M2.2.2 Modelos Supervisados y No Supervisados

Programa Big Data y Business Intelligence

Enrique Onieva
enrique.onieva@deusto.es
https://twitter.com/EnriqueOnieva
https://www.linkedin.com/in/enriqueonieva/

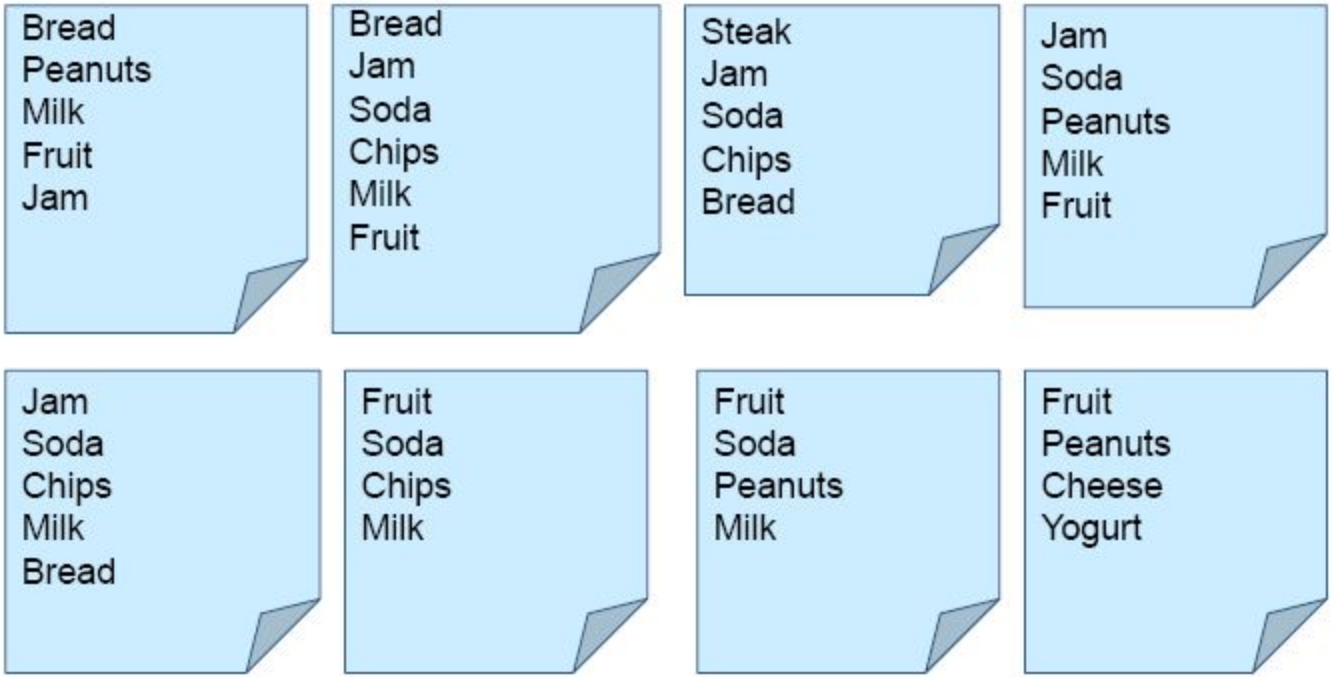
Deusto
Facultad de Ingeniería
Ingeniaritza Fakultatea

Reglas de Asociación

- Introducción a métodos de asociación
- ¿Qué es una regla de asociación?
- ¿Cómo evalúo la calidad de una regla de asociación?
- Métodos para establecer reglas de asociación
- El algoritmo Apriori

Ejemplo clásico

- Cesta de la compra:
  - Encontrar parejas de productos que suelen aparecer juntos en la cesta



Métodos de Asociación

- Tratan de encontrar patrones frecuentes, asociaciones, correlaciones dentro de conjuntos de elementos u objetos en bases de datos
- Aplicaciones
  - Análisis de la cesta de la compra
  - Marketing cruzado
  - Diseño de catálogos
  - …

Extracción de reglas

- Dado un conjunto de transacciones
  - Encontrar reglas que predicen la aparición de un elemento basándose en la aparición de otros
- Ejemplos
  - {Bread} → {Milk}
  - {Soda} → {Chips}
  - {Bread} → {Jam}
  - …

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

Definiciones

- Itemset
  - Conjunto de uno o más ítems
    - {milk, bread, jam}
  - K-itemset: un itemset con exáctamente k elementos
- Soporte
  - Porcentaje de transacciones que contienen un itemset
    - Soporte({milk, bread}) = 3
    - Soporte({soda, chips}) = 3
- Itemset frecuente
  - Un itemset cuyo soporte es superior a determinado umbral

¿Qué es una regla?

- Implicación de la forma X⟶ Y, siendo X e Y itemsets
  - {bread} ⟶ {milk}
- Evaluación de la bondad de una regla
  - Soporte: porcentaje de transacciones que contienen ambos itemsets X e Y
  - Confianza: mide con qué frecuencia elementos de Y aparecen junto a X en una transacción

$$s = \frac{\sigma(\{\text{Bread, Milk}\})}{\#\text{ of transactions}} = 0.38 \qquad c = \frac{\sigma(\{\text{Bread, Milk}\})}{\sigma(\{\text{Bread}\})} = 0.75$$

¿Qué es una regla?

- Tenemos dos bases de transacciones
  - En ambos casos, si miramos la regla "Bread→ Milk" tenemos:
    - Soporte = 0.5
      - (En el 50% de los casos aparecen juntos)
    - Confianza = 0.71
      - (El 71% de las veces que aparece Bread, aparece Milk)
  - ¿Son ambas reglas igual de "buenas"?
    - ¿Cuál debería de ser "mejor"?
    - ¿Cómo se podría medir?

Base de Transacciones 1			
ticket 1	Bread	Milk	----
ticket 2	Bread	Milk	----
ticket 3	Bread	Milk	----
ticket 4	Bread	Milk	----
ticket 5	Bread	Milk	----
ticket 6	Bread	----	----
ticket 7	Bread	----	----
ticket 8	----	----	----
ticket 9	----	----	----
ticket 10	----	----	----

Base de Transacciones 2			
ticket 1	Bread	Milk	----
ticket 2	Bread	Milk	----
ticket 3	Bread	Milk	----
ticket 4	Bread	Milk	----
ticket 5	Bread	Milk	----
ticket 6	Bread	----	----
ticket 7	Bread	----	----
ticket 8	----	Milk	----
ticket 9	----	Milk	----
ticket 10	----	Milk	----

¿Qué es una regla?

- Una regla de asociación (X → Y) se evalúa en función de
  - Soporte: nos indica cómo de frecuentemente aparecen dos elementos en la misma transacción
    - Es decir qué porcentaje de transacciones "soportamos" con esa regla
  - Confianza: nos indica cómo de fiable es la regla. Mide cómo de frecuentemente, que aparezca X implica Y
  - "Lift": Nos da una idea de si la confianza de esa regla se debe a la "casualidad"
    - Si es igual a 1, indica que las probabilidades de aparición de X e Y son independientes.
    - Si es mayor que 1, indica que ciertamente, el que aparezca X implica que aparecerá Y

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$$

¿Qué es una regla?

- Si miramos la regla "Bread→ Milk" tenemos:
  - Soporte = 0.5
    - (En el 50% de los casos aparecen juntos)
  - Confianza = 0.71
    - (El 71% de las veces que aparece Bread, aparece Milk)
  - "Lift" en la Base 1 = 1.43
  - "Lift" en la Base 2 = 0.89
    - En la Base 2, la confianza se debe a que Milk es un item muy frecuente

Base de Transacciones 1			
ticket 1	Bread	Milk	----
ticket 2	Bread	Milk	----
ticket 3	Bread	Milk	----
ticket 4	Bread	Milk	----
ticket 5	Bread	Milk	----
ticket 6	Bread	----	----
ticket 7	Bread	----	----
ticket 8	----	----	----
ticket 9	----	----	----
ticket 10	----	----	----

Base de Transacciones 2			
ticket 1	Bread	Milk	----
ticket 2	Bread	Milk	----
ticket 3	Bread	Milk	----
ticket 4	Bread	Milk	----
ticket 5	Bread	Milk	----
ticket 6	Bread	----	----
ticket 7	Bread	----	----
ticket 8	----	Milk	----
ticket 9	----	Milk	----
ticket 10	----	Milk	----

¿Qué es una regla?

- Si miramos la regla “Bread→ Milk”:
  - Soporte = 0.5
  - Confianza = 0.71
  - “Lift“ en la Base 1 = 1.43
  - “Lift“ en la Base 2 = 0.89
- Si miramos la regla “Milk→ Bread”
  - Soporte = 0.5
  - Confianza en la Base 1 = 1.00
  - Confianza en la Base 2 = 0.63
  - “Lift“ en la Base 1 = 1.43
  - “Lift“ en la Base 2 = 0.89
- ¿Cuál es mejor en qué Base?

Base de Transacciones 1			
ticket 1	Bread	Milk	----
ticket 2	Bread	Milk	----
ticket 3	Bread	Milk	----
ticket 4	Bread	Milk	----
ticket 5	Bread	Milk	----
ticket 6	Bread	----	----
ticket 7	Bread	----	----
ticket 8	----	----	----
ticket 9	----	----	----
ticket 10	----	----	----

Base de Transacciones 2			
ticket 1	Bread	Milk	----
ticket 2	Bread	Milk	----
ticket 3	Bread	Milk	----
ticket 4	Bread	Milk	----
ticket 5	Bread	Milk	----
ticket 6	Bread	----	----
ticket 7	Bread	----	----
ticket 8	----	Milk	----
ticket 9	----	Milk	----
ticket 10	----	Milk	----

¿Cuál es el objetivo?

- Dado un conjunto de transacciones:
  - Encontrar reglas que superen unos umbrales establecidos de soporte y confianza

{Bread, Jam} ⇒ {Milk} s=0.4 c=0.75
{Milk, Jam} ⇒ {Bread} s=0.4 c=0.75
{Bread} ⇒ {Milk, Jam} s=0.4 c=0.75
{Jam} ⇒ {Bread, Milk} s=0.4 c=0.6
{Milk} ⇒ {Bread, Jam} s=0.4 c=0.5

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

Métodos para extraer reglas

- Se suelen basar en variaciones de un enfoque en dos pasos:
  - Generar aquellos itemsets frecuentes
    - Con soporte mayor que un determinado umbral
  - Generación de reglas
    - Generar reglas con alta confianza haciendo uso de los itemsets frecuentes obtenidos en el paso anterior
    - Cada regla es una partición binaria de un itemset frecuente
    - Abajo tenemos todas las particiones binarias del itemset {Bread, Jam, Milk}

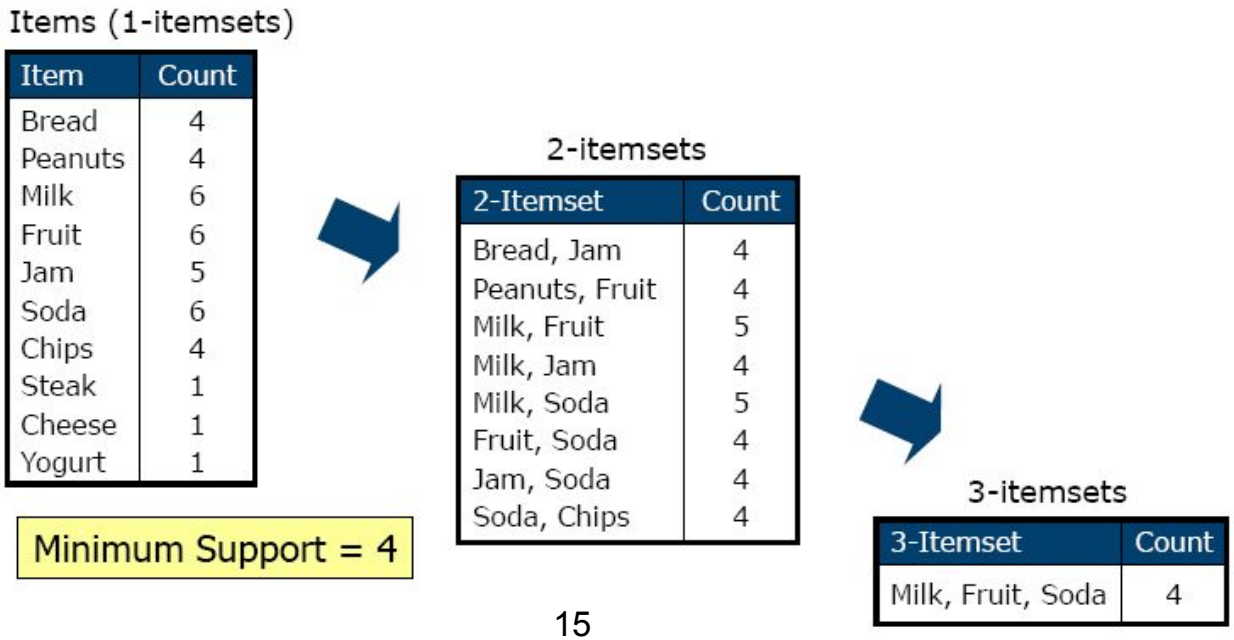
{Bread, Jam} ⇒ {Milk} s=0.4 c=0.75
{Milk, Jam} ⇒ {Bread} s=0.4 c=0.75
{Bread} ⇒ {Milk, Jam} s=0.4 c=0.75
{Jam} ⇒ {Bread, Milk} s=0.4 c=0.6
{Milk} ⇒ {Bread, Jam} s=0.4 c=0.5

Generación de Itemsets

- Hay muchos itemsets candidatos a explorar (2^N-1)
  - Para 3 productos {ABC}
    - {A}, {B}, {C}, {AB}, {AC}, {BC}, {ABC}
  - Para 4 productos {ABCD}
    - {A}, {B}, {C}, {D}, {AB}, {AC}, {AD}, {BC}, {BD}, {CD}, {ABC}, {ABD}, {ACD}, {BCD}, {ABCD}
- Es intratable cuando los elementos crecen
  - Para 25 productos → 33554431
  - Para 100 productos → 1.2676506e+30

Generación de Itemsets

- El algoritmo Apriori
  - Comenzar con itemsets de tamaño k=1, ir incrementando el valor de k de 1 en 1, descartando aquellos itemsets que no cumplan un soporte mínimo



Generación de Reglas

- Dado un itemset frecuente, encontrar reglas que tienen una mínima confianza
  - Para 3 Productos {ABC}
    1. A → BC
    2. B→AC
    3. C→ AB
    4. AB→ C
    5. AC→ B
    6. BC→ A
  - Es intratable cuando el número de elementos crece
    - Para 25 productos → 33554430 (1 menos que antes)
- Al igual que antes, el algoritmo Apriori deja de explorar alternativas que presentan baja confianza

¿Qué valor de soporte mínimo?

- Si el soporte mínimo es muy grande
  - Podemos perder itemsets que incluyan elementos de interés, pero poco frecuentes (productos caros)
- Si el soporte mínimo es muy bajo
  - Es computacionalmente muy difícil de calcular (gran número de itemsets)
- Heurística
  - Comenzar con un soporte alto, e ir reduciendo hasta obtener un número de reglas adecuado
  - Hasta que sea computacionalmente factible ejecutar el algoritmo

Enrique Onieva
enrique.onieva@deusto.es
https://twitter.com/EnriqueOnieva
https://www.linkedin.com/in/enriqueonieva/