Bases de Datos Masivas (11088)
Departamento de Ciencias Básicas

TRABAJO PRÁCTICO V: Minería de datos

PARTE 01: Arboles de decisión (J48)

Introducción:

En este trabajo se implementa el primero de una serie de algoritmos que se presentarán durante la materia para realizar Minería de datos: los árboles de decisión J48.

En primer lugar, se utilizarán las nociones de entropía y ganancia de información introducidas en clase a efectos de generar un árbol de decisión a partir de un dataset.

Luego, se utilizará el lenguaje Python, puntualmente la librería Scikit-Learn, con el objetivo de resolver problemas de la disciplina, los cuales son una combinación ejercicios clásicos de minería de datos complementados con ejercicios propuestos por el equipo docente.

Consignas:

1. A partir del dataset presentado a continuación, y teniendo en cuenta las fórmulas de entropía y ganancia de información calcule y diagrame el árbol de decisión que le permita decidir si comer asado o no en función del clima:

PRONÓSTICO	TEMPERATURA	HUMEDAD	VIENTO	ASADO
Soleado	Calor	alta	leve	no
Soleado	Calor	alta	fuerte	no
Nublado	Calor	alta	leve	si
Lluvioso	templado	alta	leve	si
Lluvioso	Frío	normal	leve	si
Lluvioso	Frío	normal	fuerte	no
Nublado	Frío	normal	fuerte	si
Soleado	templado	alta	leve	no

Soleado	Frío	normal	leve	si
Lluvioso	templado	normal	leve	si
Soleado	templado	normal	fuerte	si
Nublado	templado	alta	fuerte	si
Nublado	Calor	normal	leve	si
Lluvioso	templado	alta	fuerte	no

2. Trabaje con el dataset de Scikit Learn "wine":
    a. Utilice el metadata que provee la librería, ¿Cuál es el tema que aborda el dataset?
    b. Genere el árbol de decisión que permita clasificar los diferentes tipos de vino utilizando un muestreo con proporciones de 80% para entrenamiento y 20% para testeo.
    c. Explore la solución dada y las posibles configuraciones para obtener un nuevo árbol que clasifique "mejor". Documente las conclusiones.
3. Ahora, analice el archivo zoo.csv:
    a. Genere el árbol de decisión que permita inferir el tipo de animal en función de sus características. Explique someramente que resultado se obtiene en términos del árbol y en términos de la eficiencia del mismo.
        - ¿Varía ese resultado si se elimina el atributo "animal"? ¿Por qué?
        - Cuantos niveles posee el árbol generado? ¿Qué atributos debemos modificar si deseamos realizar una poda del mismo? Modifique esos atributos para que el árbol generado conste de 4 niveles. ¿Afecta la eficiencia de la clasificación esta modificación?
4. Se provee la base de datos de los pasajeros del famoso barco que se hundiera en su viaje inaugural (archivo titanic-en.csv) con los siguientes atributos y valores posibles:
    - `Class {"1st","2nd","3rd","crew"}`

```
-  Age {"adult","child"}
-  Sex {"male","female"}
-  Survived {"yes","no"}
```

Genere el árbol de clasificación, explore la solución dada y las posibles alternativas para obtener un nuevo árbol que clasifique "mejor".

5. Un Banco de Portugal realizó una campaña de marketing en busca de clientes de plazos fijos basada en llamados telefónicos. Se provee el dataset real (bank-full.csv) con más 45000 instancias y el detalle (bank-names.txt) de los datos registrados de cada una de las personas contactadas por la entidad bancaria.
    a. Realice las tareas necesarias para poder procesar el dataset en Scikit-Learn.
    b. Luego, genere el árbol de decisión, y optimice los resultados, con el objetivo de explicar cuáles son las características más importantes que permiten identificar a una persona que accederá o no al plazo fijo. Documente los resultados.
6. Se requiere clasificar a los aspirantes 2014 que abandonaron la Universidad al primer cuatrimestre del 2015 en función de las características socio-económicas recolectadas en una encuesta realizada al momento de inscripción y volcada en el dataset encuesta_universitaria.csv que aún no está depurado.
    a. Divida el dataset en dos partes, con proporciones del 80% y 20% del total de las instancias del dataset.
    b. Utilice el primer conjunto de instancias para generar 10 árboles diferentes con distintas configuraciones y conserve los resultados.
    c. Luego, genere un nuevo árbol aplicando, a partir de los resultados, un sistema de votación en el que el valor de la variable objetivo del árbol resultante estará dado por el valor más frecuente adoptado por los árboles originales.
    d. Valide el árbol resultante con el 20% restante del dataset original.

e. Documente las acciones realizadas y el conocimiento más importante generado a partir del árbol de decisión resultante.

7. Guarde los archivos resultantes de las actividades prácticas en una carpeta denominada tp0301-<legajo> que a su vez tenga un directorio por cada uno de los puntos de este trabajo, comprima la carpeta y envíelo al equipo docente.

Medidas de evaluación para técnicas de clasificación:

En función de la clasificación realizada, complete las siguientes actividades:

a. **Accuracy.**
    1. Ahora, calcule el accuracy de ambos modelos.
    2. ¿Cómo se interpreta la métrica anterior?
    3. ¿Qué aporta el accuracy?

b. **Recall/Precision.**
    1. Calcule las métricas recall y precisión para ambos modelos.
    2. ¿Cuál es la diferencia entre ambas?
    3. ¿Qué aspectos aborda cada una?

c. **Matiz de confusión**: ¿En qué casos el modelo clasifica mal?

Referencias sugeridas:

Data Mining: Practical Machine Learning Tools and Techniques

http://www.cs.waikato.ac.nz/ml/weka/book.html

Machine Learning, Chapter 3. Tom M. Mitchell, McGraw Hill, 1997.

Introducción a la Minería de Datos. Hérnandez Orallo & otro. Prentice Hall. 2008.