

Aula 12 — Validação de agrupamento

Mineração de Dados

Universidade Federal do ABC

Introdução

CV Externos

CV Internos

CV Relativos

Valor interessante?

CRÉDITOS

- ▶ Este material consiste de adaptações e extensões dos originais elaborados por Eduardo R. Hruschka e Ricardo J. G. B. Campello

COMENTÁRIO SOBRE VALIDAÇÃO DE AGRUPAMENTO

The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.

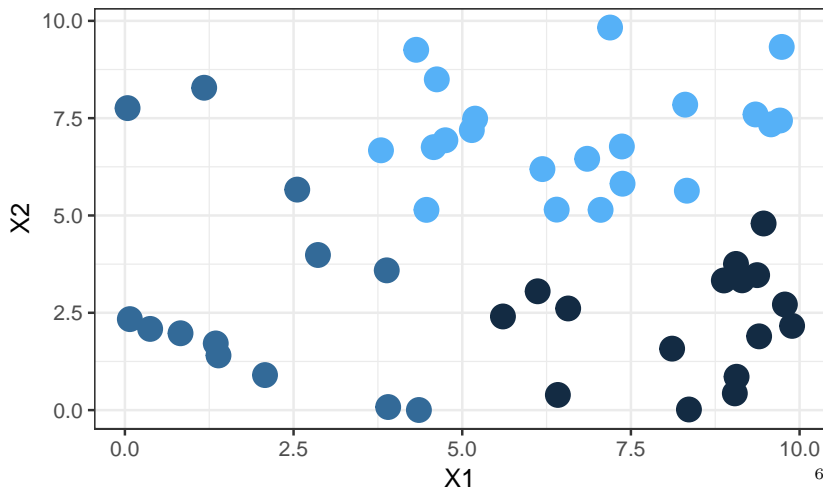
- Jain and Dubes, Algorithms for Clustering Data, 1988

VALIDAÇÃO DE AGRUPAMENTO

- ▶ *Validação* é um termo que se refere de forma ampla aos diferentes procedimentos para avaliar de maneira objetiva e quantitativa os resultados de análise de agrupamento
- ▶ Cada um desses procedimentos pode nos ajudar a responder uma ou mais questões do tipo:
 - ▶ Encontramos grupos de fato ?
 - ▶ Qual a qualidade (relativa ou absoluta) dos grupos encontrados ?
 - ▶ Qual é o número natural / mais apropriado de grupos ?

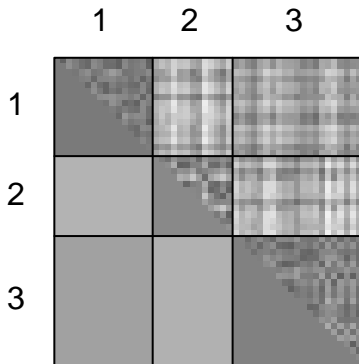
EXISTEM GRUPOS?

- ▶ Ao executarmos um algoritmo de agrupamento particional **sempre** teremos uma partição
 - ▶ e se os dados forem completamente aleatórios?



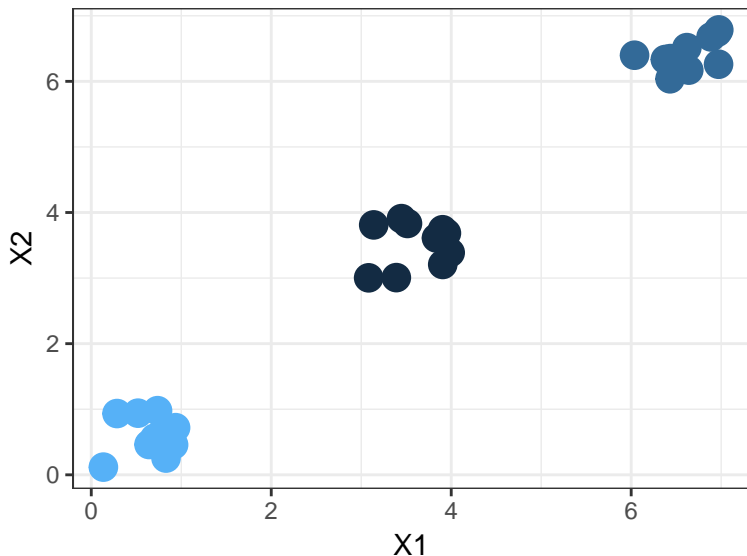
EXISTEM GRUPOS?

- ▶ Considerando nossa premissa de agrupamento, poderíamos ordenar a matriz de distância de acordo com os rótulos de grupos obtidos
 - ▶ objetos do mesmo grupo *devem* ser mais próximos entre si do que em relação aos outros grupos




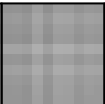
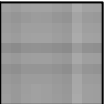






EXISTEM GRUPOS?

- Como ficaria em uma base de dados com alguma estrutura?



EXISTEM GRUPOS?

- Como ficaria em uma base de dados com alguma estrutura?

	1	2	3
1			
2			
3			

VALIDAÇÃO DE AGRUPAMENTO

- ▶ A maneira quantitativa com que se dá um procedimento de validação é alcançada através de algum tipo de índice
 - ▶ *Índice ou Critério de Validade* (de agrupamento)
- ▶ Tais índices / critérios podem ser de três tipos
 - ▶ **Externos:** 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
 - ▶ **Internos:** Avalia o grau de compatibilidade entre a estrutura de grupos sob avaliação e os dados, usando apenas os próprios dados
 - ▶ **Relativos:** Avaliam qual dentre duas ou mais estruturas de grupos é melhor sob algum aspecto. Tipicamente são critérios internos capazes de quantificar a qualidade relativa

Introdução

CV Externos

CV Internos

CV Relativos

Valor interessante?

INTRODUÇÃO

- ▶ Embora o problema de *clustering* seja não supervisionado, em alguns cenários o resultado de agrupamento desejado pode ser conhecido. Por exemplo:
 - ▶ Reconhecimento visual dos clusters naturais (bases 2D, 3D)
 - ▶ Especialista de domínio
 - ▶ Bases geradas sinteticamente com distribuições conhecidas
 - ▶ *Benchmark* data sets
 - ▶ Bases de classificação sob a hipótese que classes são clusters
- ▶ Medem o nível de compatibilidade entre uma partição obtida e uma partição de referência dos mesmos dados

CRITÉRIOS DE VALIDADE EXTERNOS

- ▶ Existem vários critérios externos na literatura:
 - ▶ *Rand Index*
 - ▶ Jaccard
 - ▶ *Rand Index* Ajustado
 - ▶ Fowlkes-Mallows
 - ▶ Estatística Γ
 - ▶ *Normalized Mutual Information*

Rand Index

- ▶ O critério que veremos é baseado na comparação de pares de objetos das partições em questão
- ▶ Por conveniência, adotaremos a seguinte terminologia:
 - ▶ grupos da partição de referência (golden truth) → classes
 - ▶ grupos da partição sob avaliação → clusters (grupos)
- ▶ Podemos então definir as grandezas de interesse:
 - ▶ **a:** No. de pares que pertencem à mesma classe e ao mesmo *cluster*
 - ▶ **b:** No. de pares que pertencem à mesma classe e a *clusters* distintos
 - ▶ **c:** No. de pares que pertencem a classes distintas e ao mesmo *cluster*
 - ▶ **d:** No. de pares que pertencem a classes e *clusters* distintos

Rand Index

- ▶ Número de pares de objetos:
 - ▶ a: Pertencem à mesma classe e ao mesmo *cluster*
 - ▶ b: Pertencem à mesma classe e a *clusters* distintos
 - ▶ c: Pertencem a classes distintas e ao mesmo *cluster*
 - ▶ d: Pertencem a classes e *clusters* distintos

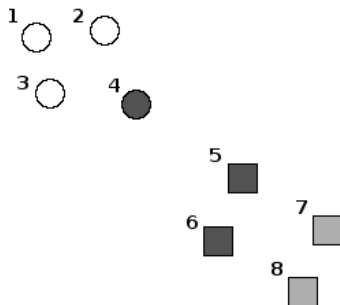


Figura 1: Classe: forma, *Clusters*: cor. Figura por Lucas Vendramin

Rand Index

► $a = 5$ $b = 7$ $c = 2$ $d = 14$

$$RI = \frac{5 + 14}{(5 + 7 + 2 + 14)} = 0,6785$$

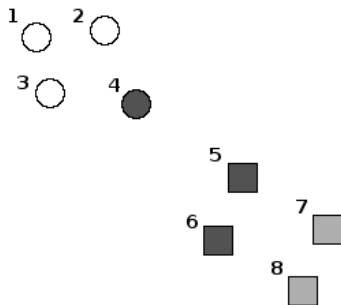


Figura 2: Classe: forma, *Clusters*: cor. Figura por Lucas Vendramin

Introdução

CV Externos

CV Internos

CV Relativos

Valor interessante?

INTRODUÇÃO

- ▶ Na prática, normalmente não se dispõe de uma partição de referência
 - ▶ temos apenas os dados e o resultado a ser avaliado
- ▶ Critérios que avaliam a estrutura de grupos obtida utilizando apenas os próprios dados são denominados critérios internos de validade de agrupamento
 - ▶ Já vimos um exemplo ao estudar o *k-means* – SSE:

$$J = \sum_{n=1}^N \sum_{k=1}^K \mu_{nk} \|\mathbf{x}_n - \bar{\mathbf{x}}_k\|^2$$

Introdução

CV Externos

CV Internos

CV Relativos

Valor interessante?

INTRODUÇÃO

- ▶ A aplicação de um ou mais algoritmos usualmente retorna múltiplas soluções que precisam ser comparadas
 - ▶ Algoritmos hierárquicos
 - ▶ Múltiplas execuções de k-means

INTRODUÇÃO

- ▶ O termo *critério relativo* se refere a uma classe particular de critérios 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
 - ▶ Para números de grupos distintos, os valores de SSE não são comensuráveis e o critério, portanto, não é relativo

INTRODUÇÃO

- ▶ Critérios relativos no contexto amplo definido anteriormente são mais flexíveis, pois:
 - ▶ Podem ser utilizados como critérios de otimização
 - ▶ Também podem ser utilizados como *stopping rules*
- ▶ Existem dezenas de tais critérios na literatura
- ▶ Estudos apontam alguns deles como superiores em algumas classes de problemas comuns na prática
 - ▶ Para problemas em geral, no entanto, não há qualquer garantia que um dado critério será o mais apropriado

INTRODUÇÃO

- ▶ Cada critério computa sua forma de compromisso entre:
 - ▶ **coesão:** distâncias entre objetos do mesmo grupo
 - ▶ **separação:** distâncias entre objetos de grupos distintos

CRITÉRIO DAVIES-BOULDIN

- ▶ Cada grupo tem seu critério valor de compromisso (D_k):
 - ▶ Coesão do grupo: $\bar{d}_k = \frac{1}{N_k} \sum_{n=1}^N \mu_{nk} \|\mathbf{x}_n - \bar{\mathbf{x}}_k\|$
 - ▶ N_k é o número de objetos no grupo k
 - ▶ Separação em relação ao grupo c : $d_{k,c} = \|\bar{\mathbf{x}}_k - \bar{\mathbf{x}}_c\|$
 - ▶ Compromisso entre coesão e separação desses dois grupos:
 $D_{k,c} = (\bar{d}_k + \bar{d}_c)/d_{k,c}$
 - ▶ Compromisso geral do grupo: $D_k = \max_{k \neq c} D_{k,c}$
- ▶ Compromisso geral da partição:

$$DB = \frac{1}{K} \sum_{k=1}^K D_k$$

CRITÉRIO DA LARGURA DE SILHUETA

- ▶ $SWC = \textit{Silhouette Width Criterion}$
- ▶ Silhueta (i-ésimo objeto):
 - ▶ $a(i)$: distância média do i-ésimo objeto ao seu *cluster*
 - ▶ $b(i)$: distância do i-ésimo objeto ao *cluster* vizinho mais próximo

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- ▶ Versão Original: $a(i)$ e $b(i)$ são calculados como a distância média do i-ésimo objeto a todos os demais objetos do *cluster* em questão

CRITÉRIO DA LARGURA DE SILHUETA

- ▶ $SWC = \frac{1}{N} \sum_{n=1}^N s(n)$
- ▶ Propriedade Favorável: $SWC \in [-1, +1]$
- ▶ Qual o valor de s para um *singleton*?

SILHUETA SIMPLIFICADA (SSWC)

- Silhueta Simplificada: $a(i)$ e $b(i)$ são calculados como a distância do i -ésimo objeto ao centróide do *cluster* em questão

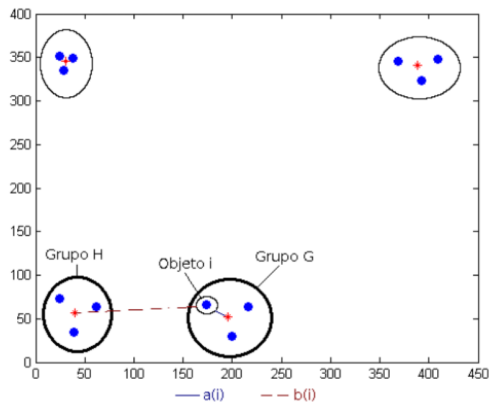
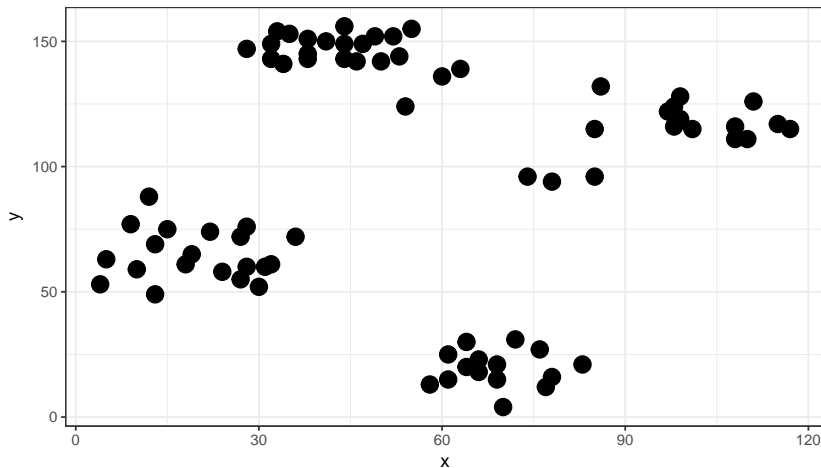


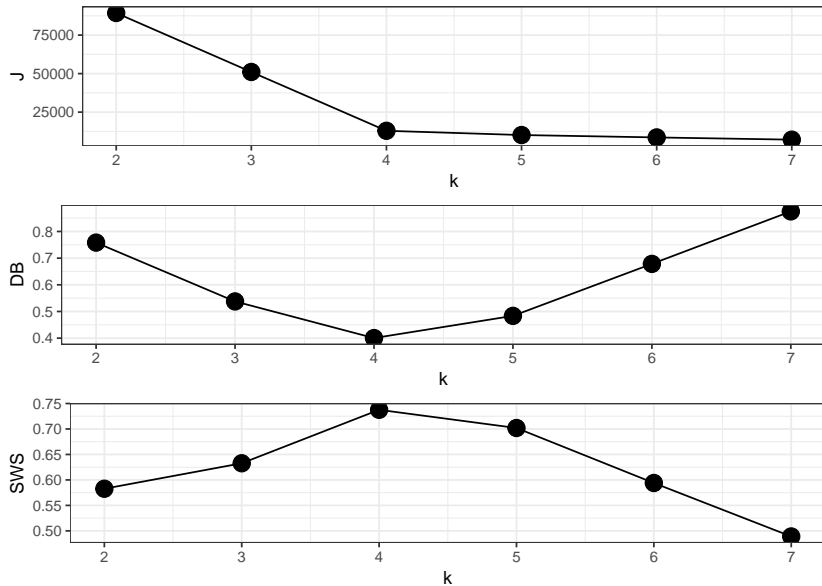
Figura 3: Figura por Lucas Vendramin

EXEMPLO

- Relembrando a subjetividade do problema:
 - Quantos grupos abaixo...? Quatro? Cinco? Seis?



EXEMPLO



Introdução

CV Externos

CV Internos

CV Relativos

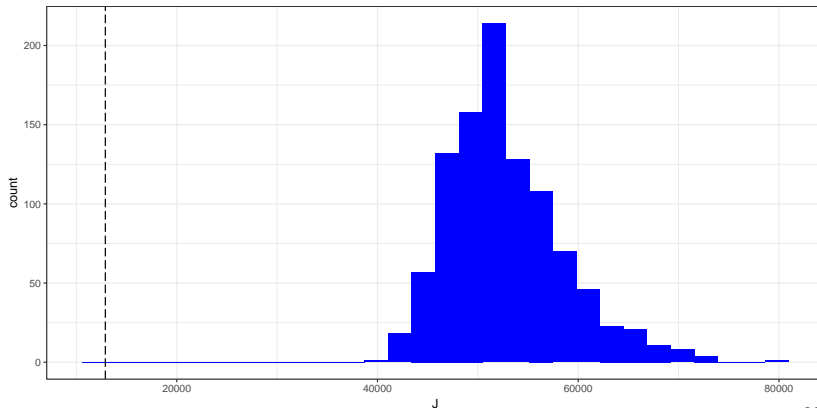
Valor interessante?

AVALIANDO UM CRITÉRIO

- ▶ Se tivermos um valor 0,5 em um critério, isso é bom?
- ▶ Podemos utilizar testes estatísticos para nos auxiliar
 - ▶ Geramos diferentes valores a partir de dados aleatórios
 - ▶ Quanto mais *atípico* um valor, maior a probabilidade de termos uma estrutura de grupos nos dados
 - ▶ Relativamente difícil de interpretar

AVALIANDO UM CRITÉRIO

- ▶ Exemplo
 - ▶ Comparar o valor de J obtido (1.2881051×10^4) considerando o mesmo número de grupos em dados aleatórios distribuídos **no mesmo espaço** dos dados
 - ▶ Mesmo número de objetos, 1000 repetições



REFERÊNCIAS

- ▶ Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. Introduction to Data Mining, (First Edition). **Capítulo 8.**
- ▶ Jain, A. K. & Dubes, R. C., Algorithms for Clustering Data, **Capítulo 4.** Prentice Hall, 1988
- ▶ Milligan, G. W. & Cooper, M. C. “An Examination of Procedures for Determining the Number of Clusters in a Data Set”, Psychometrika, Vol. 50, No. 2, 159-179, 1985
- ▶ Vendramin, L. , Campello, R. J. G. B. , Hruschka, E. R. “Relative Clustering Validity Criteria: A Comparative Overview” Statistical Analysis and Data Mining, Wiley, Vol. 3, p. 209-235, 2010