

CONSIDERAÇÕES IMPORTANTES SOBRE AGRUPAMENTO

Cristiane Neri Nobre

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

Vamos tentar entender esta fórmula?

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

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

Para isso vamos considerar a seguinte pergunta:

A palavra CANA se parece mais com LANA ou CANOA?

$$D\left(\frac{\text{CANA}}{\text{LANA}}\right) = 1 + 0 + 0 + 0 = 1$$

$$D\left(\frac{\text{CANOA}}{\text{CANOA}}\right) = 0 + 0 + 0 + 1 + 1 = 2$$

Questões importantes para algoritmos de agrupamento


Vamos tentar entender esta fórmula?

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

$$\text{dist}_{\text{Cat}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^m \text{overlap}(x_{i,r}, x_{j,r})$$

Para isso vamos considerar a seguinte pergunta:

A palavra CANA se parece mais com LANA ou CANOA?

mais parecidos 

$$D\left(\frac{\text{CANA}}{\text{LANA}}\right) = 1 + 0 + 0 + 0 = 1$$

$$D\left(\frac{\text{CANA}}{\text{CANOA}}\right) = 0 + 0 + 0 + 1 + 1 = 2$$

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 $[0,1]$

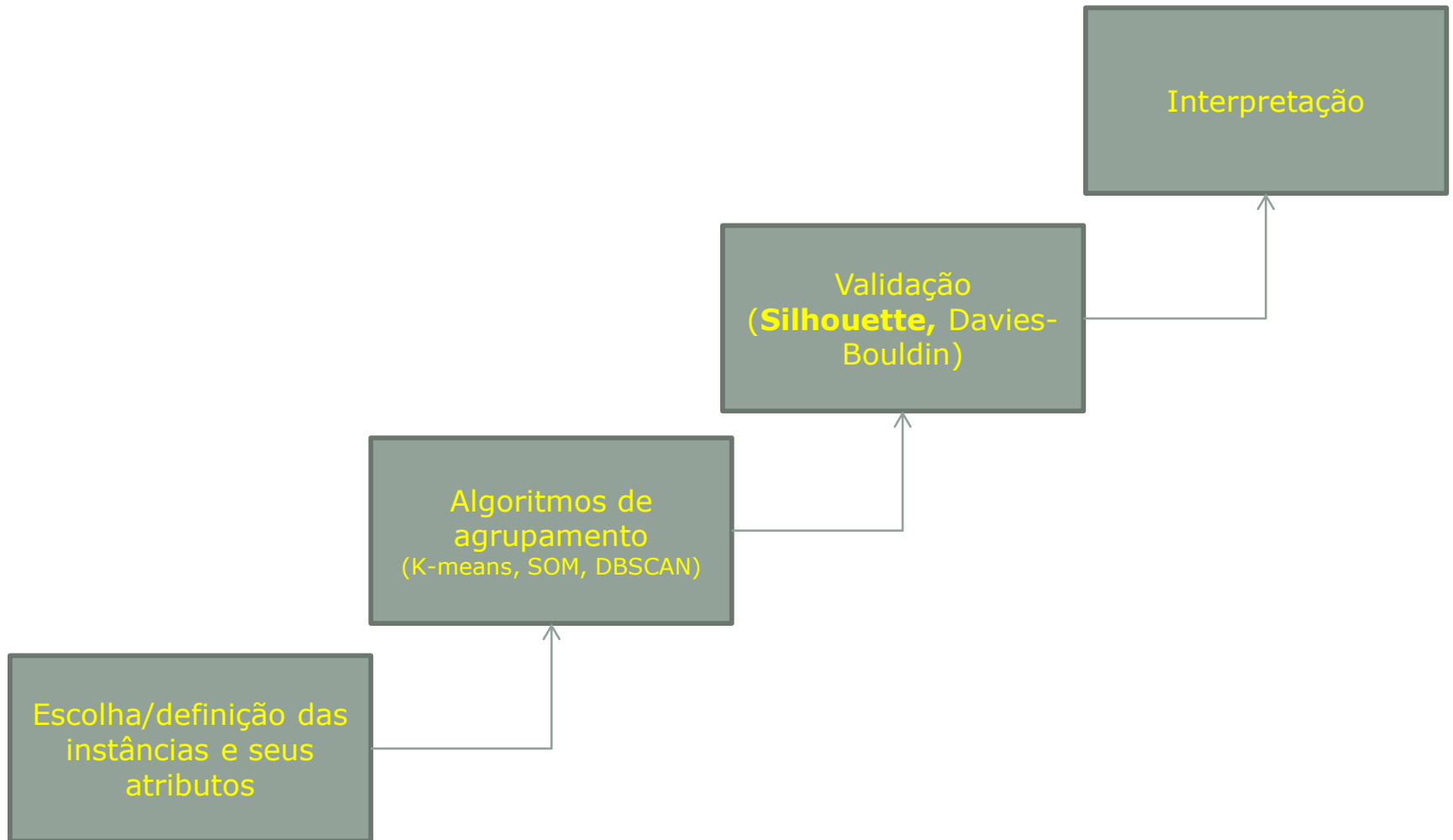
De forma a evitar ruídos, é também comum:

1. Dividir pelo desvio-padrão ao invés do intervalo ou
2. 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

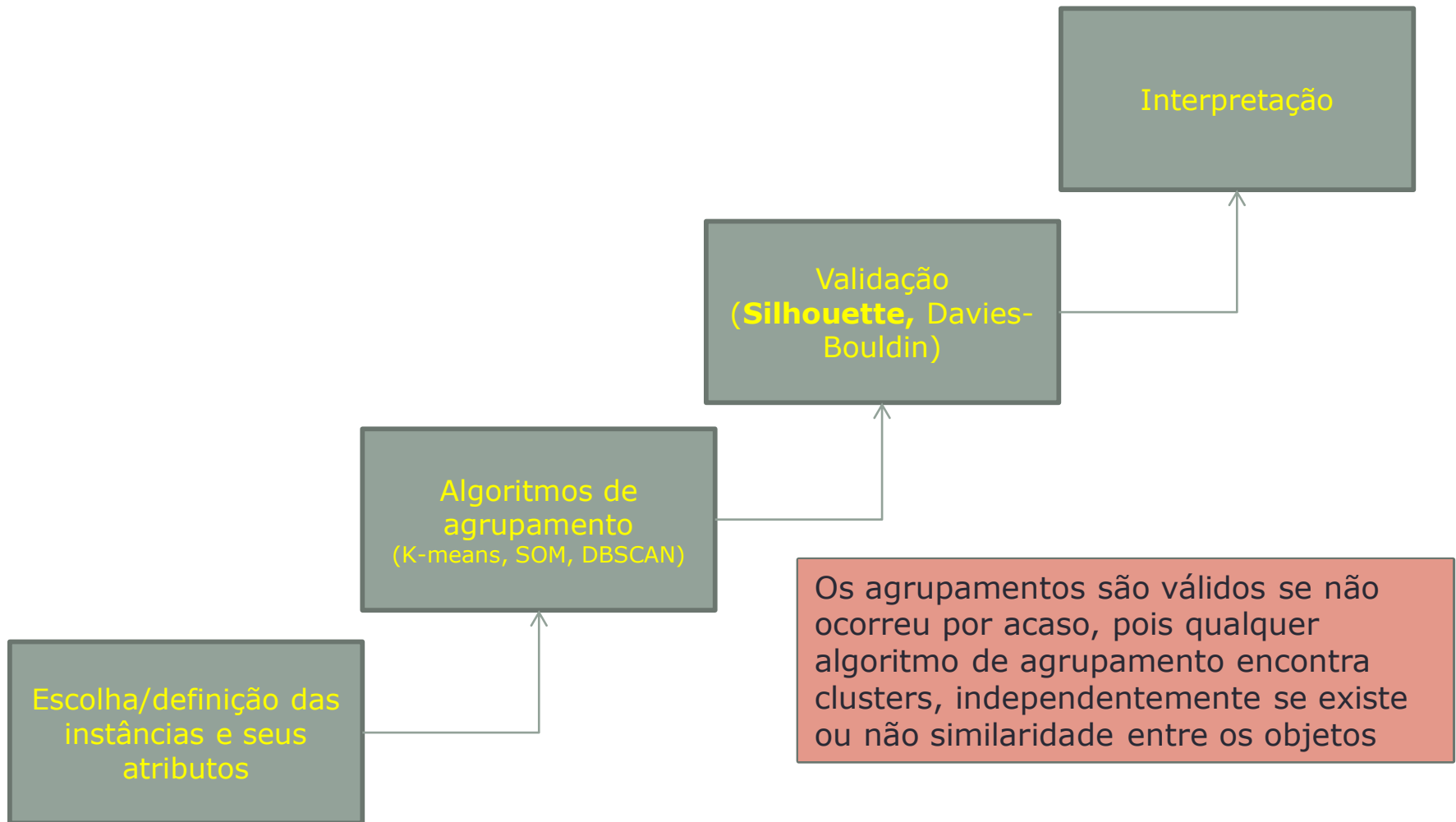
Limitações do K-means

- 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 centroides 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:



Referências:

- Capítulo 11 do livro
- Katti Faceli et al.
Inteligência Artificial, Uma
abordagem de Aprendizado
de Máquina, LTC, 2015.



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 links/textos:

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](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.ijret.org/pdf/121888.pdf>