UNIDAD TEMÁTICA 4: Algoritmos No Lineales

Trabajo de Aplicación 3 Arboles de Decisión

Ejercicio1. Comparación de rendimientos al variar los parámetros

(50 min desarrollo – 18 min presentación gráficos)

Analizar el rendimiento (exactitud) del AD sobre un conjunto de entrenamiento y un conjunto de test, separados, variando los parámetros de profundidad máxima, cantidad de elementos mínimo en página y para división, mínima ganancia para división, poda / no poda

1. Utilizar el conjunto de datos **"eReader_training"**
2. Separar con un operador "Split" en forma apropiada dos subconjuntos, uno de "training" y el otro de "prueba". Cuidar que la partición se produzca siempre de la misma manera, y que sea aleatoria y estratificada si corresponde (analizar las estadísticas de los datos de entrada para esto)
3. Armar modelo con AD, comenzando por establecer la línea de base:
    a. Criterio: ganancia de información
    b. No preprunning
    c. No prunning
    d. Profundidad : -1 (todo lo que el algoritmo arme).
4. Entrenarlo con el dataset de training, y luego evaluar la performance de predicción sobre este mismo dataset
5. Evaluar la performance del modelo sobre el dataset de test.
6. Registrar los valores de ambos vectores de performance.
7. Repetir estos pasos para cada configuración diferente del AD:
    a. Al menos 4 valores diferentes de cantidad mínima de elementos para división
    b. Al menos 4 valores diferentes de cantidad mínima de elementos en las hojas
8. Realizar una planilla electrónica para recolectar todos estos datos
9. Producir gráficos a partir de los datos obtenidos, que brinden una visión general del comportamiento del modelo al variar los parámetros.

Ampliación:

Realizar el mismo ejercicio para los datasets **"cardiac_training**" y "**titanic".** Analizar los resultados y explicar las diferencias halladas.

PARÁMETROS DEL OPERADOR DTREE DE RM (a analizar detalladamente)

1. CRITERIO (Information_gain, Gini, etc..)
2. Apply Pruning (si/ no)
3. Apply Preprunning
    a. Ganancia mínima
    b. Tamaño de hoja mínimo
    c. Tamaño mínimo para división