Entrega Práctica 5, Grupo 9

miércoles, 19 de mayo de 2021 19:49

1. Soporte y Cobertura

El archivo **Globos.xlsx** contiene datos referidos a un experimento psicológico. De cada instancia se sabe el color del globo, el tamaño, si fue inflado por un adulto o un niño y si se estira o no. El objetivo es construir un modelo predictivo para determinar si un globo permanecerá inflado o no (atributo **Inflado?**).

a) Calcular el soporte (para reglas, equivalente a la tasa de acierto o accuracy) y la cobertura de cada una de las siguientes reglas, por separado, para **Globos.xlsx**:

R1: SeEstira? = Si y Edad=Adulto -> Inflado? = Si
R2: SeEstira? = Si y Edad=Niño -> Inflado? = No
R3: SeEstira? = No -> Inflado? = No

$$Sop(X \Rightarrow Y) = \frac{|X \cap Y|}{|D|}$$

$$Cob(X \Rightarrow Y) = \frac{|X|}{|D|}$$

Regla	Soporte	Cobertura
R1	4/16 = 0,25	4/16 = 0,25
R2	2/16 = 0,125	4/16 = 0,25
R3	8/16 = 0,5	8/16 = 0,5

b) Las tres reglas en conjunto ¿cubren todo el conjunto de entrenamiento? ¿Por qué es importante eso?

Si, cubren todo el conjunto.
Es importante cubrir todo el conjunto por varias razones. La primera y fundamental es que el algoritmo que crea las reglas termina cuando todo el conjunto está cubierto, por lo que si no se cubrió no se puede dar por terminado el algoritmo y no tendríamos un modelo como tal.
Fuera del algoritmo, si el modelo no está cubierto se podrían dar casos en donde la entrada no se pueda clasificar con las reglas existentes, y tendríamos que elegir otro tipo de algoritmo default para cubrir estos casos, por ejemplo el tipo ZeroR.

c) Calcular la tasa de acierto del conjunto de las 3 reglas para el conjunto de datos

La tasa es 14/16, solo se equivoca con la entrada de la fila 3 y 7, en donde se estira, la edad es niño y está inflado.

2. Algoritmos de generación de reglas de clasificación: ZeroR, OneR, Prism

a) A partir de los datos del archivo **Globos.xlsx** determine, utilizando los métodos **ZeroR y OneR,** el conjunto de reglas de clasificación que permita predecir si un globo dado permanecerá inflado o no (atributo ***Inflado?***).

	ZeroR	OneR	PRISM
Reglas	Como ponemos la regla acá? Si True => Inflado = No o Inflado = No		

La resolución de este ejercicio debe realizarse de forma manual. En ambos casos deberá detallar los cálculos realizados para determinar las reglas indicadas.

En el caso de OneR, realizar una tabla que incluya la tasa de acierto para cada regla posible de cada atributo

Además, hacer una tabla con el error total de cada atributo.

ZeroR
Deriva a cualquier atributo a Inflado = No, ya que es la clase mayoritaria.

OneR
Para obtener las reglas no hicimos ningún cálculo, simplemente miramos en la tabla la cantidad de elementos que cumplían con la regla. Lo mismo para el total, es sumar el numerador de cada error parcial y el denominador.

Atributo	Reglas	Error	Total
Color	Si (Color = Amarillo) entonces Inflado = No	4/8	6/16
	Si (Color = Rojo) entonces Inflado = No	2/8	
Tamaño	Si (Tamaño = Chico) entonces Inflado = No	3/6	6/16
	Si (Tamaño = Mediano) entonces Inflado = No	0/4	
	Si (Tamaño = Grande) entonces Inflado = No	3/6	
Se_estira?	Si (Se_estira? = Si) entonces Inflado = Si	2/8	2/16
	Si (Se_estira? = No) entonces Inflado = No	0/8	
Edad	Si (Edad = Niño) entonces Inflado = No	2/8	6/16
	Si (Edad = Adulto) entonces Inflado = No	4/8	

b) Repetir el punto a) utilizando PRISM.

La resolución de este ejercicio debe realizarse de forma manual. Detallar los cálculos realizados para determinar las reglas indicadas. Comenzar con la clase Inflado=Si. Para cada iteración de PRISM en una clase particular, armar una tabla indicando cada regla posible (incluyendo todos los atributos en la misma tabla) y el error asociado

PARTE 2

3. Métricas de evaluación de reglas

b) En base a los datos de la tabla **enfermedad.xslx**, calcule manualmente el soporte, interés, y confianza de las siguientes:

Regla	Soporte	Cobertura	Confianza	Interés
Enfermedad --> Antecedentes	2/10	5/10	(2/10)/(5/10) = 0,4	(2/10)/((5/10)*(2/10)) = 2
Enfermedad --> Adulto	4/10	5/10	(4/10)/(5/10) = 0,8	(4/10)/((5/10)*(7/10)) = 1,1429
Adulto --> Enfermedad	4/10	7/10	(4/10)/(7/10) = 0,5714	(4/10)/((7/10)*(5/10)) = 1,1429
Trabajo pesado --> Adulto	3/10	5/10	(3/10)/(5/10) = 0,6	(3/10)/((5/10)*(7/10)) = 0,8571

Responda:

1. ¿Por qué es interesante la regla **Enfermedad → Antecedentes**, si bien su soporte y confianza son bajos?

Porque si bien no hay muchas personas con antecedentes, todas las que tienen antecedentes tienen enfermedad, lo que indica que son atributos altamente dependientes. Sin embargo, como la regla es simétrica no nos indica la dirección correcta de la implicación, que en este caso sería Antecedentes --> Enfermedad, en vez de Enfermedad --> Antecedentes.

2. ¿Porqué varía la confianza entre **Adulto → Enfermedad** y **Enfermedad→Adulto**? ¿Varía el soporte en esos casos?

Porque la confianza no es simétrica, depende del soporte del antecedente en el divisor. El soporte no varía porque es un atributo simétrico.

3. ¿Que significa que **Trabajo Pesado → Adulto** tenga una cobertura y confianza relativamente altas, pero un interés bajo?

La cobertura y confianza son relativamente altas porque hay varios ejemplos en donde se realiza trabajo pesado y gran parte del mismo está hecho por un adulto, pero que la relación entre los atributos sigue siendo negativa, por lo que no tiene un gran interés