ISLab

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Regras de Associação para Extração de Conhecimento

Aprendizagem e Extração de Conhecimento

Perfil Sistemas Inteligentes @ MiEI/4º – 1º Semestre

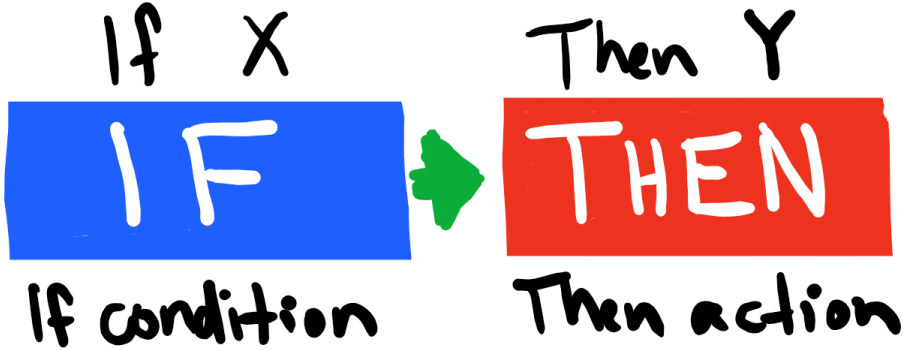
Cesar Analide, Filipe Gonçalves, Paulo Novais

ISLab
Synthetic Intelligence Lab

Definições

- Regras de Associação são declarações da forma

  SE <condição> ENTÃO <conclusão>



- A descoberta de Regras de Associação é uma técnica de aprendizagem sobre como “coisas” se relacionam.
- Regras de Associação permitem conhecer a força da relação entre ocorrências de grandes *datasets*.

ISLab
Synthetic Intelligence Lab

Objetivos

- **Encontrar padrões frequentes**, associações, correlações ou estruturas ocasionais em conjuntos de dados (*datasets*);
- A descoberta de **Regras de Associação** é usada para encontrar **elementos que ocorrem conjuntamente** em *datasets*;
- Definição de **regras de relacionamento** (correlação(*) ou implicação(**)) entre elementos que ocorrem em comum.

(*) correlação: medição do relacionamento (e respetiva direção) entre duas variáveis aleatórias (correlação de Pearson);

(**) implicação: relacionamento de causa e consequência; condição e ação; antecedente e consequente (implicação lógica);

ISLab
Synthetic Intelligence Lab

Exemplos

- Entender hábitos de consumo pela descoberta de associações ou correlações entre diferentes produtos;

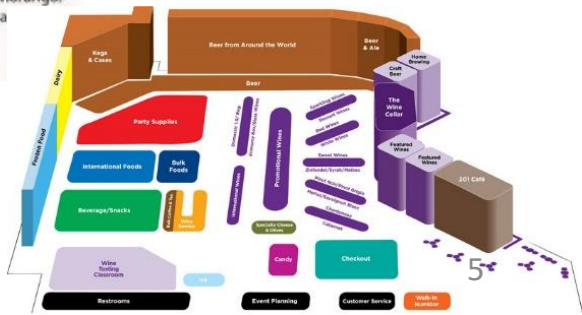
- Se um consumidor compra uma dúzia de ovos, a probabilidade de que também compre leite é de 80%:
  - Como interpretar esta regra?

- Exemplo "canónico" das Fraldas e da Cerveja:
  - "Homens entre os 30 e os 40 anos de idade fazendo compras entre as 17h e as 19h de sexta-feira que compram fraldas, apresentam uma grande probabilidade de comprar, também, cerveja".
  - Conhecendo-se a relação entre a compra de fraldas e de cerveja aconselha-se o vendedor a:
    - Colocar os produtos próximos para potenciar ainda mais a sua venda?
    - Afastar os produtos, dispondo "no caminho" do consumidor outros produtos que tenha interesse em vender/promover?
    - Oferecer melhores condições (descontos, promoções) na compra de apenas 1 dos produtos?
    - Combinar os dois em 1 só produto com diferentes características?

ISLab
Synthetic Intelligence Lab

Aplicações

- Situações em que se deverá ponderar a aplicação de técnicas de DM para a descoberta de Regras de Associação:
  - Análise de hábitos de consumo (*market basket analysis*): descobrir grupos ou conjuntos de produtos que o consumidor adquire;
  - *Cross-marketing*: promoção conjunta de produtos de diferentes empresas;
  - Definição de catálogos de compras/ *product clustering*: grupos de produtos por época (*season*);
  - Organização de loja/ *store layout*: disposição dos produtos em loja;
  - Deteção de fraudes (dependente da frequência das ações); "Quem quer ser milionário? (UK)"
  - etc.



ISLab
Synthetic Intelligence Lab

Estrutura

- Estrutura típica de Regras de Associação:

  Antecedente → Consequente [suporte, confiança]

  - O **suporte** e a **confiança** são medidas de interesse definidas pelo utilizador;
  - **Suporte** é uma medida de utilidade:
    - mede a frequência com que os elementos surgem no conjunto de dados;
    - um valor elevado significa que uma grande parte do *dataset* está representado por esses elementos;
  - **Confiança** é uma medida de certeza:
    - corresponde à percentagem de ocorrências do antecedente juntamente com o consequente.

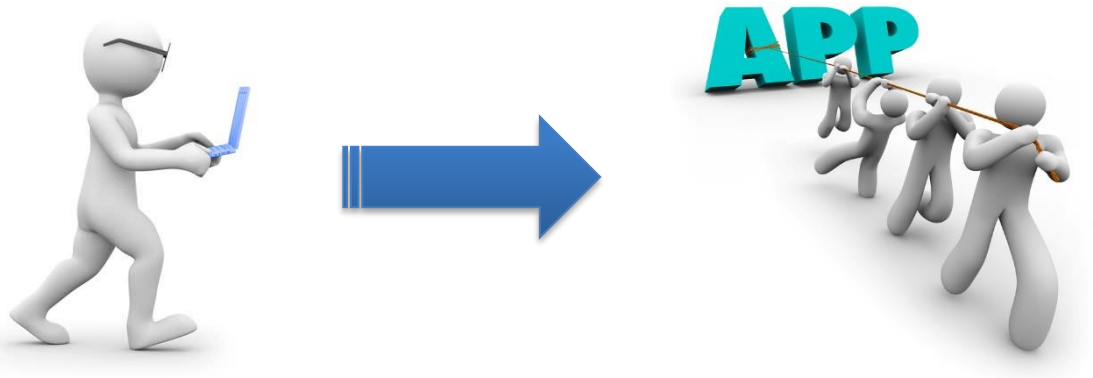
ISLab
Synthetic Intelligence Lab

Exemplo de uma Regra de Associação

- Por exemplo, para a seguinte Regra de Associação:

  computador → gescliente [2%, 60%]

  - Suporte = 2%:
    - em 2% das relações existentes no *dataset* existe uma forte relação entre a compra de um computador e a compra do pacote de software "gescliente";
  - Confiança = 60%:
    - 60% dos clientes que compraram um computador (antecedente) também compraram o pacote de *software* "gescliente" (consequente).



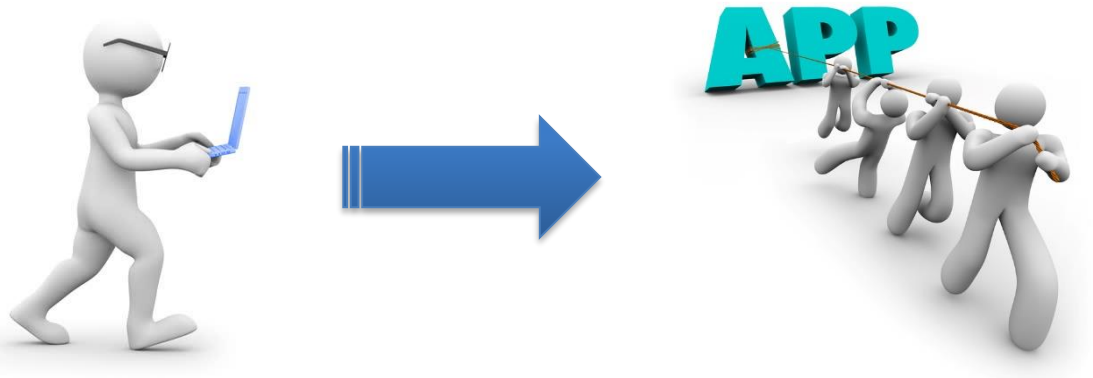
ISLab
Synthetic Intelligence Lab

Interpretação e uso de uma Regra de Associação

- Para a Regra de Associação:

  computador → gescliente [2%, 60%]

  - O **consequente** (*software* "gescliente") pode ser usado para determinar de que forma se pode agir para potenciar as suas vendas;
  - O **antecedente** (computador) pode ser utilizado para antever as consequências da falta de *stock* do produto ou da sua "retirada das estantes";
  - A **relação** entre antecedente e consequente permite considerar a oferta de outros produtos juntamente com o computador para aumentar as vendas.



ISLab
Synthetic Intelligence Lab

Pesquisa de Regras de Associação

- Considere-se o exemplo ao lado em que se descrevem algumas situações relacionadas com as condições atmosféricas, com o intuito de prever a possibilidade de realizar um encontro de uma modalidade desportiva que se realiza ao "ar livre".

(exemplo adaptado de WEKA v.3.6.11)

ID	ASPETO	TEMPERA-TURA	HUMIDADE	VENTO	JOGAR
1	SOL	29	85	FALSO	NÃO
2	SOL	27	90	VERDADEIRO	NÃO
3	NUVENS	28	86	FALSO	SIM
4	CHUVA	21	96	FALSO	SIM
5	CHUVA	20	80	FALSO	SIM
6	CHUVA	18	70	VERDADEIRO	NÃO
7	NUVENS	17	65	VERDADEIRO	SIM
8	SOL	22	95	FALSO	NÃO
9	SOL	20	70	FALSO	SIM
10	CHUVA	24	80	FALSO	SIM
11	SOL	24	70	VERDADEIRO	SIM
12	NUVENS	22	90	VERDADEIRO	SIM
13	NUVENS	27	75	FALSO	SIM
14	CHUVA	22	91	VERDADEIRO	NÃO

ISLab
Synthetic Intelligence Lab

Pesquisa de Regras de Associação

- Discretização de atributos:
  (Temperatura e Humidade)
  - valores contínuos tornam a procura de associações entre variáveis (quase) impossível.

ID	ASPETO	TEMPERA-TURA	HUMIDADE	VENTO	JOGAR
1	SOL	QUENTE	ELEVADA	FALSO	NÃO
2	SOL	QUENTE	ELEVADA	VERDADEIRO	NÃO
3	NUVENS	QUENTE	ELEVADA	FALSO	SIM
4	CHUVA	MODERADO	ELEVADA	FALSO	SIM
5	CHUVA	FRIO	NORMAL	FALSO	SIM
6	CHUVA	FRIO	NORMAL	VERDADEIRO	NÃO
7	NUVENS	FRIO	NORMAL	VERDADEIRO	SIM
8	SOL	MODERADO	ELEVADA	FALSO	NÃO
9	SOL	FRIO	NORMAL	FALSO	SIM
10	CHUVA	MODERADO	NORMAL	FALSO	SIM
11	SOL	MODERADO	NORMAL	VERDADEIRO	SIM
12	NUVENS	MODERADO	ELEVADA	VERDADEIRO	SIM
13	NUVENS	QUENTE	NORMAL	FALSO	SIM
14	CHUVA	MODERADO	ELEVADA	VERDADEIRO	NÃO

ISLab
Synthetic Intelligence Lab

Pesquisa de Regras de Associação

- Antecedente → Consequente

  Humidade=Normal, Vento=Falso → Jogar=Sim
  Suporte = 4/14 (29%); Confiança = 4/4 (100%)

ID	ASPETO	TEMPERA-TURA	HUMIDADE	VENTO	JOGAR
1	SOL	QUENTE	ELEVADA	FALSO	NÃO
2	SOL	QUENTE	ELEVADA	VERDADEIRO	NÃO
3	NUVENS	QUENTE	ELEVADA	FALSO	SIM
4	CHUVA	MODERADO	ELEVADA	FALSO	SIM
5	**CHUVA**	**FRIO**	**NORMAL**	**FALSO**	**SIM**
6	CHUVA	FRIO	NORMAL	VERDADEIRO	NÃO
7	NUVENS	FRIO	NORMAL	VERDADEIRO	SIM
8	SOL	MODERADO	ELEVADA	FALSO	NÃO
9	**SOL**	**FRIO**	**NORMAL**	**FALSO**	**SIM**
10	**CHUVA**	**MODERADO**	**NORMAL**	**FALSO**	**SIM**
11	SOL	MODERADO	NORMAL	VERDADEIRO	SIM
12	NUVENS	MODERADO	ELEVADA	VERDADEIRO	SIM
13	**NUVENS**	**QUENTE**	**NORMAL**	**FALSO**	**SIM**
14	CHUVA	MODERADO	ELEVADA	VERDADEIRO	NÃO

ISLab
Synthetic Intelligence Lab

Pesquisa de Regras de Associação

- Antecedente → Consequente

Humidade=Normal, Vento=Falso → Jogar=Sim
Suporte = 4/14 (29%); Confiança = 4/4 (100%)

Aspeto=Sol, Humidade=Elevada → Jogar=Não
Suporte = 3/14 (21%); Confiança = 3/3 (100%)

ID	ASPETO	TEMPERA-TURA	HUMIDADE	VENTO	JOGAR
1	**SOL**	**QUENTE**	**ELEVADA**	**FALSO**	**NÃO**
2	**SOL**	**QUENTE**	**ELEVADA**	**VERDADEIRO**	**NÃO**
3	NUVENS	QUENTE	ELEVADA	FALSO	SIM
4	CHUVA	MODERADO	ELEVADA	FALSO	SIM
5	CHUVA	FRIO	NORMAL	FALSO	SIM
6	CHUVA	FRIO	NORMAL	VERDADEIRO	NÃO
7	NUVENS	FRIO	NORMAL	VERDADEIRO	SIM
8	**SOL**	**MODERADO**	**ELEVADA**	**FALSO**	**NÃO**
9	SOL	FRIO	NORMAL	FALSO	SIM
10	CHUVA	MODERADO	NORMAL	FALSO	SIM
11	SOL	MODERADO	NORMAL	VERDADEIRO	SIM
12	NUVENS	MODERADO	ELEVADA	VERDADEIRO	SIM
13	NUVENS	QUENTE	NORMAL	FALSO	SIM
14	CHUVA	MODERADO	ELEVADA	VERDADEIRO	NÃO

ISLab
Synthetic Intelligence Lab

Pesquisa de Regras de Associação

- Antecedente → Consequente

Humidade=Normal, Vento=Falso → Jogar=Sim
Suporte = 4/14 (29%); Confiança = 4/4 (100%)

Aspeto=Sol, Humidade=Elevada → Jogar=Não
Suporte = 3/14 (21%); Confiança = 3/3 (100%)

Temp.=Frio, Humidade=Normal → Jogar=Sim
Suporte = 4/14 (29%); Confiança = 3/4 (75%)

ID	ASPETO	TEMPERA-TURA	HUMIDADE	VENTO	JOGAR
1	SOL	QUENTE	ELEVADA	FALSO	NÃO
2	SOL	QUENTE	ELEVADA	VERDADEIRO	NÃO
3	NUVENS	QUENTE	ELEVADA	FALSO	SIM
4	CHUVA	MODERADO	ELEVADA	FALSO	SIM
5	**CHUVA**	**FRIO**	**NORMAL**	**FALSO**	**SIM**
6	**CHUVA**	**FRIO**	**NORMAL**	**VERDADEIRO**	**NÃO**
7	**NUVENS**	**FRIO**	**NORMAL**	**VERDADEIRO**	**SIM**
8	SOL	MODERADO	ELEVADA	FALSO	NÃO
9	**SOL**	**FRIO**	**NORMAL**	**FALSO**	**SIM**
10	CHUVA	MODERADO	NORMAL	FALSO	SIM
11	SOL	MODERADO	NORMAL	VERDADEIRO	SIM
12	NUVENS	MODERADO	ELEVADA	VERDADEIRO	SIM
13	NUVENS	QUENTE	NORMAL	FALSO	SIM
14	CHUVA	MODERADO	ELEVADA	VERDADEIRO	NÃO

ISLab
Synthetic Intelligence Lab

Pesquisa de Regras de Associação

- Antecedente → Consequente

Humidade=Normal, Vento=Falso → Jogar=Sim
Suporte = 4/14 (29%); Confiança = 4/4 (100%)

Aspeto=Sol, Humidade=Elevada → Jogar=Não
Suporte = 3/14 (21%); Confiança = 3/3 (100%)

Temp.=Frio, Humidade=Normal → Jogar=Sim
Suporte = 4/14 (29%); Confiança = 3/4 (75%)

Humidade=Elevada → Jogar=Sim
Suporte = 7/14 (50%); Confiança = 3/7 (43%)

ID	ASPETO	TEMPERA-TURA	HUMIDADE	VENTO	JOGAR
1	**SOL**	**QUENTE**	**ELEVADA**	**FALSO**	**NÃO**
2	**SOL**	**QUENTE**	**ELEVADA**	**VERDADEIRO**	**NÃO**
3	**NUVENS**	**QUENTE**	**ELEVADA**	**FALSO**	**SIM**
4	**CHUVA**	**MODERADO**	**ELEVADA**	**FALSO**	**SIM**
5	CHUVA	FRIO	NORMAL	FALSO	SIM
6	CHUVA	FRIO	NORMAL	VERDADEIRO	NÃO
7	NUVENS	FRIO	NORMAL	VERDADEIRO	SIM
8	**SOL**	**MODERADO**	**ELEVADA**	**FALSO**	**NÃO**
9	SOL	FRIO	NORMAL	FALSO	SIM
10	CHUVA	MODERADO	NORMAL	FALSO	SIM
11	SOL	MODERADO	NORMAL	VERDADEIRO	SIM
12	**NUVENS**	**MODERADO**	**ELEVADA**	**VERDADEIRO**	**SIM**
13	NUVENS	QUENTE	NORMAL	FALSO	SIM
14	**CHUVA**	**MODERADO**	**ELEVADA**	**VERDADEIRO**	**NÃO**

ISLab
Synthetic Intelligence Lab

Tipos de Regras de Associação

- Regras baseadas nos tipos de valores:
  - Booleanas ou Lógicas:
    - quando a regra de associação representa à presença ou ausência de valores:

      computador → gescliente
  - Quantitativas:
    - quando a regra define associações entre valores quantitativos dos atributos, tipicamente delimitados por intervalos:
    - idade(“30..39”) ∧ ordenado(“2500..5000”) → compra(HDReady)

(exemplos adaptados de “Data Mining: Concepts and Techniques”)

ISLab
Synthetic Intelligence Lab

Tipos de Regras de Associação

- Regras baseadas nas dimensões dos dados:
  - Uni-dimensional:
    - quando a regra de associação se refere a valores ou atributos em uma única dimensão, diz-se uma regra uni-dimensional:

      compra(computador) $\rightarrow$ compra(gescliente)
    - a única dimensão presente nesta (reescrita da) regra é “compra”.
  - Se a regra incluísse outras dimensões (tipo de cliente ou data da compra), dir-se-ia multi-dimensional.

(exemplos adaptados de “Data Mining: Concepts and Techniques”)

ISLab
Synthetic Intelligence Lab

Tipos de Regras de Associação

- Regras baseadas no nível de abstração dos dados:
  - Multi-nível:
    - quando a regra de associação relaciona valores a níveis de abstração diferentes diz-se multi-nível:

      idade(“30..39”) → compra(portátil)

      idade(“30..39”) → compra(computador)
    - os valores “portátil” e “computador” referem-se, nitidamente, a diferentes níveis de abstração.

(exemplos adaptados de “Data Mining: Concepts and Techniques”)

ISLab
Synthetic Intelligence Lab

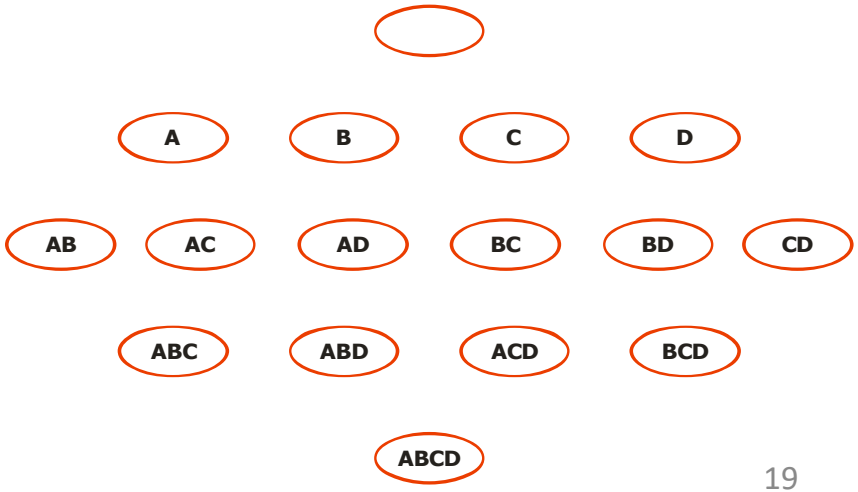
Tipos de Regras de Associação

- Outros/Extensões:
  - *Maxpattern*: padrões frequentes máximos (*maximal frequent pattern*):
    - *Maxpattern* é um padrão frequente P tal que o seu super-padrão Q não é frequente;
  - *Frequent closed itemset*;
  - Utilizam-se extensões a regras de associação com o intuito de reduzir a frequência com que surgem valores em algumas das regras encontradas.

ISLab
Synthetic Intelligence Lab

Algoritmo Apriori

- Algoritmo de pesquisa de Regras de Associação baseadas em aproximações booleanas;
- Utiliza informação conhecida *a priori* sobre a frequência dos dados para desenvolver a pesquisa;
- A procura de regras desenvolve-se iterativamente, numa pesquisa por níveis de conjuntos de dados (*itemsets*).



ISLab
Synthetic Intelligence Lab

Algoritmo Apriori: Princípios

- Princípio basilar:
  - uma relação é frequente se tem um valor de suporte elevado:
  - qualquer subconjunto de um conjunto de valores (*itemset*) frequente é, também, frequente:
    - Uma relação contendo **{cerveja, fraldas, tremoços}**, também contém a relação **{cerveja, tremoços}**;
    - Se **{cerveja, fraldas, tremoços}** é frequente, então, a relação **{cerveja, tremoços}** também é frequente.

ISLab
Synthetic Intelligence Lab

Algoritmo Apriori: Princípios

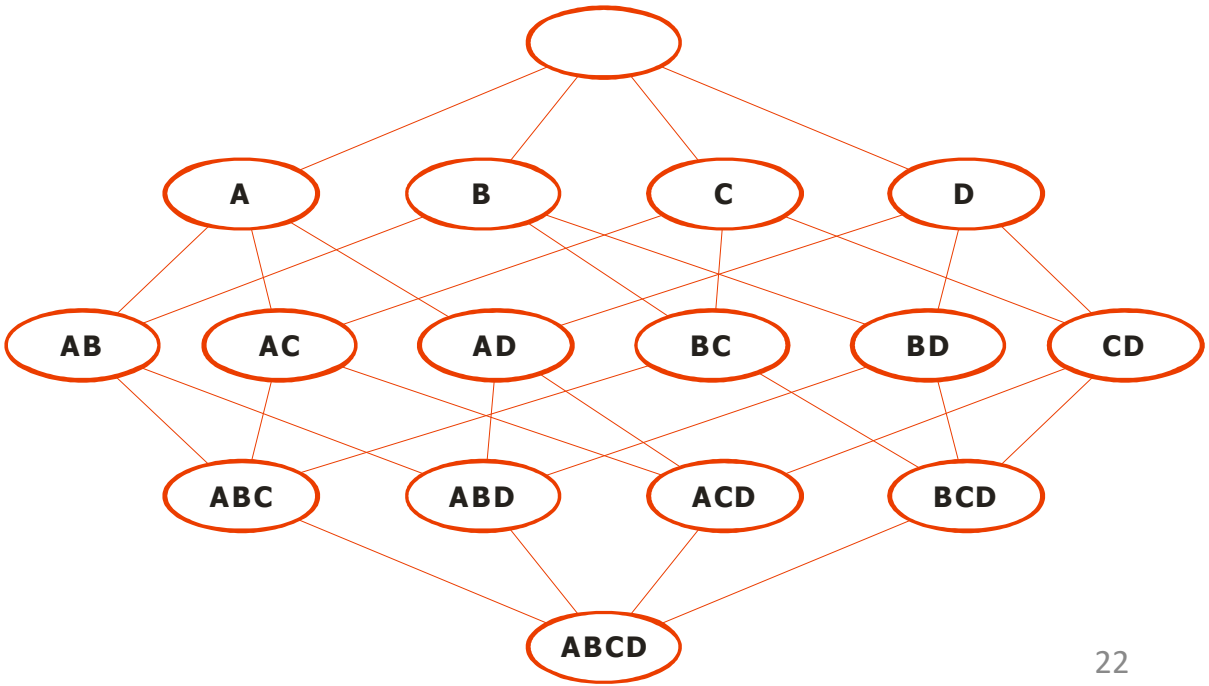
- Princípio basilar:
  - uma relação é frequente se tem um valor de suporte elevado:
  - qualquer subconjunto de um conjunto de valores (*itemset*) frequente é, também, frequente:
    - Uma relação contendo **{cerveja, fraldas, tremoços}**, também contém a relação **{cerveja, tremoços}**;
    - Se **{cerveja, fraldas, tremoços}** é frequente, então, a relação **{cerveja, tremoços}** também é frequente.

- Princípio complementar:
  - nenhum superconjunto de um conjunto de valores (*itemset*) não frequente deve ser pesquisado;
  - se um conjunto de valores V não é frequente, então, se lhe juntarmos um atributo mais, $V \cup A$, não poderá ser mais frequente;
  - este princípio permite evitar a pesquisa de um largo espectro de análise de combinações de atributos/valores.

ISLab
Synthetic Intelligence Lab

Algoritmo Apriori: Metodologia

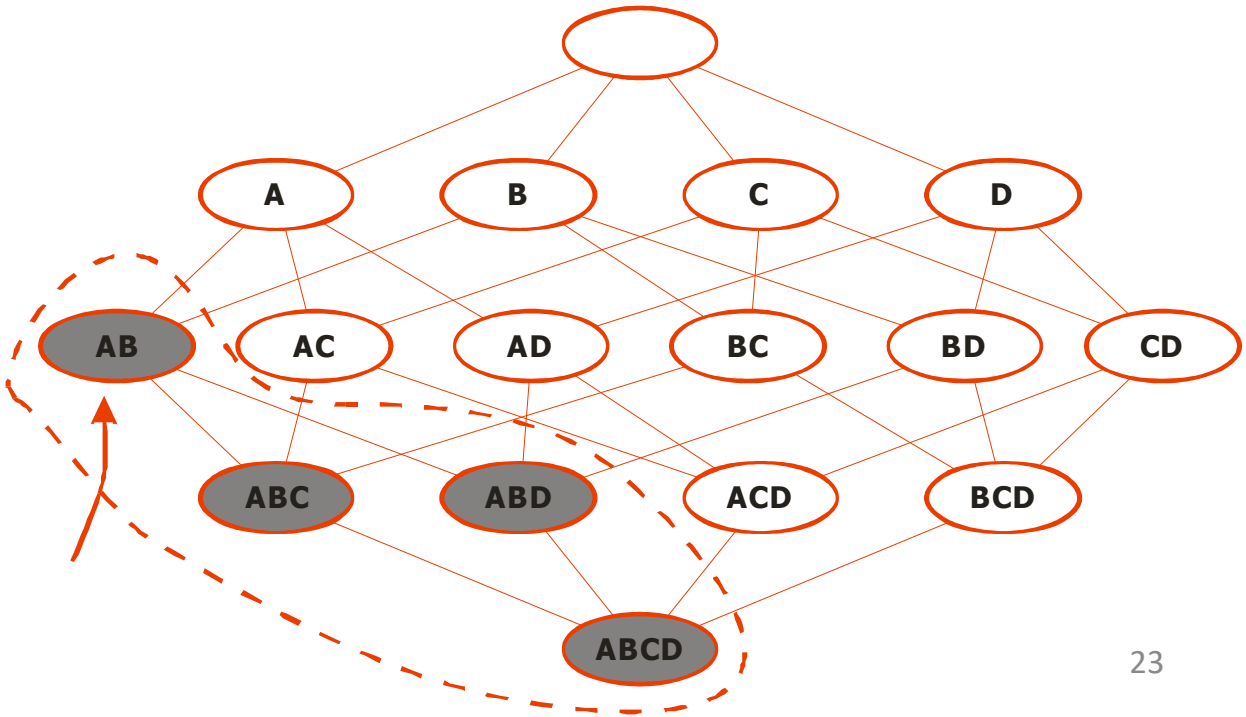
- Considerando **todas** as combinações:



ISLab
Synthetic Intelligence Lab

Algoritmo Apriori: Metodologia

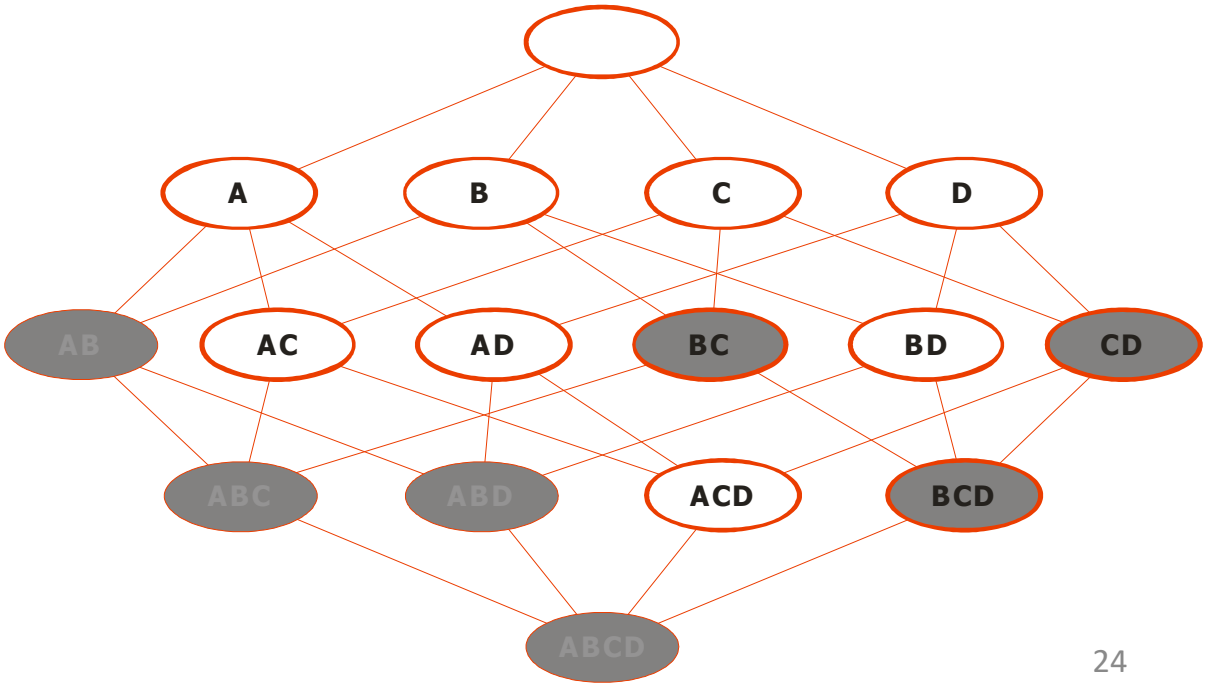
- Se a relação **AB não é frequente**, então nenhuma das super-relações poderá ser frequente:



ISLab
Synthetic Intelligence Lab

Algoritmo Apriori: Extensões

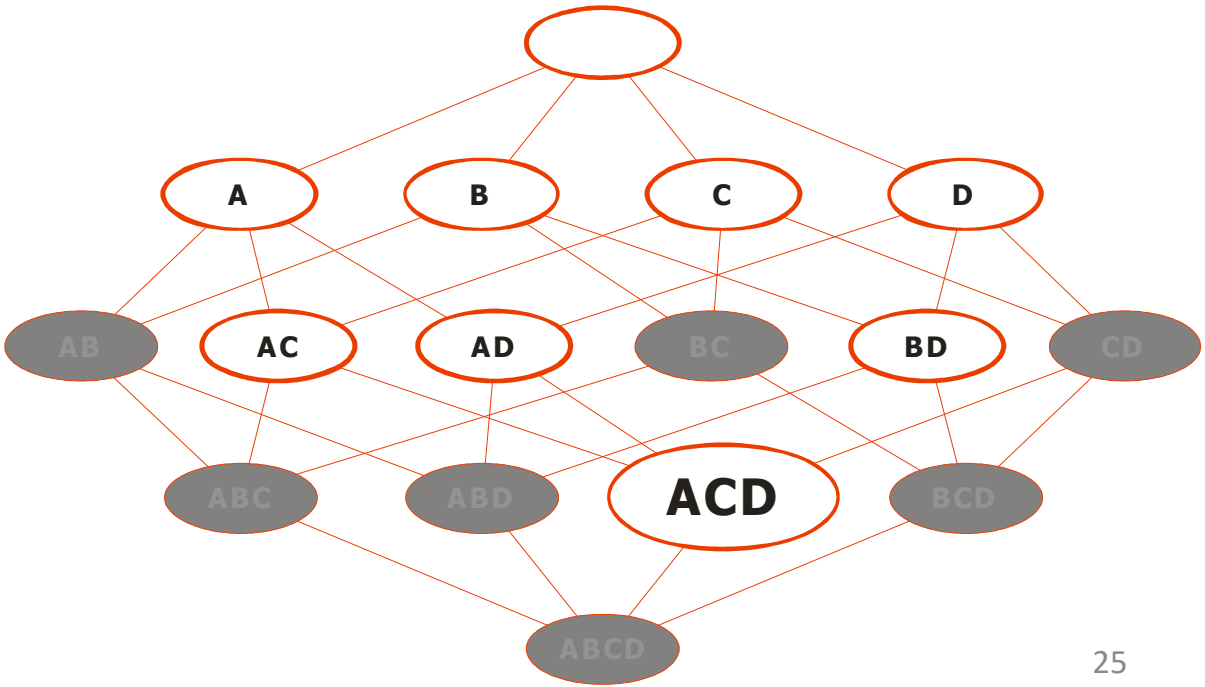
- Reduzir a quantidade de relações com base na confiança, medida pelo **interesse** de cada candidato:



ISLab
Synthetic Intelligence Lab

Algoritmo Apriori: como gerar Regras?

- Dada uma relação (*itemset*) frequente, ACD, qual a regra que deve ser gerada:
  - A → C,D
  - A,C → D
  - A,D → C
  - C → A,D
  - C,D → A
  - D → A,C
  - {} → A,C,D



ISLab
Synthetic Intelligence Lab

Medidas de Interesse de Regras de Associação

- Serão todas as Regras de Associação descobertas úteis para serem presentes ao utilizador final?
- Medidas subjetivas:
  - Surpresa, novidade, executáveis, etc.;
- Medidas objetivas:
  - *Support/suporte*;
  - *Confidence/confiança*;
  - *Conviction/convicção*;
  - *Lift/correlação*;
  - *Leverage/potência*;
  - *Coverage/amplitude*.

ISLab
Synthetic Intelligence Lab

Medidas de Interesse Subjetivas

- Uma Regra de Associação é útil se for:
  - **Inesperada** ou **Surpreendente** para o utilizador;
  - **Novidade** relativamente ao conhecimento anterior;
  - **Executável**, permitindo que se faça "alguma coisa" com ela;
  - etc.

ISLab
Synthetic Intelligence Lab

Medidas de Interesse Subjetivas

- Uma Regra de Associação é útil se for:
  - **Inesperada** ou **Surpreendente** para o utilizador;
  - **Novidade** relativamente ao conhecimento anterior;
  - **Executável**, permitindo que se faça "alguma coisa" com ela;
  - etc.

Como medir estes fatores?

ISLab
Synthetic Intelligence Lab

Medidas de Interesse Objetivas

- Cálculo de medidas objetivas:
  - Suporte;
  - Confiança (conficende/accuracy);
  - Correlação: Lift($A \rightarrow C$) = Sup(A,C) / (Sup(A) x Sup(C));
  - Convicção: Conviction($A \rightarrow C$) = (Sup(A) x Sup($\neg$C)) / Sup(A,C);
  - Potência: Leverage($A \rightarrow C$) = Sup(A,C) – Sup(A) x Sup(C);
  - Amplitude: Coverage($A \rightarrow C$) = Sup(A);
  - etc.

ISLab
Synthetic Intelligence Lab

Medidas de Interesse Objetivas

- Cálculo de medidas objetivas:
  - Suporte;
  - Confiança (conficende/accuracy);
  - Correlação: Lift($A \rightarrow C$) = Sup(A,C) / (Sup(A) x Sup(C)):
    - é uma medida da importância da associação que é independente do valor de suporte;
    - Lift = 1: a ocorrência do antecedente é independente da ocorrência do consequente;
    - Lift > 1: existe uma dependência entre o antecedente e o consequente, permitindo prever o consequente;
    - Lift < 1: o antecedente e o consequente são substitutos; um tem uma influência negativa no outro;
  - Convicção: Conviction($A \rightarrow C$) = (Sup(A) x Sup($\neg C$)) / Sup(A,C);
  - Potência: Leverage($A \rightarrow C$) = Sup(A,C) – Sup(A) x Sup(C);
  - Amplitude: Coverage($A \rightarrow C$) = Sup(A);
  - etc.

ISLab
Synthetic Intelligence Lab

Medidas de Interesse Objetivas

- Cálculo de medidas objetivas:
  - Suporte;
  - Confiança (conficende/accuracy);
  - Correlação: Lift($A \rightarrow C$) = Sup(A,C) / (Sup(A) x Sup(C));
  - Convicção: Conviction($A \rightarrow C$) = (Sup(A) x Sup($\neg C$)) / Sup(A,C):
    - é uma medida da implicação;
    - mede a expectativa de o antecedente ocorrer sem o consequente; (probabilidade de a regra prever incorretamente)
    - Conviction = 1: o antecedente e o consequente não estão relacionados;
  - Potência: Leverage($A \rightarrow C$) = Sup(A,C) – Sup(A) x Sup(C);
  - Amplitude: Coverage($A \rightarrow C$) = Sup(A);
  - etc.

ISLab
Synthetic Intelligence Lab

Medidas de Interesse Objetivas

- Cálculo de medidas objetivas:
  - Suporte;
  - Confiança (conficende/accuracy);
  - Correlação: Lift($A \rightarrow C$) = Sup(A,C) / (Sup(A) x Sup(C));
  - Convicção: Conviction($A \rightarrow C$) = (Sup(A) x Sup($\neg$C)) / Sup(A,C);
  - Potência: Leverage($A \rightarrow C$) = Sup(A,C) – Sup(A) x Sup(C):
    - também conhecida por Piatetsky/Shapiro;
    - é uma medida da proporção dos exemplos adicionais sob o âmbito do antecedente e do consequente;
    - parecido com Lift: como?
    - Lift consegue encontrar associações mais fortes em relações menos frequentes;
    - Leverage dá mais importância a relações com fortes valores de suporte;
  - Amplitude: Coverage($A \rightarrow C$) = Sup(A);
  - etc.

ISLab
Synthetic Intelligence Lab

Medidas de Interesse Objetivas

- Cálculo de medidas objetivas:
  - Suporte;
  - Confiança (conficende/accuracy);
  - Correlação: Lift($A \rightarrow C$) = Sup(A,C) / (Sup(A) x Sup(C));
  - Convicção: Conviction($A \rightarrow C$) = (Sup(A) x Sup($\neg$C)) / Sup(A,C);
  - Potência: Leverage($A \rightarrow C$) = Sup(A,C) – Sup(A) x Sup(C);
  - Amplitude: Coverage($A \rightarrow C$) = Sup(A):
    - é o suporte;
  - etc.

ISLab
Synthetic Intelligence Lab

Referências bibliográficas

- Data Mining: Concepts and Techniques
  Jiawei Han, Micheline Kamber
- Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations
  Ian Witten, Eibe Frank

ISLab

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Aprendizagem e Extração de Conhecimento

Perfil Sistemas Inteligentes @ MiEI/4º – 1º Semestre

Cesar Analide, Filipe Gonçalves, Paulo Novais