Aula 09 - Regras de associação

Na aula anterior, vimos sobre como funciona o agrupamento, um tipo de algoritmo de aprendizagem de máquina não supervisionada. As regras de associação também são um tipo de aprendizagem não supervisionada.
Mas qual o objetivo das regras de associação?
Bem, usamos regras de associação quando estamos tentando identificar padrões que são comuns a itens que pertencem a um grande volume de dados. Imagine por exemplo, uma plataforma de streaming de vídeo como a Netflix. Existem certos padrões no consumo de filmes ou séries nessa plataforma. Imagine que, alguém que assiste a um filme dos Thor também assiste ao filme dos vingadores, estes filmes estão associados de alguma forma, por algum padrão, que faz com que quem assiste ao filme do Thor também assista ao filme dos vingadores.
Isso nos dá a seguinte regra, o filme do Thor implica no filme dos vingadores. Dizemos que: {Thor} → {Vingadores}. Essa é a nossa regra de associação, generalizando, {a} → {b}. Isto é, um conjunto de itens 'a' que implica em um conjunto de itens 'b'. Também podemos dizer que há uma forte relação entre os itens 'a' e 'b'. Nesse caso, como temos o filme do Thor implicando no filme dos vingadores, dizemos que o filme do Thor é o **Antecedente** e o filme dos vingadores é o **Consequente**. Imagine agora que em um supermercado a venda de molho de tomate aumenta se o molho de tomate está perto do macarrão. Dizemos que a venda de molho de tomate está associada à venda de macarrão. Faz sentido se pensarmos que o molho de tomate é muito usado para preparar o macarrão. Nesse caso, o macarrão implica na venda de molho de tomate, então o macarrão é o antecedente e o molho de tomate é o consequente.

O algoritmo a priori é o principal algoritmo para minerar regras de associação. Seu objetivo é gerar regras, através de análise de dados, que podem definir ou prever um evento. Como já citado no exemplo do molho de tomate e macarrão. É algo de grande interesse para os mercados e supermercados entender esses padrões de associação que fará com que seus produtos sejam mais vendidos.
Para entendermos o funcionamento do algoritmo a priori é necessário que antes entendamos alguns conceitos:
Transações: O que vem a ser o conceito de transações? A gente pode dizer que uma transação é uma ocorrência registrada de um conjunto de itens. Por exemplo, filmes que alguém assistiu, livros que alguém leu, itens de compras em um carrinho. Então imagine que estamos falando de um aplicativo para realizar compras. A compra de uma determinada quantidade de produtos representa uma transação, enquanto que os produtos que são comprados são os itens.
Itemset: É a frequência de vezes que um item aparece em uma base de dados. E para calcular a frequência de um item, dividimos quantas vezes ele aparece no registro pela quantidade total de registros.
Tomemos como exemplo a seguinte tabela de filmes que foram assistidos por 6 pessoas. O 1 representa que o usuário assistiu o filme e o 0 não assistiu.

Filme	Thor Ragnarok	Capitão américa guerra civil	vingadores guerra infinita	vingadores ultimato
1	0	0	1	0
2	0	1	0	1
3	0	0	0	1
4	1	1	1	1
5	0	1	1	0
6	1	1	1	0
7	1	1	1	1
8	1	0	1	1
9	1	1	1	0
10	1	1	0	1

Então para encontrar a frequência, por exemplo, das pessoas que assistiram ao filme Thor Ragnarok é o total de pessoas que assistiram dividido pelo total de pessoas então: Thor: 6/10 = 60%. Isso significa que o filme do Thor foi assistido em 60% das transações registradas. Este é o nosso suporte. E aí podemos definir regras, para definir se estas regras são relevantes devemos definir um suporte mínimo. Então podemos dizer que a frequência que um item aparece nas transações somente irá ser relevante se o suporte mínimo for de 30%, do contrário, não vamos considerar para a criação das nossas regras. Definir este suporte mínimo faz com que o nosso algoritmo não gere muitas regras que não são necessárias. Além disso, precisamos criar regras que sejam confiáveis e para isso usamos o cálculo de confiança.

Confiança: É a indicação de quão frequente uma regra vai ocorrer. Quanto maior a confiança maior a chance de a regra no dataset seja válida. Ou seja, se um filme ‘a’ foi assistido qual a chance de um filme ‘b’ ser assistido.

E como calculamos a nossa confiança?
Dividimos o número de registros com item A e B pelo total de registros com item A.
Vamos levar em consideração os filmes Thor ragnarok e vingadores guerra infinita. O que faremos é verificar a quantidade de vezes que Thor ragnarok e vingadores guerra infinita aparecem juntos.

Filme	Thor Ragnarok	Capitão américa guerra civil	vingadores guerra infinita	vingadores ultimato
1	0	0	1	0
2	0	1	0	1
3	0	0	0	1
4	1	1	1	1
5	0	1	1	0
6	1	1	1	0
7	1	1	1	1
8	1	0	1	1
9	1	1	1	0
10	1	1	0	1

Verificamos que são 5 vezes que aparecem juntos. Isto quer dizer que 5 pessoas que assistiram ao filme Thor Ragnarok assistiram ao filme vingadores guerra infinita. Dividimos esse número, pelo número de registro de pessoas que assistiram ao filme do Thor Ragnarok. Ou seja, Se Thor Ragnarok então vingadores guerra infinita: Confiança: 5/6 = 0.83%.
Isso significa que a confiança de que essa regra seja verdadeira, isto é, que alguém que tenha assistido ao filme do Thor também tenha assistido ao filme dos vingadores guerra infinita é de 83%. Em resumo o algoritmo faz uma análise dos dados e a partir dos parâmetros de confiança e suporte gera uma determinada quantidade de regras que podem ser analisadas posteriormente.