Market Basket Anaysis

4ºAno M, AN,FZ,EN-MEC,EN-AEL

V 1.0, V.Lobo & R.Moura, EN 2021

Market Basket Analysis (regras de associação)

Victor Lobo e Ricardo Moura

4.º ano dos cursos tradicionais da Escola Naval

1

Regras de Associação

- Genericamente, é o estudo do "O que vai com o quê?"
- Também chamadas de *Market Basket Analysis*
  - *Origem no estudo das bases de dados das compras efetuadas por clientes de modo a determinar a dependência entre diferentes itens*

2

Regras de Associação

- Estabelecer padrões habituais e associação ou correlação entre itens. (variáveis maioritariamente discretas)
  - Algoritmo à priori
  - Entender os hábitos do 'consumidor'
- Pode ser utilizado para promover um produto (ou até mesmo aumentar o preço)
- Definir posição de um produto no supermercado
  - Colocar produtos cuja associação foi estabelecida em localizações próximas.

3

Regras de Associação

Market Basket Analysis – Quem compra fraldas tem uma enorme tendência para comprar cerveja! (Thomas Blischok, 1992)

- Padrão encontrado em 50 lojas.

Na década de 90, poderia ser habitual as mães que estariam a tomar conta dos seus filhos pedirem para comprar fraldas e, no percurso, estes decidiam comprar cervejas.

Decisão possível: Separar os dois itens para locais onde constariam itens de valores elevados que levariam os jovens pais a comprar tais produtos.

4

Regras de Associação

- Como descrever a informação histórica:
  - Através de expressões "If-then":

$$A \,\&\, B \implies C$$

(sendo a qualidade destas avaliada por medidas estatísticas)

- Inclui-se todas as regras?
- Como se procede à definição destas regras?
- Existem regras boas e más?

5

Dados das compras de pratos numa loja de artigos do lar

- Em primeiro lugar será necessário converter para dados binários.
- Exemplos de regras só para as cores {Vermelho, Branco e Verde}:
  - Se {Vermelho, Branco} então {verde}
  - Se {Vermelho, Verde} então {Branco}
  - Se {Branco, Verde} então {Vermelho}
  - Se {Branco} então {Vermelho, Verde}
  - Se {Vermelho} então {Branco, Verde}
  - Se {Verde} então {Vermelho,Branco}

Registo número	Cores dos pratos comprados			
1	Vermelho	Branco	Verde	
2	Branco	Laranja		
3	Branco	Azul		
4	Vermelho	Branco	Laranja	
5	Vermelho	Azul		
6	Branco	Azul		
7	Branco	Laranja		
8	Vermelho	Branco	Azul	Verde
9	Vermelho	Branco	Azul	
10	Amarelo			

Só três itens e já demasiadas regras!!!!

6

Market Basket Anaysis

4ºAno M, AN,FZ,EN-MEC,EN-AEL

V 1.0, V.Lobo & R.Moura, EN 2021

Dados das compras de pratos numa loja de artigos do lar

- Terminologia para as regras
  - "Se Vermelho e Branco, então Verde"
    - Antecedente: Vermelho e Branco
    - Consequente: Verde
- Suporte de uma regra: % (ou número) de ocorrências onde o antecedente e consequente aparecem nos dados
  - Para a regra acima, teremos o valor de 20% ou 2.

Registo número	Cores dos pratos comprados			
1	Vermelho	Branco	Verde	
2	Branco	Laranja		
3	Branco	Azul		
4	Vermelho	Branco	Laranja	
5	Vermelho	Azul		
6	Branco	Azul		
7	Branco	Laranja		
8	Vermelho	Branco	Azul	Verde
9	Vermelho	Branco	Azul	
10	Amarelo			

7

Dados das compras de pratos numa loja de artigos do lar

- Algoritmo Apriori [Agrawal & Srikant 94]
  - Definir um critério – suporte mínimo a considerar
  - Criar lista de todos os conjuntos de apenas um elemento que respeitam o critério
  - Usar lista anterior para gerar lista de todos os conjuntos de dois elementos que respeitem o critério
  - Continuar da mesma forma até obter a lista de conjuntos com k-elementos.

Registo número	Cores dos pratos comprados			
1	Vermelho	Branco	Verde	
2	Branco	Laranja		
3	Branco	Azul		
4	Vermelho	Branco	Laranja	
5	Vermelho	Azul		
6	Branco	Azul		
7	Branco	Laranja		
8	Vermelho	Branco	Azul	Verde
9	Vermelho	Branco	Azul	
10	Amarelo			

8

Dados das compras de pratos numa loja de artigos do lar

- Critério: Suporte mínimo 2

Itens	Suporte
Vermelho	5
Branco	8
Azul	5
Laranja	3
Verde	2
Vermelho, Branco	4
Vermelho, Azul	3
Vermelho, Verde	2
Branco, Azul	4
Branco, Laranja	3
Branco, Verde	2
Vermelho, Branco, Azul	2
Vermelho, Branco e Verde	2

Registo número	Cores dos pratos comprados			
1	Vermelho	Branco	Verde	
2	Branco	Laranja		
3	Branco	Azul		
4	Vermelho	Branco	Laranja	
5	Vermelho	Azul		
6	Branco	Azul		
7	Branco	Laranja		
8	Vermelho	Branco	Azul	Verde
9	Vermelho	Branco	Azul	
10	Amarelo			

9

Dados das compras de pratos numa loja de artigos do lar

- Confiança: % de antecedentes que também implicam ocorrência de consequente

$$\frac{\#\,(antecedente\ \&\ consequente)}{\#\,antecedente}$$

- A confiança da regra "Se azul então branco" é de

  4/5 = 80%

- A confiança da regra "Se Branco então azul" é de

  4/8 =1/2 =50%

Registo número	Cores dos pratos comprados			
1	Vermelho	Branco	Verde	
2	Branco	Laranja		
3	Branco	Azul		
4	Vermelho	Branco	Laranja	
5	Vermelho	Azul		
6	Branco	Azul		
7	Branco	Laranja		
8	Vermelho	Branco	Azul	Verde
9	Vermelho	Branco	Azul	
10	Amarelo			

Suporte(antecedente e/ou consequente) alto implica confiança alta.

10

Medidas de Performance alternativas

- Lift (mede a independência entre A e B: 1 ⇒ independência)

$$Lift(A \Rightarrow B) = \frac{Confiança(A \Rightarrow B)}{suporte(B)}$$

- Convicção (grau de implicação: 1 ⇒ independência)

$$Convic(A \Rightarrow B) = \frac{1 - suporte(B)}{1 - confiança(A \Rightarrow B)}$$

- Leverage (número de casos adicionais obtidos em relação ao esperado/independência)

$$Leverage(A \Rightarrow B) = suporte(A \cup B) - suporte(A) \times suporte(B)$$

- Reliability ()

$$Reliability(A \Rightarrow B) = confiança(A \Rightarrow B) - suporte(B)$$

- Teste de Qui-quadrado, etc.

11

Dados não categóricos

(numéricos, ordinais, etc.)

- Convém fazer um pré-processamento com os seguintes cuidados:
  - Ao discretizar ou a transformar em binário, podemos gerar demasiadas regras.
  - Itens de intervalos pequenos podem implicar suportes demasiado pequenos e intervalos grandes leva a confiança demasiado baixa.
  - Leva a perda de informação por se estar a agregar valores num mesmo intervalo!

12

Market Basket Anaysis

4ºAno M, AN,FZ,EN-MEC,EN-AEL

V 1.0, V.Lobo & R.Moura, EN 2021

Dados não categóricos

(numéricos, ordinais, etc.)

- Dados gerados aleatoriamente aparentemente têm regras de associação interessantes!
- Quantas mais regras produzirmos, maior o perigo de não extrairmos informação relevante!
- Regras baseadas em grande número de registos estão sujeitas a um menor perigo da chamada associação proveniente do acaso!

13

O que fazer com a informação

- Charles P. Elkan no KDNuggets (site dedicado a Machine Learning) 97:35
  - O Wal-Mart sabe que dos clientes que compram bonecas da Barbie, 60% têm tendência a comprar também um de três tipos de chocolates. O que pode fazer o Wal-Mart com esta informação?
  - Resposta do chefe de Comércio do Wal-Mart: "Não faço a mínima ideia!"
- Respostas interessantes:
  - Alterar a forma dos chocolates na sua produção para o formato de Barbies.
  - Aumentar o preço das Barbies e oferecer o chocolate de graça.
  - Juntar um produto que venda mal e oferecer um desconto neste se fizerem a compra conjunta de uma Barbie e de um desses chocolates.

14