Modelos de Soporte No Supervisado

Market Basket Association Analysis

Market Basket Association Analysis

- Nos permite identificar patrones en las compras de los clientes.

- Idealmente, nos gustaría responder preguntas como:

    - ¿Qué productos tienden a comprarse juntos?
    - ¿Qué productos pueden beneficiarse de la promoción?
    - ¿Cuáles son las mejores oportunidades de venta cruzada?

El ejemplo clásico de cerveza y pañales (una leyenda urbana, de hecho).

Market Basket Association Analysis: Transaction Data

- Una tienda vende un gran conjunto de productos.
- Una transacción (basket) $t \subseteq \boldsymbol{P}$

- es un conjunto de productos comprado por un cliente en un momento determinado.

- El conjunto $\text{set } \boldsymbol{T} = \{t\}$ de transacciones a menudo se codifica como una matriz binaria escasa (sparse binary matrix).

Market Basket Association Analysis

	p_1	p_2	p_3	p_4
t_1	0	1	1	1
t_2	1	0	0	1
t_3	0	1	1	0

Market Basket Association Analysis: Association Rules

- El método más popular es MBA.

- Genera reglas de la forma. $A \rightarrow B$

- A y B son conjuntos de productos disjuntos arbitrarios

$$(\text{often } |B| = 1)$$

- $A \rightarrow B$ implica que si A<ocurre en una canasta en particular entonces B debería ocurrir en esa canasta también.

$$\{\text{peanut butter, jelly}\} \rightarrow \{\text{bread}\}$$

Market Basket Association Analysis: Filtering the Rules

- Solo se genera un subconjunto reducido de reglas interesantes.
- $R(\boldsymbol{T}, s, c)$ es el conjunto de reglas obtenidas de T con

  Minimum support $\boldsymbol{s} \in [\mathbf{0}, \mathbf{1}]$.

  Minimum confidence $\boldsymbol{c} \in [\mathbf{0}, \mathbf{1}]$.

- Sea A el conjunto de transacciones que contiene cada producto en A (lo mismo para B).
- Entonces R (T, s, c) es el conjunto de reglas A → B tales que:

$$\hat{P}(A \cup B) = \frac{|\boldsymbol{A} \cap \boldsymbol{B}|}{|\boldsymbol{T}|} \geq s \quad \text{and} \quad \hat{P}(B|A) = \frac{|\boldsymbol{A} \cap \boldsymbol{B}|}{|\boldsymbol{A}|} \geq c$$

Minimum support ***Minimum confidence***

Market Basket Association Analysis: A priori Algorithm

Dado T, podemos generar R(T, s, c) de manera muy eficiente.

Algoritmo APRIORI:

1. Identifique los conjuntos de productos frecuentes (FIS- Frequent Item Sets) tal que |C|>=s
2. Por un valor razonablemente alto (***s)***, el número total de FIS debe ser pequeño.
3. C es frecuente ⇔ cualquier subconjunto de C es también frecuente.
4. FIS de tamaño K contiene K subconjuntos de tamaño K-1 que también son FIS.
5. Para cualquier elemento frecuente establecido C y cualquier p ∈ C, genere la regla

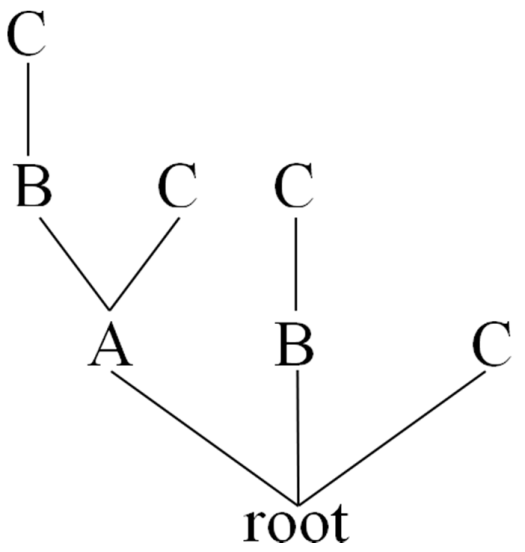
$$C - \{p\} \rightarrow \{p\} \text{ if}$$

$$\hat{P}(\{p\}|C - \{p\}) = \frac{|C|}{|C - \{p\}|} \geq c$$

Efficient Identification of FIS

6. El FIS con K elementos se almacena en una estructura de árbol.
7. El árbol se extiende con los conjuntos candidatos que contienen K+1 elementos.
8. Son la unión de dos FIS con k elementos y un padre común.
9. Una única pasada T elimina los candidatos C de modo que |C|<s.

- El costo computación está determinado por el número de conjuntos de elementos candidatos
- Si (**s)** es muy bajo el número de candidatos crece exponencialmente



Other Interest Measures

- Se utiliza un gran soporte para mantener baja la cantidad de reglas encontradas.
- Pero esto elimina las reglas potencialmente interesantes.
- En la práctica, generar un gran número de reglas es inevitable.
- Se pueden usar medidas de interés adicionales para filtrar las reglas.
- El interés es a menudo la "desviación de la independencia".
- La sustentación **(lift)** de una regla A → B se define como

$$L(A \rightarrow B) = \frac{\hat{P}(A \cup B)}{\hat{P}(A)\hat{P}(B)}$$

- Si A y B son perfectamente independientes, entonces L(A → B) = 1.

*En términos práticos lift = (predicted rate / average rate)

Other Interest Measures

- R incluye una implementación del agoritmo Apriori en el paquete "***arules"***.
- El cuál está basado en el eficiente código C desarrollado por Christian Borgelt (2002).
- El paquete "***arulesViz"*** permite visualizar reglas de asociación.

```
> library(arules)
> library(arulesViz)
> data("Groceries")
> Groceries
transactions in sparse format with
 9835 transactions (rows) and
 169 items (columns)
>
```

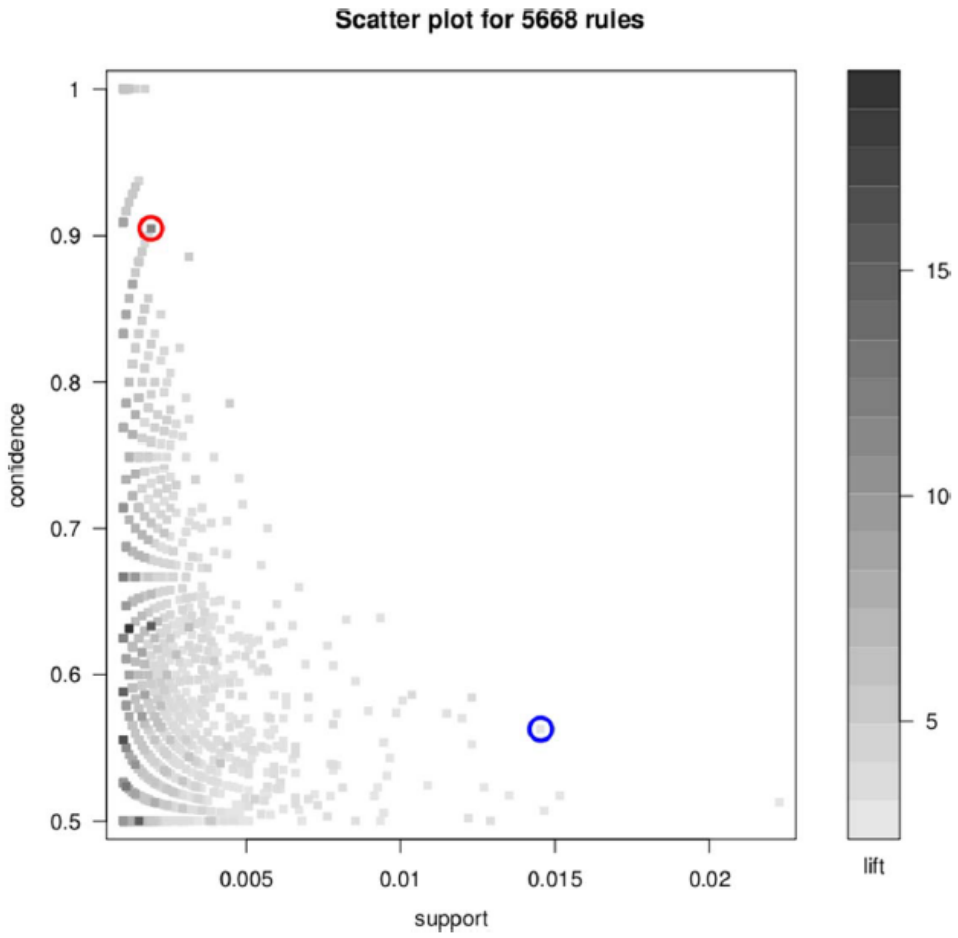
Other Interest Measures

```
> rules <- apriori(Groceries, parameter = list(support = 0.001, confidence = 0.5))

parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target
        0.5    0.1    1 none FALSE            TRUE   0.001      1     10  rules
   ext
 FALSE

algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].
sorting and recoding items ... [157 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.03s].
writing ... [5668 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
>
```

Other Interest Measures

```
> inspect(head(sort(rules, by = "lift"), 7))
  lhs                          rhs                  support confidence     lift
1 {Instant food products,
   soda}                    => {hamburger meat} 0.001220132  0.6315789 18.99565
2 {soda,
   popcorn}                 => {salty snack}    0.001220132  0.6315789 16.69779
3 {flour,
   baking powder}           => {sugar}          0.001016777  0.5555556 16.40807
4 {ham,
   processed cheese}        => {white bread}    0.001931876  0.6333333 15.04549
5 {whole milk,
   Instant food products} => {hamburger meat} 0.001525165  0.5000000 15.03823
6 {other vegetables,
   curd,
   yogurt,
   whipped/sour cream}      => {cream cheese }  0.001016777  0.5882353 14.83409
7 {processed cheese,
   domestic eggs}           => {white bread}    0.001118454  0.5238095 12.44364
>
```

Other Interest Measures



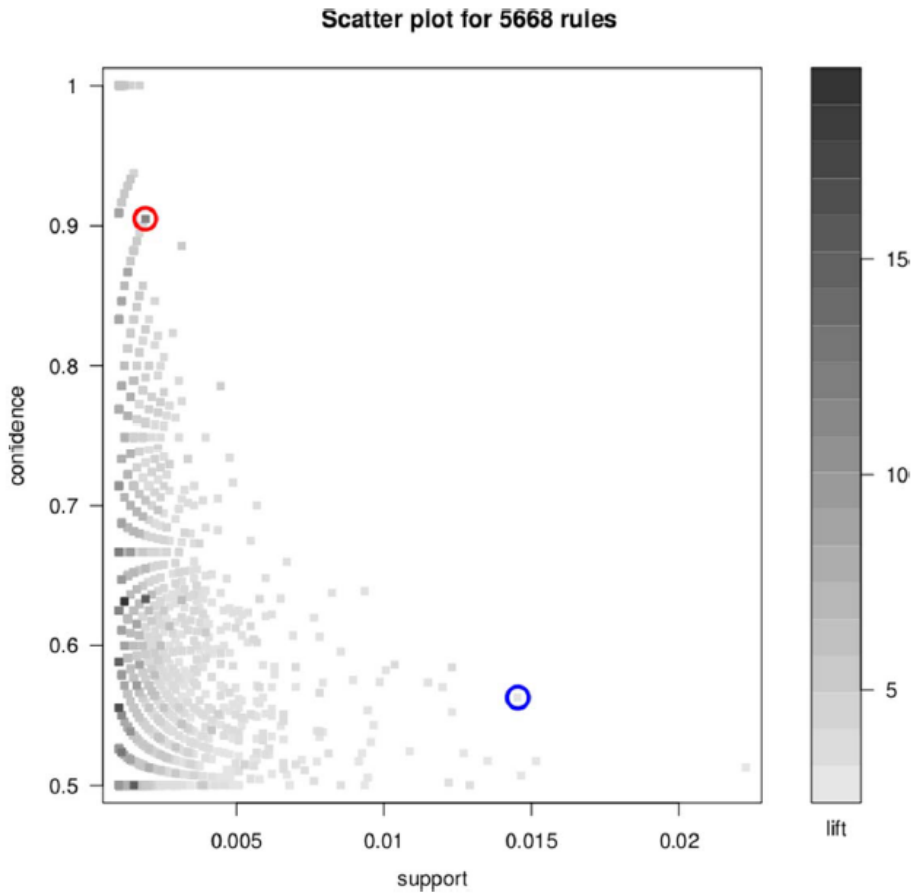
- {liquor, red/blush wine} → {bottled beer}

  Support: 0.001931876
  Confidence: 0.9047619
  Lift: 11.23527

- {root vegetables, yogurt}→ {whole milk}

  Support: 0.01453991
  Confidence: 0.5629921
  Lift: 2.203354

Other Interest Measures



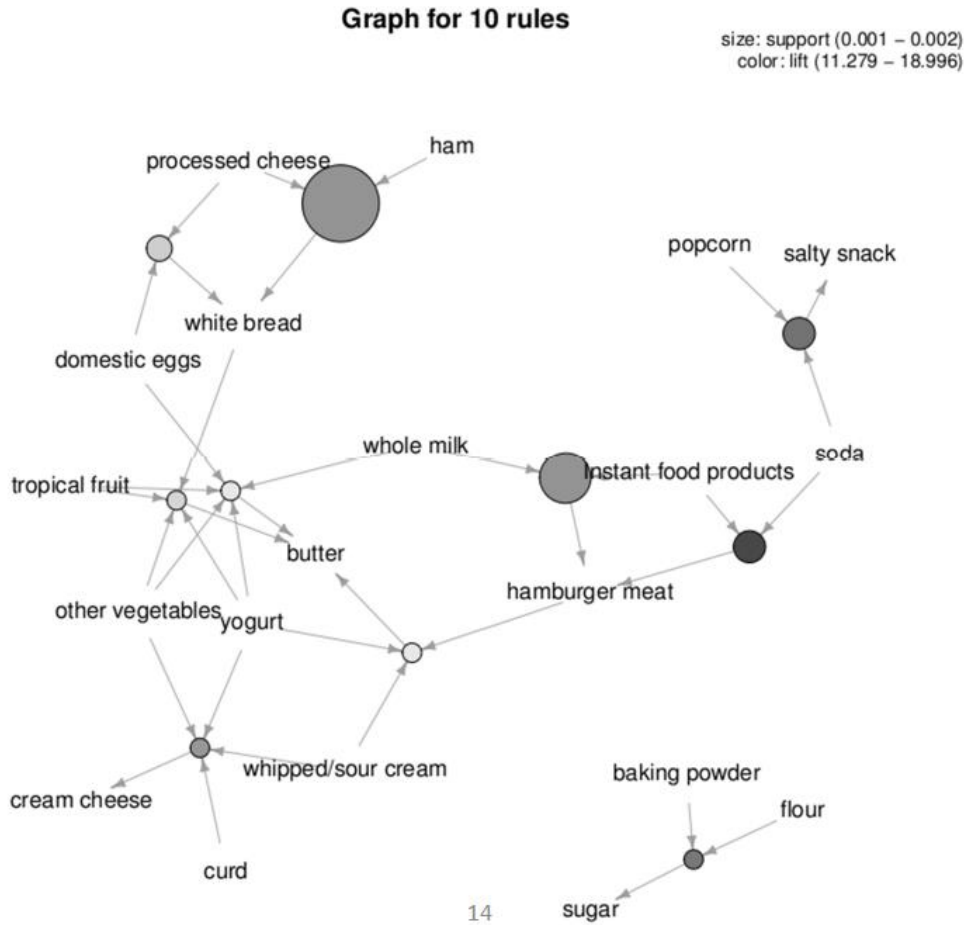
- {liquor, red/blush wine} → {bottled beer}

  Support: 0.001931876
  Confidence: 0.9047619
  Lift: 11.23527

- {root vegetables, yogurt}→ {whole milk}

  Support: 0.01453991
  Confidence: 0.5629921
  Lift: 2.203354

Other Interest Measures



Other Interest Measures

Ventajas y desventajas de usar Reglas de asociación

Ventajas:

1. Computacionalmente eficiente (¡siempre y cuando **(s)** es grande!).
2. Las reglas individuales son fáciles de interpretar.
3. Método bien investigado.

Desventajas:

1. No está claro cómo elegir ***s*** y ***c***.
2. El número de reglas obtenidas a menudo es muy grande.
3. Difícil aislar patrones interesantes. Ningún procedimiento ampliamente aceptado.
4. A veces la mayoría de las reglas obtenidas son triviales.
5. Falta de un modelo probabilístico riguroso para los datos.