TAREA I de COMP6315: Data Mining

Fecha de entrega: Jueves 6 de septiembre , 2018

Puntaje: 50 puntos

Datasets:
Air Presure System Failures in Scania Trucks (aps_failure_training_set.csv)
ISTANBUL+STOCK+EXCHANGE

Los datos están disponibles en la UCI Machine learning repository
(https://archive.ics.uci.edu/ml/index.php) y en kaggle.com

Usar Python, R o RapidMiner para responder a las siguientes preguntas (SI usa R bajar la librería data.table, que tiene una función fread que es 10 veces mas rápido que read.table):

1.(25) El conjunto de datos Air Presure System (APS) Failures in Scania Trucks contiene información acerca del sistema de Presion de Aire de los trucks Scania. Hay dos clases, que se debe a fallas relacionadas del Sistema APS la positiva y la negativa que se debe a falla por otras razones. Hay información faltante que aparece con "na". La clase es la primera columna de la tabla.
a) (5) Hacer un reporte de la información faltante. Incluyendo visualization
b) (5) Eliminar todas las columnas que tienen mas del 25% de "na" y mas del 50% de ceros
c) (7) Calcular el percentil del 99% l (Top 99%) de cada columna que aun quedan y eliminar las filas (instancias) de la tabla que tienen valores que exceden por lo menos a uno de estos percentiles.
d) (8) Aplicar imputación usando la media o mediana y el metodo knn de imputación con k=3 vecinos para sustituir los datos faltantes.

2. (25 pts)
a) (3) Normalizar la data Estambul para tener los datos en el intervalo [0,1].
b) (4) Discretizar todas las columnas de la data usando dos métodos de discretizacion
c) (5) Hacer un boxplot y un histograma(10 intervalos) de los datos de la columna EU
d) (8) Insertar al azar al conjunto de datos Estambul n 5% y un 10% of missing values. No inserte missing values en la columna "date" que corresponde a fecha en que se tomaron los datos. Luego, imputar los missing values usando inputacion por la media, mediana y knn usando 5 vecinos mas cercanos.

e)(5) Calcular el valor

MSE=suma(valor verdadero-valor imputado)^2/numero de records de la tabla

y basado en este valor comparar los metodos de imputacion usados en d).