Descoberta de Padrões com Regras de Associação

ESTAT0016 – Tópicos Especiais em Estatística (Introdução à Apredizagem de Máquina)

Prof. Dr. Sadraque E.F. Lucena

UNIVERSIDADE FEDERAL DE SERGIPE

Aprendizagem Não Supervisionada

- As Regras de Associação são uma técnica usada em Aprendizagem Não Supervisionada.
  - Neste tipo de aprendizagem não há exemplos com rótulos pré-determinados.
  - O objetivo não é fazer predição, mas sim descobrir padrões significativos e *insights* a partir dos dados.

Formalmente:

- Temos um conjunto de observações $X = (x_1, \ldots, x_N)$ com função de densidade conjunta $P(X)$, e nosso objetivo é inferir as propriedades dessa função $P(X)$ sem a assistência de um supervisor.

Regras de Associação

- As regras de associação surgiram para permitir a análise de comportamento de consumo a partir da identificação de associação entre diferentes itens em grandes conjuntos de transações.
  - Esse tipo de estudo é chamado de análise de cesta de mercado (*makert basket analysis*) ou anális de afinidade (*affinity analysis*).
  - Ele é utilizado em diferentes contextos, como varejo, marketing e planejamento de produtos.
- As regras de associação buscarão descobrir padrões que ocorram com frequência significativa no conjunto de dados.
- Esses padrões revelam relações entre produtos que permitem idenitifcar padrões de comportamento.

Análise da Cesta de Mercado

- Considere um conjunto de dados no qual cada linha representa as compras de uma pessoa em um supermercado e cada coluna indica se ela comprou ou não um produto específico.

	Produto					
Transação	**Pão**	**Leite**	**Cerveja**	**Fralda**	**Ovos**	**Refrigerante**
T1	1	1	1	0	0	0
T2	1	0	1	1	1	0
T3	0	1	1	1	0	1
T4	1	1	1	1	0	0
T5	1	1	0	1	0	1

Análise da Cesta de Mercado

Definições

- **Transação:** cada compra feita por um cliente em uma loja ou estabelecimento.
- **Item:** produto que pode ser comprado em uma transação.
- **Conjunto de itens *(itemset)*:** coleção de dois ou mais itens que são adquiridos juntos em uma transação.
  - Na tabela anterior:
    - a transação T1 contem o conjunto de intes *{pão, leite, cerveja}*;
    - a transação T3 contém o cojunto de itens *{leite, cerveja, fralda, refrigerante}*.
  - Note que um conjunto de itens é uma lista de produtos distintos que foram comprados juntos em uma única transação, sem considerar a quantidade de cada item.

Regras de Associação

- As regras de associação descrevem quais grupos de itens tendem a ocorrer juntos de acordo com os dados.
- Eles são representados usando o formato SE-ENTÃO:
  - O lado esquerdo (SE) especifica um conjunto de itens (ou eventos) que foram adquiridos simultaneamente em uma transação. Ele denominado *antecedente* e pode conter mais de um item.
  - O lado direito (ENTÃO) indica um item (ou evento) adicional que também foi comprado junto com o conjunto de itens (ou eventos) anteriores. Ele é denominado *consequente* e contém apenas um único item.
- Exemplos de *Antecedente*→*Consequente*:
  - *{pão}→{leite}*
  - *{cerveja, refrigerante}→{carne}*

Regras de Associação

- As regras de associação podem ser classificadas em três categorias:
  - **Acionáveis:** regras que fornecem insights claros e úteis que podem ser aplicados.
    - *Exemplo:* Uma regra indicando que clientes que compram leite condensado também adquirem achocolatado sugere uma preferência por brigadeiro. Como resultado, uma loja poderia colocar esses itens próximos.
  - **Triviais:** regras que oferecem insights que já são bem conhecidos por quem é da área.
    - *Exemplo:* Uma regra que mostra que clientes que compram canetas frequentemente também compram cadernos não fornece realmente novos insights significativos.
  - **Inexplicáveis:** regras que desafiam uma explicação racional, necessitam de mais pesquisas para serem compreendidas.
    - *Exemplo:* Descobrir que clientes que compram sapatos têm mais probabilidade de também comprar canetas desafia uma explicação racional e requer mais pesquisas para serem compreendidas.

Identificando Regras Fortes

- Para determinar quais regras de associação são úteis, é importante sabermos quantas combinações de itens são possíveis no conjunto de dados.
- Suponha que um conjunto de dados possui p itens distintos possíveis.
  - Como o lado esquerdo de uma regra não pode ser vazio, ele pode conter entre 1 e $p - 1$ itens (desde que o lado direito da regra não esteja vazio).
  - Suponha que o número de itens do lado esquerdo seja k e o número de itens do lado direito seja j.
  - Então, o número total de regras de associação que podem tomados desses p itens é

$$\sum_{k=1}^{p-1} \left[\binom{p}{k} \times \sum_{j=1}^{p-k} \binom{p-k}{j} \right] = 3^p - 2^{p+1} + 1$$

- Se tivermos 6 itens distintos, podemos criar $3^6 - 2^7 + 1 = 602$ regras diferentes.

Identificando Regras Fortes

- Ao invés de avaliarmos uma quantidade enorme de regras, podemos considerar apenas as regras baseadas em conjuntos de itens que ocorrem regularmente, conhecidos como *conjuntos de itens frequentes*. Para isso usamos algumas métricas.

Suporte

- O *suporte* ou *cobertura* de um conjunto de itens é a probabilidade do conjunto de itens estar contida em uma transação no conjunto de dados. Ou seja, para um conjunto de itens X,

$$Suporte(X) = P(X) = \frac{número\ de\ transações\ contendo\ o\ conjunto\ de\ itens\ X}{Total\ de\ transações}$$

- No nosso exemplo, temos:
  - $Suporte(\{cerveja, leite\}) = \frac{2}{3} = 0{,}6$
  - $Suporte(\{cerveja, leite, fralda\}) = \frac{2}{5} = 0{,}4$

Identificando Regras Fortes

- Como o suporte é baseado no conjunto de itens, todas as regras derivadas do mesmo conjunto têm o mesmo suporte, como é o caso das regras
  - *{cerveja, leite}→{fralda}*
  - *{cerveja, fralda}→{leite}*
  - *{fralda, leite}→{cerveja}*

  que derivam do conjunto de itens *{cerveja, leite, fralda}*.
- Ao calcular o suporte de cada conjunto de itens, podemos definir um limite mínimo para avaliar a utilidade das regras, reduzindo assim o número de regras a serem examinadas.

Identificando Regras Fortes

Confiança

- A *confiança* ou *acurácia* de uma regra é a probabilidade de que um item apareça no conjunto de transações, dado que outro conjunto de itens também apareceu.
- Em outras palavras, a confiança é a proporção de transações que contêm todos os itens de um antecedente (parte esquerda da regra) e consequente (parte direita da regra) em relação às transações que contêm apenas o antecedente.
- Ou seja,

$$Confiança(X \rightarrow Y) = \frac{Suporte(X, Y)}{Suporte(X)}$$

- No nosso exemplo, temos:
  - $Confiança(\{cerveja, leite\} \rightarrow \{fralda\}) = \frac{2}{3} = 0{,}67$

  (67% dos que compraram cerveja e leite, também compraram fralda)

Identificando Regras Fortes

Lift

- *Lift* é a razão entre a confiança da regra e a frequência de ocorrência do item consequente.

$$Lift(X \rightarrow Y) = \frac{Confiança(X \rightarrow Y)}{Suporte(X, Y)}$$

- Em outras palavras, o lift nos diz quantas vezes mais frequentemente os itens de interesse aparecem juntos do que se fossem independentes um do outro.
  - Lift > 1: associação mais frequente do que o esperado
  - Lift < 1: associação menos frequente do que o esperado
  - Lift = 1: associação independente

Identificando Regras Fortes

Lift

- No nosso exemplo, temos:
  - $Lift(\{cerveja, leite\} \rightarrow \{fralda\}) = \frac{0{,}67}{0{,}80} = 0{,}84$

(quem compra leite e cerveja é 0,84 vezes provável de comprar fralda)

O Algoritmo Apriori

- O processo de encontrar o conjunto de itens frequentes requer a geração de todos os conjuntos de itens para avaliar e determinar quais são frequentes e quais regras derivadas são fortes.
- Isso pode ser computacionalmentne custoso, pois em um conjunto de dados com p itens distintos, existem $2^p - 1$ possíveis conjuntos de itens.
- Uma mercearia pequena que vendesse apenas 50 itens teria, existem $2^50 - 1 = 1{,}1259 \times 10^{15}$ conjuntos de itens possíveis para avaliar (pouco mais de 1 quatrilhão).
- Para minimizar o custo computacional desse processo, podemos usar o algoritmo apriori.

O Algoritmo Apriori

- O algoritmo apriori foi introduzido pela primeira vez por Rakesh Agrawal e Ramakrishnan Srikant em 1993 e recebe seu nome pelo fato de usar conhecimento prévio sobre as propriedades de conjuntos de itens frequentes no processo de geração.
- Etapas do algoritmo:
  1. São gerados conjuntos de itens com apenas um item.
  2. Cada conjunto é verificado se atende ao limimte mínimo de suporte estabelecido pelo usuário. Se um conjunto tem suporte abaixo do mínimo, ele e seus superconjuntos são descartados (poda apriori).
  3. Em seguida são gerados conjuntos com 2 itens, baseando-se apenas nos conjuntos de itens frequentes da etapa anterior.
  4. Esses conjuntos de itens são avaliados para verificar se atendem ao limite de suporte. Os conjunots que não atendem são podados.

O Algoritmo Apriori

- Etapas do algoritmo:

  5. O processo segue até que até que não haja conjuntos de itens adicionais a serem gerados e avaliados.

- Outra abordagem popular é a abordagem de crescimento de padrões frequentes (FP-growth).
- Esta abordagem utiliza uma estrutura semelhante a uma árvore para armazenar informações que facilitam a identificação dos conjuntos de itens que são frequentes.

- Vejamos como utiliar o algoritmo Apriori no R.

FIM