Práctica 2
Reglas de Asociación y Patrones Secuenciales

Objetivo

Desarrollar un Jupyter-Notebook por entregable (2 en total) que permitan implementar algoritmos para la obtención de relgas de asociación y patrones secuenciales.

Herramientas

- Lenguaje de programación: Python
- Librerías: Numpy, pandas, scikit-learn, matplotlib
- Entorno de gestion de librerías: Anaconda
- Editor: Jupyter

Información de la entrega

La entrega se realizará a través de la tarea LAB2 disponible en la página Canvas de la asignatura.
Consistirá en un fichero comprimido (.zip,.tar.gz) con nombre LAB02-GRUPOxx.zip que contendrá:

1- Un Jupyter-Notebook por cada entregable (archivos con extensión .ipynb)
2- La memoria del laboratorio se entregará integrada en el Notebook de manera que explique y complemente el código entregado
3- El código entregado tiene que ser funcional, correcto y completo.

Las entregas que no se ajusten exactamente a este formato NO SE EVALUARÁN.

Rúbrica

Código

El valor de cada entregable para la nota final de la práctica se indica en el enunciado, así como el valor de cada uno de los apartados.
Todos los aspectos de programación se dan por supuestos.
El código debe ser:

- Funcional: debe ejecutar sin errores y el resultado debe ser el esperable
- Original: el código no puede ser una copia de trabajos publicados en Internet o de otros compañeros. Grupos con código igual serán suspendidos.
- No redundante: se penalizará el código que no sea útil o redundante
- Comentado: es obligatorio incluir comentarios en el código, en su justa medida
- Gráficas: deben incluir todos los datos que sean necesarios

Memoria

La memoria estará incluida en los Jupyter-Notebook que se entreguen de manera que complementen el código entregado. La redacción debe ser clara y correcta ortográfica y gramaticalmente. Debe incluir la justificación de cada paso que se realice para la resolución de los problemas planteados.

Entregable 1 – Reglas de Asociación

Este entregable vale 8 puntos de la nota final de la práctica 2.

El dataset radio.csv recoge qué música prefieren escuchar los usuarios de una emisora de radio. Se quiere utilizar esta información para diseñar un sistema de recomendación de música basado en los intereses/gustos de los usuarios. Para hacer esto, vamos a aplicar el algoritmo Apriori para estudiar los datos y sacar conclusiones mediante las funcionalidades de la librería *mlextend*.

1.1 (1 punto) Estudia brevemente el dataset (número de usuarios, número de atributos, categorías más frecuentes para cada atributo/columna, etc.) y realiza el preprocesamiento para generar el dataset necesario para aplicar el algoritmo Apriori a este conjunto de datos.

1.2 (4 puntos) Aprende a manejar la librería *mlextend* y su implementación del algoritmo Apriori.
1.2.1 Obtén los itemsets frecuentes para k=1. Para esto es necesario obtener el soporte de los itemset. Crea una función que utilice métodos de la librería *mlextend* de manera que dado un itemset devuelva su soporte.
1.2.2. Crea una función que devuelva los itemsets frecuentes candidatos y su soporte para k>=2
1.2.3. Crea una función que repita el paso 1.2.2. hasta que no se generen nuevos itemsets frecuentes
1.2.4. Crea una función que muestre todas las reglas posibles con su confianza
1.2.5. Lista todas las reglas que sean de alta confianza (p.ej. con valor >=0.5)
1.2.6. Crea funciones que permitan lo siguiente:
    * devolver todas las reglas que contengan un antecesor dado
    * devolver todas las reglas que tengan una confianza mayor (o igual) que un umbral mínimo dado
    * devolver todas las reglas que tengan un lift en una de las tres franjas de estudio (<1, >1, =1)
1.2.7. Utiliza al menos dos representaciones gráficas para mostrar las reglas obtenidas e interpretar los datos

1.3 (3 puntos) Prueba a aplicar sobre el conjunto de datos radio.csv al menos 3 configuraciones diferentes de soporte y confianza. ¿Qué diferencias observas al variar las configuraciones de soporte y de confianza? ¿Qué tipo de reglas desaparecen? ¿Por qué?
¿Qué diferencias hay entre las reglas de alta confianza y aquellas con valores de lift>1? ¿coinciden? ¿por qué?
Estudia e interpreta la regla que hayas obtenido con la confianza más alta mediante la aplicación de una tabla de contingencia.

Entregable 2 – Patrones Secuenciales

Este entregable vale 2 puntos de la nota final de la práctica 2.

El conjunto de datos spotify_top10.csv contiene el ranking diario de las 10 canciones más reproducidas en Spotify cada día por país. De esta manera, para cada día tenemos recogida una secuencia de canciones (donde el orden sí importa) por cada país que se tiene en cuenta. Analizando esta información se pueden extraer conclusiones sobre cómo evoluciona la popularidad de las canciones. Este estudio se puede llevar a cabo aplicando el algoritmo Generalized Sequential Patterns utilizando la implementación de éste disponible en la librería *gsppy*.

Prueba al menos dos configuraciones de soporte diferentes.

Para una de ellas, interpreta algunos de los patrones secuenciales que te resulten curiosos.

Fíjate que tienes que preprocesar los datos de manera que agrupes por día las canciones creando una lista de canciones por día y por país que serán nuestras transacciones a procesar con el GSP.

Recursos:

Librería mxlextend:

Podéis encontrar su documentación (específicamente relacionaca con el algoritmo A priori) en el siguiente enlace: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/

Librería gsppy:

Podéis encontrar su documentación en el siguiente enlace: https://pypi.org/project/gsppy/