CURSO: CMP 0575 - TÓPICOS 2 (DATA MINING)
COLEGIO: POLITÉCNICO
Semestre: Primer Semestre 2018/2019 NRC: 1068

Tarea 5: Ejercicio usando el procesamiento de los datos y la clasificación

Problema:

1. Dado el subconjunto de variables obtenidas como resultado de la tarea de selección de características (proyecto 4). Se desea:
    - Conformar un *dataset* de entrenamiento (*train*), validación (*validation*) y prueba, a partir del *dataset* original *Madelon.*
    - Aplicar la tarea de normalización de los datos para los conjuntos reducidos de entrenamiento (*train*), validación (*validation*) y prueba (*test*).
    - Aplicar la tarea de entrenamiento, validación y prueba (clasificación) usando el algoritmo ***k-nearest neighbors*** (*k-nn*).
    - El algoritmo debe ser empleado con dos medidas de distancia diferentes (*Euclidiana, Mahattan*). Por tanto, existirán dos resultados de clasificación para el conjunto de prueba.
    - Es obligatorio mostrar la trazabilidad del método durante la ejecución del programa:
        i. *datasets* empleados (*train, validation* y *test*) normalizados,
        ii. resultados de clasificación obtenidos por las diferentes medidas de distancias y con qué valor de ***k*** empleado. El resultado debe superar **95%** de rendimiento para considerarse como una buena clasificación.
        iii. Selección óptima del valor de ***k*** basado en un gráfico que muestre el ***accuracy*** obtenido (eje Y) por el clasificador a medida que varía el valor de k (eje X). Optimizar en el intervalo k=1..30.
    - Cargar al D2L los códigos implementados (fichero compactado que incluye el ejecutable ej: el .JAR de java) dentro del plazo de entrega.

Nota Importante: Esta tarea depende de la realización del proyecto 4. La no obtención de un conjunto reducido de variables conlleva a la aplicación del algoritmo ***k-nn*** sobre el *dataset* completo *Madelon*, lo cual es totalmente ineficiente. Dicha ineficiencia equivale a una penalización del 40% del valor de la tarea (4 puntos).