Mineração de padrões frequentes

Fabrício J. Barth

fabricio.barth@gmail.com

Setembro de 2016

Objetivos

Os objetivos desta aula são:

- Apresentar e discutir métodos para identificar associações úteis em grandes bases de dados (transacionais) usando medidas estatísticas simples, e;
- Apresentar e discutir todas as etapas necessárias para executar uma análise de *market basket*.

Sumário

- A ideia geral do *market basket analysis*.
- Algoritmo Apriori: mineração de itens frequentes.
- Definição de suporte, confiança e *lift*.
- Interpretando as regras.
- Visualização das regras.
- Referências e leituras adicionais.

Resultado esperado de uma *market basket analysis*

$$\{\text{pãozinho, pão de queijo}\} \rightarrow \{\text{suco de laranja}\}$$

- A regra acima representa a seguinte informação: se uma pessoa compra pãozinho e pão de queijo então existe uma possibilidade desta pessoa comprar também suco de laranja.
- Os itens indicados ente $\{\}$ fazem parte de um mesmo *itemset*.

Cenários de uso

- Algoritmos de regras de associação são geralmente utilizados em problemas de *market basket analysis*. Exemplos de *market basket analysis* são:
  - ⋆ Que produtos devem ser incluídos os excluídos de um estoque a cada mês.
  - ⋆ Propaganda cruzada entre produtos.
  - ⋆ Modificação física ou lógica de produtos dentro de categorias de produtos.
  - ⋆ Programas promocionais: incentivo de compra de múltiplos produtos.

Além disso, pode-se utilizar algoritmos de regras de associação em outros cenários:

- ⋆ Busca por padrões de sequência de DNA e proteinas que ocorrem frequentemente em dados sobre câncer.
- ⋆ Identificação por padrões de compra em transações fraudulentas.
- ⋆ Desenvolvimento de sistemas de recomendação.
- ⋆ Clickstream analysis.

- Regras de associação são utilizadas para procurar por conexões *"interessantes"* entre um grande número de variáveis.
- Pessoas são capazes de gerar tais *insights*, mas geralmente é necessário um nível de experiência bem alto no domínio da aplicação e muito tempo pensando sobre o problema.

Algoritmo Apriori: mineração de itens frequentes

- Dado:
  - ⋆ um conjunto $A = \{a_1, \cdots, a_m\}$ de itens,
  - ⋆ uma tabela $T = (t_1, \cdots, t_n)$ de transações sobre A,
  - ⋆ um número β_{min} que $0 < \beta_{min} \leq 1$, o **suporte mínimo**.
- Objetivo 1:
  - ⋆ encontrar o conjunto de **itens frequentes**, tais que o **suporte** de cada conjunto de itens é maior ou igual ao β_{min} definido pelo usuário.

Exemplo de transações

	Itens
1	{a,d,e}
2	{b,c,d}
3	{a,c,e}
4	{a,c,d,e}
5	{a,e}
6	{a,c,d}
7	{b,c}
8	{a,c,d,e}
9	{b,c,e}
10	{a,d,e}

0 itens	1 item	2 itens	3 itens
{}: 10	{a}: 7	{a,c}: 4	{a,c,d}: 3
	{b}: 3	{a,d}: 5	{a,c,e}: 3
	{c}: 7	{a,e}: 6	{a,d,e}: 4
	{d}: 6	{b,c}: 3	
	{e}: 7	{c,d}: 4	
		{c,e}: 4	
		{d,e}: 4	

Figure 1: Um banco de dados de transações, com 10 transações, e a enumeração de todos os conjuntos de itens frequentes usando o suporte mínimo = 0,3

Algoritmo Apriori: mineração de itens frequentes

- Objetivo 2:
  - ⋆ encontrar o conjunto de regras de associação com **confiança** maior ou igual que um mínimo definido pelo utilizador.

Suporte e Confiança

- O **suporte** de um conjunto de itens Z, $suporte(Z)$, representa a porcentagem de transações na base de dados que contêm os itens de Z.
- A **confiança** de uma regra de associação $A \rightarrow B$, $confianca(A \rightarrow B)$, é dado por:

$$confianca(A \rightarrow B) = \frac{Suporte(A \wedge B)}{Suporte(A)} \qquad (1)$$

Exemplos de confiança

- Se suporte($\{$pão, ovos, leite$\}$) $= 0.15$ e suporte($\{$pão, ovos$\}$) $= 0.15$ então confianca($\{$pão, ovos$\} \rightarrow \{$leite$\}$) $= 1$.
- Se suporte($\{$pão, ovos$\}$) $= 0.15$ e suporte($\{$pão$\}$) $= 0.6$ então confianca($\{$pão$\} \rightarrow \{$ovos$\}$) $= 0.25$.

Exemplo de regras geradas

Premises	Conclusion	Support	Confidence
b	c	0.300	1
e, d	a	0.400	1
e	a	0.600	0.857
a	e	0.600	0.857
d	a	0.500	0.833
a, d	e	0.400	0.800

Figure 2: Regras extraídas com confiança maior que 0.8

Confiança

- Uma confiança alta indica que uma regra $(X \rightarrow Y)$ é mais interessante ou mais confiável, baseada no dataset analisado.

- No entanto, o fato de apenas analisar $X \wedge Y$ e X, sem analisar Y pode gerar alguns problemas.

Exemplo

Considere 1.000 transações, onde:

- leite ocorre em 400
- pão ocorre em 900
- manteiga ocorre em 300
- leite e pão ocorrem em 300
- manteiga e leite ocorrem em 300

Sendo assim:

- $confianca(\{leite\} \rightarrow \{pao\}) = \frac{0,3}{0,4} = 0,75$
- $confianca(\{leite\} \rightarrow \{manteiga\}) = \frac{0,3}{0,4} = 0,75$
- Pão é algo que ocorre com muita frequência neste dataset.
- Esta informação não é levada em consideração pela $confianca(\{leite\} \rightarrow \{pao\})$.
- Talvez, esta correlação seja apenas uma coincidência.

Lift ou coeficiente de interesse

$$Lift(X \rightarrow Y) = \frac{Suporte(X \wedge Y)}{Suporte(X) \times Suporte(Y)} \qquad (2)$$

- *Lift* ou coeficiente de interesse: um valor de *lift* para uma regra $(A \rightarrow B)$ superior a 1 indica que A e B acontecem mais frequentemente juntos do que o esperado, isso significa que a ocorrência de A tem um efeito positivo sobre a ocorrência de B.

Exemplos

- $lift(\{leite\} \rightarrow \{pao\}) = \frac{0{,}3}{0{,}4 \times 0{,}9} = 0,834$
- $lift(\{leite\} \rightarrow \{manteiga\}) = \frac{0{,}3}{0{,}4 \times 0{,}3} = 2,5$

Assim, fica claro que a ocorrência de $leite$ tem um efeito positivo sobre a ocorrência da $manteiga$. Mas isto não se aplica ao $leite$ e pao.

Medida Lift

Dada uma regra de associação $A \rightarrow B$, esta medida indica o quanto mais freqüente torna-se B quando ocorre A.

- Se $Lift(A \rightarrow B) = 1$, então A e B são independentes.
- Se $Lift(A \rightarrow B) > 1$, então A e B são positivamente dependentes.
- Se $Lift(A \rightarrow B) < 1$, A e B são negativamente dependentes.

Esta medida varia entre 0 e ∞ e possui interpretação simples: **quanto maior o valor de $Lift$, mais interessante a regra, pois A aumenta B.**

Exemplo básico de uso

Exemplo Básico sobre Regras de Associação

Exemplo: *Grocery Store*

Exemplo usando um dataset de uma *Grocery Store*

Pontos fortes e fracos

- **Fortes**:
  - ⋆ É facilmente aplicável em um volume grande de dados transacionais.
  - ⋆ Resultados no formato de regras é fácil de compreender.
  - ⋆ É útil na descoberta de padrões implícitos em bases de dados.

- **Fracos**:
  - ⋆ Não é muito útil para bases pequenas.
  - ⋆ Às vezes é difícil separar *insights* de senso comum.
  - ⋆ É fácil gerar conclusões incorretas a partir de padrões aleatórios.

Material de **consulta**

- Capítulo 5 do livro EMC Education Services, editor. Data Science and Big Data Analytics: Discovering, Analysing, Visualizing and Presenting Data. John Wiley & Sons, 2015.

- Fabrício Barth. Mineração de regras de associação em servidores Web com RapidMiner[a].
- Gonçalves. Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas. INFOCOMP Journal of Computer Science, 2005, 4, 26-35.

[a] http://fbarth.net.br/materiais/webMining/webUsageMining.pdf

- Data Mining Algorithms in R - Apriori Algorithm. http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_Apriori_Algorithm. Acessado em 13 de junho de 2013.
- RDataMining.com: Association Rules. http://www.rdatamining.com/examples/association-rules. Acessado em 13 de junho de 2013.

Próximas etapas

- Exercícios, e;
- Projeto!