Practica 4

lunes, 10 de mayo de 2021 10:46

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, verifique 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

Sin realizar los cálculos:

d	
f	
a	
b	
c	
e	

Haciendo los calculos:

d	0
f	0
a	0.4690
b	0.9710
c	0.9710
e	1.5835

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)
4/8	2/8	2/8

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)	1.5

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*

	Entropía
Fuerte	0
Normal	1
Débil	1

Fuerte:
P(A) = 1
P(B) = 0
P(C) = 0

Normal:
P(A) = 0
P(B) = 0.5
P(C) = 0.5

Débil:
P(A) = 0
P(B) = 0.5
P(C) = 0.5

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

	Entropía
Voluntad	0.5
Inteligencia	0.6926
Alineamiento	0.25

Voluntad:

	Entropía
Fuerte	0
Normal	1
Débil	1

	PA
Fuerte	4/8
Normal	2/8
Débil	2/8

Inteligencia:

	Entropía
Alta	0.9235
Baja	0

	PA
Alta	6/8
Baja	2/8

Alta:
P(A) = 4/6
P(B) = 2/6
P(C) = 0

Baja:
P(A) = 0
P(B) = 0
P(C) = 1

Alineamiento:

	Entropía
NN	0
NB	0
CN	1
CM	0
LB	0
LM	0

	PA
NN	1/8
NB	1/8
CN	2/8
CM	1/8
LB	2/8
LM	1/8

NN:
P(A) = 1
P(B) = 0
P(C) = 0

NB:
P(A) = 0
P(B) = 0
P(C) = 1

CN:
P(A) = 0.5

P(B) = 0
P(C) = 0.5

CM:
P(A) = 1
P(B) = 0
P(C) = 0

LB:
P(A) = 0
P(B) = 1
P(C) = 0

LM:
P(A) = 1
P(B) = 0
P(C) = 0

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	Ganancia
Voluntad	0.5	1
Inteligencia	0.6926	0.8074
Alineamiento	0.25	1.25

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	1.1876	0.3124

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

Corte	Valor desorden
1.5	1.2051
2.5	1.1937
3.5	1.2975
4.5	1.1876
5.5	1.362

El mejor atributo para una nueva rama del árbol es el alineamiento.

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	1.5	$0.\widehat{6}$
Inteligencia	0.8113	0.9952
Alineamiento	2.5	0.5
Nivel	0.8113	0.3851

Elegiría la inteligencia.
El atributo elegido cambió porque ahora compensamos el hecho de que un atributo pueda tener muchos valores
El atributo anterior dividía con pocos atributos en cada rama

4. **Construcción de árboles**

a) Construya manualmente, a partir de los datos la hoja **Train** del archivo **trabajos_ej4.xlsx** 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.

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.

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

RAIZ

Proporción de ejemplos de cada clase:

P(SI)	P(NO)
9/16	7/16

Entropía total E del conjunto:

Entropía(E)	0,988699

Para el cálculo de la entropía usamos la siguiente fórmula:

$$Entropia(E) = \sum_{i=1}^{n} - p_i \, log_2(p_i)$$

- (9/16) * log2(9/16) - (7/16) * log2(7/16) = 0,988699

Entropía y ganancia de cada atributo:
Para calcular la ganancia usamos la fórmula
Ganancia(E,A) = Entropia(E) – Entropia(E,A)

Atributo	Entropía(E,A)	Ganancia(E,A)
Titulo universitario	0,809028	0,179671

Experiencia en el cargo	0,772783	0,215916
Cantidad de Trabajos Anteriores	0,935096	0,053603
Trabaja actualmente	0,806781	0,181918

Ganancia(E,Título universitario) = 0,988699 - 0,809028 = 0,179671
Ganancia(E,Experiencia en el cargo) = 0,988699 - 0,772783 = 0,215916
Ganancia(E,Cantidad de Trabajos Anteriores) = 0,988699 - 0,935096 = 0,053603
Ganancia(E,Trabaja actualmente) = 0,988699 - 0,806781 = 0,181918

El atributo de mayor ganancia es "Experiencia en el cargo", por lo cual va a ser la raíz.

Para calcular la entropía de cada atributo usamos la siguiente fórmula:

$$Entropia(E,A) = \sum_{v \in V_a} \frac{|E_v|}{E} Entropia(E_v)$$

Título universitario
Probabilidades:
Aclaración: P(SI) es P(Obtiene trabajo = SI), lo mismo para P(NO)

	P(SI)	P(NO)
SI	3/3	0/3
NO	6/13	7/13

Entropía(E_{SI}) = - (3/3) * log2(3/3) = 0
Entropía(E_{NO}) = - (6/13) * log2(6/13) - (7/13) * log2(7/13) = 0,995727

Entropía(E_{SI})	0
Entropía(E_{NO})	0,995727

	Proporción
SI	3/16
NO	13/16

Entropía(E,Título universitario) = 3/16 * 0 + 13/16 * 0,995727 = 0,809028

Entropía(E,Título universitario)	0,809028

Experiencia en el cargo
Probabilidades:

	P(SI)	P(NO)
BAJA	3/5	2/5
MEDIA	5/6	1/6
ALTA	1/5	4/5

Entropía(E_{BAJA}) = - (3/5) * log2(3/5) - (2/5) * log2(2/5) = 0,970951
Entropía(E_{MEDIA}) = - (5/6) * log2(5/6) - (1/6) * log2(1/6) = 0,650022
Entropía(E_{ALTA}) = - (1/5) * log2(1/5) - (4/5) * log2(4/5) = 0,721928

Entropía(E_{BAJA})	0,970951

Entropía(E_{MEDIA})	0,650022
Entropía(E_{ALTA})	0,721928

	Proporción
BAJA	5/16
MEDIA	6/16
ALTA	5/16

Entropía(E,Experiencia en el cargo) = 5/16 * 0,970951 + 6/16 * 0,650022 + 5/16 * 0,721928 = 0,772783

Entropía(E,Experiencia en el cargo)	0,772783

Cantidad de Trabajos Anteriores

Corte	Valor entropía
7	0,935096
8.5	0,953186

Como el valor de entropía del corte en 7 es menor, vamos a utilizar ese:

Entropía(E,Cantidad de Trabajos Anteriores)	0,935096

Corte = 7

	P(SI)	P(NO)
<=7	4/9	5/9
>7	5/7	2/7

Entropía($E_{<=7}$) = - (4/9) * log2(4/9) - (5/9) * log2(5/9) = 0,991076
Entropía($E_{>7}$) = - (5/7) * log2(5/7) - (2/7) * log2(2/7) = 0,863121

Entropía($E_{<=7}$)	0,991076
Entropía($E_{>7}$)	0,863121

	Proporción
<=7	9/16
>7	7/16

Entropía(E,Cantidad de Trabajos Anteriores con corte en 7) = 9/16 * 0,991076 + 7/16 * 0,863121 = 0,935096

Entropía(E,Cantidad de Trabajos Anteriores con corte en 7)	0,935096

Corte = 8.5

	P(SI)	P(NO)
<=8.5	8/13	5/13
>8.5	1/3	2/3

Entropía($E_{<=8.5}$) = - (8/13) * log2(8/13) - (5/13) * log2(5/13) = 0,961237
Entropía($E_{>8.5}$) = - (1/3) * log2(1/3) - (2/3) * log2(2/3) = 0,918296

Entropía($E_{<=8.5}$)	0,961237
Entropía($E_{>8.5}$)	0,918296

	Proporción
<=8.5	13/16
>8.5	3/16

Entropía(E,Cantidad de Trabajos Anteriores con corte en 8.5) = 13/16 * 0,961237 + 3/16 * 0,918296 = 0,953186

Entropía(E,Cantidad de Trabajos Anteriores con corte en 8.5)	0,953186

Trabaja actualmente

Probabilidades:
Aclaración: P(SI) es P(Obtiene trabajo = SI), lo mismo para P(NO)

	P(SI)	P(NO)
SI	1/5	4/5
NO	8/11	3/11

Entropía(E_{SI}) = - (1/5) * log2(1/5) - (4/5) * log2(4/5) = 0,721928
Entropía(E_{NO}) = - (8/11) * log2(8/11) - (3/11) * log2(3/11) = 0,845350

Entropía(E_{SI})	0,721928
Entropía(E_{NO})	0,845350

	Proporción
SI	5/16
NO	11/16

Entropía(E,Trabaja actualmente) = 5/16 * 0,721928 + 11/16 * 0,845350 = 0,806781

Entropía(E,Trabaja actualmente)	0,806781

RAMA ALTA

Proporción de ejemplos de cada clase:

P(SI)	P(NO)
1/5	4/5

Entropía total E del conjunto:
- (1/5) * log2(1/5) - (4/5) * log2(4/5) = 0,721928

Entropía(E)	0,721928

Atributo	Entropía(E,A)	Ganancia(E,A)

Título universitario	0	0,721928
Cantidad de Trabajos Anteriores	0,649022	0,072906
Trabaja actualmente	0,550978	0,170950

Ganancia(E,Título universitario) = 0,721928 - 0 = 0,721928
Ganancia(E,Cantidad de Trabajos Anteriores) = 0,721928 - 0,649022 = 0,072906
Ganancia(E,Trabaja actualmente) = 0,721928 - 0,550978 = 0,170950

El atributo de mayor ganancia es "Título universitario", por lo que va a subdividir la rama ALTA.

Título universitario

Probabilidades:
Aclaración: P(SI) es P(Obtiene trabajo = SI), lo mismo para P(NO)

	P(SI)	P(NO)
SI	1/1	0/1
NO	0/4	4/4

Entropía(E_{SI}) = - (1/1) * log2(1/1) = 0
Entropía(E_{NO}) = - (4/4) * log2(4/4) = 0

Entropía(E_{SI})	0
Entropía(E_{NO})	0

	Proporción
SI	1/5
NO	4/5

Entropía(E,Título universitario) = 1/5 * 0 + 4/5 * 0 = 0

Entropía(E,Título universitario)	0

Cantidad de Trabajos Anteriores

El valor de corte es 7.5, ya que los valores numéricos son 6 y 9.

Entropía(E,Cantidad de Trabajos Anteriores)	0,649022

	P(SI)	P(NO)
<=7.5	1/4	3/4
>7.5	0/1	1/1

Entropía($E_{<=7.5}$) = - (1/4) * log2(1/4) - (3/4) * log2(3/4) = 0,811278
Entropía($E_{>7.5}$) = - (1/1) * log2(1/1) = 0

Entropía($E_{<=7.5}$)	0,811278
Entropía($E_{>7.5}$)	0

	Proporción
<=7.5	4/5
>7.5	1/5

Entropía(E,Cantidad de Trabajos Anteriores con corte en 7.5) = 4/5 * 0,811278 + 1/5 * 0 = 0,649022

Trabaja actualmente

Probabilidades:
Aclaración: P(SI) es P(Obtiene trabajo = SI), lo mismo para P(NO)

	P(SI)	P(NO)
SI	0/2	2/2
NO	1/3	2/3

Entropía(E_{SI}) = - (2/2) * log2(2/2) = 0
Entropía(E_{NO}) = - (1/3) * log2(1/3) - (2/3) * log2(2/3) = 0,918296

Entropía(E_{SI})	0
Entropía(E_{NO})	0,918296

	Proporción
SI	2/5
NO	3/5

Entropía(E,Trabaja actualmente) = 2/5 * 0 + 3/5 * 0,918296 = 0,550978

Entropía(E,Trabaja actualmente)	0,550978

RAMA MEDIA

Proporción de ejemplos de cada clase:

P(SI)	P(NO)
5/6	1/6

Entropía total E del conjunto:
- (5/6) * log2(5/6) - (1/6) * log2(1/6) = 0,650022

Entropía(E)	0,650022

Atributo	Entropía(E,A)	Ganancia(E,A)
Título universitario		
Cantidad de Trabajos Anteriores		
Trabaja actualmente		

Ganancia(E,Título universitario) = 0,721928 - 0 =
Ganancia(E,Cantidad de Trabajos Anteriores) = 0,721928 - 0,649022 =
Ganancia(E,Trabaja actualmente) = 0,721928 - 0,550978 =

El atributo de mayor ganancia es "", por lo que va a subdividir la rama MEDIA.

Título universitario

Probabilidades:
Aclaración: P(SI) es P(Obtiene trabajo = SI), lo mismo para P(NO)

	P(SI)	P(NO)
SI	0	0

NO	5/6	1/6

Este atributo se descarta porque no tiene ejemplos si el aspirante tiene título.

Cantidad de Trabajos Anteriores

Corte	Valor entropía
7	
8.5	

Como el valor de entropía del corte en X es menor, vamos a utilizar ese:

Entropía(E,Cantidad de Trabajos Anteriores)	

Corte = 7

	P(SI)	P(NO)
<=7		
>7		

Entropía($E_{<=7}$) = - (4/9) * log2(4/9) - (5/9) * log2(5/9) =
Entropía($E_{>7}$) = - (5/7) * log2(5/7) - (2/7) * log2(2/7) =

Entropía($E_{<=7}$)	
Entropía($E_{>7}$)	

	Proporción
<=7	
>7	

Entropía(E,Cantidad de Trabajos Anteriores con corte en 7) = 9/16 * 0,991076 + 7/16 * 0,863121 =

Entropía(E,Cantidad de Trabajos Anteriores con corte en 7)	

Corte = 8.5

	P(SI)	P(NO)
<=8.5		
>8.5		

Entropía($E_{<=8.5}$) = - (8/13) * log2(8/13) - (5/13) * log2(5/13) =
Entropía($E_{>8.5}$) = - (1/3) * log2(1/3) - (2/3) * log2(2/3) =

Entropía($E_{<=8.5}$)	
Entropía($E_{>8.5}$)	

	Proporción
<=8.5	
>8.5	

Entropía(E,Cantidad de Trabajos Anteriores con corte en 8.5) = 13/16 * 0,961237 + 3/16 * 0,918296 =

Entropía(E,Cantidad de Trabajos Anteriores con corte en 8.5)	

Trabaja actualmente

Probabilidades:

Aclaración: P(SI) es P(Obtiene trabajo = SI), lo mismo para P(NO)

	P(SI)	P(NO)
SI	0/2	2/2
NO	1/3	2/3

Entropía(E_{SI}) = - (2/2) * log2(2/2) = 0

Entropía(E_{NO}) = - (1/3) * log2(1/3) - (2/3) * log2(2/3) = 0,918296

Entropía(E_{SI})	0
Entropía(E_{NO})	0,918296

	Proporción
SI	2/5
NO	3/5

Entropía(E,Trabaja actualmente) = 2/5 * 0 + 3/5 * 0,918296 = 0,550978

Entropía(E,Trabaja actualmente)	0,550978

RAMA BAJA