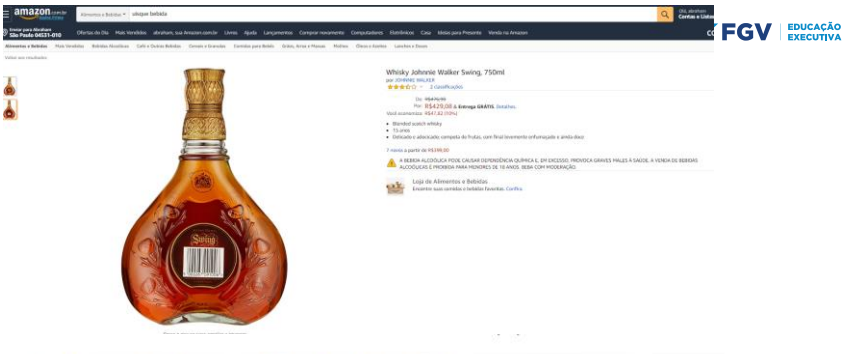
MBA Big Data e Business Intelligence

RA - Regras de Associação

Prof. Abraham **Laredo** Sicsu

2

Conceitos e Aplicações



Clientes que visualizaram este item também visualizaram

Whisky Johnnie Walker Green Label 750ml
★★★★★ 2
R$249,00

Whisky Johnnie Walker Ultimate 18, 18 Anos, 750ml
★★★★★ 1
R$379,74

Whisky Johnnie Walker Blue Label 750ml
★★★★☆ 43
R$888,95

Whisky Dimple Golden Seletion 1L
R$458,75

4

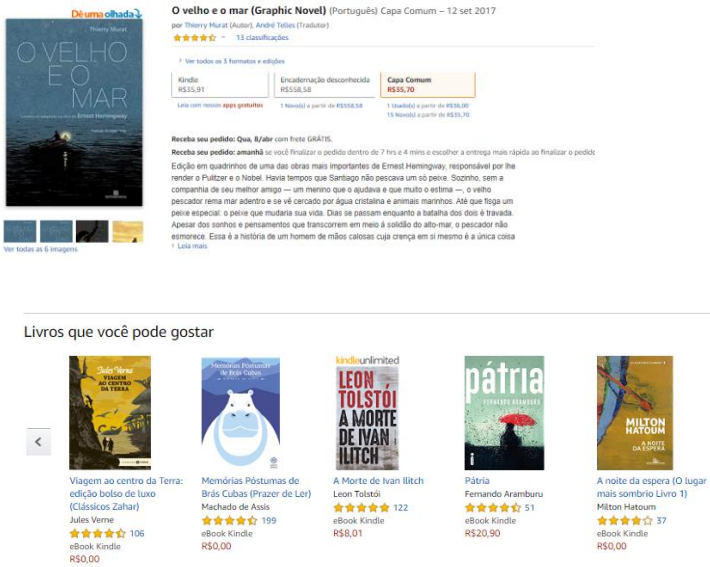
Nomes

- Objetivo: descobrir relações ocultas entre itens (ex. Produtos, pessoas, palavras, sintomas...) em grandes bases de dados
- O que "vai com o que"?
  - Sites A, D e X são frequentemente consultados "simultaneamente"
  - **"SE...ENTÃO...":** Clientes que compraram XXX também compraram YYYY
- Outros nomes : *Affinity rules* / *Market basket analysis*
- Regras de associação são algoritmos **não supervisionados**.
  - Verificação de validade ou de interesse é complexa
- Outros métodos de recomendação serão vistos em cursos futuros

Exemplo de aplicação

- Regras para estratégias de marketing:
  - Colocação dos produtos “associados” em gôndolas próximas
  - Cross selling
  - Malas diretas customizadas por cliente, etc
  - Sistema de recomendação na internet
    - Amazon➔ “*frequently bought together*”
- Análise das ações da interface / consultas do usuário em um ou mais sites
- Cientista médico deseja pesquisar que sintomas aparecem conjuntamente
- Conjunto de cursos online adquiridos por acadêmicos
- Recomendações de filmes ou músicas em canais de streaming
- Planejamento de estoques de uma revendedora: partes que ocorrem simultaneamente em reparos de autos

Amazon.com



Definições

- **I={ i_1, i_2, …,i_d} conjunto de itens**

- **Itemset** : Um conjunto com um ou mais itens
  - Exemplo : {Milk, Bread, Diaper}
- **k-itemset**
  - Itemset com k indivíduos
  - {Milk, Bread} 2-itemset

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Cola
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Cola

- **Transação t** : um conjunto de itens (t =1,…,T)
- **TID**: identificação da transação
- **TDB: Transaction database** : tabela de transações

8

Representação binária das transações

Matriz de transações

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Cola
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Cola

Binárias assimétricas: presença (1) mais importante que ausência (0)

Matriz binária de transações

TID	Bread	Milk	Diapers	Beer	Eggs	Coke
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Estamos ignorando quantidade de itens vendidos e os preços pagos (ver no fim)

9

Regras de associação (RA)

FGV EDUCAÇÃO EXECUTIVA

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Cola
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Cola

Exemplos

{Diaper} → {Beer},

{Milk, Bread} → {Eggs, Coke},

{Beer, Bread} → {Milk},

Regras de associação (RA)

FGV EDUCAÇÃO EXECUTIVA

- **Expressão do tipo X ➔ Y onde X e Y são dois itemsets**

{Milk, Bread} → {Eggs, Coke}

- Se **X** ocorre, então **Y** ocorre (*com certa probabilidade*)
- **X** e **Y** são disjuntos (**X** ∩ **Y** =∅)
- X denominado **Antecedente, Condição, Premissa**, **Corpo**
- Y denominado **Consequente, Conclusão , Cabeça**
- (**LHS ➔ RHS**) notação no pacote do R
- Implicação representa ocorrência; **não é causalidade**
- Regras de associação são baseadas em probabilidades

FGV EDUCAÇÃO EXECUTIVA

Medidas
Avaliando a "qualidade" de uma regra

SUPORTE de um *itemset*

FGV EDUCAÇÃO EXECUTIVA

- Suporte (**s**) de um **itemset X**
  - Proporção de transações que contem o itemsets (antecedente + consequente)
  - **s** baixo → itemset, em geral, de pouco interesse (caviar???)
  - **s** estima P(**X**)

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Cola
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Cola

{Milk, Diaper} ➔ Beer

$$s = \frac{freq(\text{Milk}, \text{Diapers}, \text{Beer})}{\text{T}} = \frac{2}{5} = 0.4$$

T = número total de transações

13

CONFIANÇA de uma regra de associação

FGV EDUCAÇÃO EXECUTIVA

- Confiança (**c**)
- Regra: **X → Y**
- Frequência de **Y** em transações que contem **X**
- Probabilidade condicional de **Y** dado **X** (dado que...) $\frac{\mathbf{Prob(X \& Y)}}{\mathbf{Prob(X)}}$

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Cola
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Cola

- **{Milk, Diaper} → Beer**

$$c = \frac{freq(\text{Milk}, \text{Diapers}, \text{Beer})}{freq(\text{Milk}, \text{Diapers})} = \frac{2}{3} = 0.67$$

- **Calcular c(Milk → Bread)**

14

Interpretação

FGV EDUCAÇÃO EXECUTIVA

Regra de associação

{ laptop } => { mouse} [Suporte =20%, confiança =60%]

- Em 20% das transações houve compra conjunta de **mouse & laptop**
- Dentre os clientes que compraram laptop, 60% compraram mouse

15

Representação tabular

A ➔ B

	B	Não B	
A	f_{11}	f_{10}	f_{1+}
Não A	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	T

- f_{11} : número de transações que contem A e B
- f_{10} : número de transações que contem A e não contem B
- f_{01} : número de transações não contem A e contem B
- f_{00} : número de transações que não contem nem A e nem B
- T : número total de transações

$$\sup(A \rightarrow B) = \frac{f_{11}}{T} \qquad conf(A \rightarrow B) = \frac{f_{11}}{f_{1+}}$$

16

Representação tabular

A ➔ B : {Uisque, Caviar} ➔ {pistaches}

	pistaches	Não pistaches	
{Uisque ,caviar}	40	110	150
não {Uisque ,caviar}	20	30	50
	60	140	200

- Sup{uísque, caviar, pistaches} = 40 /200 ➔ 20%

	pistaches	Não pistaches	
{Uisque ,caviar}	40	110	150
não {Uisque ,caviar}	20	30	50
	60	140	200

Qual o suporte de pistaches?

- Conf [{uísque, caviar} ➔ {pistaches}] = 40 /150 ➔ 26,7%

	pistaches	Não pistaches	
{Uisque ,caviar}	40	110	150
não {Uisque ,caviar}	20	30	50
	60	140	200

17

Gerando regras de associação

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Cola
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Cola

Exemplos :

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Note que

- Todas as regras acima se originam do mesmo itemset: **{Milk, Diaper, Beer}**
- Regras originadas de um mesmo itemset tem mesmo suporte; confianças podem variar

18

Propriedades

1) $\text{conf}[\{i_1,i_2\} \rightarrow \{i_3, \ldots, i_k\}] \leq \text{conf}[\{i_1,i_2,i_3\} \rightarrow \{i_4, \ldots, i_k\}]$

- A confiança nunca decresce quando deslocamos um item da esquerda para direita (LHS aumenta → confiança aumenta}
- **A regra com menor confiança extraída de um itemset contem apenas um item em seu lado esquerdo**

2) $X=\{i_1,i_2,\ldots i_k\} \rightarrow s(X) \leq \min [\, s(i_j) \,, j=1,\ldots,k]$

- O suporte de um k-itemset é menor ou igual que o menor suporte de um de seus itens
  - sup {A,B, C} ≤ sup{A} sup {A,B, C} ≤ sup{B} sup {A,B, C} ≤ sup{C}
- Regras envolvendo todos os itens de um mesmo itemset possuem o mesmo suporte. Confiança pode diferir

19

Parâmetros comuns para seleção de RA

- **Itemsets frequentes**

  s > **minsup** (parâmetro pré definido pelo usuário)

- **Regras fortes**

  s > **minsup** (parâmetro pré definido pelo usuário)

  c > **minconf** (parâmetro pré definido pelo usuário)

- **Adiante veremos outros parâmetros**

Escolha de minsup

Decisão difícil:

- **minsup muito grande**
  - transações interessantes porém pouco frequentes serão ignoradas
  - Itens pouco frequentes podem ser de interesse por motivos econômicos
    - Joias
    - Caviar....
    - **Literatura apresenta técnicas especiais para "infrequente itemsets"**
- **minsup pouco pequeno**
  - Inclusão de um número muito grande de regras de associação
    - Algumas irrelevantes, outras redundantes
  - Requisitos de processamento e memória crescem demasiadamente
  - *Podemos ter itemsets relacionando itens muito frequentes contendo itens raros (ex. {Leite e Caviar}). Estes padrões são denominados "cross-products patterns". Em geral conduzem a regras de associação espúrias pois a correlação entre estes itens é pequena (veremos adiante)*

Decisões com base em *minconf*

FGV EDUCAÇÃO EXECUTIVA

- Se o antecedente e o consequente tem alto grau de suporte (muito frequentes), podemos ter alta confiança **c** mesmo se os itens forem independentes
  - A → B; se A e B independentes → Prob (A e B)=Prob(A) . Prob(B)
  - Confiança = $\frac{Prob(A).Prob(B)}{Prob(A)} = Prob(B)$
    - ***conf*** não leva em conta o suporte do RHS (*consequente*)
  - Medida ***lift*** apresentada a seguir considera o suporte do RHS
  - Independência entre A e B: Regra de associação não faria sentido!
  - O valor de ***conf*** pode ser alto mesmo se correlação entre A e B for negativa

Confiança : medida nem sempre interessante

FGV EDUCAÇÃO EXECUTIVA

	B	Não B	
A	150	50	200
Não A	750	50	800
	900	100	1000

Regra A→B

- ***conf*** (A→B) = 150/200 = 0,75 alta.....mas $\widehat{P}(B) = 0,90$ →
- se uma pessoa consome A, a probabilidade de que consuma B cai de 90% para 75% →
- recomendar B apenas para quem A não é interessante
- ***conf*** (A→B) não leva em consideração a $\widehat{P}(B)$ *probabilidade do RHS*

23

Lift: Outra medida para avaliação de uma RA

FGV EDUCAÇÃO EXECUTIVA

Gosto mais de olhar desta forma

$$Lift(A,B) = \frac{conf(A \rightarrow B)}{\sup(B)} = \frac{\sup(A,B)}{\sup(A)\,\sup(B)}$$

- Medida simétrica (A → B ou B→A tem o mesmo lift)

{Bread}=>{Beer}

Sup (Bread)=4/5=0,8

Sup (Beer)=3/5=0,6

Sup (Bread, Beer) = 2/5 = 0,4

Lift (Bread=>Beer)=0,4/0,48 = 0,83

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Cola
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Cola

Comentários

24

Lift: Outra medida para avaliação de uma RA

FGV EDUCAÇÃO EXECUTIVA

$$Lift = \frac{conf(A \rightarrow B)}{\sup(B)} = \frac{\sup(A,B)}{\sup(A)\,\sup(B)}$$

- Lift avalia o relacionamento entre A e B
- Se A e B forem independentes, sup(A,B)= sup(A).sup(B) → ***Lift=1*** → não faz sentido derivar uma RA
- ***Lift < 1*** → conf (A→B) < sup (B) → A e B negativamente relacionados →
  - ✓ A terá efeito negativo sobre B → regra não é útil *per se*
    - **No exemplo acima, 60% dos clientes consomem Beer**
    - **A conf({Bread}=>{Beer}) = 2/4=0,5 , ou seja, se restringimos a quem consome Bread, consumo de Beer cai para 50%**
- ***Lift > 1*** → conf (A→B) > sup (B) → A e B positivamente relacionados →
  - ✓ A terá efeito positivo sobre B → A e B itens complementares → regra é util

Cálculo do *lift*

- Retomando exemplo anterior

$$Lift = \frac{conf(A \rightarrow B)}{\sup(B)}$$

	B	Não B	
A	150	50	200
Não A	750	50	800
	900	100	1000

Regra A→B

sup (B) = 90% ***conf*** (A→B) = 0,75 altos .....

$$\textit{Lift}\ (A \rightarrow B) = \frac{0{,}75}{0{,}9} = 0{,}83 \quad (menor\ que\ 1)$$

- ***Lift*** mostra a capacidade da regra em encontrar bons **consequentes** (LHS)

26

Lift pode não ser conveniente para avaliar RA

	B	Não B	
A	880	50	930
Não A	50	20	70
	930	70	1000

	D	Não D	
C	20	50	70
Não C	50	880	930
	70	930	1000

Lift (A,B)= 0,88 / (0,93×0,93) = 1,02

Lift (C,D) = 0,02/ (0,07×0,07) = **4,08**

conf (A→B) = 0,88/0,93 = **0,95**

conf (C→D) = 0,02/0,07 = 0,29

lift não mostra que a relação (A,B) é mais interessante que (C,D)

conf seria melhor métrica!!!

27

Outra métrica: ϕ-Coefficient

- ϕ: análogo ao coeficiente de correlação para variáveis quantitativas
- Regra X➔ Y

	Y	Não Y	
X	f_{11}	f_{10}	f_{1+}
Não X	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	T

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$$

ϕ Coefficient is the same for both tables

28

Outra métrica: ϕ-Coefficient

- ϕ: análogo ao coeficiente de correlação para variáveis quantitativas

	Y	$\overline{Y}$	
X	60	10	70
$\overline{X}$	10	20	30
	70	30	100

	Y	$\overline{Y}$	
X	20	10	30
$\overline{X}$	10	60	70
	30	70	100

$$\phi = \frac{0.6 - 0.7 \times 0.7}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238 \qquad \phi = \frac{0.2 - 0.3 \times 0.3}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238$$

ϕ Coefficient is the same for both tables

Cliente & transação

Respostas no fim

Baseado em :
Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, "Introduction to Data Mining", Pearson Addison Wesley, 2006,

Transações	clientes	Itens adquiridos
1	Joao	a d e
2	Joao	a b c e
3	Maria	a b d e
4	Maria	a c d e
5	Lucia	b c e
6	Lucia	b d e
7	Lucia	c d
8	Pedro	a b c
9	Pedro	a d e
10	Jaci	a b e

a) Transforme em matriz binária considerando as transações independentemente do cliente

b) Calcular o suporte, confiança e lift para **{b,d} → {e}** e para **{e} →{b,d}** tendo como base as transações

c) Tratando cada cliente como uma transação: construa uma matriz binária onde um item recebe o valor 1 se foi adquirido alguma vez pelo cliente; 0 caso contrário. Calcular o suporte, confiança e lift para **{b,d} → {e}** e para **{e} →{b,d}** tendo como base as os clientes

d) Compare os resultados obtidos em (b) e (c). Discuta. Qual o mais interessante?

30

Exercício

- As RA podem ser selecionadas eliminando (podando) as que tiverem baixo suporte (ou baixa confiança ou lift<1,0)

- Gere aleatoriamente 5000 tabelas do tipo (dica: Excel). Os valores em cada uma das quatro caselas f_{ij} devem ser valores inteiros entre 0 e 1000

	B	Não B	
A	f_{11}	f_{10}	
Não A	f_{01}	f_{00}	

- Para cada uma calcule suporte, confiança, lift e correlação ϕ
  - Verifique se há uma relação entre as métricas duas a duas (dica: utilize gráficos de dispersão)
  - Remova regras com suporte abaixo de 0,40 e compare o comportamento das demais métricas entre regras podadas e não podadas (utilize box-plots)
  - Repita podando quando a confiança é inferior a 0,50

31

Geração das regras de associação

32

Algoritmos para Regras de Associação

Abordagem básica :

Passo 1: Geração de itemsets frequentes ($s \geq min\_sup$)

- Parte que requer maior esforço computacional
- Na prática podemos ter milhares de itens e milhões de transações
- Importante ter um algoritmo que pode (elimine) gradativamente itemsets que não serão uteis a posteriori (algoritmo *APRIORI*)

Passo 2: Geração das regras de associação (dentre os itemsets frequentes, selecionar aqueles com $c \geq min\_conf$, por exemplo)

33

RA sem uso de algoritmo apropriado

Método "força bruta"

- Listar todas as possíveis regras de associação
- Calcular o suporte e confiança de todas as regras
- Eliminar as que não satisfazem min_sup e min_conf

- Inviável!!!!!! Computacionalmente proibitivo

Se d= # itens , via análise combinatória calculamos o número R de possíveis regras de associação $R = 3^d - 2^{d+1} + 1$

Numero de associações

- Se temos **d** itens ➔ número total de itemsets $= \sum_{i=1}^{d} \binom{d}{i} = 2^d - 1$ (exclui vazio)
- Pode-se provar (Tan et al., 2006) que o número de possíveis regras de associação quando temos d itens é igual a $3^d - 2^{d+1} + 1$ (verifique para d=4 identificando primeiro os LHS e depois os possíveis RHS associados a cada LHS)

d	itemsets	regras potenciais
2	3	2
3	7	12
4	15	50
5	31	180
6	63	602
7	127	1.932
8	255	6.050
9	511	18.660
10	1.023	57.002
20	1.048.575	3.484.687.250
30	1.073.741.823	205.888.984.611.002
40	1.099.511.627.775	12.157.663.260.033.700.000
50	1.125.899.906.842.620	717.897.985.440.053.000.000.000

Qual deve ser o valor de d na Amazon.com?

35

Algoritmos para mineração de regras

- Há grande número de algoritmos
  - Estratégias computacionais e requisitos de memória variam, mas....

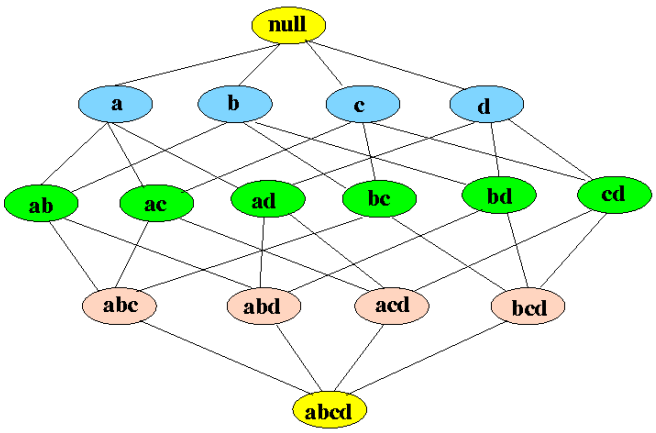
As regras de associação resultantes para min_sup e min_conf pré definidos serão sempre as mesmas

- Vamos estudar o algoritmo mais popular **: o *Apriori Algorithm***

Itemset lattice 4 itens

d itens ➔ 2^d -1 possíveis itemsets (excluindo conjunto vazio)

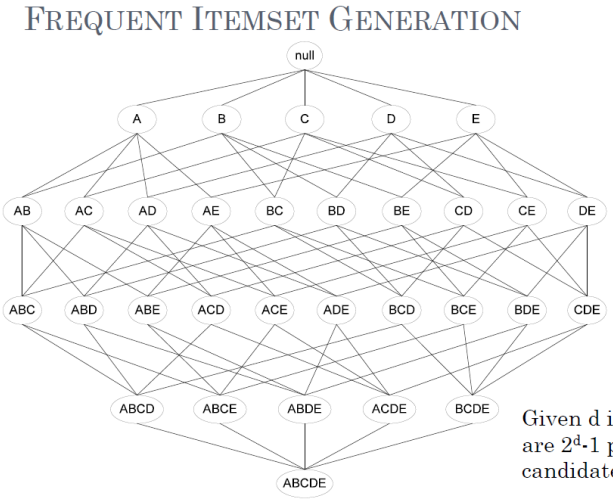
Itemset lattice: representação gráfica dos itemsets



https://www.datacamp.com/community/tutorials/market-basket-analysis-r#code

Itemset lattice 5 itens

FGV EDUCAÇÃO EXECUTIVA



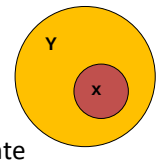
38

Princípio do algoritmo *Apriori*

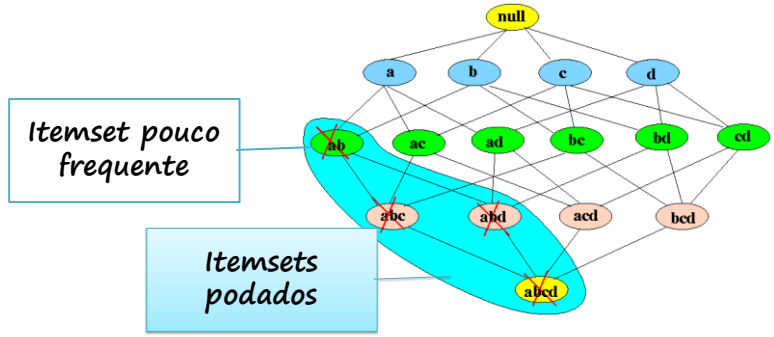
FGV EDUCAÇÃO EXECUTIVA

Propriedade do suporte

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$



- Se [*Y contem X*] e [*X não é frequente*] ➔ Y não é frequente
- Se um itemset é frequente, todos seus subconjuntos são frequentes
- Se um itemset não é frequente, todos os itemsets que o contem não são frequentes (**pequenininho=ruim ➔ grandão=ruim**)



https://www.datacamp.com/community/tutorials/market-basket-analysis-r#code

Aplicação

TRANSAÇÕES
A,B
B,C,D,E
A,C,D
B,C,D

ITEMSET	SUP
A	2
B	3
C	3
D	3
E	1

1

ITEMSET	SUP
A,B	1
A,C	1
A,D	1
B,C	2
B,D	2
C,D	3

2

ITEMSET	SUP
B,C,D	2

- vamos definir (por exemplo) **minsup = 40%** (*ou seja, 2 ocorrências pois T=4*)
- Em (1) eliminamos todos os itemsets que contem E
- Em (2) eliminamos itemsets que contem AB, AC, AD
- Geramos 5 + 6 + 1 = 12 itemsets (**com d=5 seria 31 se não podássemos**)
- Itemsets frequentes: {A}, {B}, {C}, {D}, **{B,C}, {B,D}, {C,D}, {B,C,D}**

40

Algoritmo Apriori – Geração de itemsets

- Atravessa o *lattice itemset* um nível a cada rodada

- A cada nível:
  - a cada rodada **gera** os candidatos itemset a partir dos itemsets gerados na etapa anterior
    - Há diversos algoritmos para acelerar o processo de geração (Tan et al.pp339)
  - **Testa** candidatos utilizando o critério min_sup
  - **Poda** os itemsets os pouco frequentes

41

Geração das *regras de associação*

Geração de itemsets frequentes ≠ geração de regras de associação

Propriedade útil para reduzir busca de regras :

A confiança de regras de associação geradas a partir do mesmo itemset tem a seguinte propriedade: seja o itemset X= {A,B,C,D}:

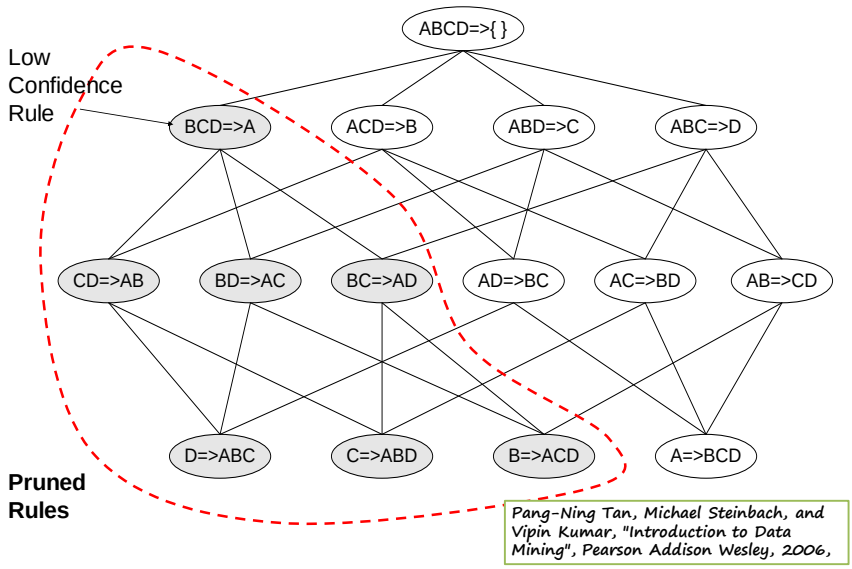
$$conf(ABC \rightarrow D) \geq conf(AB \rightarrow CD) \geq conf(A \rightarrow BCD)$$

Aumentando o número de itens do consequente → diminui a confiança (propriedade anti-monotónica)

Implicação – vide slide seguinte

Geração das regras de associação

Lattice of rules (≠ lattice de itemsets)



Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, "Introduction to Data Mining", Pearson Addison Wesley, 2006,

Exercício

FGV EDUCAÇÃO EXECUTIVA

TRANSAÇÕES
A,B
B,C,D,E
A,C,D
B,C,D

- Gerar as regras de associação para a matriz de transações abaixo.
- Utilizar **minsup=40% e minconf=80%**
- Anteriormente (slide 38) obtivemos:

  Itemsets frequentes: {A}, {B}, {C}, {D}, **{B,C}, {B,D}, {C,D} , {B,C,D}**

Itemsets frequentes	regras de associação
{B,C}	**B➔C C➔B**
{B,D}	
{C,D}	
{B,C,D}	

- **Que RA's Você selecionaria?**

44

FGV EDUCAÇÃO EXECUTIVA

Aplicação com R

45

Geração de dados no formato *"transaction"*

FGV EDUCAÇÃO EXECUTIVA

- **Formato necessário para rodar o algoritmo**

- ***Caso 1: matriz binária para transactions***

TID	Bread	Milk	Diapers	Beer	Eggs	Coke
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

```
#Seja  "a.mat"  o nome da matriz binária
# gerando a matriz de transações para pode rodar o algoritmo
>library(arules)
>a.tr=as( a.mat,  "transactions")
# a.tr estará no formato transactions
                                    # aplicação no caso CharlesBookClub
```

46

Geração de dados no formato *"transaction"*

FGV EDUCAÇÃO EXECUTIVA

- ***Caso 2: lista em Excel para transactions***

BASKETS
citrus fruit,semi-finished bread,margarine,ready soups
tropical fruit,yogurt,coffee
whole milk
pip fruit,yogurt,cream cheese ,meat spreads
other vegetables,whole milk,condensed milk,long life bakery product
whole milk,butter,yogurt,rice,abrasive cleaner
rolls/buns

```
# formato Excel: cada linha uma transação; itens separados por vírgulas
> GL=groceries list  #simplifica digitação
# transformar em matriz  .csv  para preparar o próximo passo
> write.csv(GL, "GL_basket.csv", quote=FALSE, row.names=TRUE )
> tr <- read.transactions('GL_basket.csv', format = 'basket', sep=',')
> summary(tr)
                    # aplicação no caso groceries list ➔ script R:  groceries list.R
```

Estudo de caso 1 – Charles Book Club

The problem:
CBC sent mailings to its club members each month containing its latest offering. …. The decreasing profits led CBC to revisit their original plan of using database marketing to improve its mailing yields and to stay profitable.

"This Dataset is part of a case prepared by Ms. Vinni Bhandari, a data mining consultant and Dr. Nitin Patel, a visiting professor of Operations Research at the MIT Sloan School of Management. The case has been derived from a Case Study in Database Marketing titled 'BBB - The Bookbinders Club' prepared by Nissan Levin and Jacob Zahavi, Tel Aviv University for the Direct Marketing Educational Foundation, Inc. (March 1995)."

Estudo de caso 1 – Charles Book Club

ID#	Gender	M	R	F	FirstPurch	ChildBks	YouthBks	CookBks	DoItYBks	RefBks	ArtBks	GeoBks	ItalAtlas	ItalArt
25	1	297	14	2	22	0	1	1	0	0	0	0	0	0
29	0	128	8	2	10	0	0	0	0	0	0	0	0	0
46	1	138	22	7	56	2	1	2	0	1	0	1	0	0
47	1	228	2	1	2	0	0	0	0	0	0	0	0	0
51	1	257	10	1	10	0	0	0	0	0	0	0	0	0
60	1	145	6	2	12	0	0	0	0	0	0	0	0	0
61	1	190	16	1	16	0	0	0	0	0	0	1	0	0
79	1	187	14	1	14	1	0	0	0	0	0	0	0	0
81											0	0	0	0
90											0	0	0	0
95											0	0	0	0
100	0	320	2	3	18	0	0	0	0	0	1	0	0	0

M Monetary—Total money spent on books
R Recency—Months since last purchase
F Frequency—Total number of purchases
FirstPurch Months since first purchase

Script do R ➔ CBOOK.R

Estudo de caso 2 = GROCERIES_LEVEL2

- Transações originais estão no R (ou na planilha ***groceries list)***
- Os produtos foram agrupados em categorias. LEVEL2 representa essas categorias ➔ trabalharemos com script ***groceries level2.R***

> data(Groceries) #já está na biblioteca do R em formato de transactions
> head(Groceries@itemInfo,15)

	labels	level2	level1
1	frankfurter	sausage	meat and sausage
2	sausage	sausage	meat and sausage
3	liver loaf	sausage	meat and sausage
4	ham	sausage	meat and sausage
5	meat	sausage	meat and sausage
6	finished products	sausage	meat and sausage
7	organic sausage	sausage	meat and sausage
8	chicken	poultry	meat and sausage
9	turkey	poultry	meat and sausage
10	pork	pork	meat and sausage
11	beef	beef	meat and sausage
12	hamburger meat	beef	meat and sausage
13	fish	fish	meat and sausage
14	citrus fruit	fruit	fruit and vegetables
15	tropical fruit	fruit	fruit and vegetables

Diferentes níveis de hierarquia (vamos discutir adiante)

Estudo de caso 3 = OnlineRetailamostracomplicado

- Base de dados extraída do UCI
- Arquivo original muito grande ➔ demora muito para rodar
- Selecionamos uma amostra que está na planilha ***OnlineRetailamostracomplicado***
- Preparação dos dados é mais complexa devido ao formato da base original ➔ ver script ***online_complicado.R***

Alunos deverão terminar o estudo de caso

Banco de dados

- Dados para elaboração de exercícios

Banco de dados Association Rules 2020 mês dia

52

Análise e avaliação das regras de associação geradas

53

Análise das RA

- Número de itens em casos reais podem gerar milhares ou milhões de regras de associação
  - Imagine o número de itens comercializados pela
    - –Amazon,
    - –Lojas Americanas
    - –Carrefour
    - –Magalu....
- Analista deve verificar a forma de eliminar regras que não sejam muito interessantes para a tomada de decisões.

54

Critérios para avaliação das regras - I

Medidas objetivas de grau de interesse ("*interestingness*")

- Uso de indicadores quantitativos
- Baseados em estatísticas
- Medidas: suporte, confiança, lift, correlação…
  - Literatura apresenta grande número de medidas
- Ordena regras de acordo com medida escolhida
- *Risco (um problema sério): não dependem do contexto do problema (domain-independent)*

Medidas objetivas de grau de interesse

- Na literatura podem ser encontradas diversas medidas para avaliar uma regra de associação
- Nem sempre essas diferentes medidas levam às mesmas decisões quanto à utilidade / interesse de uma RA.
- Tan et al. Compararam 10 regras de associação diferentes E1,...,E10
- Tabela seguinte apresenta resultados comparando as regras com diferentes medidas (há muitas outras) (1: mais interessante, 10 menos interessante)

Table 6.14. Rankings of contingency tables using the symmetric measures given in Table 6.11.

	ϕ	α	κ	I	IS	PS	S	ζ	h
E_1	1	3	1	6	2	2	1	2	2
E_2	2	1	2	7	3	5	2	3	3
E_3	3	2	4	4	5	1	3	6	8
E_4	4	8	3	3	7	3	4	7	5
E_5	5	7	6	2	9	6	6	9	9
E_6	6	9	5	5	6	4	5	5	7
E_7	7	6	7	9	1	8	7	1	1
E_8	8	10	8	8	8	7	8	8	7
E_9	9	4	9	10	4	9	9	4	4
E_{10}	10	5	10	1	10	10	10	10	10

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, "Introduction to Data Mining", Pearson Addison Wesley, 2006,

56

Critérios para avaliação das regras - II

- **Argumentos subjetivos de interesse**
- Ordenam regras de acordo com crenças e interesses do analista
- Risco: crenças e interesses dos analistas podem variar de acordo com experiências, interesses, objetivos, área de atuação e ao longo do tempo em função de novos aprendizados
- (Silberschatz & Tuzhilin,1996) O grau de interesse em uma regra depende de:
  - (***unexpected pattern***) Uma regra pode ser considerado interessante se contradiz expectativa do analista / surpreende o analista
    - Pão → Margarina : óbvio não revela nada novo!
    - Fraldas → Cerveja: interessante, relação inesperada!!!
  - (***actionable pattern***) Uma regra é interessante se puder ser aplicada / implementada de forma vantajosa e rentável.
  - Autores destacam que um ***unexpected pattern*** em geral é ***actionable pattern, e vice-versa.*** Será?

57

Critérios para avaliação das regras - II

- **Argumentos subjetivos de interesse**
  - Ordenam regras de acordo com **crenças** e interesses do analista
  - Risco: crenças e interesses dos analistas podem variar de acordo com experiências, interesses, objetivos, área de atuação e ao longo do tempo em função de novos aprendizados
  -

Palavrinha complicada
Literatura discute bastante este conceito

58

Cuidado com o acaso.....

- Base de dados gerados aleatoriamente podem dar origem a RA interessantes do ponto de vista dos indicadores estatísticos. Ver livro da G. Shmueli
- Em casos reais, quanto mais regras produzimos, maior o perigo de gerar regras que não fazem sentido no contexto do problema (regras espúrias) (apesar de apresentar bons indicadores estatísticos)
- Regras baseadas em grande número de itens estão menos sujeitas a serem regras espúrias

59

RA redundantes

- Algumas das regras geradas podem ser redundantes:
  - Regra 1: {a,b} → {d} é uma regra "mais geral" que Regra 2: {a,b,c} → {d}
  - Se a primeira regra tiver *confiança* maior ou igual que a segunda, a segunda é redundante
  -
- Do manual de `arules`:
  - A rule is more general if it has the same RHS but one or more items removed from the LHS ("menor LHS"). ← Regra 1 é "mais geral" que Regra 2
  - A rule is redundant if a more general rules with the same or a higher confidence exists.
    - Regra 2 é redundante se conf(Regra1)>= conf(Regra2)
  - a rule X → Y is redundant if for some X' (subset of) X,

    conf(X' -> Y) >= conf(X -> Y)
  - Other measures than confidence, e.g. improvement of lift, can be used as well.

```
inspect(rules[is.redundant(rules, measure="lift")])
rules.pruned=rules[!is.redundant(rules, measure="confidence")]
```

Redundância em CharlesBookClub

- [18] {ChildBks,DoItYBks} => {CookBks} 0.16346 0.7910 1.4878 511
- [19] {ChildBks,DoItYBks,ArtBks} => {CookBks} 0.05438 0.7906 1.4871 170

```
> inspect(rules.sorted[is.redundant(rules.sorted)])
# output em branco se não houver regras redundantes
```

- lhs rhs support confidence lift count
- **[1] {ChildBks,DoItYBks,ArtBks} => {CookBks} 0.0543826 0.7906977 1.487197 170**
- [2] {YouthBks,DoItYBks,GeoBks} => {ChildBks} 0.0537428 0.7706422 1.528571 168
- Etc

Lift <1 → produtos "competem"

FGV EDUCAÇÃO EXECUTIVA

```
                          support    confidence      lift  count
{ArtBks} => {CookBks}  0.14459373 0.5067265  0.9530848  452
```

```
  ChildBks   YouthBks    CookBks   DoItYBks     RefBks
0.50415867 0.30486244 0.53166987 0.32597569 0.26199616
   ArtBks     GeoBks  ItalAtlas    ItalArt
0.28534869 0.34133077 0.04158669 0.05534229
```

CookBks tem 53,16% de probabilidade de ser vendido

Se oferecemos somente a quem compra ArtBks, probabilidade cai para 50,67%

Tratamento de dados não binários

FGV EDUCAÇÃO EXECUTIVA

- Até agora, temos trabalhado com 1-0 (sim-não)

TID	Bread	Milk	Diapers	Beer	Eggs	Coke
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
----	----					

- Há situações em que nosso interesse envolve variáveis **qualitativas** ou **quantitativas**. Exemplo analise de padrões de consulta a redes sociais

ID	Sexo	Rede mais utilizada	Meio usual	Idade	Compra pela internet	Educação
1	M	Facebook	Celular	21	sim	sup
2	F	Instagram	Celular	19	sim	sup
3	F	Linkedin	Comput.	56	não	secund
----	----					

Variáveis qualitativas

- Transformar as k categorias de uma vari em k dummies

ID	Sexo	Rede mais utilizada	Meio usual	Idade	Compra pela internet	Educação
1	M	Facebook	Celular	21	sim	sup
2	F	Instagram	Celular	19	sim	sup
3	F	Linkedin	Comput.	56	não	secund
----	----					

ID	M	F	Facebook	instagram	Linkedin	Celular	Computador	
1	1	0	1	0	0	1	0	
2	0	1	0	1	0	1	0	Etc.
3	0	1	0	0	1	0	1	
----	----							

Variáveis qualitativas - comentários

- **Categorias com baixa frequência**
  - Provavelmente serão ignoradas por baixo suporte
  - Recomenda-se fundir com outras categorias
    - Exemplo: fundir UF em regiões, ou, fundir todas as UF poucos frequentes.....(depende o contexto do problema)

- **Categorias com alta frequência** podem gerar regras redundantes
  - {tem PC=sim, compra online=sim} → {curso=superior}
    - quase todos tem PC, não agrega muita informação
  - { compra online=sim} → {curso=superior} ← regra mais geral
    - a primeira é provavelmente redundante
  - Sugestão: remover redundantes

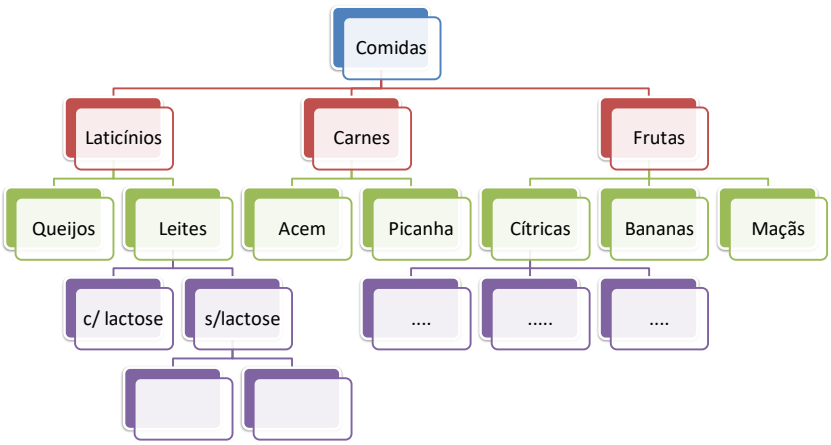
- Algumas **associações podem não fazer sentido**:
  - (UF=SP, UF=RJ, compra on line = sim}
    - não faz sentido, além do mais suporte=0%
    - ocorre pois cada UF é uma dummy independente
    - Ideal se algoritmos eliminasse dummies associadas (não conheço)

Variáveis quantitativas

- **Transformação :**
  1. discretização (mesma frequência, mesma amplitude, supervisionada..) → depois geração de dummies
  2. Geração das dummies correspondentes a cada categoria
- **Problemas**
  - Em quantas intervalos discretizar?
  - Qual a melhor forma de discretizar?
  - Intervalos muito estreitos → baixo suporte → podem ser ignorados
  - Intervalos muito largos → podem levar a menor confiança → desprezados
- **Alternativas**
  - Testar diferentes forma de discretização (número de intervalos / amplitude dos intervalos)
  - Problema: demora em obter e compara diferentes regras

Hierarquias

- Às vezes (milhares de produtos) é preferível trabalhar com diferentes níveis



Hierarquias

- Às vezes (milhares de produtos) é preferível trabalhar com diferentes níveis
- Analista deve escolher níveis adequados em função dos objetivos do problema
- Itens na base da pirâmide → podem ser pouco frequentes (suporte baixo)
- Itens no topo da pirâmide → muito gerais → pouco úteis!
  - Vejam os menus dos supermercados online
    - Comidas
    - Bebidas
    - Limpeza....
- Alternativa sugerida na literatura: colocar "pais" e filhos" como se fossem itens diferentes. Selecionar as regras convenientes
  - Número giigate de itens → complica computação / interpretação
  - Muitas regras redundantes ou sem valia
    {queijo prato, cerveja} → {laticínios}

Resposta por Transação

B,D	-->	E	
	E	~E	
BD	2	0	2
~BD	6	2	8
	8	2	10
SUP=	0,20		
CONF	1,00		
LIFT	1,25		

E	-->	B,D	
	BD	~BD	
E	2	6	8
~E	0	2	2
	2	8	10
SUP=	0,2		
CONF	0,25		
LIFT	1,25		

Resposta por cliente

B,D	-->	E	
	E	~E	
BD	4	0	4
~BD	1	0	1
	5	0	5
SUP=	0,80		
CONF	1,00		
LIFT	1,00		
E	**-->**	**B,D**	
	BD	~BD	
E	4	1	5
~E	0	0	0
	4	1	5
SUP=	0,8		
CONF	0,8		
LIFT	1		