

Laboratorio 03: Reglas de asociación

Objetivos de aprendizaje

Familiarizar al estudiante ante una situación similar a la real en donde se requiere analizar un conjunto de datos. En particular, transacciones de compra de una tienda de conveniencia (retail).

Explorar, describir y argumentar el comportamiento de distintas implementaciones de algoritmos distintos que resuelven el mismo problema. En particular, algoritmos de reglas de asociación.

Elaborar reportes académicos posteriores a la experimentación, medición y observación de fenómenos estudiados. Medición de rendimiento de dos implementaciones, elaboración de apoyo a la explicación del fenómeno, y argumentación basada en la observación, razonamiento y estudio del fenómeno.

Equipo y fecha de entrega

- Tamaño del equipo: 2 estudiantes máximo.
- Fecha de entrega: lunes 15 de marzo, 23:59
- Tipo de entrega: carga en la plataforma del aula virtual.

Entregables

- Documento en Word o PDF con una longitud máxima de 4 páginas. Tamaño de fuente: 12pt. Tipo de Fuente: times new roman. Columnas: 2. El reporte debe contener al menos las siguientes secciones: introducción, respuesta a las preguntas del taller, y conclusiones. El reporte puede contener gráficos y tablas que contribuyan a la explicación del texto.
- Código en Python/Notebook con el desarrollo del taller hola adjunto el envío.
- Adjunte un solo archivo .ZIP con los el contenido arriba mencionado.

Desarrollo

A cada uno de los grupos se le asigna un archivo de datos qué contiene las facturas de venta de una tienda de retail.

Descargue el archivo asignado a su grupo. El archivo tiene un formato de texto que se encuentra estructurado como una línea por factura de compra. La factura contiene múltiples ítems vendidos, la cantidad de elementos comprados por ítem, y el subtotal por ítem, el valor del impuesto, y el valor.

1. Lea la documentación de las reglas de asociación para los algoritmos apriori y fp-growth.
2. Construya las funciones o métodos que permita leer y *parser* en objetos de tipo line-item y factura.

Un objeto de tipo line-item de factura contiene el nombre del producto, la cantidad de ítems vendidos el producto precio por unidad, y el valor el subtotal de la compra.

Un objeto de tipo Factura contiene: el número o consecutivo de factura, la fecha o marca de tiempo, el total de la factura, el total del impuesto, la cantidad total de ítems.

3. Realice un análisis descriptivo que contenga:

- a. Generalidades del conjunto de datos. Este análisis contiene al menos: la cantidad de registros, facturas, cantidad promedio de ítems por factura, cantidad máxima de ítems encontrado en las facturas del conjunto de datos, distribución de las compras por cantidad y precio de facturas por hora.
 - b. Identifique el producto más vendido por hora y del conjunto de datos.
 - c. Compare la cantidad del producto más vendido con respecto del TOP 10.
 - d. Identifique el producto menos vendido, y su aparición de hora. Describa otras características como si se vendió solo, o en conjunto con otros productos.
4. Construya una hipótesis a partir del análisis inicial. La hipótesis orienta el sentido del análisis, en este caso se quiere identificar cuales productos son más o menos vendidos y si es posible elaborar combos entre los productos más vendidos, los productos menos vendidos y otras relaciones entre productos con la misma/distinta categoría.
5. Utilice la biblioteca xlmtd para encontrar relaciones entre productos. Construya una tabla de comparación entre las dos implementaciones de Reglas de Asociación.
 - a. Utilice la implementación de apriori,
 - b. Utilice la implementación de fp-growth.
 - c. Mida el tiempo de ejecución para cada algoritmo.
 - d. Ejecute múltiples veces las implementaciones con distintos parámetros para cada implementación, mida la ejecución y el resultado obtenido.
6. Preguntas del taller:
 - a. Liste reglas *predecibles* entre los productos vendidos $k=1$. Ejemplo: avena -> pandebono.
 - b. ¿Existen diferencias entre las ejecuciones de apriori y fp-growth?
 - c. ¿Cómo afectan los parámetros, e.g. soporte, tanto en la ejecución como en el resultado?
 - d. ¿Encuentra reglas espurias? ¿Con cuales parámetros?
 - e. Liste reglas con más de 1 producto, $k > 1$.
 - f. ¿Cómo calcula la probabilidad de que al azar se compren estos dos productos?
 - g. ¿Encuentra reglas *interesantes*?
 - h. ¿cómo explica el comportamiento del consumidor para estas reglas?