{1}{m}}$$

Família de Minkowski: Distância de Manhattan

- Caso especial de Minkowski em que $m = 1$.
- Distância sensível a *outliers*.
- Forma dos clusters é hiper-retangular.
- Independe da distribuição do conjunto de dados.

$$d_{man} = (\sum_{i=1}^{n} |x_i - y_i|)$$

Família de Minkowski: Distância euclidiana

- Ideal para clusters isolados ou compactos.
- "se dois vetores de dados não tiverem valores de atributos em comum, eles podem ter uma distância menor do que o outro par de vetores de dados contendo os mesmos valores de atributos" [Shirkhorshidi, Aghabozorgi e Wah 2015, Ricotta 2021].
  - **Par A (Sem Valores em Comum)**:

$$\mathbf{v1} = (1,4,7), \quad \mathbf{v2} = (2,5,8)$$

  - **Par B (Com Valores em Comum)**:

$$\mathbf{v3} = (1,4,7), \quad \mathbf{v4} = (1,4,10)$$

- Dominância por atributos de maior escala.
- Normalizar os dados soluciona o problema da dominância.
- Independe da distribuição do conjunto de dados.

$$d_{euc} = (\sum_{i=1}^{n} |x_i - y_i|^2)^{\frac{1}{2}}$$

Distância média

- Modificação da distância euclidiana que considera a média entre a quantidade de atributos dos vetores de dados.
- Menos sensível a *outliers*.
- Atribui igual importância a todos os atributos.
- Pode perder informação sobre a distribuição e disposição dos pontos dentro de um cluster.

$$d_{media} = (\frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|^2)^{\frac{1}{2}}$$

Distância ponderada euclidiana

- Modificação da distância euclidiana que considera o valor do peso dado ao i-ésimo atributo que está sendo analisado a cada iteração nos vetores de dados x_i e y_i.
- O cálculo desse peso é inerentemente relacionado ao conjunto de dados.
- Pesos apropriados reduzem a sensibilidade à escala (magnitude dos atributos) em relação a distância euclidiana.
- Seleção de pesos é complexo e algumas vezes baseado na intuição.

$$d_{pond} = \left(\sum_{i=1}^{n} w_i \times |x_i - y_i|^2\right)^{\frac{1}{2}}$$

Distância do cosseno

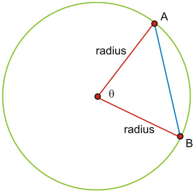
- Mais usada em descoberta de clusters em documentos por capturar a similaridade semântica entre os dados.
- Invariante à escala.
- Não possui informação sobre a magnitude dos atributos, somente do ângulo entre estes.
- Não lida com vetores ortogonais.
- Por não possuir informação sobre a distância entre os pontos, pode falhar para demonstrar a similaridade em alguns cenários.
- $||X|| * ||Y|| * cos\theta$
- Norma L^2: $||x_2|| = (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}$

$$d_{cos} = \frac{\sum_{i=1}^{n} x_i * y_i}{||x_2|| * ||y_2||}$$

Distância da corda

- Resolve problemas com a escala das medidas dos atributos.
- "Definida como o comprimento da corda que une dois pontos normalizados dentro de uma hiperesfera de raio igual a 1"([Shirkhorshidi, Aghabozorgi e Wah 2015]).
- Norma L^2: $||x_2|| = (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}$

$$d_{cord} = \left(2 - 2 * \frac{\sum_{i=1}^{n} x_i * y_i}{||x_2|| * ||y_2||}\right)^{\frac{1}{2}}$$

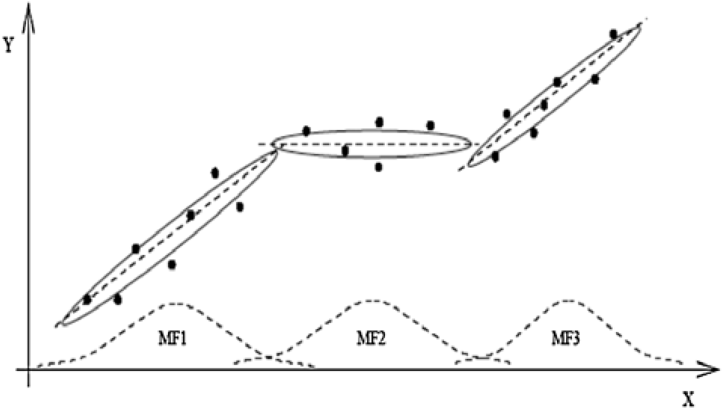


Distância de Mahalanobis

- Descoberta de cluster hiperelipsoidais por meio da análise da distância entre um ponto e uma distribuição.
- Invariante a escala.
- Comumente usada em dados multivariados.
- Depende da distribuição do conjunto de dados, mas assume que os dados seguem a distribuição normal.
- Útil para detectar *outliers* por considerar o desvio padrão entre os pontos.
- Mitiga distorções causadas por correlação lineares entre os atributos ao usar transformações por meio da matriz de covariância ou do quadrado da distância de Mahalanobis.

$$d_{mah} = ((x-y) * S^{-1} * (x-y)^T)^{\frac{1}{2}}$$

Distância de Mahalanobis



Fonte: Nayak, Sudheer e Jain 2007

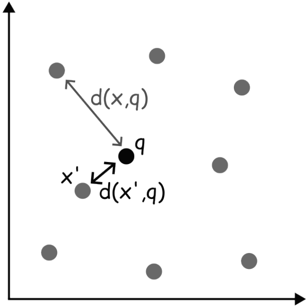
Consideração final do artigo

- Distância média obteve mais precisão nos algoritmos de agrupamento avaliados.
- Distância média converge rapidamente com k-means.
- Pearson não é recomendado para dados com baixa dimensionalidade, nem para uso em algoritmos baseados em centróides.
- Pearson é mais recomendado para dados com alta dimensionalidade e com abordagens hierárquicas.

A maldição da dimensionalidade

Problema

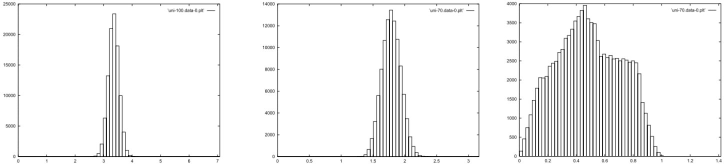
Encontrar o vizinho mais próximo x_{NN} de um ponto $q \in \mathbb{R}^d$ em um conjunto de dados $D \subset \mathbb{R}^d$, onde d é a quantidade de atributos do conjunto [Hinneburg, Aggarwal e Keim 2000].



$$x_{NN} = \{x' \in D | \forall x \in D, x \neq x' : dist(x', q) \leq dist(x, q)\}$$

A maldição da dimensionalidade

- O maior problema está na qualidade das soluções do que na representação dos dados.
- As medidas baseadas nas normas L_k tratam as dimensões de forma igual.



(a) 50 Dimensions (b) 10 Dimensions (c) 2 Dimensions

Similaridades baseadas em correlações

Pearson

- Mede a similaridade entre variáveis por meio da correlação (par-a-par) entre estas, ou seja, a força do relacionamento e a direção entre as variáveis.
- Um valor igual a 0 indica que não existe um relacionamento linear entre as variáveis.
- Um valor igual a 1 indica um relacionamento linear perfeito onde o aumento ou decremento nos valores de uma variável é refletida linearmente na outra.
- Um valor igual a (-1) indica que quando uma variável tem seu valor alto, a outra tem o seu valor baixo.

Exemplo 1: A renda tende a aumentar com a idade.

Exemplo 2: O tempo para se percorrer uma distância aumenta na medida em que uma pessoa anda mais devagar.

Pearson

- Conjunto de dados formado por p objetos dado por $D = \{v_1, v_2, ..., v_i, ..., v_p\}$ e n atributos onde o conjunto de atributos é dado por $A = \{x_1, x_2, ..., x_i, ..., x_n\}$.
- Considere x_s e x_t como atributos deste conjunto de dados.
- Considere as médias $\overline{x_s} = \frac{\sum_{i=1}^{p} v_{is}}{p}$ e $\overline{x_t} = \frac{\sum_{i=1}^{p} v_{it}}{p}$
- Os coeficientes de $P_{Coef}(x_s, x_t)$ variam entre $[-1, 1]$.

$$P_{Coef}(x_s, x_t) = \frac{\sum_{i=1}^{p} (v_{is} - \overline{x_s})(v_{it} - \overline{x_t})}{\sqrt{\sum_{i=1}^{p} (v_{is} - \overline{x_s})^2} \sqrt{\sum_{i=1}^{p} (v_{it} - \overline{x_t})^2}}$$

Dissimilaridade baseada em Pearson

A conversão da similaridade do coeficiente de Pearson em medida de distância é dada por:

$$d(x_s, x_t) = \frac{1 - P_{Coef}(x_s, x_t)}{2}$$ com intervalo entre $[0, 1]$.

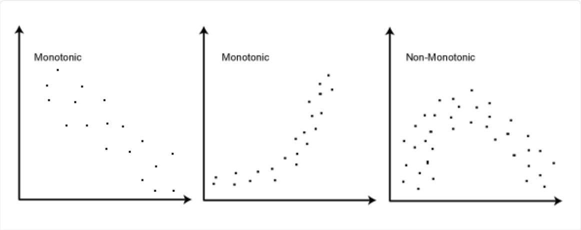
ou:

$$d(x_s, x_t) = 1 - P_{Coef}(x_s, x_t)$$ com intervalo entre $[0, 2]$.

Quando usar o coeficiente de Pearson

- As variáveis são quantitativas.
- As variáveis possuem distribuição normal (ou quase normal), sendo uma medida paramétrica .
- Não existem outliers.
- Relacionamento linear (scatterplot).

Quando um ou mais aspectos acima não forem satisfeitos ou se o relacionamento das variáveis for não linear e monotônico, escolher a medida não paramétrica rank de Spearman.



Spearman

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}}$$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

onde $d_i = \mathrm{R}(X_i) - \mathrm{R}(Y_i)$

Number	1	2	3	4	5	6	7	8	9	10
X_1	7	6	4	5	8	7	10	3	9	2
Y_1	5	4	5	6	10	7	9	2	8	1
Rank X_1	6.5	5	3	4	8	6.5	10	2	9	1
Rank Y_1	4.5	3	4.5	6	10	7	9	2	8	1
d^2	4	4	2.25	4	4	0.25	1	0	1	0

Fonte: https://www.geeksforgeeks.org/spearmans-rank-correlation/

Spearman

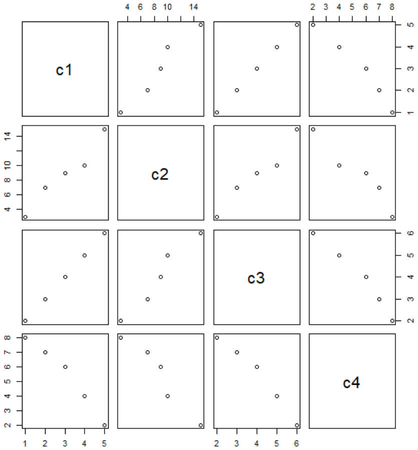
```
> dados4 <- data.frame(c1 = c(7,6,4,5,8,7,10,3,9,2), c2 = c(5,4,5,6,10,7,9,2,8,1))
> dados4
   c1 c2
1   7  5
2   6  4
3   4  5
4   5  6
5   8 10
6   7  7
7  10  9
8   3  2
9   9  8
10  2  1
>
> dados4.spearman <- cor(dados4, method = "spearman")
> dados4.spearman
      c1    c2
c1 1.000 0.875
c2 0.875 1.000
```

Exemplos

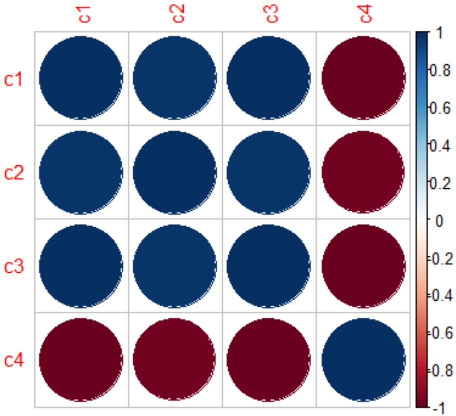
```
library("ggpubr")
library("ggplot2")
library("corrplot")

dados2 <- data.frame(c1=c(1,2,3,4,5),c2=c(3,7,9,10,15), c3 = c(2,3,4,5,6), c4 = c(8,7,6,4,2))
plot(dados2)

dados2.pearson <- cor(dados2, method = "pearson")
print(dados2.pearson, digits = 2)
corrplot(dados2.pearson)


dados3 <- data.frame(c1=c(0,16,12,20,10),c2=c(5,18,27,30,7), c3 = c(13,2,25,10,1), c4 = c(1,18,9,3,2))
plot(dados3)
dados3.pearson <- cor(dados3, method = "pearson")
print(dados3.pearson, digits = 2)
corrplot(dados3.pearson)

dados3.spearman <- cor(dados3, method = "spearman")
print(dados3.spearman, digits = 2)
corrplot(dados3.spearman)
```

Exemplos

```
> dados2
  c1 c2 c3 c4
1  1  3  2  8
2  2  7  3  7
3  3  9  4  6
4  4 10  5  4
5  5 15  6  2
> plot(dados2)
>
> dados2.pearson <- cor(dados2, method = "pearson")
> print(dados2.pearson, digits = 2)
      c1    c2    c3    c4
c1  1.00  0.97  1.00 -0.98
c2  0.97  1.00  0.97 -0.96
c3  1.00  0.97  1.00 -0.98
c4 -0.98 -0.96 -0.98  1.00
```

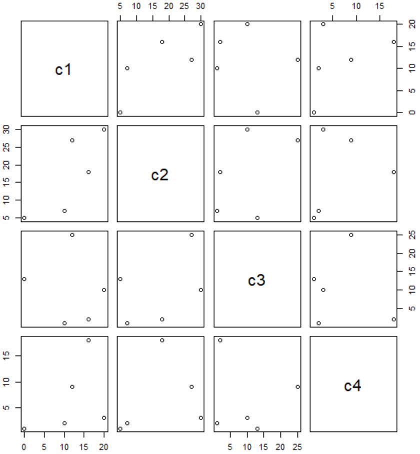


Exemplos

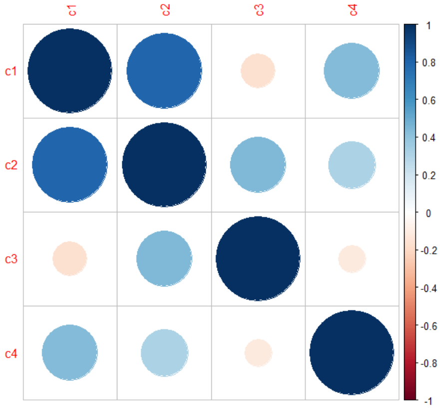


Exemplos

```
> dados3
  c1 c2 c3 c4
1  0  5 13  1
2 16 18  2 18
3 12 27 25  9
4 20 30 10  3
5 10  7  1  2
> plot(dados3)
> dados3.pearson <- cor(dados3, method = "pearson")
> print(dados3.pearson, digits = 2)
      c1   c2    c3    c4
c1  1.00 0.80 -0.17  0.44
c2  0.80 1.00  0.44  0.32
c3 -0.17 0.44  1.00 -0.11
c4  0.44 0.32 -0.11  1.00
```

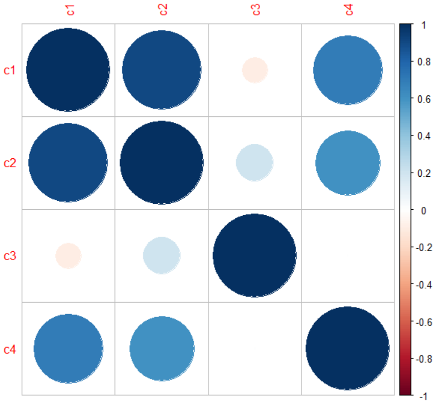


Exemplos

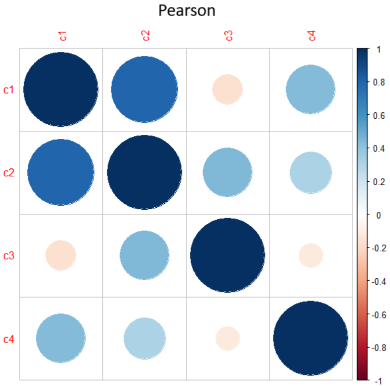


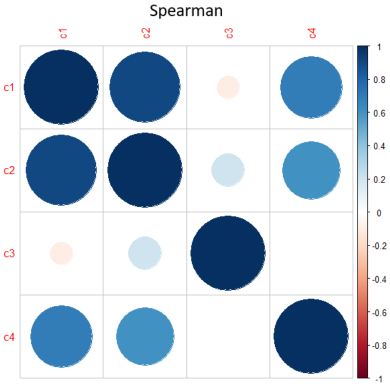
Exemplos

```
> dados3
  c1 c2 c3 c4
1  0  5 13  1
2 16 18  2 18
3 12 27 25  9
4 20 30 10  3
5 10  7  1  2

> dados3.spearman <- cor(dados3, method = "spearman")
> print(dados3.spearman, digits = 2)
     c1  c2   c3  c4
c1  1.0 0.9 -0.1 0.7
c2  0.9 1.0  0.2 0.6
c3 -0.1 0.2  1.0 0.0
c4  0.7 0.6  0.0 1.0
```



Exemplos


Pearson
c1
c2
c3
c4
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Spearman
c1
c2
c3
c4
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Exercícios

Distância Euclidiana:

1. Dados os pontos A(2, 3) e B(5, 7), calcule a distância euclidiana entre eles.
2. Dado um vetor tridimensional A = (3, 8, 4) e um vetor B = (1, 6, 2), qual é a distância euclidiana entre eles?
3. Dado um conjunto de pontos (x, y) representados por (2, 4), (7, 1), (3, 9), (6, 5), calcule as distâncias euclidianas de cada ponto em relação ao ponto de referência (4, 3).

Exercícios

Distância de Manhattan:

1. Dados os pontos A(2, 3) e B(5, 7), calcule a distância de Manhattan entre eles.
2. Dado um vetor tridimensional A = (3, 8, 4) e um vetor B = (1, 6, 2), qual é a distância de Manhattan entre eles?
3. Suponha que você está em uma cidade em uma grade e deseja ir do ponto (1, 3) ao ponto (6, 8), contando apenas movimentos para cima, para baixo, para a esquerda e para a direita. Qual é a distância de Manhattan percorrida?

Exercícios

Distância de Cosseno:

1. Dados os vetores A = (2, 3) e B = (5, 7), calcule a similaridade de cosseno entre eles.
2. Dado os vetores A = (1, 0, 2) e B = (3, 1, 0), calcule a similaridade de cosseno entre eles.
3. Dado os vetores A = (4, 1, 3) e B = (2, 2, 6), calcule a similaridade de cosseno entre eles.

Exercícios

Distância Euclidiana Média:

1. Dado um conjunto de pontos (1, 2), (4, 6), (3, 8), calcule a distância euclidiana média entre todos os pares de pontos.
2. Dado um conjunto de pontos (2, 3), (5, 7), (1, 1), qual é a distância euclidiana média entre os pontos?
3. Dado quatro pontos em um espaço tridimensional: P(1, 2, 3), Q(4, 5, 6), R(0, 1, 2) e S(3, 4, 5). Calcule a distância euclidiana média entre todos os pares de pontos.

Exercícios

Correlação de Pearson:

- Calcule a correlação de Spearman entre as variáveis e construa uma matriz de correlação com os resultados obtidos.
- Indique quais variáveis são mais diretamente e quais são inversamente correlacionadas.
- Use R ou Python para verificar os resultados.

	a1	a2	a3	a4
o1	1	3	9	2
o2	4	6	5	6
o3	9	2	3	11
o4	3	4	8	4
o5	8	9	6	9
o6	5	5	2	10

Exercícios

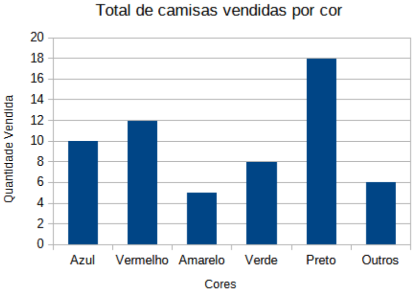
Correlação de Spearman:

- Calcule a correlação de Spearman entre as variáveis e construa uma matriz de correlação com os resultados obtidos.
- Indique quais variáveis são mais diretamente e quais são inversamente correlacionadas.
- Use R ou Python para verificar os resultados.

	a1	a2	a3	a4
o1	1	3	9	2
o2	4	6	5	6
o3	9	2	3	11
o4	3	4	8	4
o5	8	9	6	9
o6	5	5	2	10

Contextualizando

- Dificuldade em obter grupos onde os atributos são categóricos.
- K-modes, algoritmo baseado em K-means, foi proposto utilizando a medida de dissimilaridade *SimpleMatching* para descoberta de clusters em dados categóricos (anos 90).
- Uso de histogramas para análise de dados categóricos teve início em 2002.



Contextualizando

- As medidas podem ser associadas a um peso w. Considere o peso igual a 1 neste material.
- Medidas são consideradas como um dos 3 tipos para avaliar a dissimilaridade [dos Santos e Zárate 2015]:

Tipo 1: $$S_k(i,j) = \begin{cases} y, & \text{if } i=j, \text{ where } 0 \leqslant y \leqslant 1 \\ 0, & \text{if } i \neq j \end{cases}$$

Tipo 2: $$S_k(i,j) = \begin{cases} 1, & \text{if } i=j \\ y, & \text{if } i \neq j, \text{ where } 0 \leqslant y \leqslant 1 \end{cases}$$

Tipo 3: $$S_k(i,j) = \begin{cases} y, & \text{if } i=j, \text{ where } 0 \leqslant y \leqslant 1 \\ z, & \text{if } i \neq j, \text{ where } 0 \leqslant z \leqslant 1 \end{cases}$$

Medidas de similaridades para dados categóricos

Similaridade de Gower

- Útil para medir similaridade em dados heterogêneos (dados categóricos e numéricos).
- Medida de tipo 1.

$$S_k(x_{ak}, x_{bk}) = \begin{cases} 1, & \text{if } x_{ak} = x_{bk} \\ 0, & \text{if } x_{ak} \neq x_{bk} \end{cases}$$

$$S(X_a, X_b) = \sum_{k=1}^{n} \frac{\omega_k S_k(x_{ak}, x_{bk})}{n}$$

Medidas de similaridades para dados categóricos

Prós da medida de Gower:

- Adequação para Dados Heterogêneos;
- Menor sensibilidade (mas não totalmente insensível) à variação na escala dos dados;
- Fácil implementação; e,
- Valor de similaridade entre objetos intuitivo onde quanto mais perto de 1 maior a similaridade. Quanto mais perto de zero, mais dissimilaridade existe entre os objetos.

Contras da medida de Gower:

- A sensibilidade à escala de atributos é menor do que algumas métricas, porém ainda pode ser sensível frente à escala dos dados ordinais;
- Escolha de pesos para cada atributo pode ser subjetiva;
- A dimensionalidade dos dados (em atributos ou objetos) impacta diretamente na complexidade computacional; e,
- A interpretação de resultados em dados categóricos é um pouco mais limitada do que nas demais.

Medidas de similaridades para dados categóricos

- Se igual 1. Senão, 0.

Atributo	Indivíduo 1	Indivíduo 2	Igual
Fumante	1	0	0
Tem carro	1	1	1
Estudante	0	0	1
Empregado	1	0	0
Tem casa própria	0	1	0

$$Sim_{Gower} = \frac{0+1+1+0+0}{5} = 0,4$$

Ou seja, os indivíduos possuem 40% de semelhança.

Medidas de similaridades para dados categóricos

- Exemplo de uso da medida de Gower em dados heterogêneos. (Fonte: https://jamesmccaffrey.wordpress.com/2020/04/21/example-of-calculating-the-gower-distance/)
  - A faixa (*range*) é calculada pelo maior valor assumido pelo atributo menos o menor valor deste atributo.

	Age (n)	Race	Height (n)	Income (n)	IsMale	Politic
[1]	22	1	3	0.39	TRUE	moderate
[2]	33	3	1	0.34	TRUE	liberal
[3]	52	1	2	0.51	FALSE	moderate
[4]	46	6	3	0.63	TRUE	conservative
range	30	NA	2	0.29	NA	NA

Medidas de similaridades para dados binários

- Observe que para o exemplo específico, houve uma conversão da similaridade em dissimilaridade (distância).

```
      Age Race  Ht  Inc   Male  Politic
[1] = (22,  1,  3,  0.39, True, moderate)
[2] = (33,  3,  1,  0.34, True, liberal)

numeric: abs(diff) / range
non-numeric: 0 if equal, 1 if different

dist([1], [2]) =

Age:      abs((22 - 33) / 30)       = 0.367
Race:     (different)               = 1
Height:   abs((3 - 1) / 2)          = 1.000
Inc:      abs((0.39 - 0.34) / 0.29) = 0.172
IsMale:   (same)                    = 0
Politic:  (different)               = 1

 = (0.367 + 1 + 1.000 + 0.172 + 0 + 1) / 6
 = 3.539 / 6
 = 0.590
```

Medidas de similaridades para dados categóricos

Similaridade de Eskin

- Usada para medir a similaridade entre objetos com atributos categóricos.
- Menor valor alcançado quando o atributo possui somente $m = 2$ valores possíveis.
- Medida de tipo 2.
- A matriz contém proximidades entre todos os pares de objetos.
- Pode ser usado em análises de cluster hierárquicos, como o algoritmo AGNES.

Limites: $\left[\frac{2}{3}, \frac{m^2}{m^2+2}\right]$

$$S_k(x_{ak}, x_{bk}) = \begin{cases} 1, & \text{if } x_{ak} = x_{bk} \\ \frac{m_k^2}{m_k^2+2}, & \text{if } x_{ak} \neq x_{bk} \end{cases}$$

Medidas de similaridades para dados categóricos

Prós da medida de Eskin:

- Muito útil em dados binários onde a ausência ou presença de atributos é o único fator considerado;
- Adequada para análise de similaridade em variáveis categóricas e nominais; e,
- Fácil de calcular.

Contras da medida de Eskin:

- Não pode ser usada sobre dados contínuos;
- Pode apresentar baixo desempenho em atributos que possuem um número grande de possíveis valores [Boriah, Chandola e Kumar 2008];
- Pode apresentar alguns resultados inadequados por não considerar a ausência parcial ou total de atributos nos objetos a e b em análise, considerando a ausência igualmente importante; e,
- Não considera a frequência com que itens aparecem em um conjunto, sendo tratados semelhantemente.

Medidas de similaridades para dados categóricos

Eskin

Considere dois indivíduos com um atributo categórico "Cidade". O Indivíduo 1 tem o valor "São Paulo"e o Indivíduo 2 tem uma ausência de dados (representada por "?"). O atributo "Cidade"tem 5 valores, então $m_k = 5$. A fórmula de Eskin trata a ausência da mesma forma que qualquer outro valor diferente.

Indivíduo 1	**Indivíduo 2**	similaridade
São Paulo	?	$S_k = \frac{5^2}{5^2+2} = \frac{25}{27} = 0.9259$

Medidas de similaridades para dados categóricos

Similaridade de Lin

- Baseada em teoria de informação para medir a similaridade entre dois conceitos por meio da informação mútua compartilhada entre estes;
- É frequentemente usada em processamento de linguagem natural (NLP) e na recuperação de informações para avaliar a semelhança entre palavras ou termos em um contexto específico.
- A ocorrência de dois termos juntos com uma frequência maior do que o esperado então estes termos devem ser mais similares.

$$S_k(x_{ak}, x_{bk}) = \begin{cases} 2\log p_k(x_{ak}), & \text{if } x_{ak} = x_{bk} \\ 2\log(p_k(x_{ak}) + p_k(x_{bk})), & \text{if } x_{ak} \neq x_{bk} \end{cases}$$

Medidas de similaridades para dados categóricos

Distância de correspondência simples (Simple Matching)

- Considera presenças mútuas e ausências mútuas como a mesma importância, sendo portanto uma medida simétrica.
- Exemplo de uso: semelhança entre clientes baseando nos produtos que consomem e restrigem.

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Medidas de similaridades para dados categóricos

Similaridade de Jaccard

- Considera somente presenças como sendo importantes, logo é uma medida assimétrica.
- Exemplo de uso: contabilização de exames positivos para um diagnóstico de doença.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Exercício prático avaliativo (5 pontos)

- Ler o artigo “A Survey of Binary Similarity and Distance Measures"(disponível aqui).
- Selecionar 8 medidas de avaliação em dados binários.
- Criar uma matriz binária contendo 6 objetos e 10 atributos preenchidos com valores aleatórios.
- Montar a matriz de similaridade entre os objetos da matriz (Pode-se usar mapa de calor para apresentar cada matriz).
- Montar uma breve apresentação (aprox. 20 minutos) falando brevemente sobre as medidas escolhidas, definindo por exemplo se são ou não simétricas, e outras características interessantes, apresentando os resultados da aplicação das medidas na matriz.

REFERÊNCIAS

BORIAH, S.; CHANDOLA, V.; KUMAR, V. Similarity measures for categorical data: A comparative evaluation. In: SIAM. *Proceedings of the 2008 SIAM international conference on data mining*. [S.l.], 2008. p. 243–254.

dos Santos, T. R.; ZáRATE, L. E. Categorical data clustering: What similarity measure to recommend? *Expert Systems with Applications*, v. 42, n. 3, p. 1247–1260, 2015. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095741741400551X>.

HINNEBURG, A.; AGGARWAL, C. C.; KEIM, D. A. What is the nearest neighbor in high dimensional spaces? In: *26th Internat. Conference on Very Large Databases*. [S.l.: s.n.], 2000. p. 506–515.

NAYAK, P. C.; SUDHEER, K.; JAIN, S. Rainfall-runoff modeling through hybrid intelligent system. *Water Resources Research*, v. 43, 07 2007.

RICOTTA, C. From the euclidean distance to compositional dissimilarity: What is gained and what is lost. *Acta Oecologica*, v. 111, p. 103732, 2021. ISSN 1146-609X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1146609X2100031X>.

RUDD, J. M. et al. An empirical study of downstream analysis effects of model pre-processing choices. *Open journal of statistics*, Scientific Research Publishing, v. 10, n. 5, p. 735–809, 2020.

SHIRKHORSHIDI, A. S.; AGHABOZORGI, S.; WAH, T. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLOS ONE*, v. 10, p. e0144059, 12 2015.