Big Data Analytics com R

Associações

Agenda

- Implementar análise de associações e sequências utilizando R;
- Diferenciar resultados das análises de associações e sequências;
- Compreender os parâmetros opcionais para as análises de associações e sequências utilizando R.

Associações

Motivação

- O objetivo da análise de associações é encontrar relações entre itens.
- Por exemplo, relações de compras:
  - Leite, farinha e ovos são usualmente comprados juntos? Ou;
  - Caso uma pessoa compre leite e farinha, ela estaria inclinada à comprar também ovos?

Motivação

- Temos diversos casos de grandes massas de dados onde podemos aplicar esta técnica.
  - Dados de transações comerciais (pontos de venda, programas de fidelidade, plataformas de comércio eletrônico).
  - Dados de navegação na Web (Web analytics, ferramentas de busca, bibliotecas digitais, Wikis).
  - Dados genéticos (Cadeias de DNA).

Motivação

- Tamanhos típicos dos conjuntos de dados.
  - Varejista médio: de 10 a 500 grupos de produtos, com 500 - 10.000 produtos individuais.
  - Amazon: mais de 200M de produtos (2013).
  - Wikipedia: aproximadamente 5M de artigos (2015).
  - Google: 47B de páginas estimadas (2015).
  - Projeto Genoma: entre 20.000 e 25.000 genes no DNA humano, com 3B de pares-base.
  - Entre 10.000 e 10M de transações típicas (cestas de compras, sessões de usuários, observações, pacientes, etc.)

Um problema

- Quando vamos ao supermercado, normalmente levamos uma lista do que devemos comprar.
- Cada usuário possui uma lista diferente, dependendo de suas necessidades, uma pode conter itens para um jantar familiar enquanto outra produtos para um churrasco ou mesmo uma tarde de futebol entre amigos.

Um problema

- Entender estes padrões de compra pode ajudar a aumentar as vendas de diversas formas. Se encontrarmos um par de itens X e Y que são adquiridos sempre juntos podemos pensar em:
  - X e Y podem ser colocados próximos de forma que um comprados que procure um dos dois seja incentivado à comprar o outro produto?
  - Podemos aplicar descontos em um dos dois produtos apenas para maximizar o lucro da venda dos dois em conjunto?
  - Podemos direcionar a propaganda de um dos produtos aos usuários do outro para incentivar a compra?
  - Podemos combinar os dois produtos em um no produto que atenda à demanda?

Um problema

- Quando temos a percepção de determinados padrões de compras, como efetivamente definimos estas associações?
- Além de aumentar potenciais lucros, as regras de associação são usadas em outras áreas, como por exemplo na busca por diagnósticos mais precisos e na relação entre os sintomas, melhorando o cuidado com pacientes e a prescrição médica.

Definição

- A análise de regras de associação é uma técnica usada para descobrir relações entre itens.

Métricas

- Existem três métricas básicas para medir associações.

Suporte (*Support*)

- Esta métrica nos diz quão popular é um determinado conjunto de dados, medido pela proporção de transações onde o conjunto de dados aparece.
- No exemplo de conjunto de transações a seguir, o suporte para {maçã} é 4 em 8, ou seja, 50%.
- Estes conjuntos de itens podem incluir mais de um item. Ainda em nosso exemplo, o suporte para {maçã, cerveja, arroz} é 2 em 8, ou seja, 25%.

Suporte (*Support*)

Support {🍎} = $\frac{4}{8}$

Transaction 1	🍎 🍺 ⚪ 🍗
Transaction 2	🍎 🍺 ⚪
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	💉 🍺 ⚪ 🍗
Transaction 6	💉 🍺 ⚪
Transaction 7	💉 🍺
Transaction 8	💉 🍐

Suporte (*Support*)

- Se for descoberto que a venda de itens acima de determinada proporção tem um impacto significativo nos lucros, pode-se considerar esta proporção como um tipo de limite (chamado em inglês de *support threshold*).
- Pode-se então considerar os conjuntos de itens com suporte acima deste limite como significativos para o negócio.

Confiança (*Confidence*)

- Esta métrica nos mostra qual a probabilidade de compra do produto Y ao ser comprado o produto X.
- É medida como a proporção de transações com o item X nas quais o item Y também aparece.

Confiança (*Confidence*)

- No exemplo anterior, a confiança da relação {maçã -> cerveja} é 3 em 4, ou 75%.

$$\text{Confidence}\{🍎 \rightarrow 🍺\} = \frac{\text{Support}\{🍎, 🍺\}}{\text{Support}\{🍎\}}$$

Confiança (*Confidence*)

- Um problema usualmente relacionado à métrica de confiança é que ela pode não representar bem a importância de determinada associação.
- Em nosso exemplo, ela leva em consideração apenas a popularidade das maçãs e não das cervejas.
- Caso as cervejas sejam muito populares, a probabilidade de uma transação que contenha cerveja e maçãs aumenta consideravelmente, tornando a medida de confiança tendenciosa.

Lift

- Métrica que leva em consideração a popularidade dos produtos de um conjunto de itens.
- Ela mede a probabilidade de compra do produto Y quando o produto X é comprado, levando em consideração o quão popular são os produtos individualmente.

Lift

- Em nosso exemplo, o *lift* do conjunto {maçã -> cerveja} é 1, o que implica que não há associação entre estes itens.
- Um *lift* superior a 1 indica maior chance de compra de Y caso X seja comprado.
- Já um *lift* inferior a 1 indica menor chance de compra de Y caso X seja comprado.

$$\text{Lift}\{🍎 \rightarrow 🍺\} = \frac{\text{Support}\{🍎, 🍺\}}{\text{Support}\{🍎\} \times \text{Support}\{🍺\}}$$

Algoritmo *apriori*

Princípio *apriori*

- O princípio *apriori* pode reduzir o número de conjuntos de itens a examinar.
- De forma simples, o princípio *apriori* diz que se um determinado conjunto de itens é pouco frequente, então todos os seus subconjuntos de itens também serão pouco frequentes.

Princípio *apriori*

- Em outras palavras, se {cerveja} for pouco frequente no conjunto de transações, então esperamos que {cerveja, pizza} também o seja.
- Logo, para definir a lista de conjuntos de itens mais frequentes, deveremos desconsiderar tanto o conjunto {cerveja, pizza} quanto qualquer outro conjunto que contenha {cerveja}.

Encontrando itens populares

- Usando o princípio *apriori*, o número de conjuntos que deverá ser examinado pode ser "podado" e a lista de conjuntos populares pode ser determinada por meio dos seguintes passos.

Encontrando itens populares

1) Iniciar com os conjuntos contendo apenas um item, como {maçã} e {pera}.
2) Determinar o suporte para os conjuntos. Manter os conjuntos que estão dentro do limite mínimo de suporte (*minimum support threshold*) e remover os demais.
3) A partir dos conjuntos mantidos no item anterior, gerar todas as combinações possíveis de itens.
4) Repetir as etapas 2 e 3 até que não mais existam novos conjuntos de itens.

Exemplo

- Imaginemos um exemplo onde {maçã} possua um valor de suporte baixo.
- Desta forma, este conjunto será removido e todos os demais conjuntos que contenham {maçã} também serão removidos (a chamada “poda”).
- Neste caso, reduzimos o conjunto de itens em mais de 50%.

Exemplo

Observação

- O valor de suporte escolhido na etapa 2 pode ser baseado em uma análise formal ou experiência.
- Caso saiba que a venda de alguns itens acima de determinada taxa cause um impacto efetivo sobre o lucro, você poderá utilizar esta referência como limite de poda.

Usando confiança e *lift*

- Vimos que o algoritmo *apriori* utiliza a métrica chama suporte como referência.
- Porém, tanto a confiança quanto o *lift* podem ser usados.

Usando confiança e *lift*

- Vamos a um exemplo que utilize confiança. Se a regra {cerveja, chips -> maçã} tiver baixa confiança, então todas as regras que contenham {maçã} no lado direito (chamado também de *rhs* – *right hand side*) também terão baixa confiança e sofrerão então a "poda".
- Exemplo:
  - {cerveja -> maçã, chips};
  - {chips -> maçã, cerveja}.

Limitações

- Custo computacional
  - Mesmo reduzindo o número de conjuntos candidatos a cada iteração, este conjunto pode ainda ser bem grande quando o inventário em análise for muito grande ou o limite de suporte escolhido for muito baixo.
  - Uma possível solução é reduzir o número de comparações utilizando estruturas de dados avançadas (por exemplo, tabelas *hash*) para ordenar os conjuntos de forma mais eficiente.

Limitações

- Associações espúrias
  - A análise de grandes inventários pode envolver grandes configurações de conjuntos de itens e o limite de suporte poderá ser baixo para que certas associações tenham destaque.
  - Porém, ao diminuirmos o limite, poderemos aumentar o número de associações indevidas ou de difícil avaliação prática.

Referências

- http://michael.hahsler.net/
- https://algobeans.com/