Curso 2 – CD, AM e DM

Mineração de Dados

Parte 2

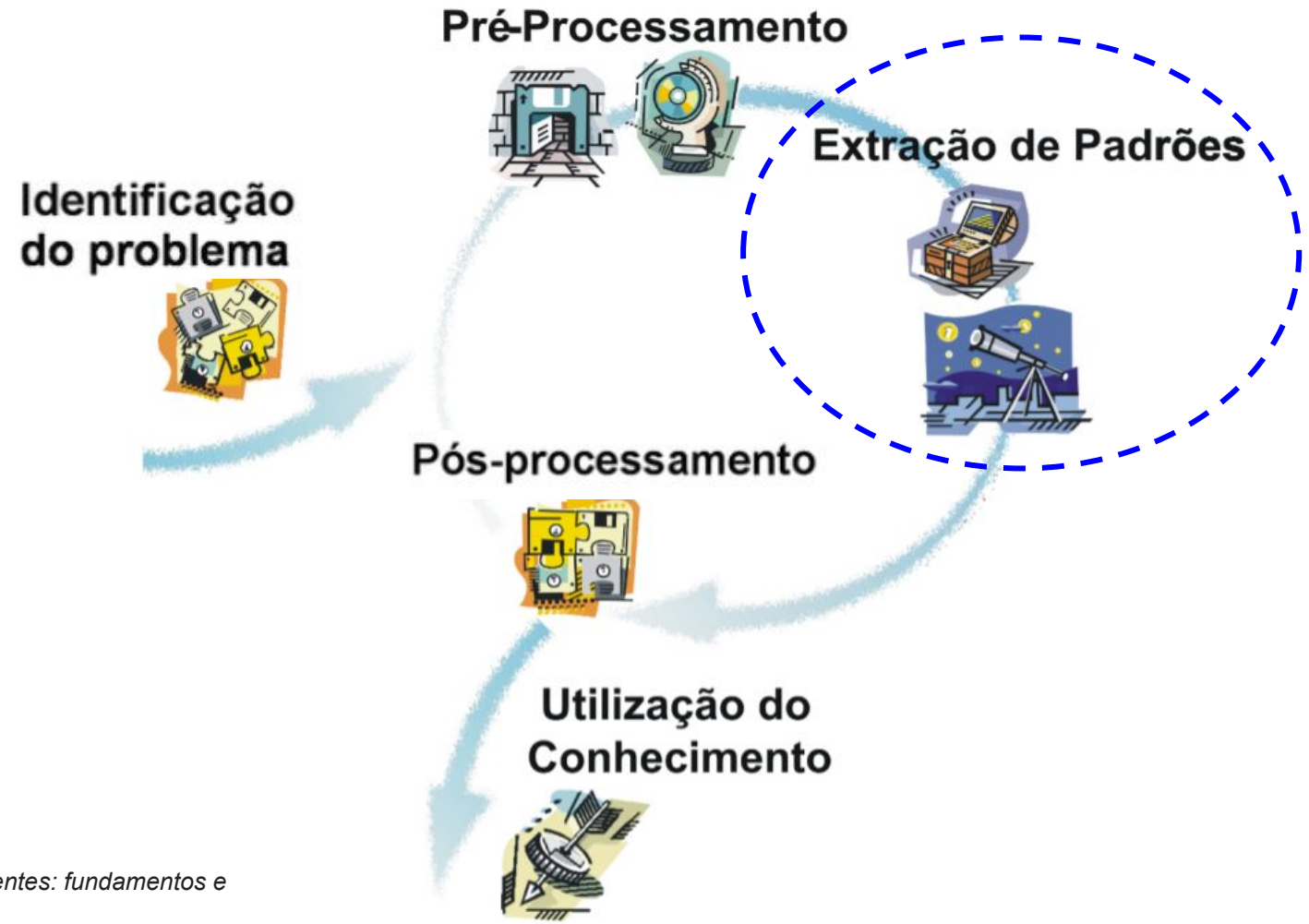
Extração de Padrões
Agrupamento de Dados
Medidas de Proximidade

Prof. Ricardo M. Marcacini
ricardo.marcacini@icmc.usp.br

Etapas do Processo de Mineração de Dados

Fonte: Rezende, S. O. *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda, 2003.

Etapas do Processo de Mineração de Dados



Fonte: Rezende, S. O. *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda, 2003.

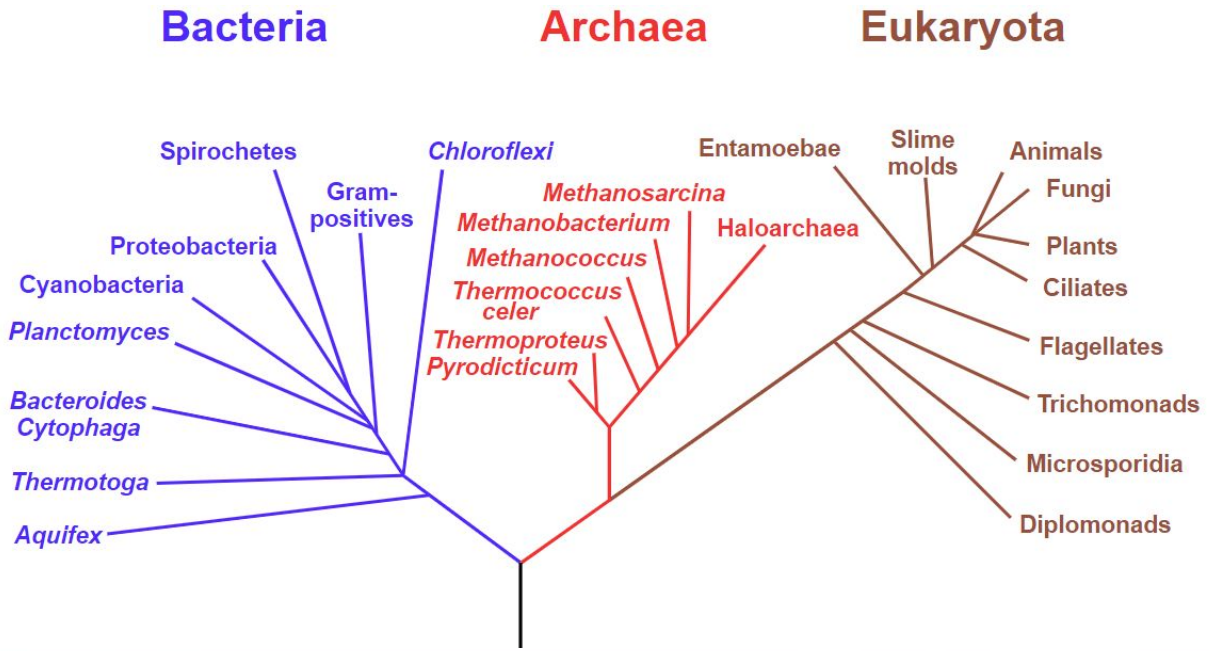
Extração de Padrões

Aprendizado não Supervisionado

- Tarefas Descritivas
- Agrupamento de Dados

Motivação

- Humanos se interessam por “organizar e agrupar”
  - Filmes: animação, dramas, comédias, terror, ...
  - Músicas: sertanejo, rock, funk, mpb, …
  - Biologia:

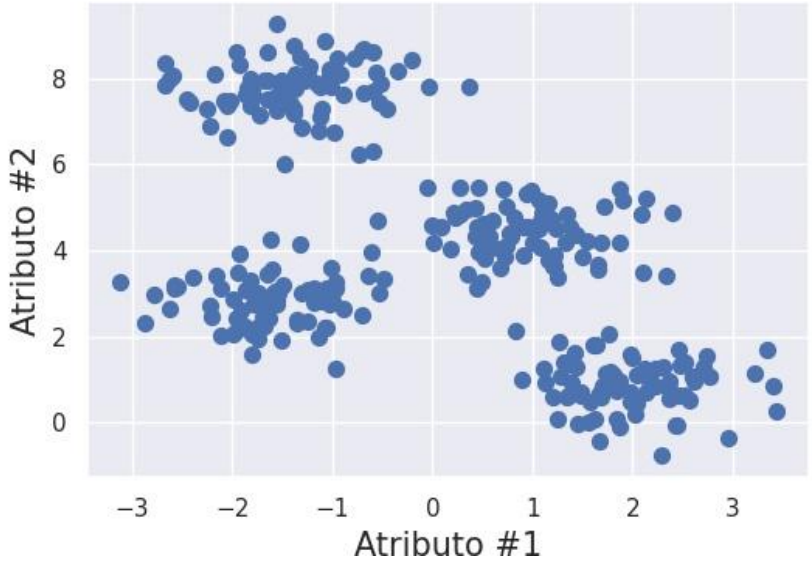


Motivação

- Humanos se interessam por “organizar e agrupar”
  - Filmes: animação, dramas, comédias, terror, ...
  - Músicas: sertanejo, rock, funk, mpb, …
  - Biologia: organização de espécies
  - Psicologia: organizar pessoas em perfis de personalidade
  - Medicina: organizar por tipos ou subtipos de doenças
  - Administração/Marketing: segmentação de clientes

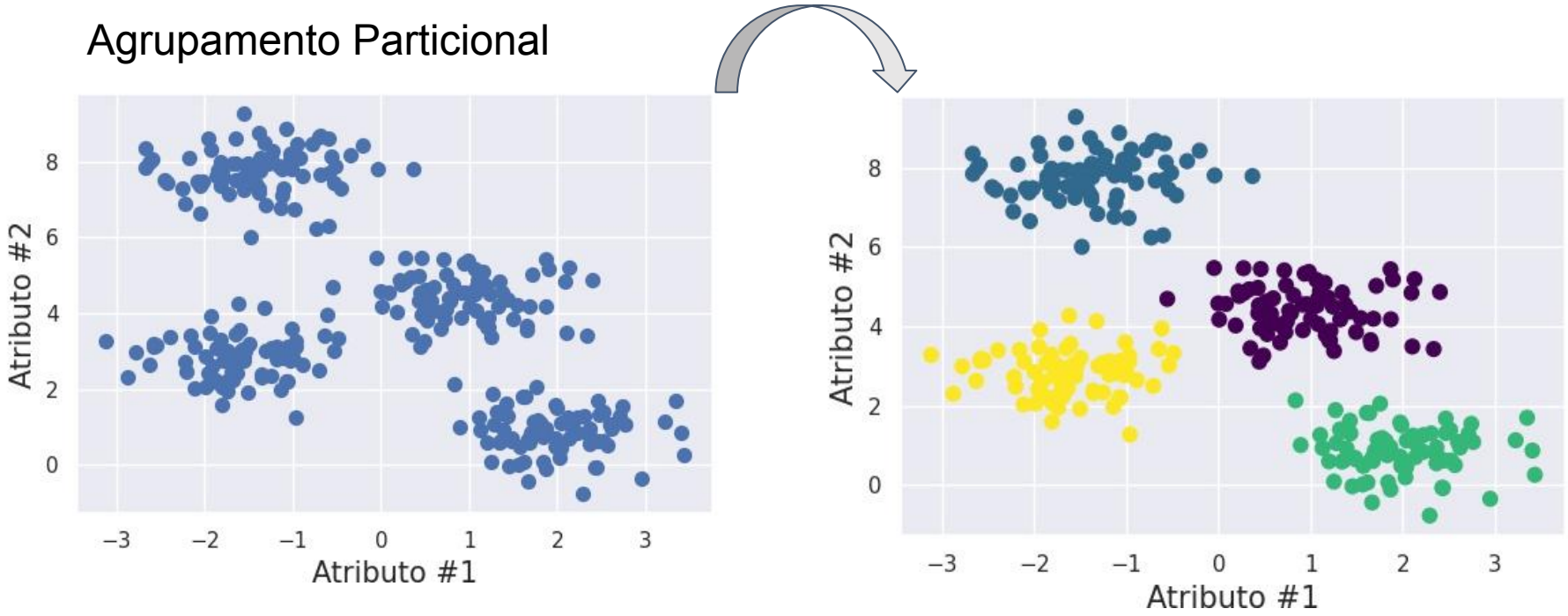
Motivação

- Humanos se interessam por “organizar e agrupar”
  - Análise Exploratória de Dados



Motivação

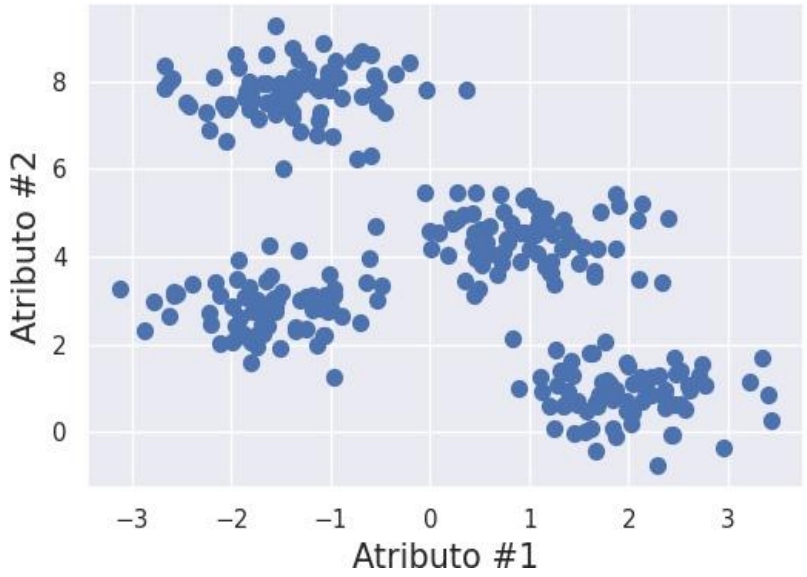
- Humanos se interessam por “organizar e agrupar”
  - Análise Exploratória de Dados

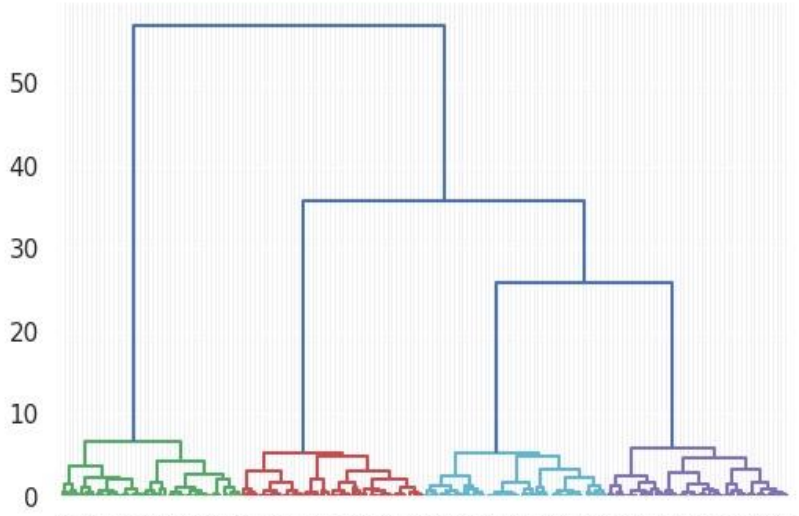


Motivação

- Humanos se interessam por “organizar e agrupar”
  - Análise Exploratória de Dados

Agrupamento Hierárquico

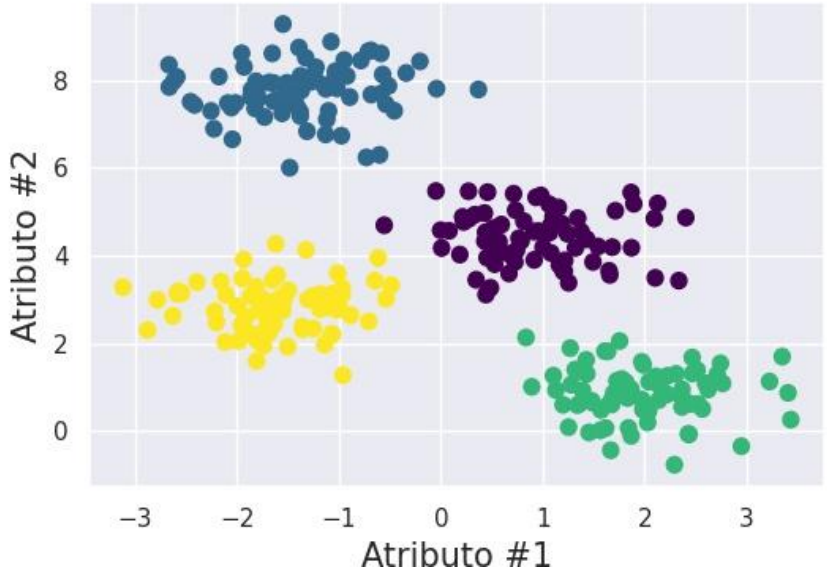




Conceitos Básicos

- O que é um *cluster*?

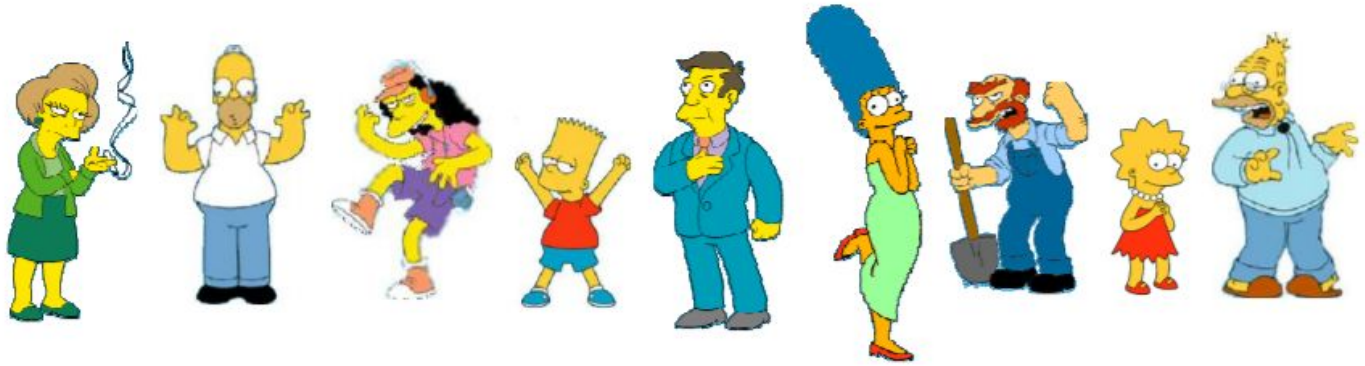
Um *cluster* (grupo) é um conjunto de objetos semelhantes.
Objetos alocados a diferentes *clusters* não são semelhantes.



- Homogeneidade: coesão interna
- Heterogeneidade: separação entre grupos

Conceitos Básicos

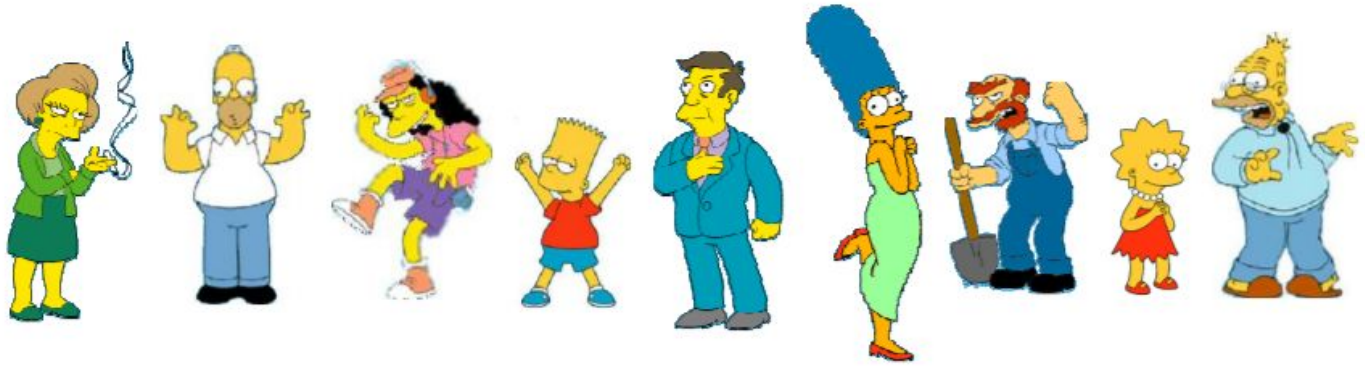
- Desafio: Vamos agrupar os seguintes personagens...



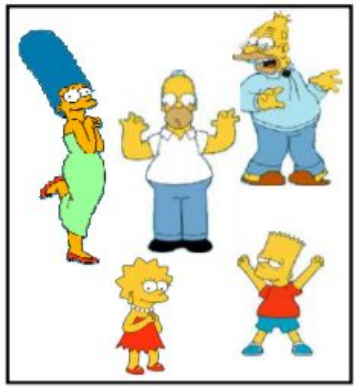
Fonte do exemplo:
Keogh, E. A Gentle Introduction to Machine Learning and A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Conceitos Básicos

- Desafio: Vamos agrupar os seguintes personagens...



Solução #1

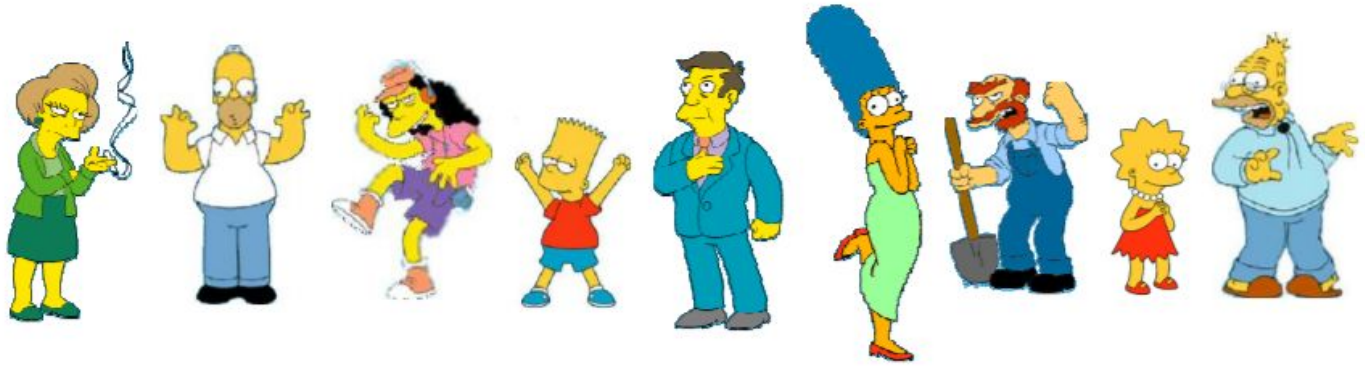


Família

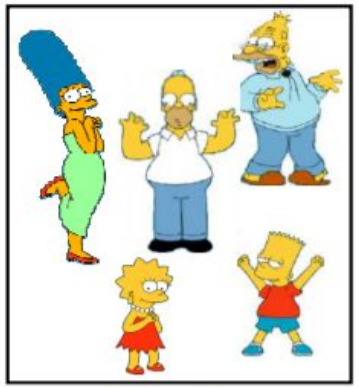
Empregados da escola

Conceitos Básicos

- Desafio: Vamos agrupar os seguintes personagens...



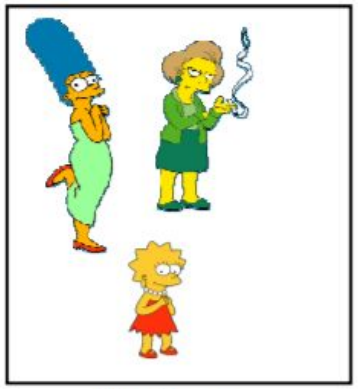
Solução #1



Família

Empregados da escola

Solução #2



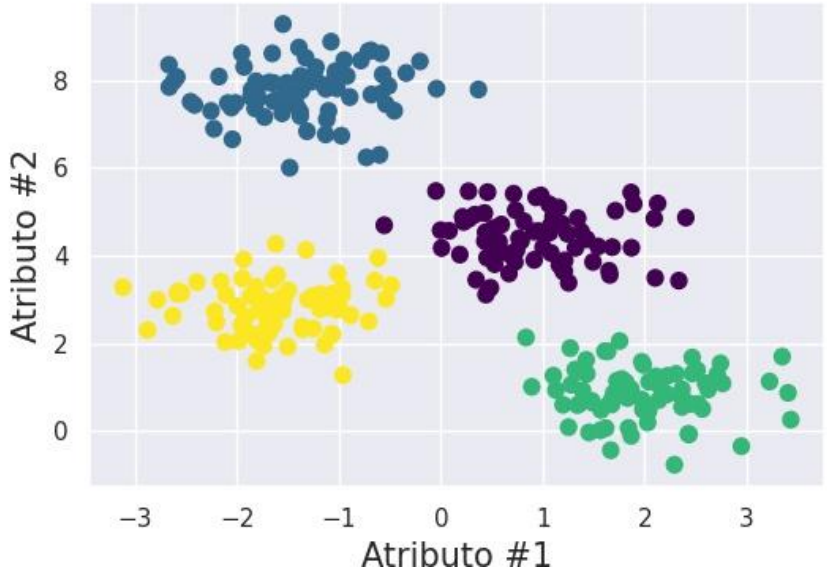
Mulheres

Homens

Conceitos Básicos

- O que é um *cluster*?

Um *cluster* (grupo) é um conjunto de objetos semelhantes.
Objetos alocados a diferentes *clusters* não são semelhantes.

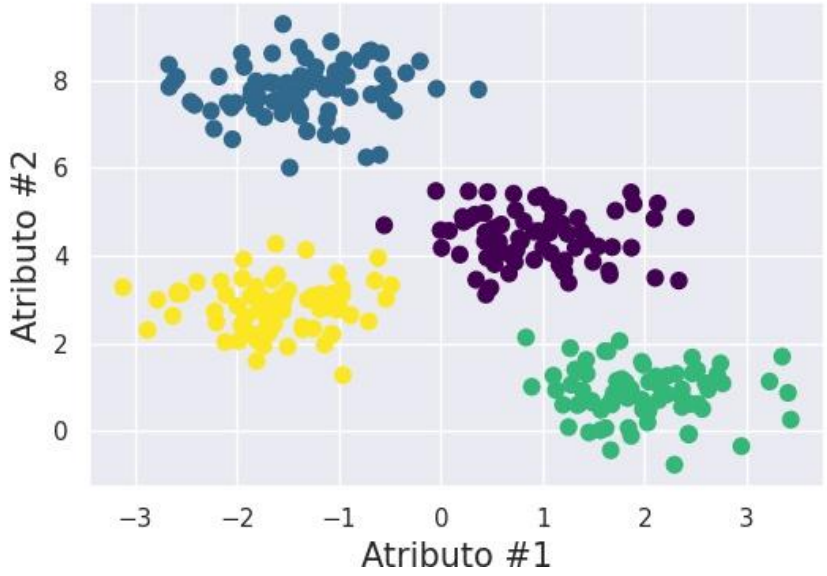


- Homogeneidade: coesão interna
- Heterogeneidade: separação entre grupos

Conceitos Básicos

- O que é um *cluster*?

Um *cluster* (grupo) é um conjunto de objetos semelhantes.
Objetos alocados a diferentes *clusters* não são semelhantes.



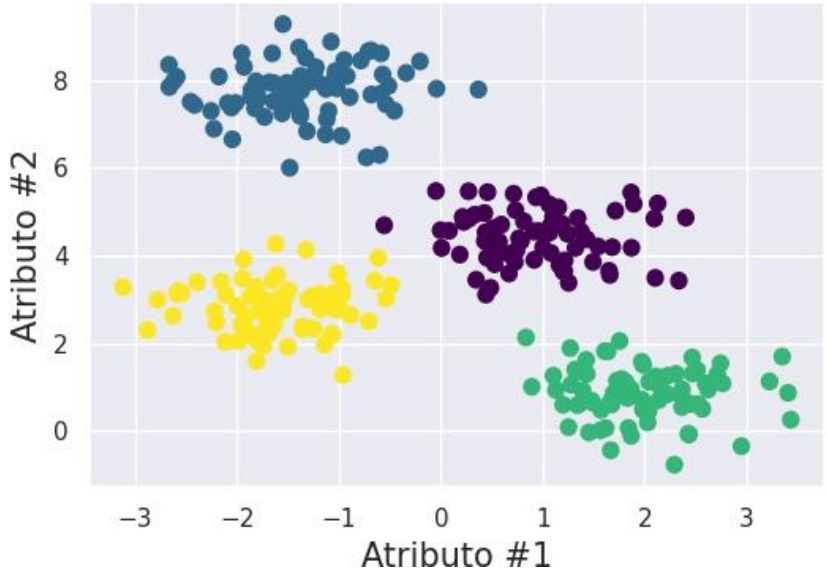
- Homogeneidade: coesão interna
- Heterogeneidade: separação entre grupos

Cluster é um conceito subjetivo!

Conceitos Básicos

- O que é um *cluster*?

Um *cluster* (grupo) é um conjunto de objetos semelhantes.
Objetos alocados a diferentes *clusters* não são semelhantes.



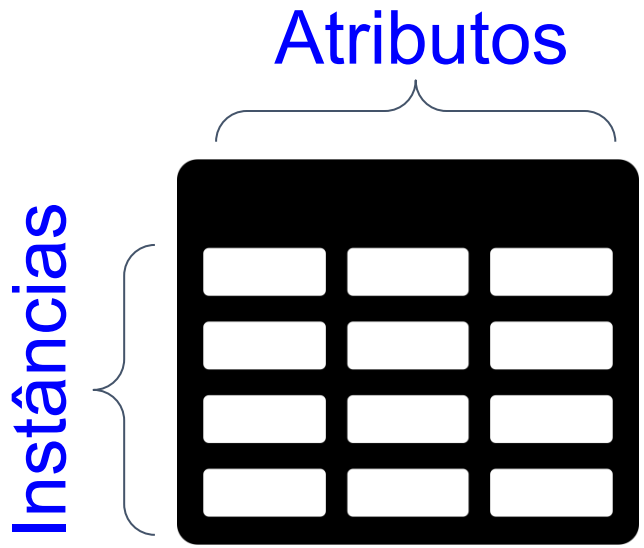
- Homogeneidade: coesão interna
- Heterogeneidade: separação entre grupos

Medidas de Similaridade e Dissimilaridade

Conceitos Básicos

- Notação:

Denotemos por $\mathbf{X}_{n \times d}$ uma matriz atributo-valor formada por n objetos da base de dados e d atributos.



$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

Conceitos Básicos

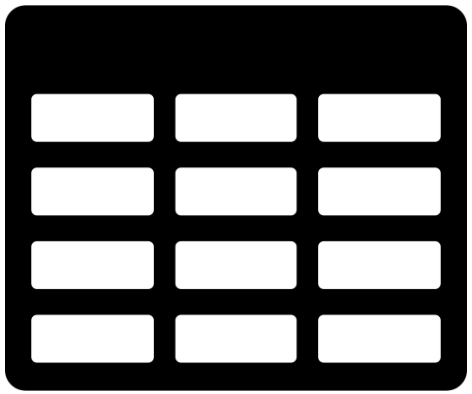
- Notação:

Por simplicidade, cada objeto será denotado por um vetor:

$$\mathbf{x}_i = [x_{i1} \cdots x_{id}]$$

Atributos

Instâncias



$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

Medidas de Proximidade

- Medida de proximidade será denotada por $d(\mathbf{x}_i, \mathbf{x}_j)$
  - Representa tanto uma similaridade quanto uma dissimilaridade (distância)

- Dissimilaridade: $d(\mathbf{x}_i, \mathbf{x}_i) = 0 \quad \forall \mathbf{x}_i$

- Similaridade: $d(\mathbf{x}_i, \mathbf{x}_i) \geq \max_j d(\mathbf{x}_i, \mathbf{x}_j) \quad \forall \mathbf{x}_i$

Medidas de Proximidade

- Medida de proximidade será denotada por $d(\mathbf{x}_i, \mathbf{x}_j)$
  - Representa tanto uma similaridade quanto uma dissimilaridade (distância)
- Propriedade desejáveis (para dissimilaridade)
  - Simetria $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i) \quad \forall \mathbf{x}_i, \mathbf{x}_j$
  - Positividade $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \quad \forall \mathbf{x}_i, \mathbf{x}_j$
  - Reflexividade $d(\mathbf{x}_i, \mathbf{x}_j) = 0 \quad \text{se, e somente se,} \quad \mathbf{x}_i = \mathbf{x}_j$
  - Desigualdade Triangular *(nas próximas aulas)*

Medidas de Proximidade

- Medida de proximidade será denotada por $d(\mathbf{x}_i, \mathbf{x}_j)$
  - Representa tanto uma similaridade quanto uma dissimilaridade (distância)
- Propriedade desejáveis (para dissimilaridade)
  - Simetria $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i) \ \ \forall \mathbf{x}_i, \mathbf{x}_j$ *Métrica!*
  - Positividade $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \ \ \forall \mathbf{x}_i, \mathbf{x}_j$
  - Reflexividade $d(\mathbf{x}_i, \mathbf{x}_j) = 0$ se, e somente se, $\mathbf{x}_i = \mathbf{x}_j$
  - Desigualdade Triangular *(nas próximas aulas)*

Distância Euclidiana

- Dados contínuos
- Métrica (propriedades desejáveis)

$$d^{E}_{(\mathbf{x}_i,\mathbf{x}_j)} = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})^2}$$

- Atenção!
  - Atributos com maiores valores e variâncias podem "dominar" os demais atributos

Distância Euclidiana

- Dados contínuos
- Métrica (propriedades desejáveis)

$$d^{E}_{(\mathbf{x}_i,\mathbf{x}_j)} = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})^2}$$

Objeto	Atributo #1	Atributo #2	Atributo #3
$\mathbf{x}_1$	1	1,5	5055
$\mathbf{x}_2$	3	2,3	5943
$\mathbf{x}_3$	2	5,4	7100
$\mathbf{x}_4$	4	3,2	8590

- Atenção!
  - Atributos com maiores valores e variâncias podem "dominar" os demais atributos

Distância Euclidiana

- Dados contínuos
- Métrica (propriedades desejáveis)

$$d^{E}_{(\mathbf{x}_i,\mathbf{x}_j)} = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})^2}$$

Objeto	Atributo #1	Atributo #2	Atributo #3
$\mathbf{x}_1$	1	1,5	5055
$\mathbf{x}_2$	3	2,3	5943
$\mathbf{x}_3$	2	5,4	7100
$\mathbf{x}_4$	4	3,2	8590

- Atenção!
  - Atributos com maiores valores e variâncias podem "dominar" os demais atributos

Distância Euclidiana

- Dados contínuos
- Métrica (propriedades desejáveis)

$$d^{E}_{(\mathbf{x}_i,\mathbf{x}_j)} = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})^2}$$

Objeto	Atributo #1	Atributo #2	Atributo #3
x1	1	1,5	0,6
x2	3	2,3	0,7
x3	2	5,4	0,8
x4	4	3,2	1,0

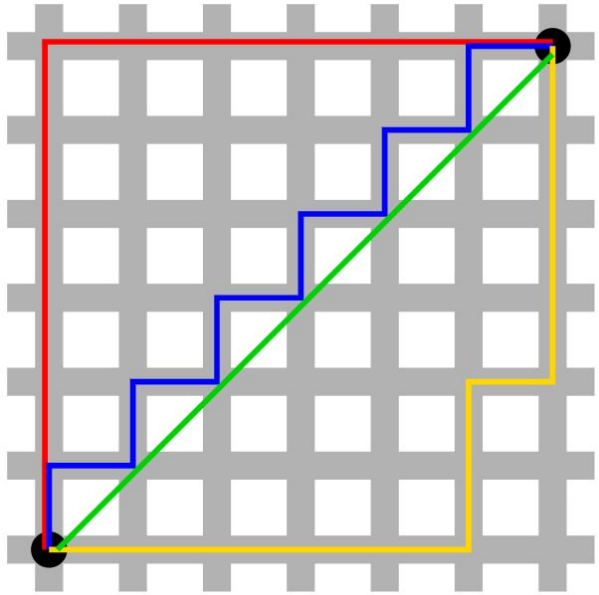
Normalizados

- Normalizar o atributo?
  - Assumimos que a importância do atributo é inversamente proporcional à variabilidade de seus valores!

Distância de Manhattan

- Dados contínuos
- Métrica (propriedades desejáveis)
- Também conhecida como *city block*

$$d^{M}_{(\mathbf{x}_i,\mathbf{x}_j)} = \sum_{k=1}^{d} |x_{ik} - x_{jk}|$$



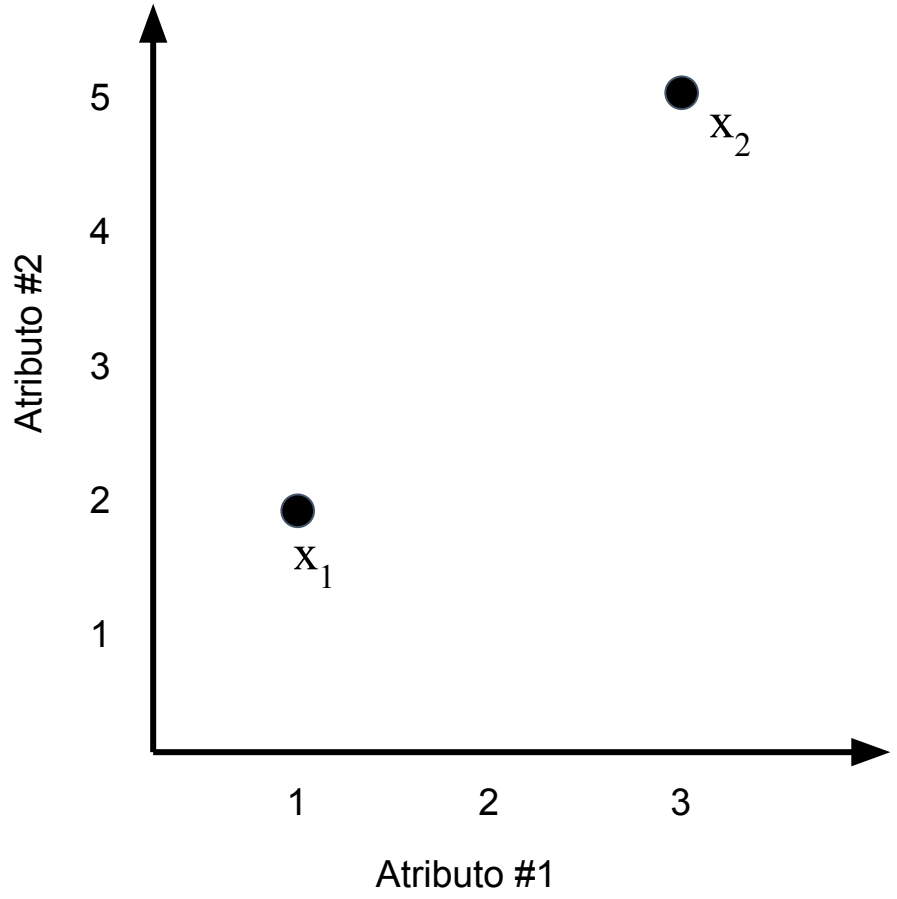
Distância Suprema

- Dados contínuos
- Métrica (propriedades desejáveis)
- Atributo com diferença máxima entre dois objetos

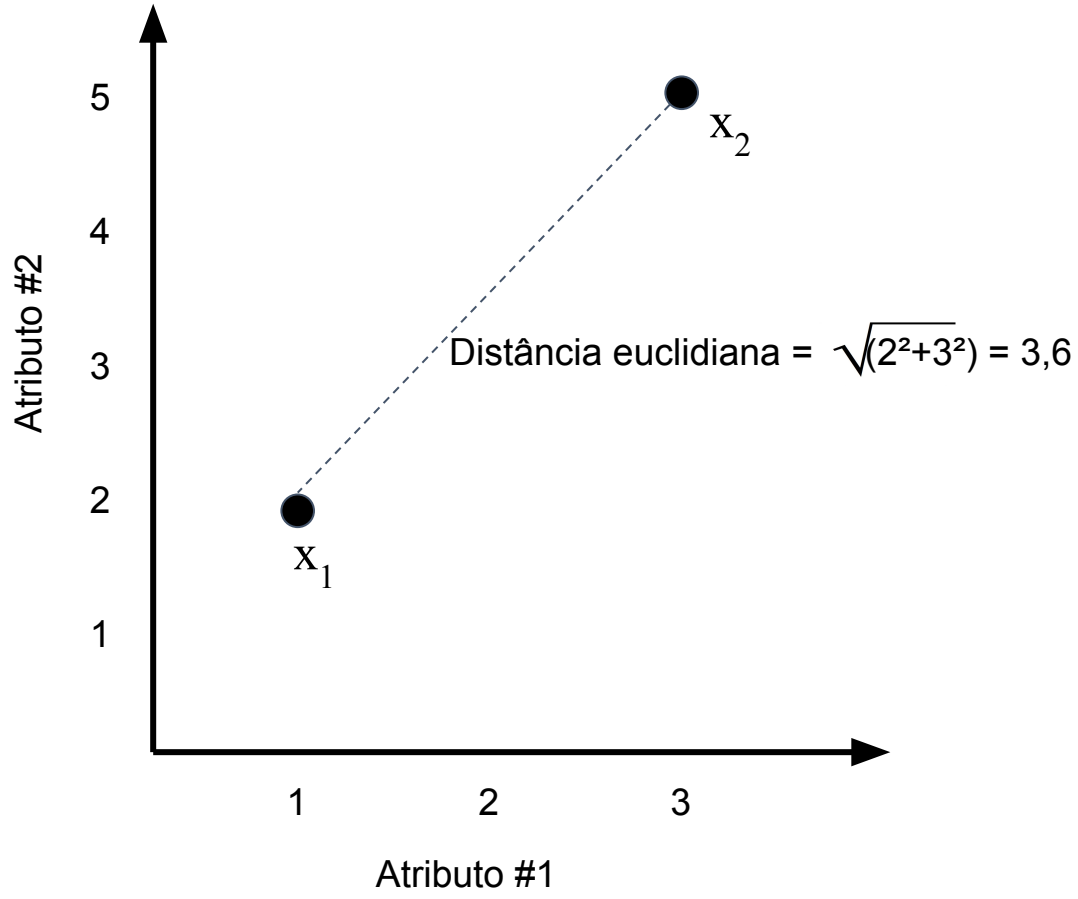
$$d^{S}_{(\mathbf{x}_i,\mathbf{x}_j)} = \max_{1 \leq k \leq d} |x_{ik} - x_{jk}|$$

Nota: Também conhecida como distância de Chebyshev

Exemplo

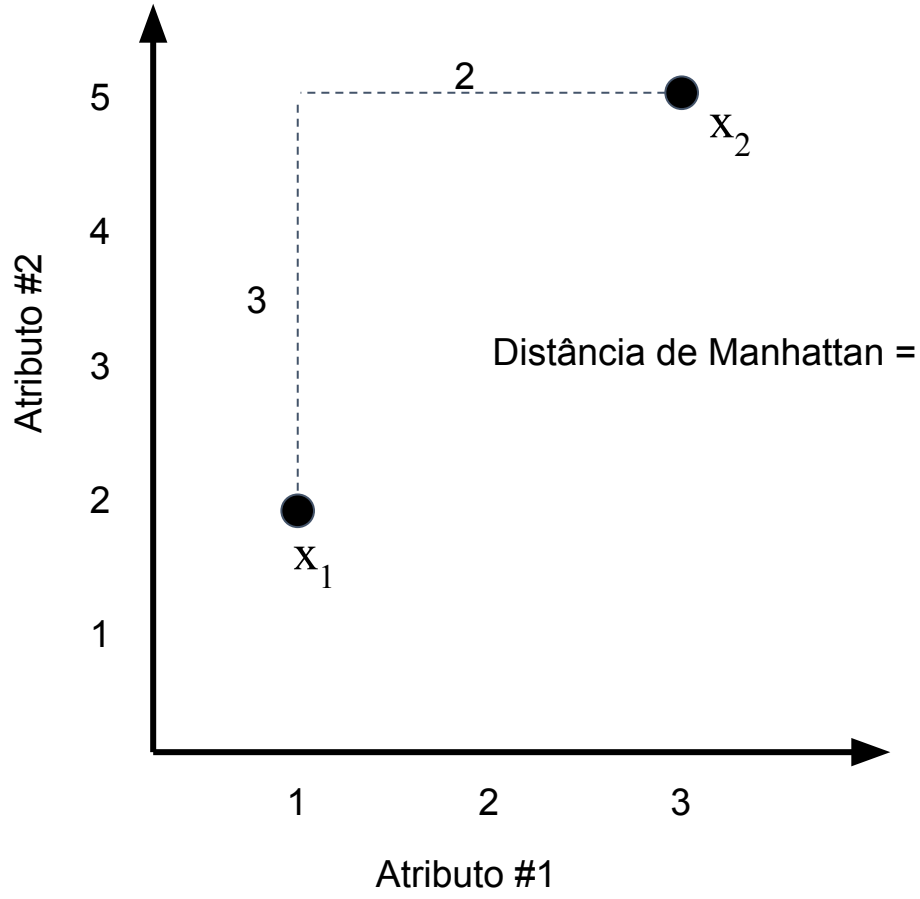


Exemplo



$$d^{E}_{(\mathbf{x}_i,\mathbf{x}_j)} = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})^2}$$

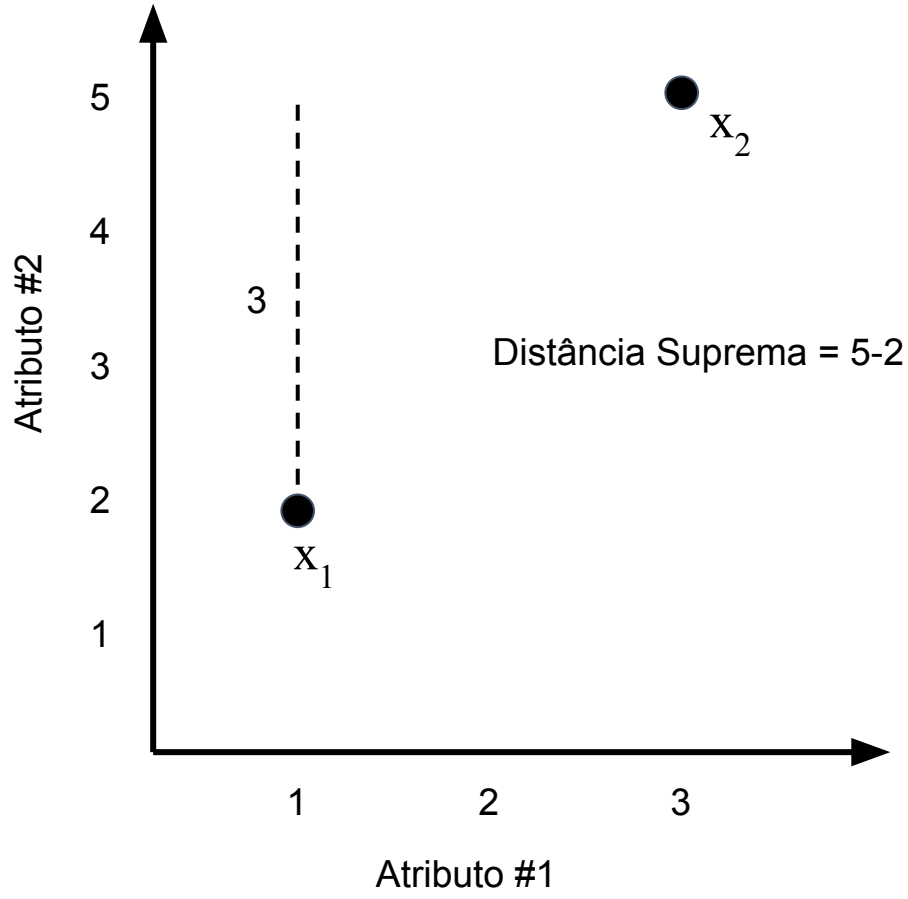
Exemplo



Distância de Manhattan = 2+3 = 5

$$d^{M}_{(\mathbf{x}_i,\mathbf{x}_j)} = \sum_{k=1}^{d} |x_{ik} - x_{jk}|$$

Exemplo



Distância Suprema = 5-2 = 3

$$d^{S}_{(\mathbf{x}_i,\mathbf{x}_j)} = \max_{1 \leq k \leq d} |x_{ik} - x_{jk}|$$

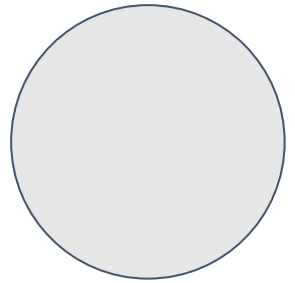
Distância de Minkowski

- Para $p=2$: Distância Euclidiana
- Para $p=1$: Distância de Manhattan
- Para $p \longrightarrow \infty$: Distância Suprema

$$d^{p}_{(\mathbf{x}_i,\mathbf{x}_j)} = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^{p} \right)^{\frac{1}{p}}$$

Distância de Minkowski

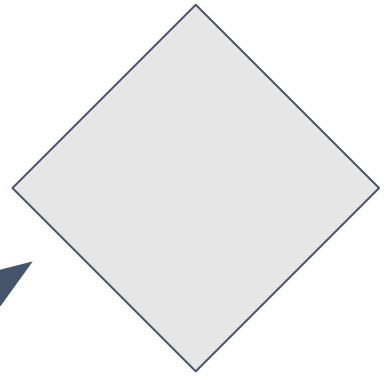
- Para $p=2$: Distância Euclidiana
- Para $p=1$: Distância de Manhattan
- Para $p \longrightarrow \infty$: Distância Suprema

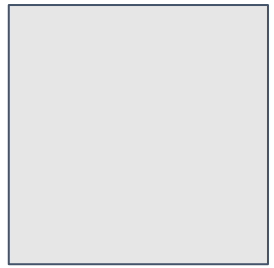


$$d^{p}_{(\mathbf{x}_i,\mathbf{x}_j)} = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^{p} \right)^{\frac{1}{p}}$$

Distância de Minkowski

- Para $p=2$: Distância Euclidiana
- Para $p=1$: Distância de Manhattan
- Para $p \longrightarrow \infty$: Distância Suprema





$$d^{p}_{(\mathbf{x}_i,\mathbf{x}_j)} = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^{p} \right)^{\frac{1}{p}}$$

Distância de Mahalanobis

- Considera o grau de interdependência entre atributos

$$d^{C}_{(\mathbf{x}_i,\mathbf{x}_j)} = (\mathbf{x}_i - \mathbf{x}_j)^{\top}\mathbf{C}^{-1}(\mathbf{x}_i - \mathbf{x}_j)$$

- Permite capturar *clusters* de formatos mais variados
- Alto custo computacional:
  - Cálculo da matriz de covariância
  - Cálculo da inversa da matriz de covariância

Nota: o assunto de covariância já ministrado pela Profa. Roseli.

Distância de Mahalanobis

- Considera o grau de interdependência entre atributos

$$d^{C}_{(\mathbf{x}_i,\mathbf{x}_j)} = (\mathbf{x}_i - \mathbf{x}_j)^{\top} \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

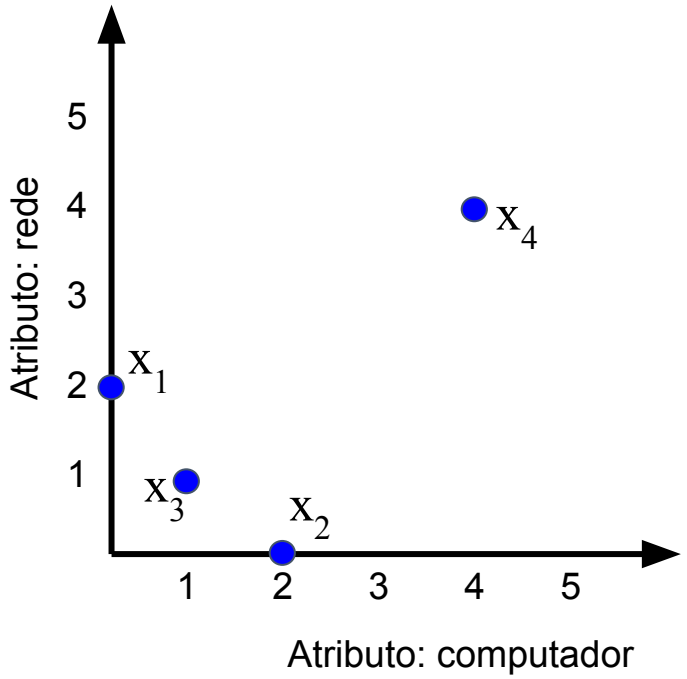
- Permite capturar *clusters* de formatos mais variados
- Alto custo computacional:
  - Cálculo da matriz de covariância
  - Cálculo da inversa da matriz de covariância

Nota: o assunto de covariância já foi ministrado pela Profa. Roseli.

Similaridade de Cosseno

- Medida de correlação
- Muito comum em dados de bioinformática e textos

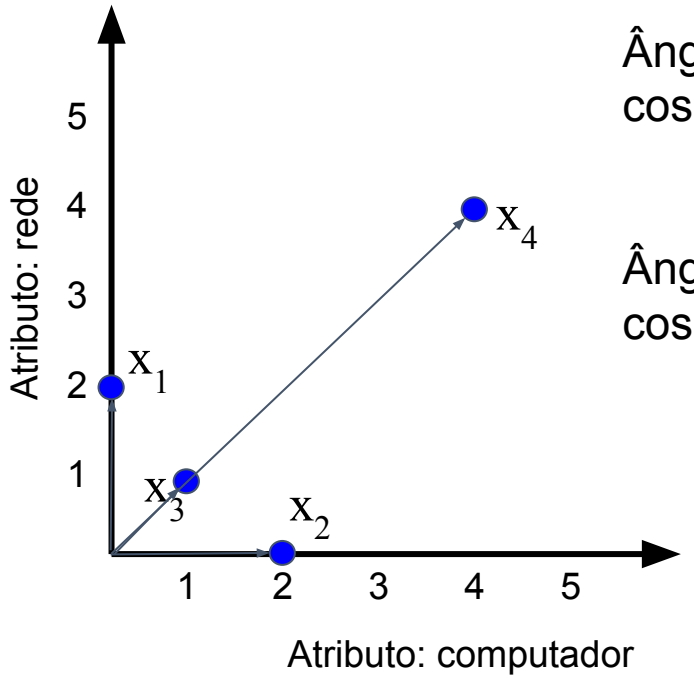
Objeto	Atributo: rede	Atributo: computador
x_1	2	0
x_2	0	2
x_3	1	1
x_4	4	4



Similaridade de Cosseno

- Medida de correlação
- Muito comum em dados de bioinformática e textos

Objeto	Atributo: rede	Atributo: computador
x_1	2	0
x_2	0	2
x_3	1	1
x_4	4	4



Ângulo entre x_3 e x_4 é 0°
cos(0°) = 1

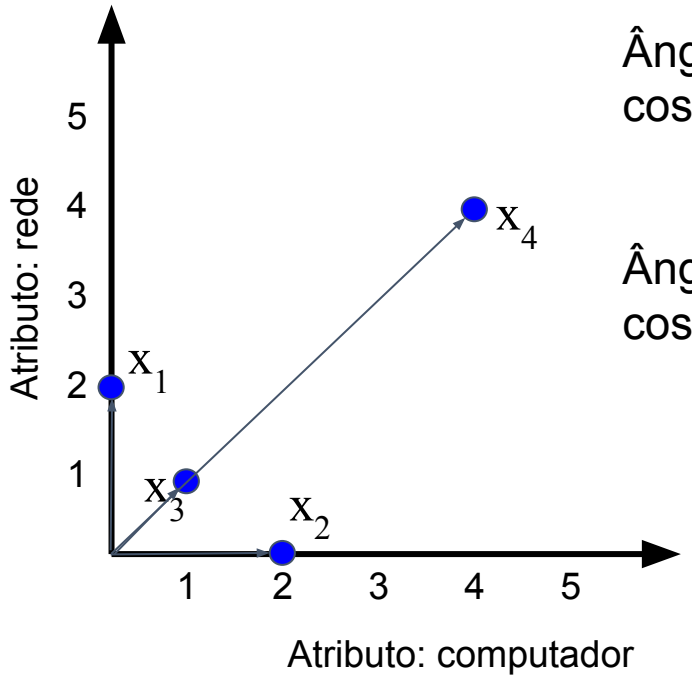
Ângulo entre x_1 e x_2 é 90°
cos(90°) = 0

Similaridade de Cosseno

- Medida de correlação
- Muito comum em dados de bioinformática e textos

Objeto	Atributo: rede	Atributo: computador
x_1	2	0
x_2	0	2
x_3	1	1
x_4	4	4

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



Ângulo entre x_3 e x_4 é 0°
cos(0°) = 1

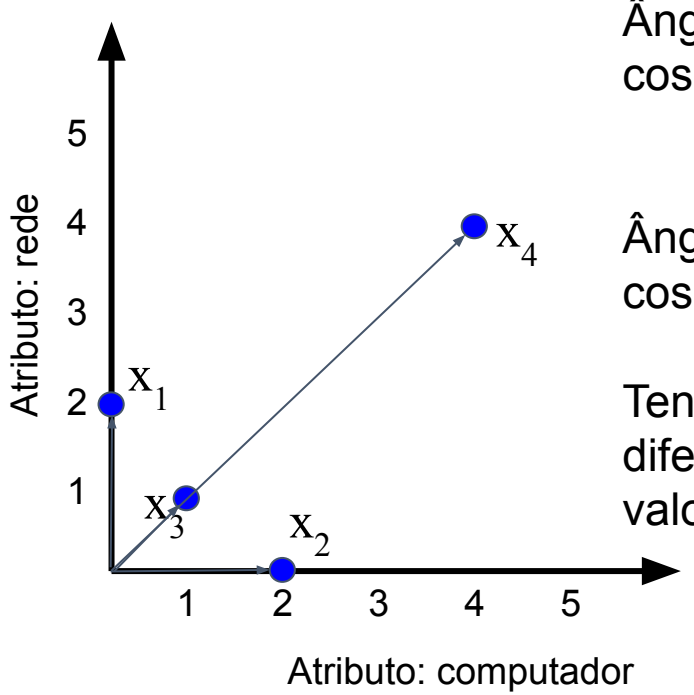
Ângulo entre x_1 e x_2 é 90°
cos(90°) = 0

Similaridade de Cosseno

- Medida de correlação
- Muito comum em dados de bioinformática e textos

Objeto	Atributo: rede	Atributo: computador
x_1	2	0
x_2	0	2
x_3	1	1
x_4	4	4

$$d^{cos}_{(\mathbf{x}_i,\mathbf{x}_j)} = 1 - cos_{(\mathbf{x}_i,\mathbf{x}_j)}$$



Ângulo entre x_3 e x_4 é 0°
cos(0°) = 1

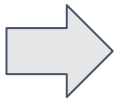
Ângulo entre x_1 e x_2 é 90°
cos(90°) = 0

Tende a não capturar diferenças de magnitude entre valores de atributos!

Medidas para Dados Binários

- Motivação: dados nominais podem ser transformados para binários (binarização)

Instância	País	Estado Civil
x_1	Brasil	Casado
x_2	Brasil	Solteiro
x_3	França	Solteiro
x_4	França	Casado



Instância	País:Brasil	País:França	Estado Civil: Casado	Estado Civil: Solteiro
x_1	1	0	1	0
x_2	1	0	0	1
x_3	0	1	0	1
x_4	0	1	1	0

Nota: o assunto sobre transformação de dados por binarização já foi ministrado pela Prof. Roseli.

Medidas para Dados Binários

- Considere dois objetos com atributos binários:
  - $x_i = [1\ 1\ 0\ 1\ 1\ 0]$
  - $x_j = [0\ 1\ 0\ 1\ 0\ 1]$
- Primeiro, vamos calcular uma tabela de contingência:

objeto x_i \ objeto x_j	1	0
1	n_{11}	n_{10}
0	n_{01}	n_{00}

$n_{11} = 2$

$x_i = [1\ \underline{1}\ 0\ \underline{1}\ 1\ 0]$

$x_j = [0\ \underline{1}\ 0\ \underline{1}\ 0\ 1]$

$n_{10} = 2$

$x_i = [\underline{1}\ 1\ 0\ 1\ \underline{1}\ 0]$

$x_j = [\underline{0}\ 1\ 0\ 1\ \underline{0}\ 1]$

Medidas para Dados Binários

- Considere dois objetos com atributos binários:
  - $x_i = [1\ 1\ 0\ 1\ 1\ 0]$
  - $x_j = [0\ 1\ 0\ 1\ 0\ 1]$
- Primeiro, vamos calcular uma tabela de contingência:

objeto x_i \ objeto x_j	1	0
1	n_{11}	n_{10}
0	n_{01}	n_{00}

$n_{01} = 1$

$x_i = [1\ 1\ 0\ 1\ 1\ \underline{0}]$

$x_j = [0\ 1\ 0\ 1\ 0\ \underline{1}]$

$n_{00} = 1$

$x_i = [1\ 1\ \underline{0}\ 1\ 1\ 0]$

$x_j = [0\ 1\ \underline{0}\ 1\ 0\ 1]$

Medidas para Dados Binários

- Considere dois objetos com atributos binários:
  - $x_i = [1\ 1\ 0\ 1\ 1\ 0]$
  - $x_j = [0\ 1\ 0\ 1\ 0\ 1]$
- Após calcular a tabela de contingência:

objeto x_i \ objeto x_j	1	0
1	n_{11}	n_{10}
0	n_{01}	n_{00}

- Coeficiente de Casamento Simples
(Similaridade)

$$sm(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11} + n_{00}}{n_{11} + n_{00} + n_{10} + n_{01}}$$

(Transformação para Dissimilaridade)

$$d^{sm}_{(\mathbf{x}_i, \mathbf{x}_j)} = 1 - sm(\mathbf{x}_i, \mathbf{x}_j)$$

Medidas para Dados Binários

- Coeficiente de Casamento Simples é uma medida para atributos simétricos: presença e ausência têm a mesma importância!
- Em alguns problemas, a presença é mais importante que a ausência. Exemplo: sintomas de doenças.
  - $x_1 = [1\ 1\ 0\ 0\ 0\ 0]$
  - $x_2 = [1\ 1\ 0\ 0\ 0\ 1]$
  - $x_3 = [0\ 0\ 0\ 0\ 0\ 0]$

Note que o Coeficiente de Casamento Simples irá falhar nesse exemplo, considerando 3 objetos que apresentam (1) ou não (0) seis sintomas para uma determinada doença.

Medidas para Dados Binários

- Coeficiente de Casamento Simples é uma medida para atributos simétricos: presença e ausência têm a mesma importância!
- Em alguns problemas, a presença é mais importante que a ausência. Exemplo: sintomas de doenças.
  - $x_1 = [1\ 1\ 0\ 0\ 0\ 0]$
  - $x_2 = [1\ 1\ 0\ 0\ 0\ 1]$
  - $x_3 = [0\ 0\ 0\ 0\ 0\ 0]$

Note que o Coeficiente de Casamento Simples irá falhar nesse exemplo, considerando 3 objetos que apresentam (1) ou não (0) seis sintomas para uma determinada doença.

Precisamos de uma medida para atributos assimétricos!

Medidas para Dados Binários

- Coeficiente de *Jaccard* (assimétrico)
- Foco nos casamentos do tipo 1-1
- Desconsidera casamentos do tipo 0-0

(Similaridade)

$$jac(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

(Transformação para Dissimilaridade)

$$d^{jac}_{(\mathbf{x}_i, \mathbf{x}_j)} = 1 - jac(\mathbf{x}_i, \mathbf{x}_j)$$

objeto x_i \ objeto x_j	1	0
1	n_{11}	n_{10}
0	n_{01}	n_{00}

Medidas de Proximidade

- Estudamos diferentes medidas de proximidade
- Como selecionar a medida mais apropriada?
  - Primeiro, verificar seu tipo de dados!
  - Mesmo assim, ainda são muitas opções de medidas!
  - Não há uma resposta definitiva para essa pergunta :(

Lembre-se! *Cluster* é um conceito subjetivo.
Difícil definir a noção de semelhança.

Medidas de Proximidade

- Estudamos diferentes medidas de proximidade
- Como selecionar a medida mais apropriada?
  - Primeiro, verificar seu tipo de dados!
  - Mesmo assim, ainda são muitas opções de medidas!
  - Não há uma resposta definitiva para essa pergunta :(

> *"A escolha da medida de dis(similaridade) é importante para aplicações, e a melhor escolha é freqüentemente obtida via uma combinação de experiência, habilidade, conhecimento e sorte!"*

Gan, G., Ma, C., Wu, J., Data Clustering: Theory, Algorithms, and Applications, SIAM Series on Statistics and Applied Probability, 2007.

Medidas de Proximidade

- Estudamos diferentes medidas de proximidade
- Como selecionar a medida mais apropriada?
  - Primeiro, verificar seu tipo de dados!
  - Mesmo assim, ainda são muitas opções de medidas!
  - Não há uma resposta definitiva para essa pergunta :(

Tarefa Descritiva

Análise Exploratória de Dados

Gan, G., Ma, C., Wu, J., Data Clustering: Theory, Algorithms, and Applications, SIAM Series on Statistics and Applied Probability, 2007.

Bibliografia

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.

Tan, P.N.; Steinbach, M.; Karpatne, A.; Kumar, V. (2016). *Introduction to Data Mining (2nd Edition)*. Pearson.