PRACTICA 4 – ÁRBOLES DE CLASIFICACIÓN

Material de Lectura: Capítulo 11 del Libro Introducción a la Minería de Datos de Hernández Orallo

Ejercicios a entregar: 4a, 4c, 6 completo

1. **Concepto de entropía**
   Dado un conjunto de datos con 3 clases (A, B y C), la siguiente tabla presenta distintas distribuciones de ejemplos para cada clase. Ordene de menor a mayor las distribuciones de *a, b, c, d, e y f* en términos de su **Entropía**. Intente primero hacer este ejercicio sin realizar cálculos. Luego, verifíque el resultado numéricamente.

	P(A)	P(B)	P(C)
a	0.9	0.1	0
b	0.4	0.6	0
c	0.6	0	0.4
d	1	0	0
e	0.33	0.33	0.33
f	0	1	0

2. **Cálculo de entropía**
   En base al siguiente conjunto de datos del archivo **personajes.xlsx**:

   a) Calcule la proporción de ejemplos de cada clase.

P(A)	P(B)	P(C)

   b) En base a la distribución de clases del punto a), calcule la **Entropía** total E del conjunto de datos, es decir, el nivel de entropía base sin considerar ningún atributo:

Entropía(E)	

3. **Cálculo de entropía de un atributo**

   a) En base al conjunto de datos del ejercicio 2, calcule la **Entropía** de los distintos valores del atributo *Voluntad*

b) Calcule la entropía de los atributos *Voluntad, Inteligencia* y *Alineamiento*.

c) Calcule la **Ganancia de Información** de estos atributos. Si tuviera que elegir uno de los atributos para crear una rama del árbol en base a la **Ganancia de Información**, ¿cuál preferiría?

Atributo	Entropía(E,A)	Ganancia(E,A)
Voluntad		
Inteligencia		
Alineamiento		

d) Calcule la **Entropía** y la **Ganancia de Información** del atributo **numérico** *Nivel*. Tenga en cuenta que deberá aplicar el algoritmo específico para atributos numéricos visto en la clase de árboles.

Atributo	Entropía(E,A)	Ganancia(E,A)
Nivel		

Considerando ahora los cuatro atributos, ¿cuál elegiría para crear una nueva rama del árbol?

e) Calcule la **InfoDivisión** y **Tasa de Ganancia (Gain Ratio)** de los atributos. Si tuviera que elegir uno de los atributos para crear una rama del árbol en base a la **Tasa de Ganancia**, ¿Cuál preferiría? ¿Por qué cambió el atributo elegido? ¿Qué problema tenía el atributo elegido por la **Ganancia de Información**?

Atributo	InfoDivision(E,A)	GainRatio(E,A)
Voluntad		
Inteligencia		
Alineamiento		
Nivel		

4. **Construcción de árboles**

*De este ejercicio, deben entregarse **solamente** los incisos a y c.*

a) Construya manualmente, a partir del archivo **trabajos_ej4_train.csv** y utilizando como criterio la Ganancia de Información, el árbol de clasificación capaz de predecir si una persona obtendrá o no el trabajo según los antecedentes que posea. Indique en cada paso los valores de Entropía obtenidos y las selecciones realizadas.

   Para cada selección de atributo para dividir un nodo, incluir una tabla con tantas columnas como atributos y la ganancia de información y entropía de cada uno.

   Puede verificar los resultados obtenidos manualmente al consultar los valores devueltos por el operador **Weight by Information Gain** de RapidMiner o los scripts de Python provistos en la teoría.

b) Dibuje y explique el árbol obtenido. ¿Podría darle algún consejo a quienes quieran obtener el trabajo?

c) Calcule el accuracy del árbol en el conjunto de entrenamiento.

d) Utilice el operador **W-J48** de RapidMiner para verificar el árbol de clasificación creado en el inciso (a). Recuerde tener instalada la extensión Weka (Extensions→Marketplace→Weka Extension). Cambie el parámetro **M** del operador **W-J48** al valor **1**, de modo de quitar el requisito mínimo de tamaño de hoja, y active la opción U para obtener un árbol sin podar (**Unpruned tree**) para obtener los mismos resultados que en el inciso a).

   1. Obtenga y analice el **accuracy** del árbol utilizando los operadores *"Apply Model"* y *"Performance"* sobre los datos de entrenamiento.
   2. Explique qué significan los valores en cada una de las celdas de la matriz de confusión obtenida. ¿Cuál de ellos corresponde con la precisión del árbol para predecir "*Obtiene_trabajo=NO*"?
   3. En base al árbol obtenido, ¿cuál es el criterio utilizado para darle el trabajo a una persona?

e) Utilizando el árbol obtenido en el inciso (a), prediga los valores de la clase para los datos de **trabajos_ej4_test.csv** en forma manual. Luego, calcule el accuracy del modelo.

f) Verifique el resultado obtenido en el inciso anterior utilizando Rapidminer.

5. **Efecto de la discretización en la generación del arbol**

Aplique el operador ID3 de RapidMiner a los datos de entrenamiento del archivo **trabajos_ej5_train.csv** realizando dos tipos distintos de discretización de los atributos numéricos:

a) ***Discretize by Binning*** utilizando tres intervalos.

b) ***Discretize by User Specification*** utilizando los siguientes intervalos:
   i. "*Referencias ofrecidas* (0 a 3 -> Baja, 4 a 6 -> Media y 7 a 10 -> Alta)
   ii. *Cantidad de Trabajos Anteriores* (0 a 4 -> Pocos, 5 a 7 -> Suficientes y 8 a 10 -> Muchos)

Para cada tipo de discretización ¿Se obtuvo el mismo resultado? ¿Por qué razón? Compare los árboles obtenidos.

6. **Efecto del tamaño del árbol en la tasa de aciertos de entrenamiento y prueba**

   a) Genere un árbol de decisión ID3 para el archivo **diabetes_nominal_train**. Utilice la medida de desorden "*information gain*", "minimal gain" igual a 0.01, con "*minimal size for split*" igual a 40 y "*minimal leaf size*" igual a 20. Observe la tasa de aciertos obtenida sobre el conjunto de entrenamiento y el tamaño del árbol resultante (la cantidad de nodos). Luego aplique el árbol obtenido sobre los datos del archivo **diabetes_nominal_test.xlsx** y mida tasa de aciertos.
   Nota: para calcular la cantidad de nodos, puede contar la cantidad de filas de la descripción textual del árbol. Recomendamos copiar la descripción del árbol a un editor de texto donde le indique la cantidad de filas.

   b) Repita el inciso (a) utilizando "*minimal size for split*" igual a 100 y "*minimal leaf size*" igual a 50. Observe nuevamente el tamaño del árbol resultante y la performance obtenida en los conjuntos de entrenamiento y testeo.
   c) Repita el inciso (a) pero ahora utilizando "*minimal size for split*" igual a 300 y "*minimal leaf size*" igual a 200. Observe el tamaño del árbol resultante y la performance obtenida en ambos conjuntos de datos.
   d) ¿Qué ocurre con el tamaño y complejidad del árbol en cada caso? ¿Se observan diferencias en la performance del árbol en ambos conjuntos de datos? ¿A qué se deben estas diferencias?

Para resolver el ejercicio puede completar la siguiente tabla con la respuesta a cada inciso:

Inciso		a)	b)	c)
minimal size for split		40	100	300
minimal leaf size		20	50	200
Tamaño del árbol				
Accuracy	*Train*			
	Test			

7. Clasificación de conjuntos de datos reales con Árboles de Decisión

Vuelva a realizar los ejercicios de **Clasificación de Jugadores de Fútbol** y **Predicción de batallas Pokemon**, pero ahora con Árboles de Decisión.

a) Busque los hiperparámetros del árbol que permitan clasificar mejor en las tareas planteadas. Indique el accuracy en el conjunto de entrenamiento y de validación para todos los casos.
b) Compare el desempeño con el de los modelos de Naive Bayes entrenados anteriormente. ¿Cuál método funciona mejor?
c) Evalúe qué tan interpretables son los árboles resultantes. ¿Puede hacerlos un poco más entendibles, achicando su tamaño?