Parciales

miércoles, 16 de junio de 2021 15:37

UNIDADES	DESCUENTO	ENVIO
20	BAJO	NORMAL
29	MEDIO	NORMAL
45	BAJO	SUPERIOR
30	ALTO	SUPERIOR
30	MEDIO	NORMAL
2	BAJO	NORMAL
90	ALTO	SUPERIOR
120	ALTO	SUPERIOR
46	MEDIO	SUPERIOR

1) Indique el valor de verdad de las siguientes afirmaciones justificando su respuesta.
   a. El atributo UNIDADES presenta un único valor fuera de rango extremo.
   b. Si se discretiza en dos intervalos el atributo UNIDADES por rango y por frecuencia se obtiene el mismo resultado.

a.
Q1 = 20 + 0,5 * (29 - 20) = 24,5
Q3 = 46 + 0,5 * (90 - 46) = 68
RIC = 68 - 24,5 = 43,5

Q1 - 3*RIC
Q3 + 3*RIC

24,5 - 3*43,5 =
68 + 3*43,5 =

Rango extremos: (-inf ; -106) U (198,5 ; +inf)

Falso.

b.
118 / 2 = 59

(-inf; 61];(61; +inf]
(inf 30 30 inf

Falso.

2) Considerando los datos de la tabla, discretice el atributo UNIDADES en dos intervalos según si el valor es menor o igual que 30 (POCAS) o mayor a 30 (MUCHAS). Luego
   a) Indique cuál sería el nodo raíz del árbol generado con el método ID3 para predecir el tipo de ENVIO. Justifique su respuesta incluyendo los cálculos que haya realizado.

a)

Entropía(E) = - 4/9 * log2(4/9) - 5/9 * log2(5/9) = 0,991076059838222

Entropía(E, Unidades) = 5/9 * 0,721928094887362 + 4/9 * 0 = 0,401071163826312
Entropía(E_{POCAS}) = - 4/5 * log2(4/5) - 1/5 * log2(1/5) = 0,721928094887362
Entropía(E_{MUCHAS}) = 0
Ganancia(E, Unidades) = 0,991076059838222 - 0,401071163826312 = 0,59000489601191

Entropía(E, Descuento) = 0 + 3/9 * 0,918295834054489 + 3/9 * 0,918295834054489 = 0,612197222702993
Entropía(E_{ALTO}) = 0
Entropía(E_{MEDIO}) = - 1/3 * log2(1/3) - 2/3 * log2(2/3) = 0,918295834054489
Entropía(E_{BAJO}) = 0,918295834054489
Ganancia(E, Descuento) = 0,991076059838222 - 0,612197222702993 = 0,378878837135229

Ganancia (A,E) = Entropía(E) - Entropia(A,E)

b) Indique cuál es la primera regla para la clase (ENVIO=NORMAL) que se obtiene aplicando el método PRISM. Especifique tanto el antecedente como el consecuente de la regla.

Envío = Normal

Unidades	POCAS	4/5
	MUCHAS	0/4
Descuento	BAJO	2/3
	MEDIO	2/3
	ALTO	0/3

La que mayor precisión tiene es unidades, pero como no es perfecta se sigue ampliando:

Si Unidades = POCAS y ... entonces Envío = NORMAL

Descuento	BAJO	2/2
	MEDIO	2/2
	ALTO	0/1

La primera regla es Si Unidades = POCAS y Descuento = BAJO entonces Envío = NORMAL

c) Calcule la cobertura, el soporte, la confianza y el interés de la regla

Soporte:

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

Cobertura:

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

Confianza:

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

Interés:

$$Interes(A \rightarrow B) = \frac{Sop(A \rightarrow B)}{Sop(A) * Sop(B)}$$

La primera regla es Si Unidades = POCAS y Descuento = BAJO entonces Envio = NORMAL

Interés:

$$Interes(A \rightarrow B) = \frac{Sop(A \rightarrow B)}{Sop(A) * Sop(B)}$$

MUCHAS	90	ALTO	SUPERIOR
MUCHAS	120	ALTO	SUPERIOR
POCAS	30	ALTO	SUPERIOR
MUCHAS	45	BAJO	SUPERIOR
POCAS	2	BAJO	NORMAL
POCAS	20	BAJO	NORMAL
MUCHAS	46	MEDIO	SUPERIOR
POCAS	29	MEDIO	NORMAL
POCAS	30	MEDIO	NORMAL

c) Calcule la cobertura, el soporte, la confianza y el interés de la regla

SI (DESCUENTO=MEDIO) ENTONCES (UNIDADES=POCAS) AND (ENVIO=NORMAL)

Incluya los cálculos intermedios que haya realizado

Soporte:
2/9 = 0,2222

Cobertura:
3/9 = 0,3333

Confianza:
0,2222/0,3333 = 0,6667

Interés:
(2/9) / (3/9 * 4/9) = 1,5

d) Indique si la siguiente afirmación es verdadera o falsa. Justifique su respuesta.

"El conjunto {DESCUENTO=MEDIO, ENVIO=NORMAL} es un conjunto de ítems frecuentes si se pide un soporte mínimo de 0.2".

Soporte minimo es del 20% o de 0,2
En elementos, la cantidad mínima es 0,2 * 9 = 1,8, redondea a 2.

La conjunto tiene 2.

3) Se ha numerizado el atributo DESCUENTO de la siguiente forma: BAJO→0, MEDIO →25, ALTO→ 50

a) Luego de la numerización se calculó el coeficiente de correlación lineal entre los atributos UNIDADES y DESCUENTO y se obtuvo como resultado 0.679. ¿Cómo debe interpretarse este valor?

a) Tienen una correlación lineal debil

b) Con los ejemplos de la tabla numerizados se entrenó un perceptrón para predecir el atributo ENVIO. El modelo obtenido fue el siguiente:

Intercept: -3.065 w(UNIDADES) = 0.093 w(DESCUENTO) = 0.007

¿Cómo clasifica el perceptrón entrenado al ENVIO de un pedido de 70 UNIDADES con un DESCUENTO = BAJO?

Se clasifica como superior

c) Luego de numerizar el atributo DESCUENTO como se explicó previamente, se aplicó el algoritmo k-medias para agrupar los ejemplos de la tabla en 2 grupos con el objetivo de identificar similitudes según los valores de los atributos UNIDADES y DESCUENTO. El atributo ENVIO no fue tenido en cuenta. El modelo obtenido fue el siguiente:

Atributo	Cluster_0	Cluster_1
UNIDADES	28.857	105.0
DESCUENTO	17.857	50.0

¿A cuál de los dos grupos pertenecería un pedido de 80 UNIDADES al que se le ha realizado un 50% de descuento (DESCUENTO=50)? Justifique su respuesta.

c1 = (28 ; 17)
c2 = (105 ; 50)

x(80;50)

Distancia a c1 = (28 - 80)^2 + (17 - 50)^2 = 3793
Distancia a c2 = (105 - 80)^2 + (50 - 50)^2 = 625

d) El índice Silhouette correspondiente al agrupamiento para k=2 es 0.625 mientras que el mismo índice para k=3 es 0.435 ¿Cuál de los agrupamientos es mejor? Explique.

El mejor agrupamiento es k=2, según el índice Silhouette, ya que tiene menor cohesión y/o mayor separación que el agrupamiento con k=3

4) Indique el valor de verdad de las siguientes afirmaciones justificando su respuesta

a) Sólo puede utilizarse una matriz de confusión para analizar la performance de los modelos que emplean aprendizaje supervisado.

b) Un atributo numérico puede aparecer más de una vez como nodo de una misma rama de un árbol.

c) Es posible que una regla tenga un valor de soporte superior a su valor de cobertura.

d) El cálculo de la Ganancia de Información requiere una cantidad de cálculos mayor cuando se trabaja sobre atributos numéricos que sobre atributos nominales

a) Verdadero (con sus versiones distintas, como el clasificador binario o como se llame)
b) Verdadero
c) Falso, porque el valor de cobertura solo tiene en cuenta el antecedente de la regla, lo que hace que sea igual o mayor al soporte, que tiene en cuenta tanto el antecedente como el consecuente de la regla.
d) Si, en términois generales. Porque para calcular la ganancia de información es necesario tener la entropía. Para calcular la entropía de un nominal es ez, mientras que para un numérico tener que ordenar todos sus valores y calcular la entropía para cada separación.

Autoevaluación 9

Q1 = 2 + 0,25 * (3 - 2) = 2,25
Q3 = 6 + 0,75 * (7 - 6) = 6,75
RIC = 6,75 - 2,25 = 4,5

Bigote inferior
Q1 - 1,5 * RIC
2,25 - 1,5 * 4,5 = -4,5

Bigote inferior
Q3 + 1,5 * RIC
6,75 + 1,5 * 4,5 = 13,5

Intervalos
Regla = A --> B
(Asistencia < 70%) Y (AutoEval < 7) --> (Condicion = NO)
soporte
2/8 = 0,25

cobertura
3/8 = 0,375

confianza
(2/8) / (3/8) = 0,6667

interes
(2/8) / ((3/8) * (2/8)) = 2,6667
