Aula 12 – Validação de Agrupamentos

1001524 – Aprendizado de Máquina I
2023/1 - Turmas A, B e C
Prof. Dr. Murilo Naldi

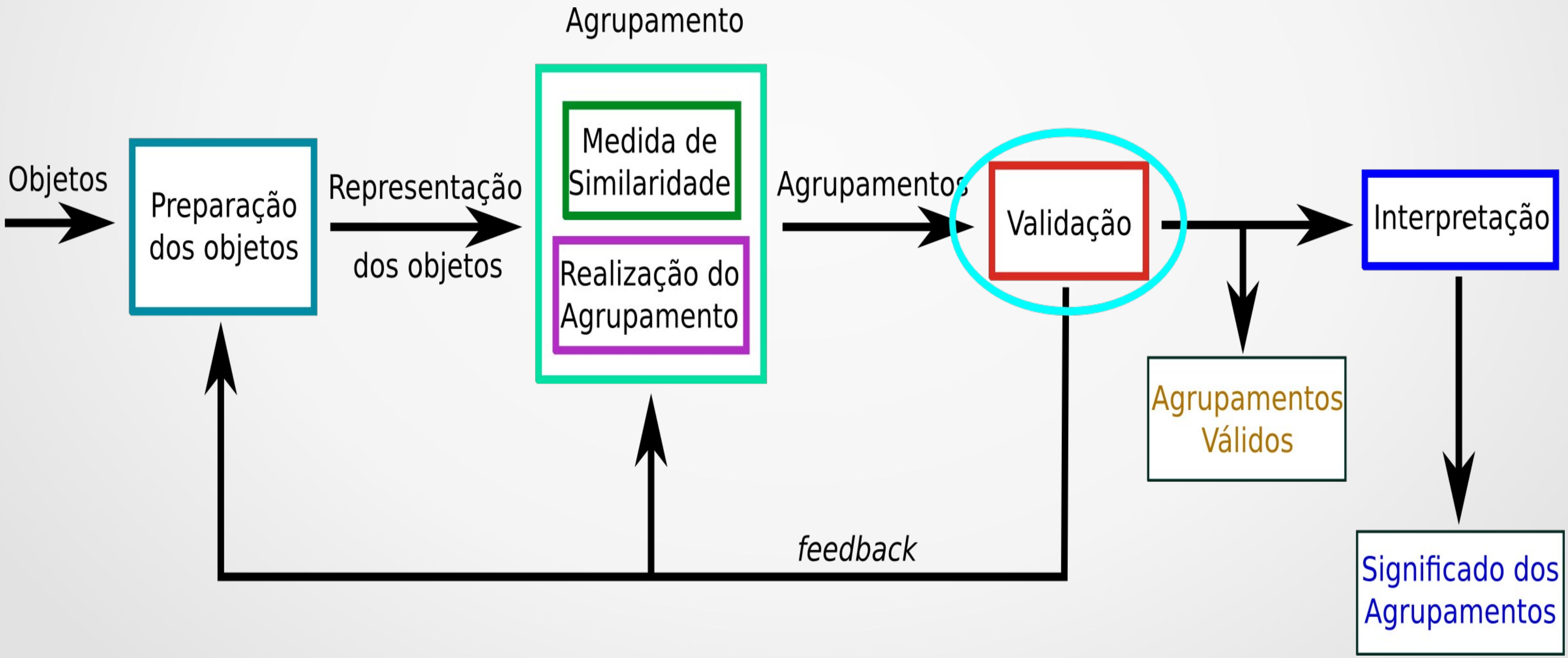
naldi@ufscar.br

Agradecimentos

- Parte do material utilizado nesta aula foi cedido pelos professores Ricardo Campello, Diego Silva e André Carvalho e, por esse motivo, o crédito deste material é deles
- Parte do material utilizado nesta aula foi disponibilizado por M. Kumar no endereço: www-users.cs.umn.edu/~kumar/dmbook/index.php

Validação

- Etapa que faz referência a avaliação da qualidade dos agrupamentos obtidos



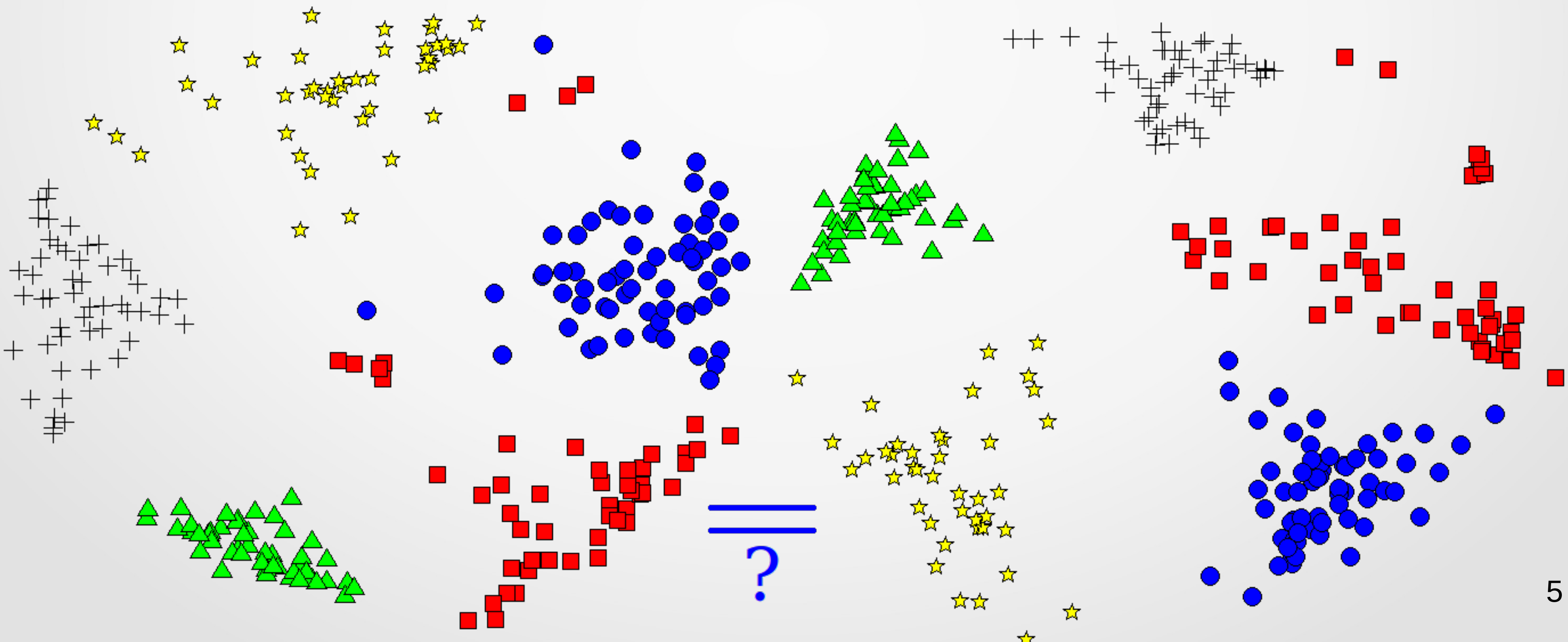
Índice de validação

- De maneira quantitativa, o procedimento de validação é feito por meio de um índice

- Tipos de índice ou critério:
  - Externos
  - Internos
  - Relativos

Índice externo

- Avalia o grau de correspondência entre a estrutura de grupos (partição ou hierarquia) sob avaliação e informação *a priori* na forma de uma solução de agrupamento esperada ou conhecida



Índice externo

- O índice externo é utilizado para verificar a qualidade de uma agrupamento de acordo com uma referência
  - Ou simplesmente para comparar dois agrupamentos
- O nome externo indica que informação externa aos dados é utilizada

Índice externo

- Alguns índices destes tipo são:
  - Rand Index
  - Jaccard
  - Rand Index Ajustado
  - Fowlkes-Mallows
  - Estatística G
  - Normalized Mutual Information

Baseados em pares

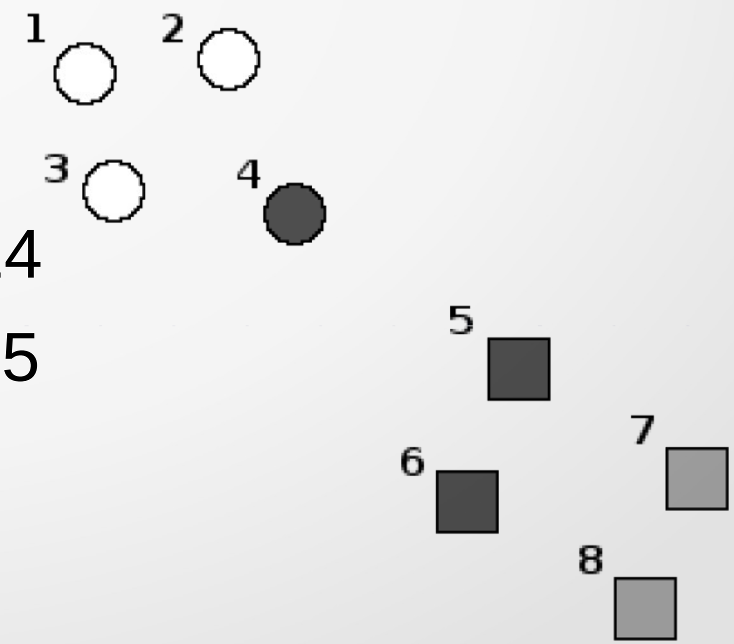
- *MM*: No. de pares que pertencem ao mesmo grupo em ambas partições
- *MD*: No. de pares que pertencem ao mesmo grupo apenas na partição 1
- *DM*: No. de pares que pertencem ao mesmo grupo apenas na partição 2
- *DD*: No. de pares que não pertencem ao mesmo grupo em ambas partições

Rand Index

- Dado por:

$$Rand(\pi_1, \pi_2) = \frac{MM + DD}{MD + DM + MM + DD}$$

- 2 Grupos
  - (Círculos e Quadrados)
- 3 Grupos
  - (Preto, Branco e Cinza)
- *MM* = 5; *DM* = 2; *MD* = 7; *DD* = 14
- *Rand*= 5+14/(5+7+2+14) = 0.6785



Limitações

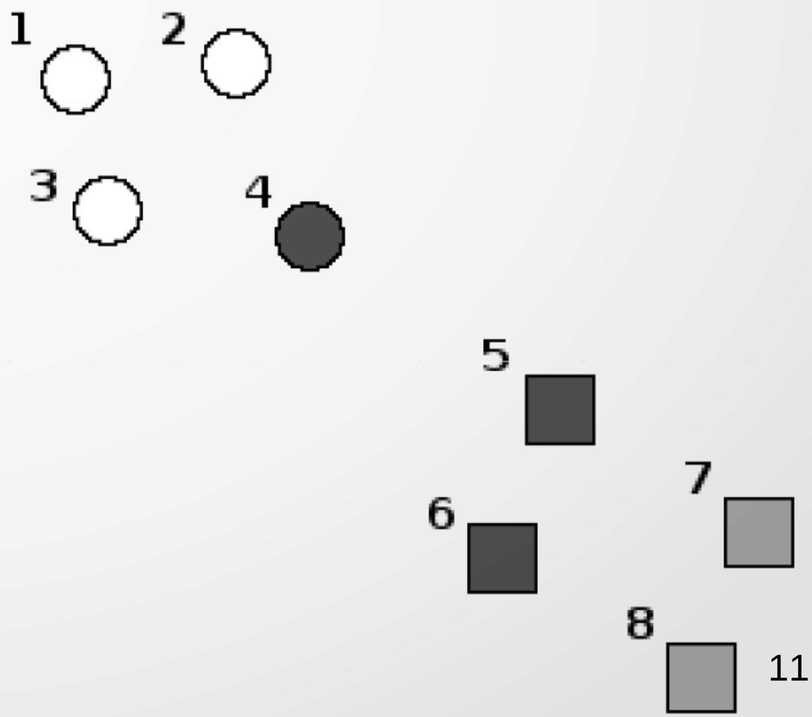
- Considera pares de objetos agregados (*MM*) e não agregados (*DD*) com a mesma importância no cálculo do índice
- Favorece a comparação de partições com níveis mais elevados de granularidade
  - Quanto mais grupos, mais pares pertencem a grupos distintos
  - Válido para qualquer partição

Índice Jaccard

- Dado por:

$$Jc(\pi_1, \pi_2) = \frac{MM}{MD + DM + MM}$$

- 2 Grupos
  - (Círculos e Quadrados)
- 3 Grupos
  - (Preto, Branco e Cinza)
- $MM = 5$; $MD = 7$; $DM = 2$
- $Jc = 5/(5+7+2) = 0.3571$



Limitações

- Para *Jaccard* e *Rand*
  - Valor esperado não é nulo para 2 partições completamente aleatórias de um conjunto de dados
- Por isso, utiliza-se uma correção, dando origem ao critério *Rand* Ajustado

$$ARI = \frac{MM - \frac{(MM+DM)(MM+MD))}{T}}{\frac{(MM+DM)+(MM+MD)}{2} - \frac{(MM+DM)+(MM+MD)}{T}}$$

Índices internos

- Normalmente não se dispõe de uma partição de referência para validar a estrutura de grupos obtida
  - temos apenas os dados e o resultado a ser avaliado...
- Índices que avaliam a estrutura de grupos obtida utilizando apenas os próprios dados são denominados internos

Exemplo

- A função objetivo do *k*-médias, o *SSE* = *Sum of Squared Errors* (variâncias intra-cluster) dado por *J*:

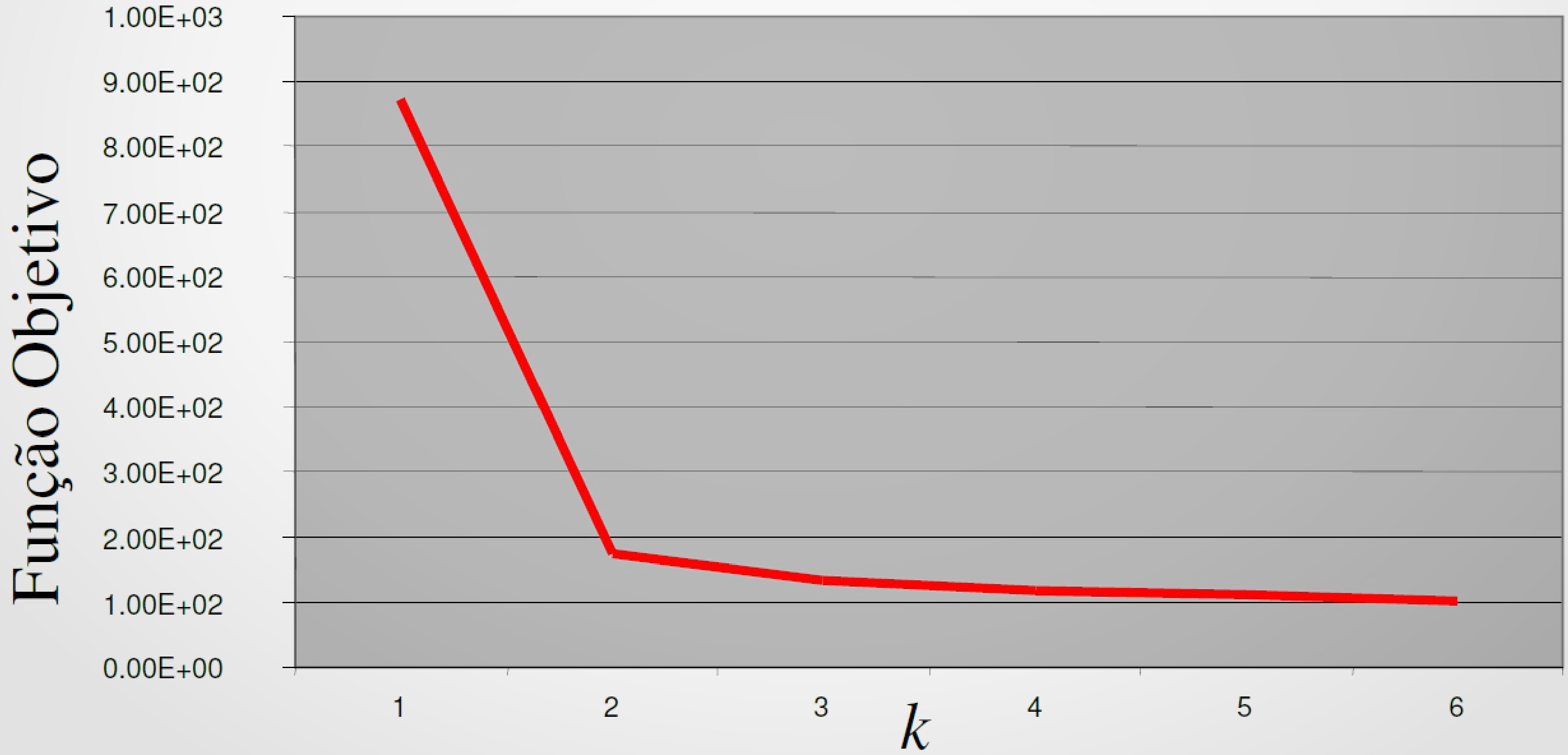
$$J = \sum_{c=1}^{k} \sum_{\mathbf{x}_j \in \mathbf{C}_c} d\left(\mathbf{x}_j, \overline{\mathbf{x}}_c\right)^2$$

onde d = Euclidiana e o centróide do c-ésimo grupo é dado por:

$$\overline{\mathbf{x}}_c = \frac{1}{\left|\mathbf{C}_c\right|} \sum_{\mathbf{x}_j \in \mathbf{C}_c} \mathbf{x}_j$$

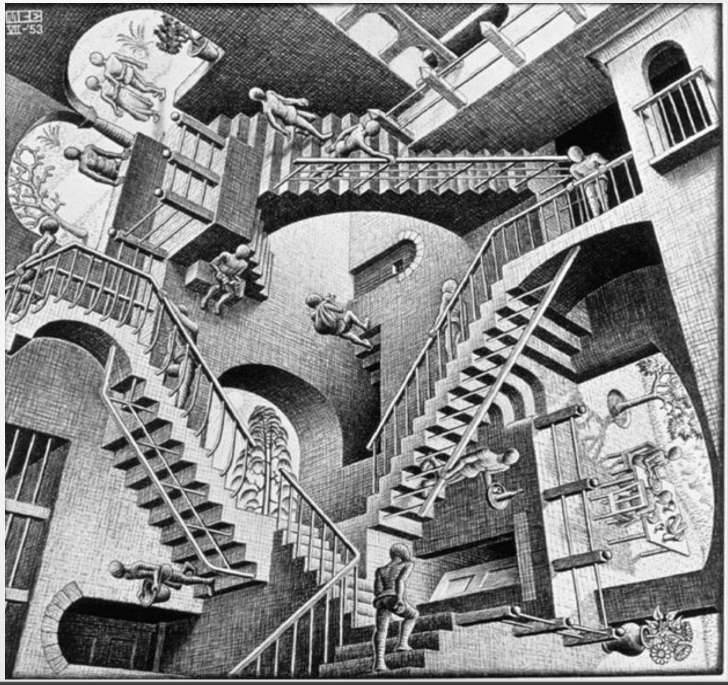
Exemplo

- Lembrando que J tende a obter valores menores a medida que k aumenta
  - Porque? (ver aula algoritmos particionais)



Índices relativos

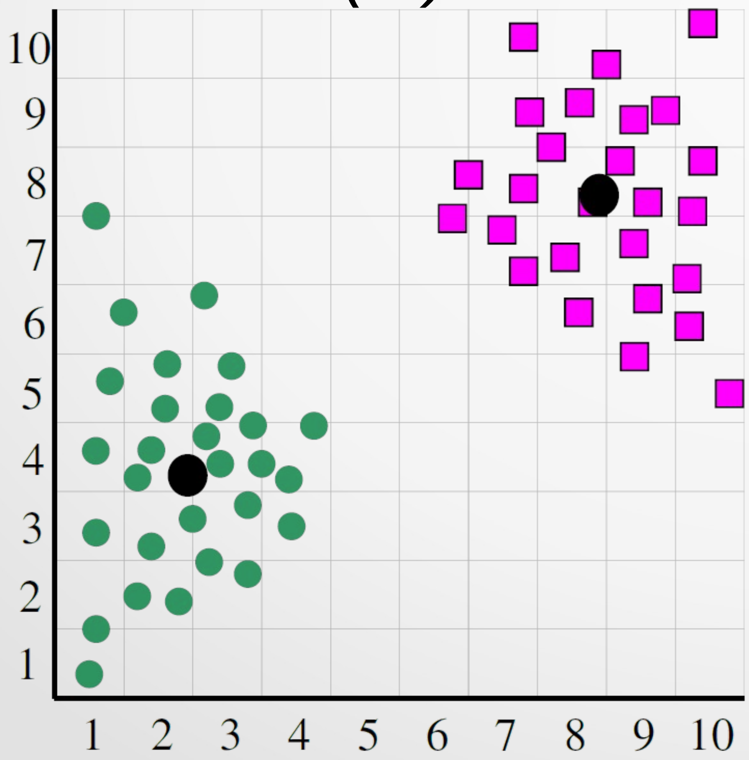
- Relativo se refere a uma classe de índices com habilidade para indicar qual a melhor dentre duas ou mais partições
- A caracterização como relativo pode não depender apenas do critério, mas eventualmente do contexto
- Por exemplo, o *SSE* é um critério relativo se as partições a serem comparadas possuem o mesmo no. de grupos



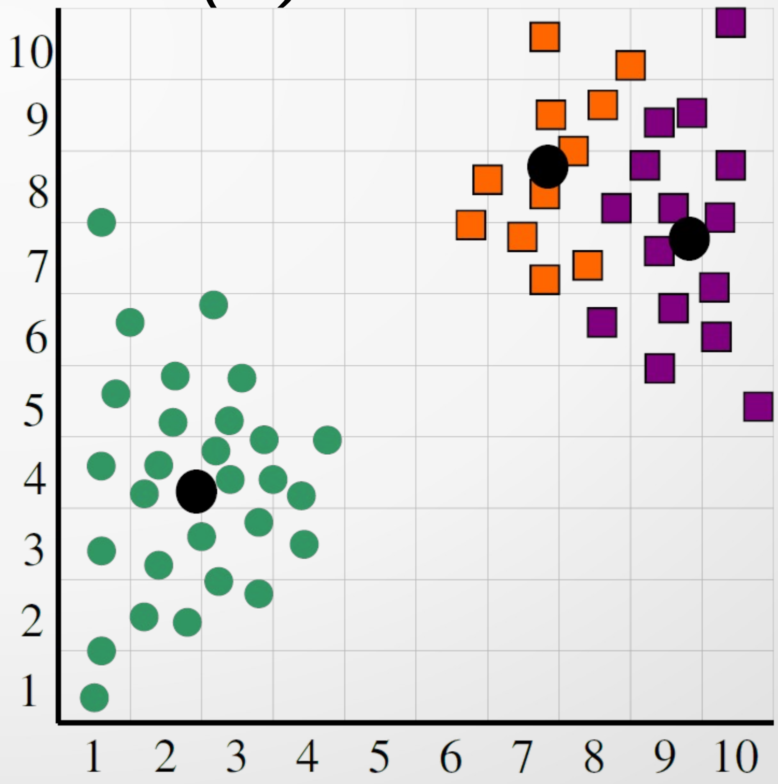
Índice relativos

- Exemplo de uso
  - Considere um índice relativo $I(\pi)$ cujo resultado é proporcional a qualidade do agrupamento

$I(\pi_1) = 23$

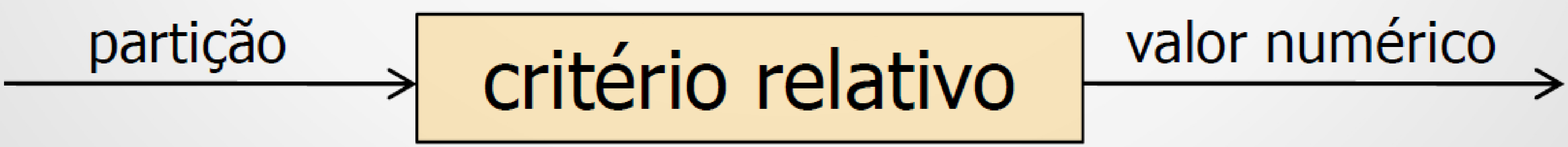


$I(\pi_2) = 11$



Índices relativos

- Índices relativos num contexto amplo são definidos aqui como aqueles capazes de:
  - Avaliar individualmente uma única partição
  - Quantificar esta avaliação através de um valor que possa ser comparado relativamente



Utilizações

- Índices relativos são mais flexíveis, pois:
  - Podem ser utilizados como critérios de otimização
    - Usado para escolher melhor partição
  - Também podem ser utilizados como regras de parada
    - Exemplo: ao se encontrar uma qualidade satisfatória
  - Além de serem usados para escolher entre partições
    - Inclusive partições diferentes em uma hierarquia

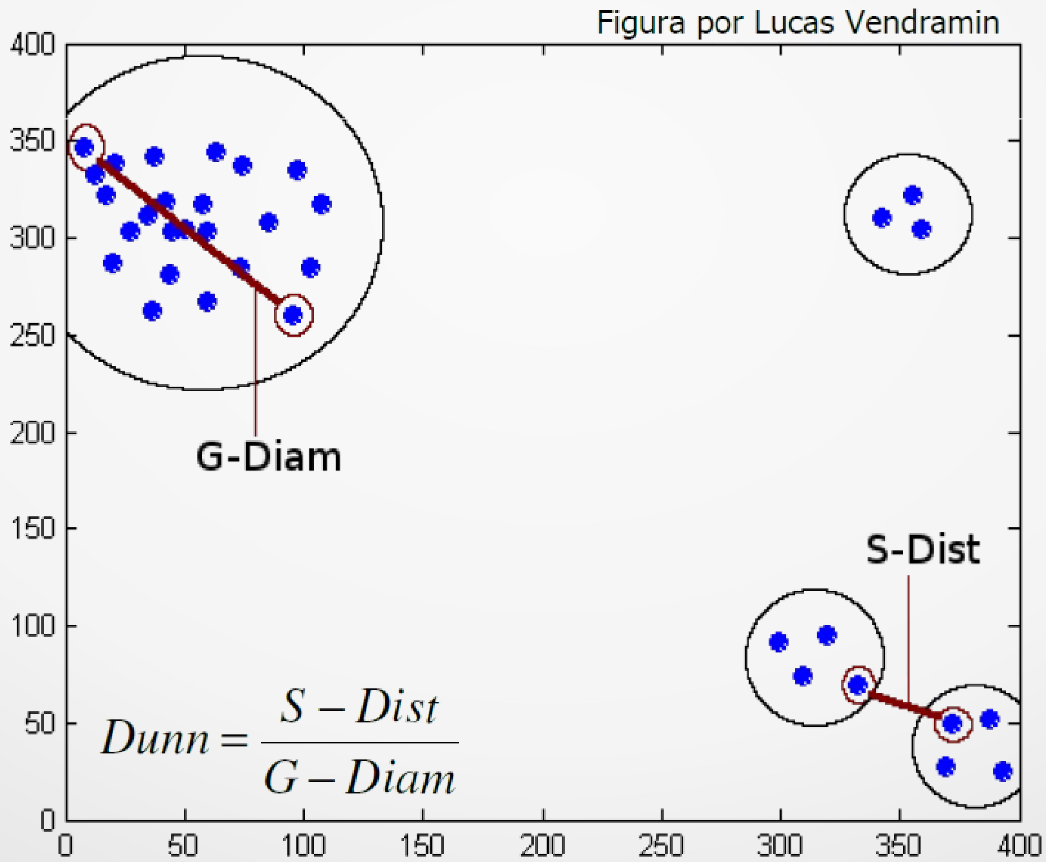
Qual escolher?

- Existem dezenas de índices na literatura
- Estudos apontam superioridade de alguns para determinados tipos de dados
  - Milligan, G.W., Cooper, M.C. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179 (1985).
  - Vendramin et. al. Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining 3 (4), 209-235

- Para problemas em geral, no entanto, não há qualquer garantia que um dado critério será o mais apropriado

Índices *Dunn*

- Razão entre a menor distância inter-grupos (S-Dist) e a maior distância intra-grupo (G-Diam)



Índice *Dunn*

- No índice original, as distâncias inter-grupos e intra-grupos são calculadas segundo ligação simples e diâmetro máximo
  - Muito sensível a ruído e *outliers*!

$$Dunn(k) = \min_{i=1,\ldots,k} \left\{ \min_{j=1,\ldots,k} \left\{ \frac{\delta(C_i, C_j)}{\max\limits_{z=1,\ldots,k} \Delta(C_z)} \right\} \right\}$$

Índice *Dunn*

- Distância de ligação simples

$$\delta(C_i, C_j) = \min_{\mathbf{x}_y \in C_i, \mathbf{x}_l \in C_j} d(\mathbf{x}_y, \mathbf{x}_l)$$

- Diâmetro máximo

$$\Delta(C_i) = \max_{\mathbf{x}_y, \mathbf{x}_l \in C_i} d(\mathbf{x}_y, \mathbf{x}_l)$$

Índices *Dunn*

- Outras distâncias intra-grupos e inter-grupos foram sugeridas no trabalho:
  - *Bezdek, J. C. & N. R. Pal (1998). Some new indexes of cluster validity. IEEE Trans. on Systems, Man and Cybernetics 28 (3), 301-315.*

com a finalidade de solucionar os problemas do índice *Dunn* original

Índice Silhueta

- Silhueta é uma característica relativa a forma com que um objeto foi agrupado
  - Quanto maior a distância do objeto para os outros grupos ($b(\mathbf{x}_i)$) e menor a distância para seu grupo ($a(\mathbf{x}_i)$), melhor será a avaliação
- A silhueta de um objeto é dada por

$$silhouette(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}}$$

Índice Silhueta

- Como calcular a silhueta dos grupos e partições?
  - É possível calcular a silhueta dos grupos de agrupamento e de toda a partição por meio da média das silhuetas de seus objetos

$$ASWC(C_j) = \sum_{i=1}^{|C_j|} \frac{silhouette(\mathbf{x}_{ji})}{|C_j|}$$

$$ASWC(\pi) = \sum_{i=1}^{n_o} \frac{silhouette(\mathbf{x}_i)}{n_o}$$

Tipos de silhueta

- Diferentes tipos de silhueta podem ser calculados, segundo os valores de ($a(\mathbf{x}_i)$) e ($b(\mathbf{x}_i)$).
- Original:
  - $a(\mathbf{x}_i)$: dissimilaridade média do i-ésimo objeto em relação aos objetos de seu grupo
  - $b(\mathbf{x}_i)$: dissimilaridade média do i-ésimo objeto em relação aos objetos do grupo mais próximo a que $\mathbf{x}_i$ não pertence

Silhueta Simplificada

- Uma versão simplificada da silhueta considera apenas a distância do objeto em relação ao centróide de seu grupo ($a(\mathbf{x}_i)$) e ao centróide do grupo mais próximo que não seja o seu ($b(\mathbf{x}_i)$).
- Não relacional, porém possui complexidade computacional linear ao invés de quadrática

Exemplo

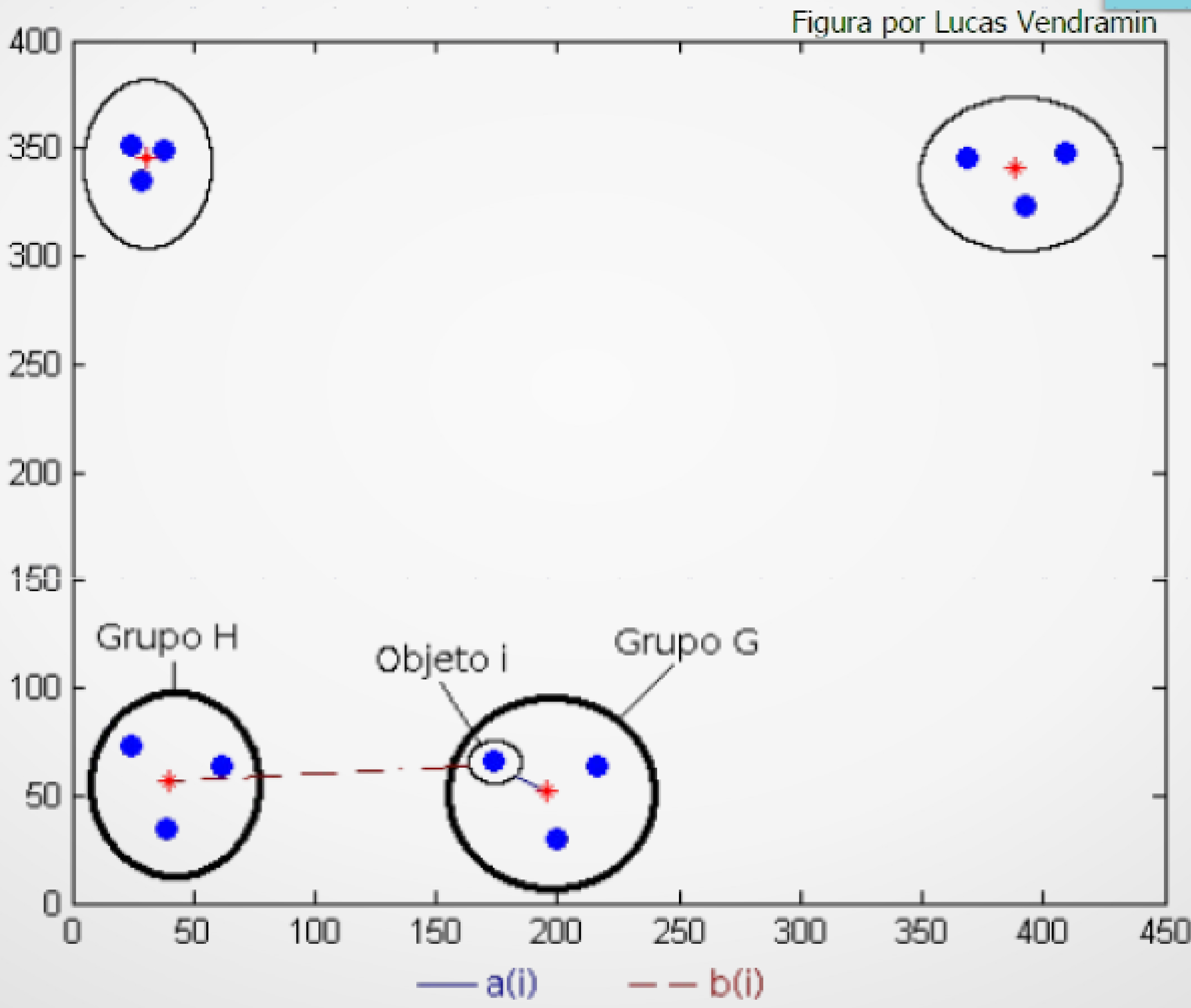
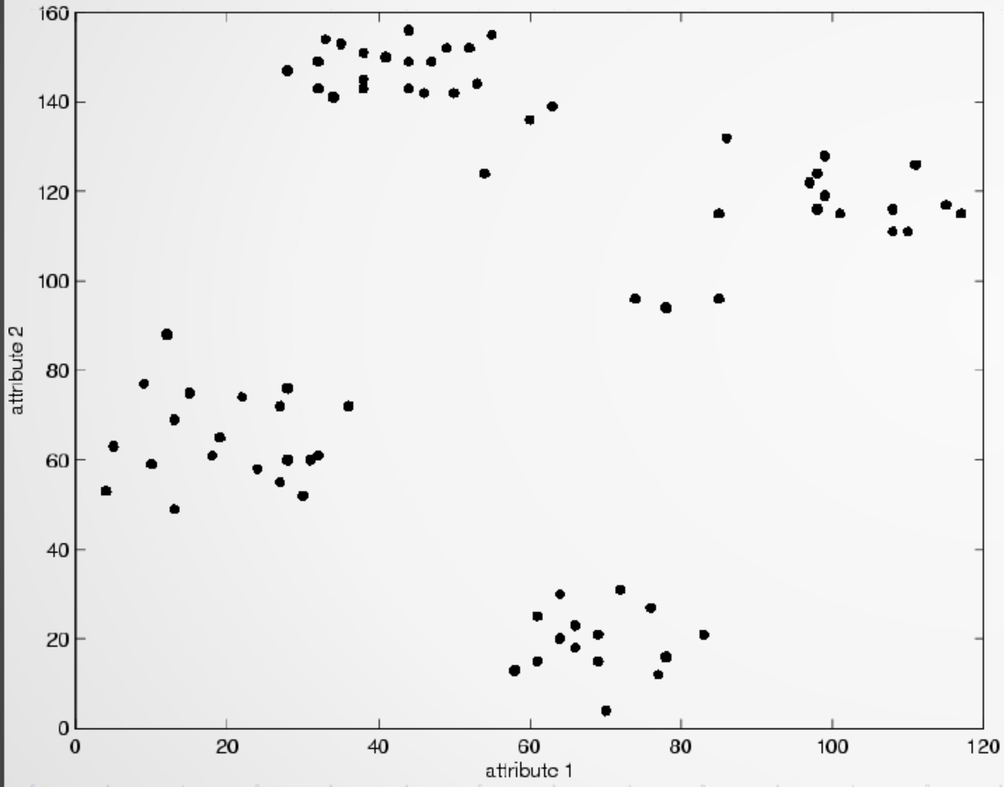
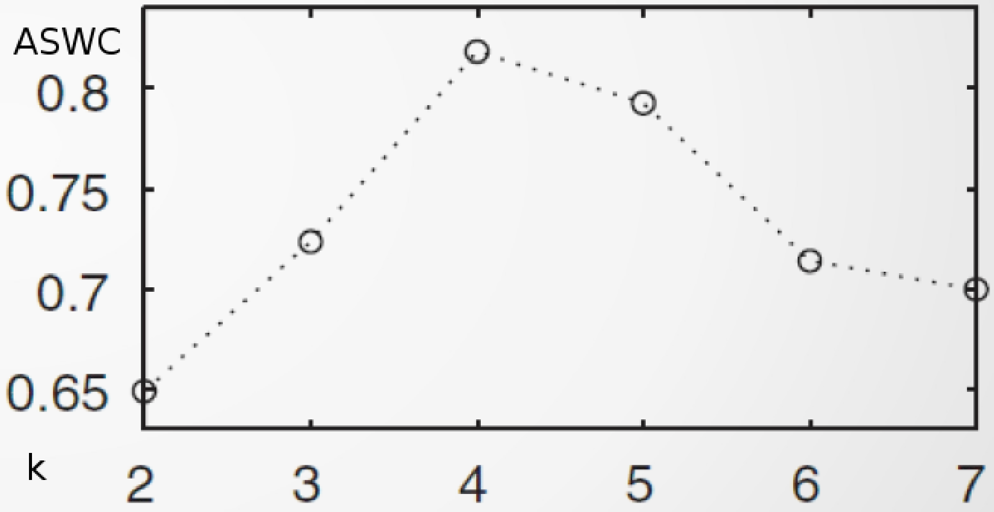


Figura por Lucas Vendramin
400
350
300
250
200
150
100
50
0
0
50
100
150
200
250
300
350
400
450
Grupo H
Objeto i
Grupo G
a(i)
b(i)

Subjetividade

- Quantos grupos possui esse conjunto de dados?



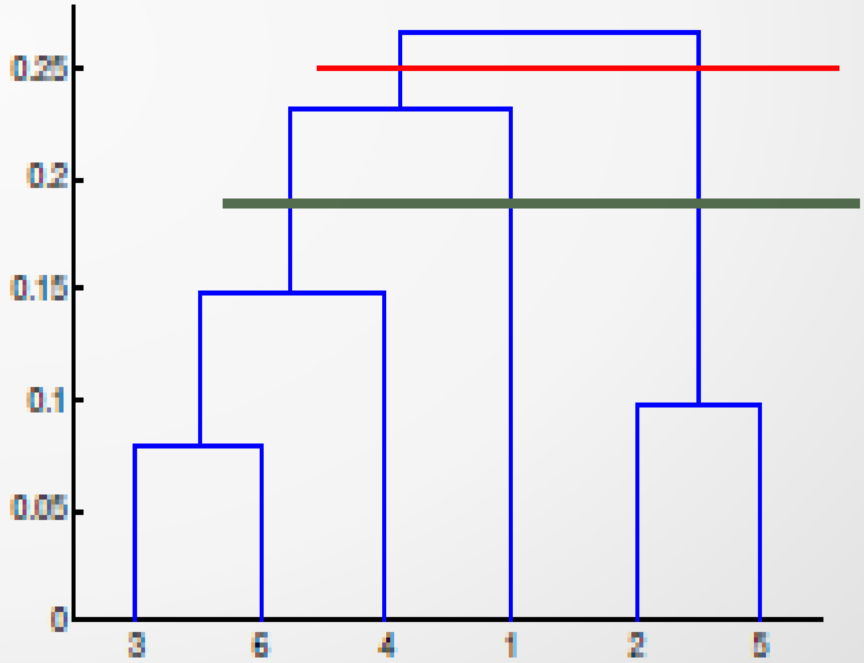


E muitos outros...

-	Criterion	Complexity
	Calinski-Harabasz (VRC)	$O(nN)$ [Eqs. (5) − (8)]
	Davies-Bouldin (DB)	$O(n(k^2 + N))$
	Dunn	$O(nN^2)$
	Silhouette Width Criterion (SWC)	$O(nN^2)$
	Alternative Silhouette (ASWC)	$O(nN^2)$
	Simplified Silhouette (SSWC)	$O(nNk)$
	Alternative Simplified Silhouette (ASSWC)	$O(nNk)$
	PBM	$O(n(k^2 + N))$
	C-Index	$O(N^2(n + log_2 N))$
	Gamma	$O(nN^2 + N^4/k])$
	G(+)	$O(nN^2 + N^4/k])$
	Tau	$O(nN^2 + N^4/k])$
	Point-Biserial	$O(nN^2)$
	$\mathrm{C}/\sqrt{\mathrm{k}}$	$O(nN)$
*	Trace(W)	$O(nN)$
*	Trace(CovW)	$O(nN)$
*	$\mathrm{Trace}(\mathrm{W}^{-1}\mathrm{B})$	$O(n^2N + n^3)$
*	|T|/|W|	$O(n^2N + n^3)$
*	Nlog(|T|/|W|)	$O(n^2N + n^3)$
*	$\mathrm{k}^2\mathrm{W}$	$O(n^2N + n^3)$
*	log(SSB/SSW)	$O(n(k^2 + N))$
*	Ball-Hall	$O(nN)$
*	McClain-Rao	$O(nN^2)$

Algoritmos hierárquicos

- Como validar algoritmos hierárquicos?
  - Basta lembrar que um agrupamento hierárquico nada mais é do que uma sequência aninhada de partições
- Podemos aplicar o critério relativo em todos os níveis da hierarquia e utilizar o resultado para comparação



Seleção de partições

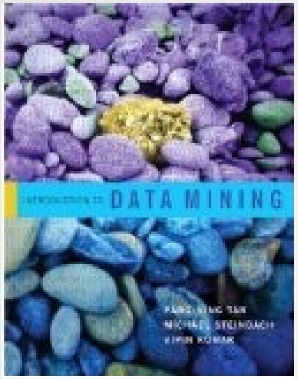
- Podemos obter partições de diferentes algoritmos ou do mesmo algoritmo e selecionar qual partição é a mais indicada pelos dados, segundo um critério de validação
  - O que dizer do resultado abaixo?

k	2	3	4	5	6	7	8
Alg1	2	2	21	23	7	12	9
Alg2	3	4	24	25	9	12	10
Alg3	1	7	30	23	7	11	9
Alg4	2	13	12	27	3	16	11
Alg5	2	12	15	24	2	17	9

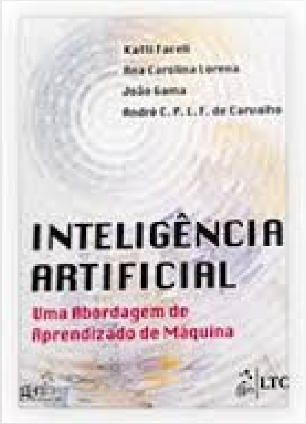
Em resumo

- Índices de validação avaliam a qualidade de um agrupamento
  - Em comparação com uma estrutura conhecida
  - Quantitativamente
  - Em relação a outro
- São frequentemente usados para selecionar agrupamentos
- Índices diferentes podem resultar em avaliações distintas

Bibliografia



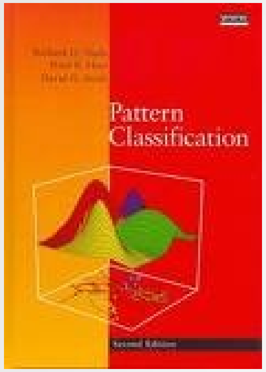
V. TAN, STEINBACH, M., KUMAR, P. Introdução ao Data Mining (Mineração de Dados). Edição 1. Ciência Moderna 2009.   ISBN 9788573937619.



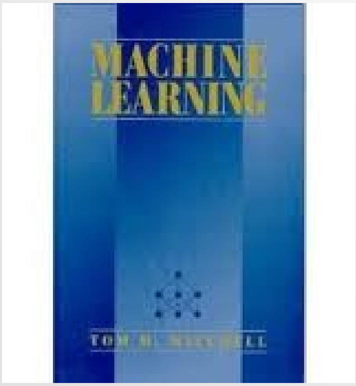
Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. Katti Faceli, Ana Carolina Lorena, João Gama, André C. P. L. F. de Carvalho. Grupo Gen 2011

Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993

Referencias



Duda, R.O., Hart, P. E. and Stork, D. G. Pattern Classification (2nd Edition). Wiley-Interscience



MITCHELL, T. Machine Learning, McGraw Hill, 1997.

Referências

- Jain, A. K. and Dubes, R. C., Algorithms for Clustering Data, Prentice Hall, 1988
- Kaufman, L., Rousseeuw, P. J., Finding Groups in Data – An Introduction to Cluster Analysis, Wiley, 2005.
- Tan, P.-N., Steinbach, M., and Kumar, V., Introduction to Data Mining, Addison-Wesley, 2006
- Wu, X. and Kumar, V., The Top Ten Algorithms in Data Mining, Chapman & Hall/CRC, 2009
- D. Steinley, K-Means Clustering: A Half-Century Synthesis, British J. of Mathematical and Stat. Psychology, V. 59, 2006