

Módulo 7

Aprendizaje de máquina no supervisado

2.1 Reglas de asociación,

Market Basket

Eduardo Espinosa Avila



Reglas de asociación, *Market Basket*

Muchos negocios acumulan grandes cantidades de datos en sus actividades diarias.

Por ejemplo, en tiendas de autoservicio recolectan datos de compras diariamente en sus recibos de compra.

El ***análisis de asociaciones*** intenta extraer ***reglas de asociación*** de interés, ocultas en esta gran cantidad de datos.

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Market Basket, definición

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Por ejemplo, de la tabla se podría obtener la regla:

$$\{\text{Diapers}\} \longrightarrow \{\text{Beer}\}.$$

Que sugiere una relación entre pañales y cerveza. Las tiendas pueden aprovechar este tipo de reglas para vender otros productos a sus clientes.

Market Basket, definición

La tabla anterior se puede presentar en *formato binario*: cada fila representa una nota, y cada columna es un producto. En cada renglón, un 1 indica presencia del producto y un 0 denota su ausencia.

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Market Basket, definición

Con ayuda de esta representación, se pueden calcular las medidas que sustentan las reglas de asociación del tipo:

$$X \longrightarrow Y$$

- *Support*: es una medida importante, dado que si su valor es bajo puede indicar que la asociación es sólo por azar y, por tanto, resulta de poco interés:

$$\frac{\sigma(X \cup Y)}{N}$$

- *Confidence*: mide la confianza de la regla. Mayor valor indica que es más probable que Y esté presente en transacciones que incluyan a X :

$$\frac{\sigma(X \cup Y)}{\sigma(X)}$$

- Para ambas, σ representa el número de ocurrencias de un conjunto en las transacciones y N es el total de filas.

Market Basket, ejemplo

Con la tabla siguiente, calculemos las medidas para la regla mostrada:

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

$$\{\text{Milk, Diapers}\} \longrightarrow \{\text{Beer}\}$$

- *Support*: filas en las que se encuentran $\{\text{Milk, Diapers, Beer}\} = 2$ y el total de filas = 5. Por lo tanto, esta medida tiene valor de $\frac{2}{5} = 0.4$
- *Confidence*: filas que contienen $\{\text{Milk, Diapers, Beer}\} = 2$ y filas que incluyen $\{\text{Milk, Diapers}\} = 3$. Por lo tanto, esta medida tiene valor de $\frac{2}{3} = 0.67$

Reglas de asociación, definición

Dado un conjunto de transacciones, encontrar todas las reglas que tengan $support \geq minsup$ y $confidence \geq minconf$, donde, $minsup$ y $minconf$ son umbrales esperados.

El número total de reglas que pueden extraerse de un conjunto de transacciones es exponencial, con respecto al número de productos. Por este motivo, la tarea suele dividirse en:

- **Generación del *Frequent Itemset*:** encontrar los conjuntos que satisfacen $minsup$.
- **Generación de reglas:** extraer las reglas con alta $confidence$, a partir de los conjuntos generados en el paso previo.

Market Basket, aplicaciones

Una vez que se han extraído las reglas, se pueden utilizar para varias tareas.

- Recomendaciones de compras
 - Productos afines
- Colocación de productos
- Detección de fraudes
- Analizar datos de encuestas
 - Gracias a que procesa datos categóricos



<https://twitter.com/lisachwinter/status/494063906625961984>

Referencias

- Tan, Pang-Ning, Steinbach, Michael, Karpatne, Anuj and Kumar, Vipin
Introduction to Data Mining (Second Edition)
/ Pang-Ning Tan, Michael Steinbach, Anuj Karpatne and Vipin Kumar --
Birmingham - Mumbai : Pearson, 2019
<https://www-users.cs.umn.edu/~kumar001/dmbook/index.php>
- Rakesh Agrawal, Tomasz Imieliński and Arun Swami
Mining association rules between sets of items in large databases
<https://dl.acm.org/doi/10.1145/170036.170072>
- Michael Hahsler, Bettina Grün, Kurt Hornik and Christian Buchta
Introduction to arules – A computational environment for mining association rules and frequent item sets
<https://mran.revolutionanalytics.com/web/packages/arules/vignettes/arules.pdf>

Contacto

Dr. Eduardo Espinosa Avila

laloea@fisica.unam.mx

Tels: 5556225000 ext. 5003

Redes sociales:

<https://twitter.com/laloea>

<https://www.linkedin.com/in/eduardo-espinosa-avila-84b95914a/>