Reglas de Asociación y Análisis de Canastas de Mercado

Jorge Gallego

Facultad de Economía, Universidad del Rosario

mayo 9 de 2017

Introducción

- Los supermercados son estratégicos al organizar sus productos
- Pocos detalles son dejados al azar
- Se agrupan para maximizar la probabilidad de compras conjuntas
- Los algoritmos de aprendizaje supervisado han sido clave para esto
- Y no solo los supermercados lo hacen

Sistemas de Recomendación

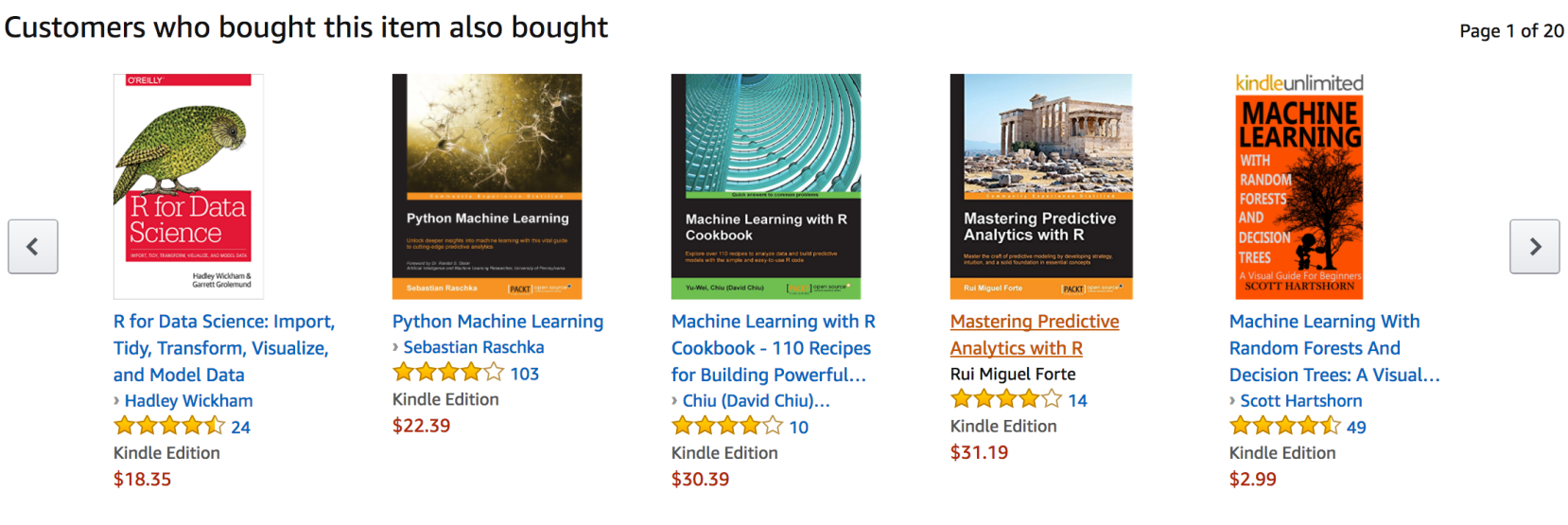
Figure: Supermercados

Sistemas de Recomendación

Figure: Vintage

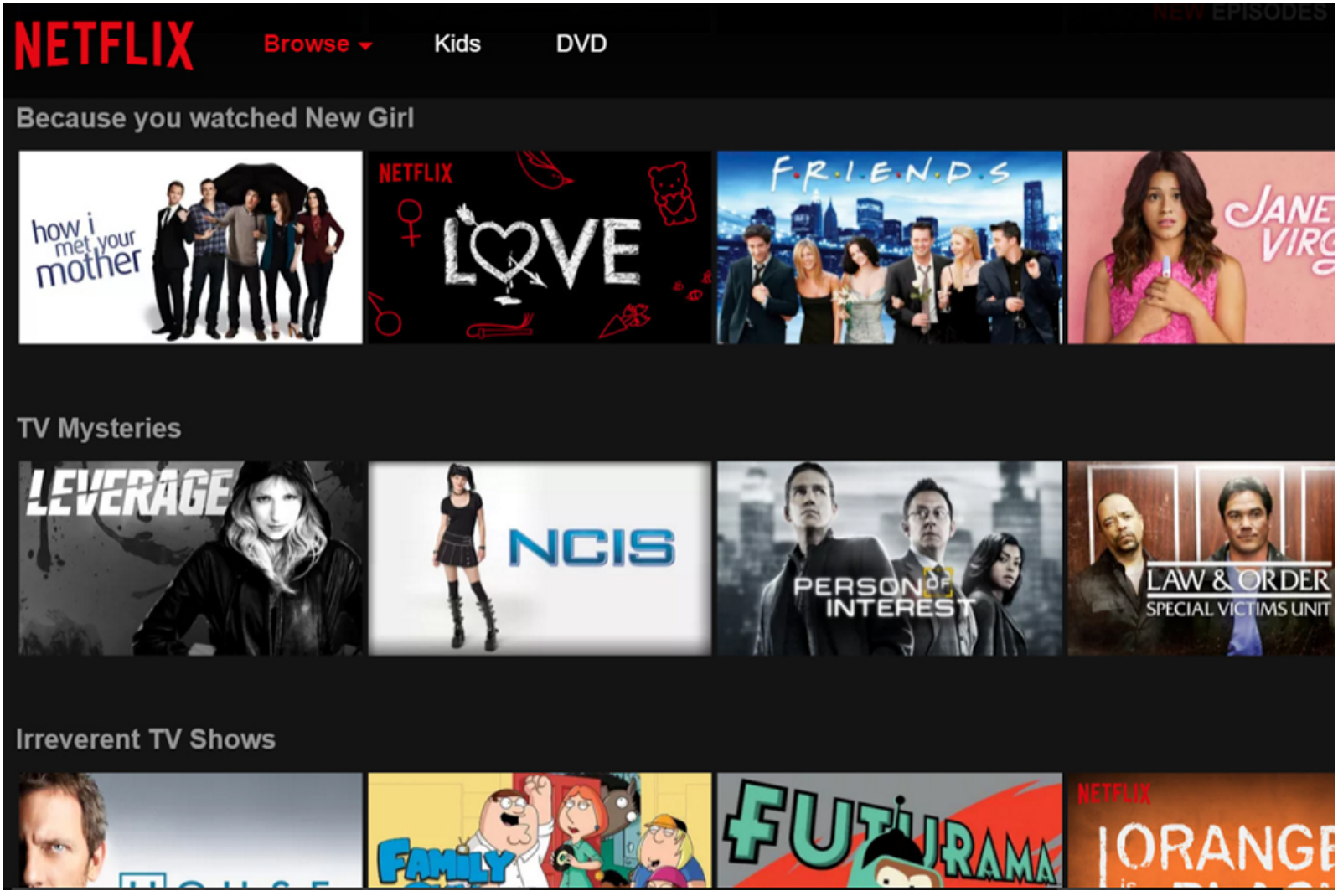
Sistemas de Recomendación

Figure: Amazon



Sistemas de Recomendación

Figure: Netflix



Introducción

- Por mucho tiempo estos sistemas se basaron en la intución de los vendedores
- Pero los datos disponibles hoy han revolucionado el tema
- Scanners de códigos, sistemas digitales de inventario, compras en línea, datos transaccionales
- Técnicas de machine learning usan estos datos para encontrar patrones
- Veremos los fundamentos del *market basket analysis*

Reglas de Asociación

- El fundamento del análisis son los items que forman parte de una transacción
- Llamaremos **itemset** a un grupo de uno o varios items
- Por ejemplo: {pan, jamón, queso}
- Construiremos reglas de asociación
- Especifican patrones encontrados en las relaciones entre items de un itemset
- Relacionan un itemset en el lado izquierdo de la relación, con otro en el lado derecho

Reglas de Asociación

- Una regla de asociación puede ser:

$$\{\text{jamón, queso}\} \rightarrow \{\text{pan}\}$$

- Significa que si se compra jamón y queso, es probable que se compre pan también
- Acá el objetivo central no es predecir sino encontrar patrones interesantes de manera no supervisada
- No es necesario entrenar un modelo como tal, ni identificado en una categoría u otra

Algoritmo Apriori

- Incluso para las máquinas es complejo analizar datos transaccionales
- Suelen ser bases de datos muy grandes
- Muchas transacciones y muchos productores transados
- El número de itemsets crece exponencialmente
- Es imposible evaluarlos a todos para encontrar patrones

Algoritmo Apriori

- Para hacer el proceso más eficiente se descartan itemsets
- Aquellos poco frecuentes, de baja probabilidad
- Por ejemplo: {aceite de motor, labial}
- Las combinaciones raras y poco importantes son ignoradas
- Esto es importante para disminuir el número de asociaciones posibles

Algoritmo Apriori

- El algoritmo más usado para reducir el número de itemsets es el Apriori
- Se utilizan creencias simples sobre la frecuencia de un itemset
- Supuesto: un itemset es frecuente si sus subconjuntos son frecuentes también
- {aceite de motor, labial} son frecuentes si {aceite de motor} y {labial} son frecuentes también
- Si alguno de los dos es infrecuente, el itemset conjunto lo será

Algoritmo Apriori

Consideremos el siguiente ejemplo:

Figure: Compras en Tienda de Hospital

Transaction number	**Purchased items**
1	*{flowers, get well card, soda}*
2	*{plush toy bear, flowers, balloons, candy bar}*
3	*{get well card, candy bar, flowers}*
4	*{plush toy bear, balloons, soda}*
5	*{flowers, get well card, soda}*

Algoritmo Apriori

- Dos patrones claros se infieren
- Quien visita a un enfermo, le compra flores y una tarjeta
- Quien visita a una nueva madre, le compra un oso y globos
- Los patrones descritos se derivan de ciertas frecuencias
- Veremos cómo el algoritmo Apriori cuantifica estas frecuencias

Medidas de Interés

- El algoritmos cuantifica qué tan "interesante" es un itemset
- Dos medidas se usan para esto: soporte (*support*) y confianza (*confidence*)
- Se establencen umbrales para estas medidas y se aplica el principio Apriori
- Para reducir el número de reglas reportadas
- Veamos cada medida en detalle

Medida de Soporte

- El soporte de un itemset mide con qué frecuencia ocurre en los datos
- Se define como

$$support(X) = \frac{count(X)}{N}$$

- donde N es el número total de transacciones y $count(X)$ es el número de transacciones que contienen a X
- Por ejemplo, $support(\{get\ well\ card, flowers\}) = 3/5 = 0.6$
- Similarmente, $support(\{candy\ bar\}) = 2/5 = 0.4$

Medida de Confianza

- La confianza de una regla mide su poder predictivo o precisión
- Se define como

$$confidence(X \rightarrow Y) = \frac{support(X, Y)}{support(X)}$$

- Es la proporción de veces que ocurren X y Y conjuntamente, sobre eventos en los que ocurre X
- Si la confianza es 1, siempre que ocurre X, ocurre Y. Luego X predice muy bien a Y

Medida de Confianza

- Por ejemplo,

$$\textit{confidence}(\text{flowers} \rightarrow \text{get well card}) = 0.6/0.8 = 0.75$$

- Con una confianza del 75%, si alguien compra flores, compra también la tarjeta
- Ojo: $\textit{confidence}(X \rightarrow Y) \neq \textit{confidence}(Y \rightarrow X)$

$$\textit{confidence}(\text{get well card} \rightarrow \text{flowers}) = 0.6/0.6 = 1$$

- Con una confianza del 100%, si alguien compra la tarjeta, compra flores también

Algoritmo Apriori

- Reglas tipo get well card $\rightarrow$ flowers son fuertes porque tienen soporte y confianza altos
- El algoritmo Apriori usa niveles mínimos de soporte y confianza para encontrar reglas fuertes
- Lo hace descartando reglas no interesantes
- El principio es que si $\{A, B\}$ son frecuentes, entonces $\{\mathrm{A}\}$ y $\{\mathrm{B}\}$ deben serlo también
- Si por ej. $\{\mathrm{A}\}$ tiene soporte bajo (según cierto umbral), se descarta y no se considera $\{\mathrm{A,B}\}$

Algoritmo Apriori

El proceso de crear reglas con el algoritmo Apriori es:

1. Identificar todos los itemsets cuyos soportes superan cierto umbral mínimo

2. Crear reglas de estos itemsets de aquellas que cumplan cierto umbral mínimo de confianza

Medida de Confianza

- La primera fase se hace de manera iterativa
- En la iteración se evalúan los itemsets de tamaño 1
- En la 2 los de tamaño 2. Y así sucesivamente
- El resultado de cada iteración es el conjunto de itemsets cuyo soporte supera el umbral

Ejemplo

- Supongamos cuatro items: A, B, C, D
- Solo A, B, C cumplen con el soporte mínimo; D se descarta en la primera iteración
- En la segunda solo se evalúan {A,B}, {A,C} y {B,C}
- Supongamos que solo {A,B} y {A,C} son frecuentes. Se descarta {B,C}
- Luego, el algoritmo se detiene y no debe evaluar {A, B, C}

Ejemplo

- Luego de esto arranca la segunda fase del algortimo
- Se forman las reglas de asociación
- Por ejemplo, $A \rightarrow B$ o $B \rightarrow A$
- Sobreviven las que superen el umbral de la medida de confianza
- Se interpretan las reglas finales

Reglas de Asociación

- ¿Cómo medir la importancia de una regla?
- El soporte y la confianza cumplen este propósito
- Pero hay otra medida: el lift:

$$lift(X \rightarrow Y) = \frac{confidence(X \rightarrow Y)}{support(Y)}$$

Reglas de Asociación

- El lift mide qué tanto es más probable que se compre Y dado que se compra X, en relación a su tasa típica de compra
- Luego mide qué tanto ganamos sobre la probabilidad de compra de Y al saber que X se compra
- Si lift es mayor que 1, sabemos que es más probable que se compre Y cuando se compra X respecto a la tasa típica
- Reglas con lift alto tienen mayor importancia

Reglas de Asociación

Conviene clasificar a las reglas finales en una de tres categorías

1. Accionables
2. Triviales
3. Inexplicables

El objetivo es encontrar reglas accionables