TÉCNICAS DE AGRUPAMENTO

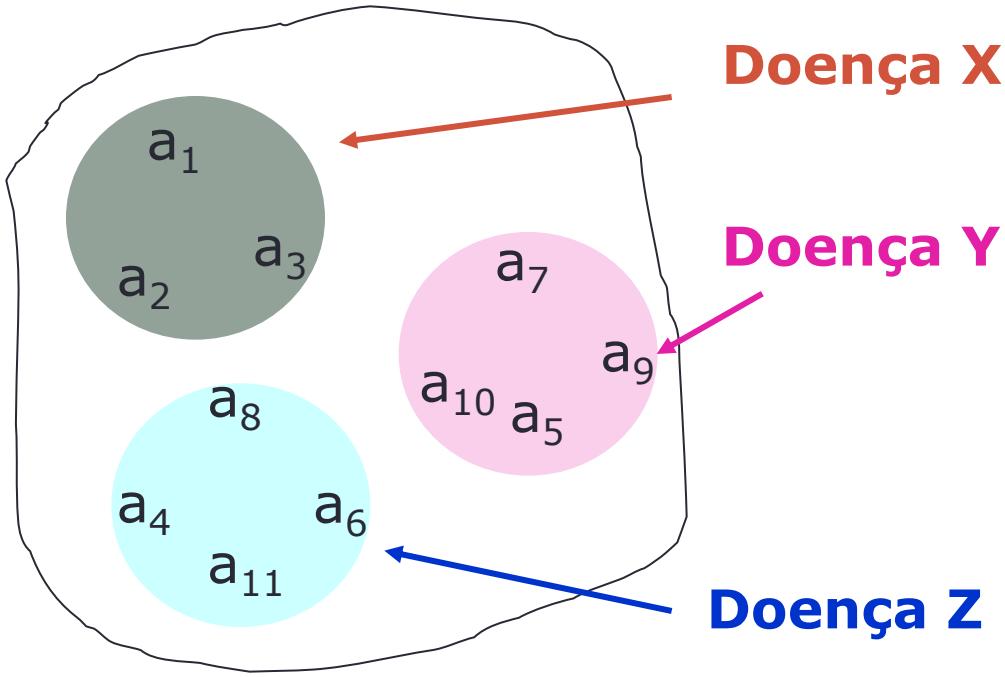
Cristiane Neri Nobre

Agrupamento - Análise de Clusters

Número de Clusters = 3

a_1	a	F	1	0	1	1
a_2	b	M	0	0	1	1
.	c	F	1	1	1	0
.						
.	d	F	1	0	0	0
	e	M	1	1	0	1

Nome Sexo Sintomas



Conceito = Doença

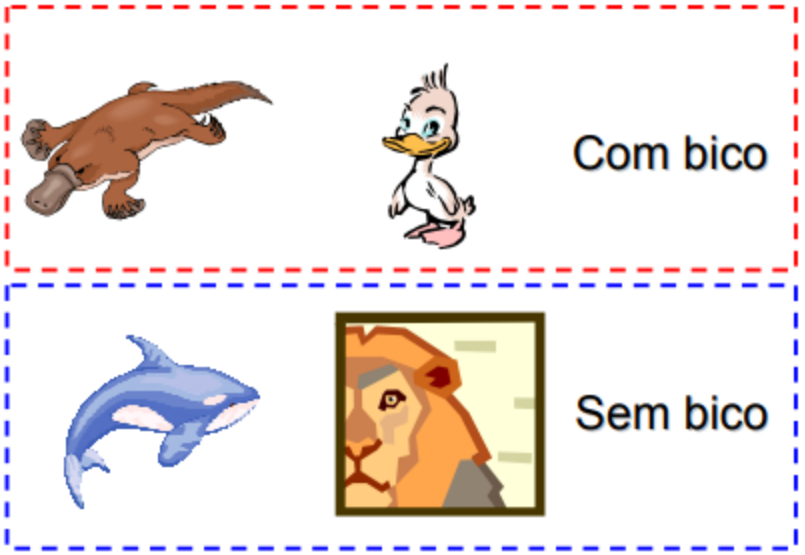
Agrupamento -Análise de Clusters

- Por exemplo, como agrupar estes animais*?

* Exemplo extraído de http://dcm.ffclrp.usp.br/~augusto/teaching/ami/AM-I-Clustering.pdf

Agrupamento -Análise de Clusters

- Por exemplo, como agrupar estes animais?

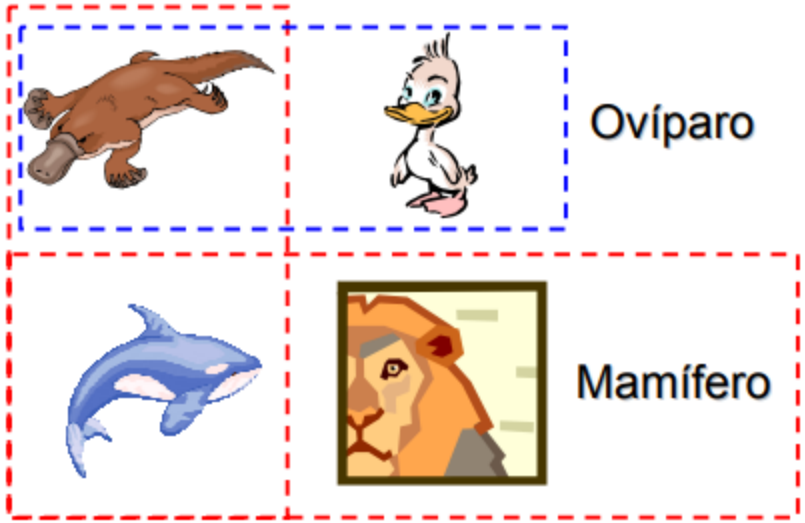


Agrupamento -Análise de Clusters

- Por exemplo, como agrupar estes animais?

Agrupamento -Análise de Clusters

- Por exemplo, como agrupar estes animais?



Agrupamento versus Classificação

- **Classificação**
  - Aprendizado **Supervisionado**
    - Amostras de treinamento são classificadas
    - Número de Classes é conhecido
  - Aprendizado por **Exemplo**

- **Agrupamento**
  - Aprendizado **Não Supervisionado**
  - Aprendizado por Observação

Agrupamento versus Classificação

- **Objetivo do agrupamento:**

Dado um conjunto de objetos descritos por múltiplos valores (atributos):

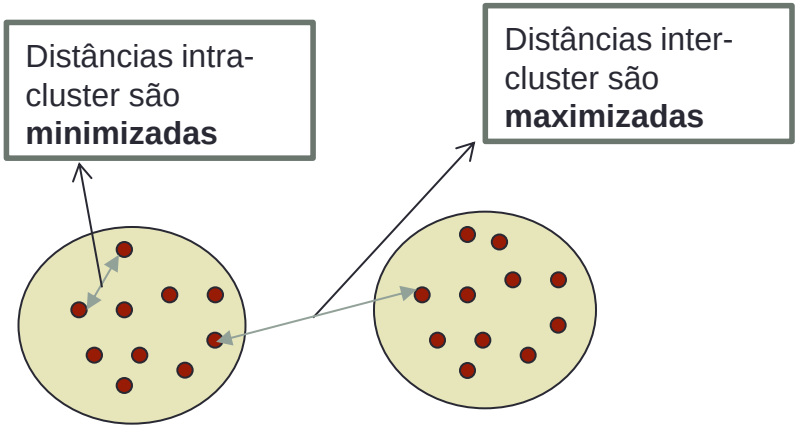
1) Atribuir grupos (clusters) aos objetos particionando-os objetivamente em grupos homogêneos de maneira a: ☐

a) Maximizar a similaridade de objetos dentro de um mesmo cluster ☐
b) Minimizar a similaridade de objetos entre clusters distintos ☐

2) atribuir uma descrição **(rótulo)** para cada cluster formado

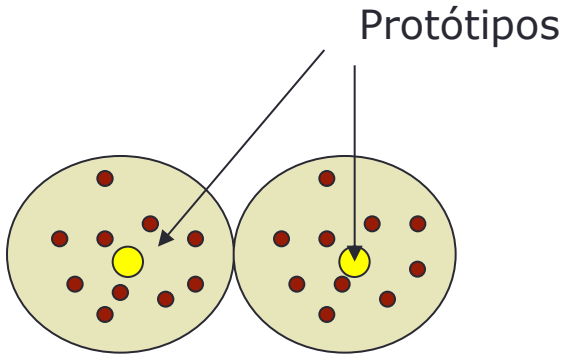
O que é um cluster ?

Como definir a noção de Cluster?



Bem separados

Um *cluster* é um conjunto de objetos no qual cada objeto está mais próximo (ou é mais similar) a objetos dentro do cluster do que qualquer objeto fora do cluster.



Baseados em Protótipos

Um *cluster* é um conjunto de objetos no qual cada objeto está mais próximo ao *protótipo que define o cluster* do que dos protótipos de quaisquer outros clusters.
Em geral: Protótipo = centróide

Como avaliar se as instâncias estão no mesmo grupo?

Pela distância!

Manhattan $$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_p - y_p|$$

Minkowski $$d(x,y) = \sqrt[m]{(x_1 - y_1)^m + (x_2 - y_2)^m + \ldots + (x_p - y_p)^m}$$

Distância Euclidiana $$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_p - y_p)^2}$$

Investigue outras distâncias!

Exercício

- Sejam X1 = (1,2) e X2 = (4,6).

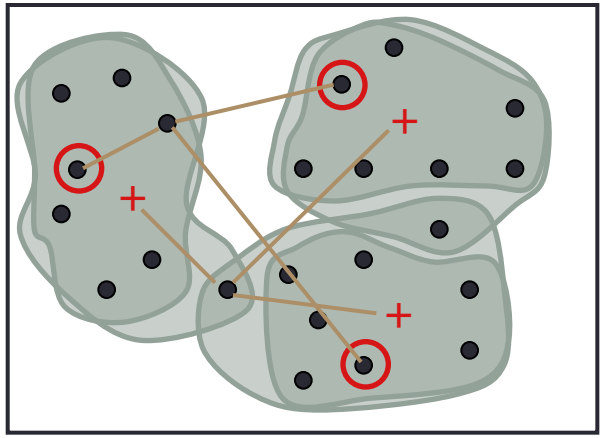
  Calcule as distâncias euclidianas, Minkowski com *m* = 3 e Manhattan entre X1 e X2.

- Ilustre no plano xy os segmentos representando tais distâncias.

Algoritmo K-means

K-means (MacQueen, 1967*) é um algoritmo de agrupamento

Exemplo K = 3



1ª Iteração

*J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297

Algoritmo K-Means

1. Selecione *k* pontos como centroides iniciais

Repetir

1. Forme *k* clusters associando cada objeto a seu centroide mais próximo
2. Recalcule o centroide de cada cluster

Até Centroides não apresentarem mudanças

Centroide = centro de gravidade do cluster

Coordenada *i* = média aritmética das coordenadas *i* de seus objetos constituintes.

Observações

- Boa parte dos clusters já convergem nos primeiros passos do algoritmo, ficando somente uma quantidade pequena de clusters que ainda modificam.

- Assim, a condição de parada do algoritmo é substituída por "*até que somente 1% dos objetos mudam de clusters*".

- Complexidade do K-means é O(n*K*I*d)
  - N = número de instâncias
  - K = Número de clusters
  - I = Número de iterações
  - d = Número de atributos

Exercício 1

Visualizar a base de dados "**Iris**" no Weka utilizando o algoritmo **Kmeans**

Deixar com que o Kmeans considere a classificação (setosa, versicolor, virgínica), e verifique os erros dos agrupamentos.

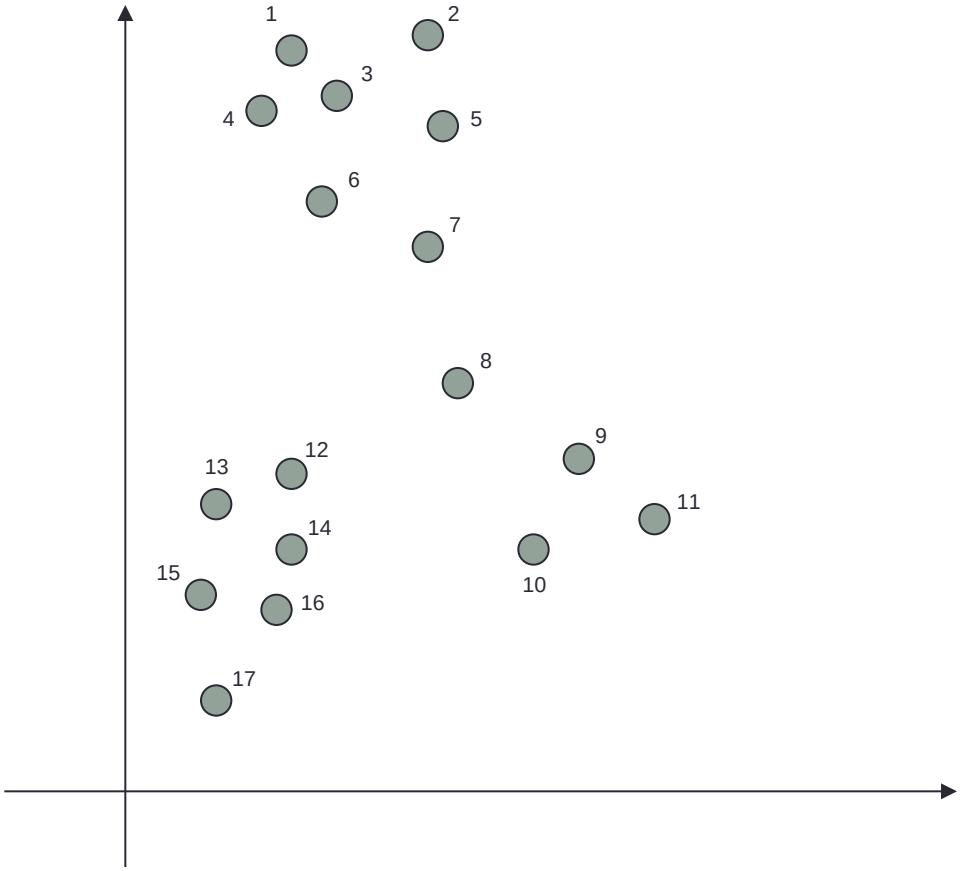
Anote os valores de acertos e erros, discuta e visualize as instâncias visualmente

Observar os resultados fornecidos pela ferramenta.

- Quem são os centróides iniciais?
- Quais são os centróides finais?

Exercício 2

Achar 3 clusters utilizando o k-means



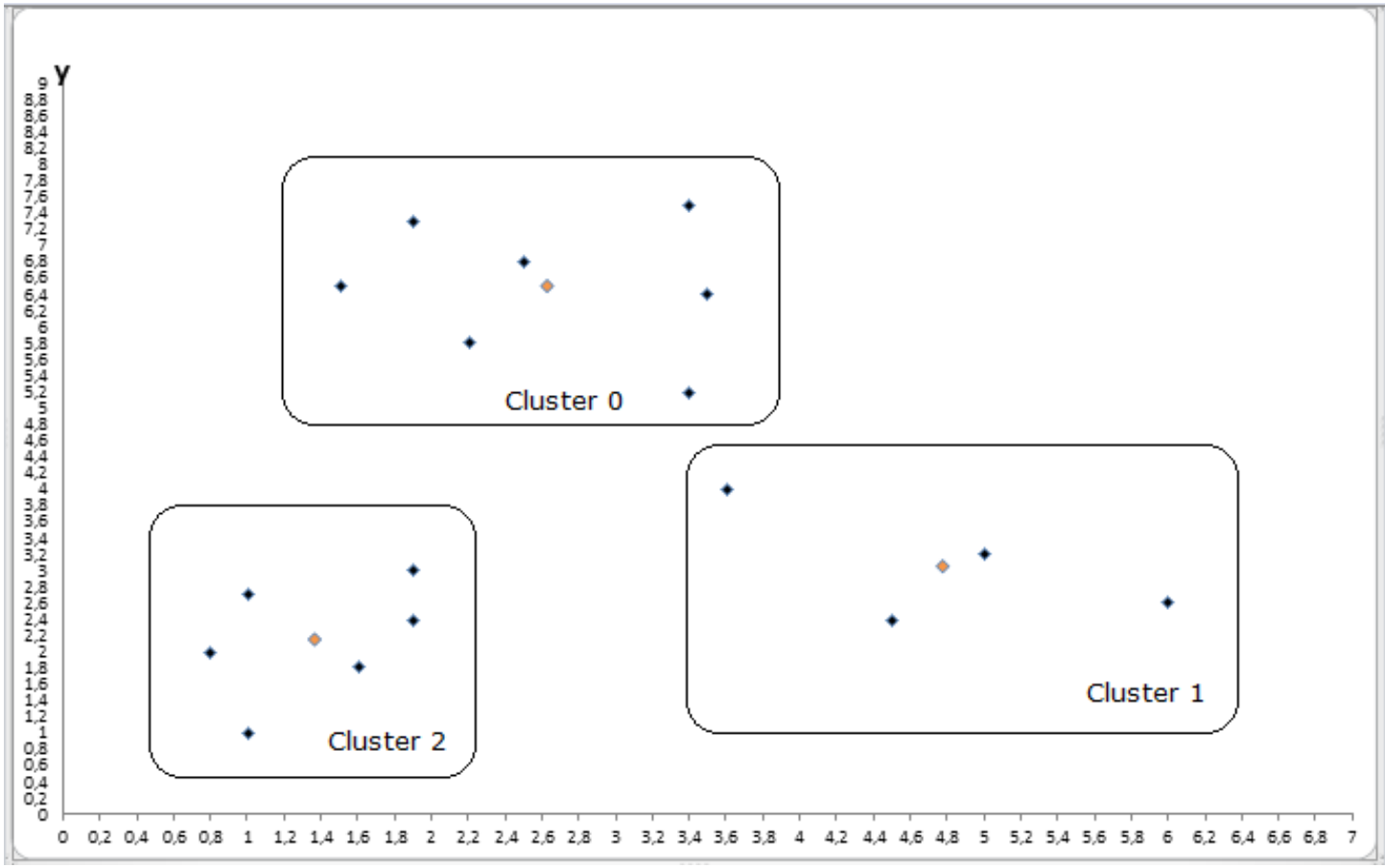
1	1,9	7,3
2	3,4	7,5
3	2.5	6,8
4	1,5	6,5
5	3,5	6,4
6	2,2	5,8
7	3,4	5,2
8	3,6	4
9	5	3,2
10	4,5	2,4
11	6	2,6
12	1.9	3
13	1	2,7
14	1.9	2,4
15	0,8	2
16	1,6	1,8
17	1	1

Exercício 2

Tente visualizar estes pontos no excel, após o agrupamento. Marque os centróides.

Saída do WEKA

1.9,7.3,cluster0
3.4,7.5,cluster0
2.5,6.8,cluster0
1.5,6.5,cluster0
3.5,6.4,cluster0
2.2,5.8,cluster0
3.4,5.2,cluster0
3.6,4,cluster1
5,3.2,cluster1
4.5,2.4,cluster1
6,2.6,cluster1
1.9,3,cluster2
1,2.7,cluster2
1.9,2.4,cluster2
0.8,2,cluster2
1.6,1.8,cluster2
1,1,cluster2



Exercício 3

Veja o vídeo:

https://www.youtube.com/watch?v=E2M_yTulcmU

E analise as limitações deste algoritmo

Métricas de avaliação

Silhouette Index

O Silhouette Index é uma medida de avaliação que avalia a coesão e a separação dos clusters, e baseia-se na diferença entre a distância média dos pontos pertencentes ao cluster mais próximo para os pontos de um grupo.

Métricas de avaliação

Silhouette Index

Para cada ponto da base de dados, x_i, é calculado o valor do silhouette index, S_i, de acordo com a Equação:

$$S_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)}$$

onde $\mu_{in}(x_i)$ é a distância média de x_i para os pontos do seu próprio cluster, e

$\mu_{out}^{min}(x_i)$ é a distância média de x_i para os pontos dos clusters mais próximo.

Métricas de avaliação

Silhouette Index

- O valor S_i de um ponto encontra-se no intervalo $[-1, 1]$. Um valor próximo a 1 indica que x_i está mais próximo dos pontos do seu próprio cluster e distante dos clusters vizinhos.

- Um valor próximo de 0 indica que x_i está próximo da fronteira de dois clusters.

- Finalmente, um valor próximo a -1 indica que x_i está próximo dos pontos que pertencem a outro cluster.

Métricas de avaliação

Silhouette Index

O sillhouette index é definido como o valor médio S_i entre todos os pontos, dado pela Equação

$$SilhouetteIndex = \frac{1}{n}\sum_{i=e}^{n} S_i$$

Métricas de avaliação

Silhouette Index

Segundo Rousseeuw (1987), para cada grupo é apresentado um valor do silhouette index, baseado na comparação da sua coesão (análise intra-cluster) e separação (com os demais grupos).

Este índice mostra quais objetos se encontram adequadamente agrupados no cluster e quais estão localizados indevidamente no cluster.

Todo o agrupamento é exibido pela combinação dos silhouettes em um único valor, permitindo uma validação da qualidade relativa dos clusters e uma visão geral da distribuição dos dados.

Métricas de avaliação

Silhouette Index

Tabela 1 – Análise do *Silhouette Index*

Valor	Significado
0.71 - 1.0	Uma estrutura forte foi encontrada
0.51 - 0.70	Uma estrutura razoável foi encontrada
0.26 - 0.50	A estrutura é fraca e pode ser artificial
< 0.25	Nenhuma estrutura substancial foi encontrada

Fonte: Adaptado de Rousseeuw (1987)

Métricas de avaliação

Investigue outras métricas!

Ex: Davies-Bouldin Index

Veja:
https://www.youtube.com/watch?v=438C3vGxYTE

Verifique também outros métodos de agrupamento:

DBSCAN

https://www.youtube.com/watch?v=Xw_ig_yzHyU

https://www.youtube.com/watch?v=s5nn4xxA2XA

Veja esta série de vídeos:

https://www.youtube.com/watch?v=7-vwS8Jh6eo

Questões importantes para algoritmos de agrupamento

1) Os atributos de entrada podem ser numéricos ou nominais

O método mais simples para atributos categóricos é o seguinte:

$$overlap(x_{i,r}, x_{j,r}) = \begin{cases} 1 & \text{se } x_{i,r} \text{ ou } x_{j,r} \text{ são desconhecidos} \\ 1 & \text{se } x_{i,r} \neq x_{j,r} \\ 0 & \text{se } x_{i,r} = x_{j,r} \end{cases}$$

$$dist_{\text{Cat}}(\mathbf{x_i}, \mathbf{x_j}) = \sum_{r=1}^{m} overlap(x_{i,r}, x_{j,r})$$

Questões importantes para algoritmos de agrupamento

2) Quando os atributos são numéricos, é importante normalizar os valores de entrada

a) Por exemplo, se uma aplicação tem apenas dois atributos A e B e A varia entre 1 e 1000 e B entre 1 e 10, então a influência de B na função de distância será sobrepujada pela influência de A.

b) Portanto, as distâncias são frequentemente normalizadas dividindo-se a distância de cada atributo pelo intervalo de variação (i.e. diferença entre valores máximo e mínimo) daquele atributo

c) Assim, a distância para cada atributo é normalizada para o intervalo [0,1]

De forma a evitar ruídos, é também comum: ☐

a) Dividir pelo desvio-padrão ao invés do intervalo ou ☐

b) "cortar" o intervalo por meio da remoção de uma pequena porcentagem (e.g. 5%) dos maiores e menores valores daquele atributo e somente então definir o intervalo com os dados remanescentes

Questões importantes para algoritmos de agrupamento

3) Conhecimento do domínio pode frequentemente ser utilizada para decidir qual método é mais apropriado

Limitações do K-means

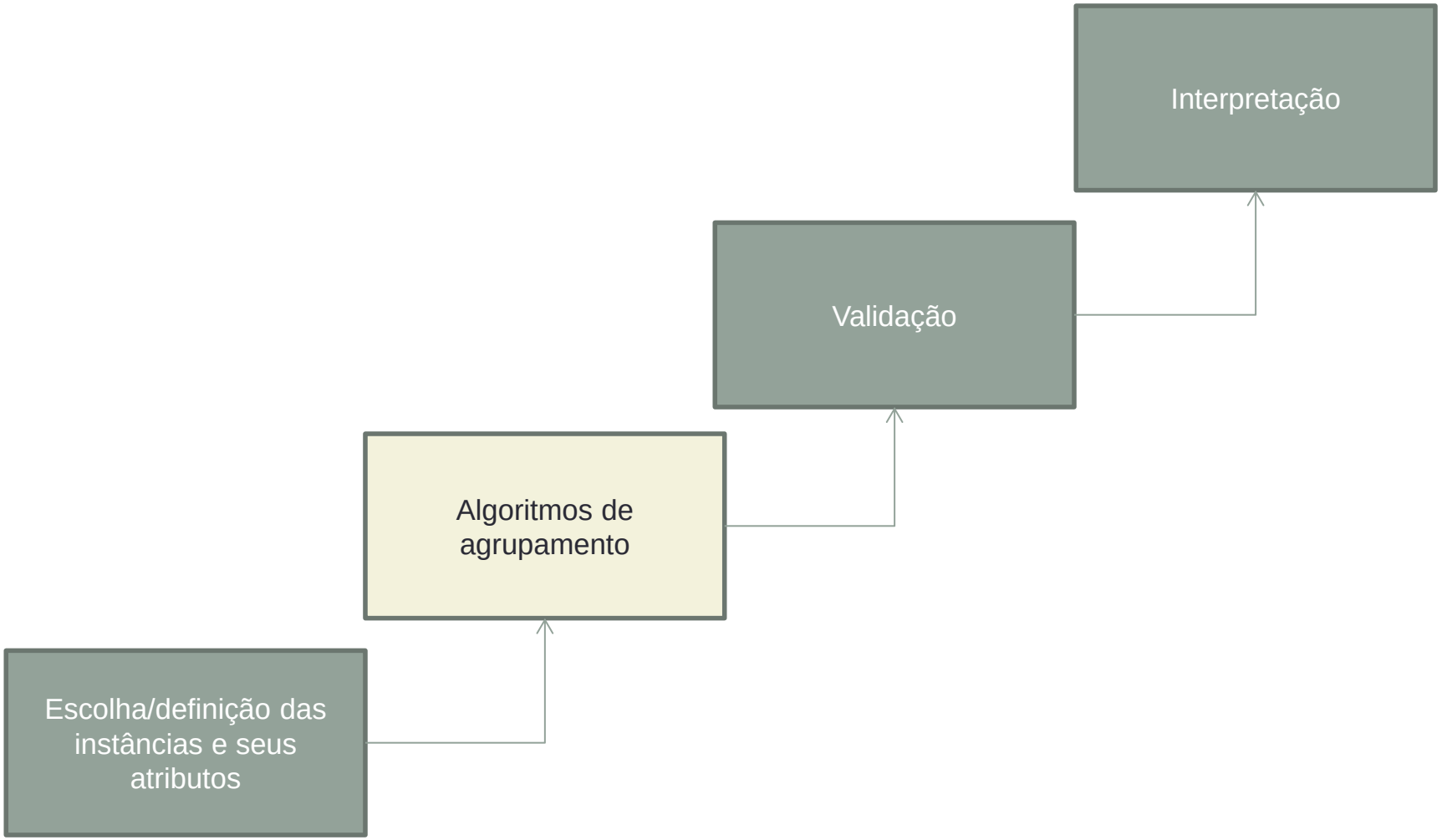
1) É necessário especificar o número de clusters. No entanto, nem sempre sabemos quantos temos

2) Os resultados podem alterar bastante, dependendo da localização do centróides iniciais

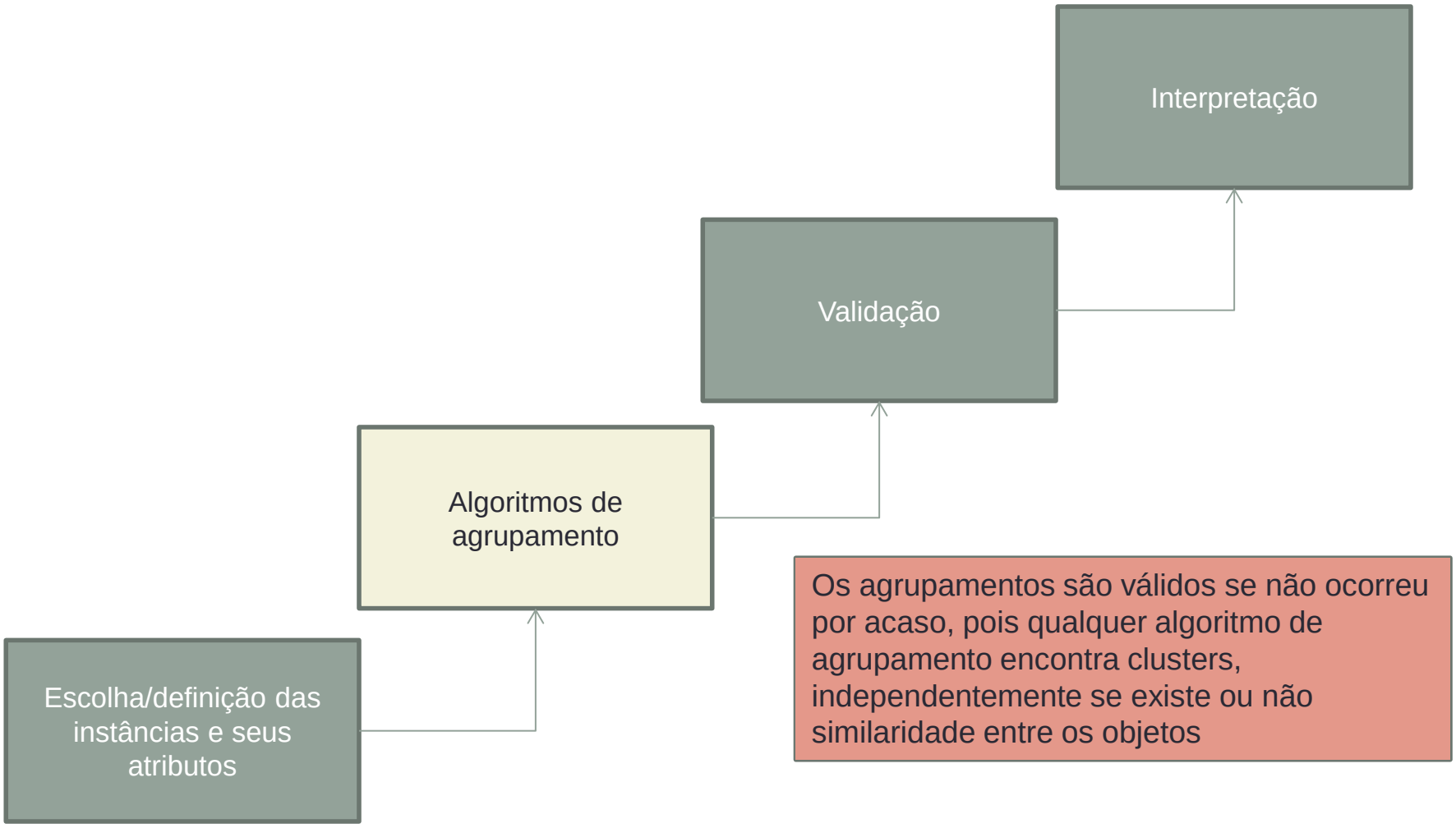
3) A análise usando Kmeans não é recomendada se você tiver muitas variáveis categóricas

4) K-means assume que os grupos são esféricos, distintos, e aproximadamente iguais em tamanho

Resumo da aplicação de algoritmos de agrupamento:



Resumo da aplicação de algoritmos de agrupamento:



Mais informações

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

http://www.rob.cs.tu-bs.de/content/04-teaching/06-interactive/Kmeans/Kmeans.html

Slides baseados nos slides:

www.deamo.prof.ufu.br/arquivos/AnalisedeClusters.ppt

ROUSSEEUW, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math., Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 20, n. 1, p. 53–65, nov. 1987. ISSN 0377-0427. Disponível em: <http://dx.doi.org/10.1016/0377-0427(87)90125-7>.

ZAKI, M. J.; MEIRA, W. Data Mining and Analysis: Fundamental Concepts and Algorithms. New York, NY, USA: Cambridge University Press, 2014. ISBN 0521766338, 9780521766333.

Artigos para leitura

https://www.researchgate.net/publication/298082409_A_Survey_on_Clustering_Techniques_for_Big_Data_Mining

http://www.ijsret.org/pdf/121888.pdf