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

Tarea 6: Ejercicio usando la clasificación con redes neuronales artificiales

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 un modelo de ANN seleccionado (sin restricción de elección).
    - El modelo de ANN empleado debe obtener un resultado de clasificación de área sobre la curva ROC promedio *avg AUC >= 0.95* (umbral heurístico de calidad para problemas de clasificación).
    - Es obligatorio mostrar la trazabilidad del método durante la ejecución del programa:
        i. *datasets* empleados (*train, validation* y *test*) normalizados,
        ii. área sobre la curva ROC (*AUC*) obtenidos durante el proceso de entrenamiento, validación y prueba.
        iii. Hacer un plot que muestre el desempeño *AUC* (eje Y) obtenido por la ANN durante el proceso de entrenamiento (iteraciones eje X). Las iteraciones se pueden optimizar en el rango n= 10..1000 (con pasos de 10 en el gráfico).
        iv. Mostrar las características del modelo ANN seleccionado (mejor topología) por el estudiante.
    - Es obligatorio realizar una presentación científica (NO FILOSÓFICA) al profesor con todos los elementos requeridos.
    - La programación del modelo ANN utilizado debe ser en el lenguaje JAVA o C++
    - 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).