Master Universitario en Sistemas Inteligentes (MUSI)
Assignatura: 11758 - Técnicas Avanzadas en Minería de Datos
Elemento de Evaluación: Práctica Final
Reglas de Asociación y Detección de Anomalías

Este elemento de evaluación consta de la resolución de dos problemas, se realizará individualmente de forma autónoma y no presencial. Todos los alumnos han de subir su resolución a Campus Extens.

Fecha límite de entrega: 31 de mayo de 2020 a las 23.55 h

Qué hay que entregar: Se tiene que entregar una memoria de la práctica (en formato html y Rmd). Así como todos aquellos ficheros que sean necesarios para obtener los resultados que mostréis en el informe de la práctica.

Dónde: en el espacio habilitado en Aula Digital.

Resolver los siguientes problemas usando R. Aseguraros de comentar detalladamente el script para facilitar su implementación.

1.- (5.0 PUNTOS) Consider the Extended BAKERY dataset

Brief outline. A bakery chain has a menu of about 40 pastry items and 10 coffee drinks. It has a number of locations in West Coast states (California, Oregon, Arizona, Nevada). The database stores information about the food/drinks offered for sale, locations, employees at each location and individual sales (receipts) at those locations. The document bakerygoods.txt offers a description of the goods. There are four canonical sets of data available for this dataset: a) 1,000 Receipts; b) 5,000 Receipts; c) 20,000 Receipts; d) 75,000 Receipts.

1. The `XXXX-out1.csv` file format is: receipt number followed by item numbers that are on that receipt (sparse vector representation).
2. The `XXXX-out2.csv` file format is: receipt number followed by 0's and 1's indicating if an item was on a given receipt (full binary vector representation).
3. The `XXXXi.csv` file format is: receipt number followed by item number and quantity (CSV version of the Items tables).

Using the data sets provided, find frequent itemsets, generate association rules and interpret the results. Present a visualization of the rules. Find redundant rules and interpret your findings. Study negative association rules and explain them.

2.- (5.0 PUNTOS) Consider the Breast Cancer Wisconsin (Diagnostic) Data Set

Brief outline. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].
Perform anomaly detection analysis using the "Breast Cancer Wisconsin (Diagnostic) Data Set" from UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets.php).
Eliminate the class label and detect anomalies using a) DBSCAN, b) Expectation Maximization, c) Local Outler Factor (LOF).
If the following convention is used for the class labels: M = malignant is an outlier, B = benign is normal, what can you say about the results you obtained? Which method is "best"?