COMPITO DI DATA MINING & ANALYTICS

19 giugno 2019 (Punteggio 17; Tempo 2h)

Esercizio 1 (punti 4)

Dato il seguente training set S:

Tipo	**Mobillità**	**Classe**
A	Si	Si
B	No	Si
B	No	No
C	No	No
C	No	No
A	Si	Si
B	No	No
C	Si	No
A	?	Si
A	No	No
B	Si	Si
C	?	Si
A	No	No
A	Si	Si
B	?	Si

a) Si calcoli l'entropia del training set rispetto all'attributo Classe

Entropia: $H(C)=-\sum_j P(c_j)\log_2 P(c_j)$

dove $P(c_j)$ è la probabilità della classe c_j.

b) Si calcoli il rapporto di guadagno dei due attributi rispetto a questi esempi di training

c) si costruisca un albero decisionale ad un solo livello per il training set dato, indicando le etichette delle foglie (numero di esempi finiti nella foglia/numero di esempi finiti nella foglia non appartenenti alla classe della foglia).

d) si classifichi l'istanza:

A	?

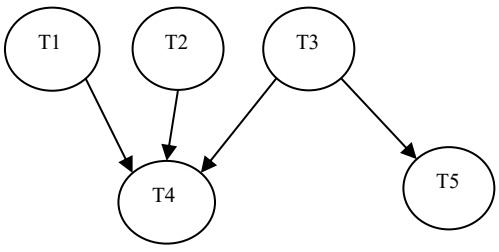
Esercizio 2 (punti 5)

Si trovino gli itemset con supporto maggiore o uguale al 50% dal database:

ID transazione	Items acquistati
1	1,2,3,5
2	1,4,5,6
3	1,4,5,6
4	2,3,4,5
5	2,3,4,5,6
6	1,2,3,4

Esercizio 3 (punti 4)

Sia data la seguente rete bayesiana



Dove tutte le variabili assumono i valori vero e falso.

Le tabelle di probabilità condizionata sono
per T1:

	T1=Falso	T1=Vero
	0.2	0.8

per T2:

	T2=Falso	T2=Vero
	0.5	0.5

per T3:

	T3=Falso	T3=Vero
	0.4	0.6

per T4:

T1	T2	T3	T4=falso	T4=vero
Falso	Falso	Falso	0.8	0.2
Falso	Falso	Vero	0.6	0.4
Falso	Vero	Falso	0.1	0.9
Falso	Vero	Vero	0.3	0.7
Vero	Falso	Falso	0.7	0.3
Vero	Falso	Vero	0.9	0.1
Vero	Vero	Falso	0.1	0.9
Vero	Vero	Vero	0.2	0.8

per T5:

T5	T3=falso	T3=vero
Falso	0.3	0.7
Vero	0.7	0.3

Si calcoli la probabilità P(~T2 |T1, ~T3,T4, T5).

Esercizio 4 (punti 4)

Dato il seguente LPAD

```
a : 0.3; b : 0.7:- c(X).

c(1).
c(2).

d:0.3.

c:- d,a.
c:- \+ d,b.
```

dove \+ indica la negazione (\+ d è vero se d è falso).
Si calcoli la probabilità di `c.`

SOLUZIONE

Esercizio 1

a) info(S)=-8/15*$\log_2$ 8/15-7/15*$\log_2$ 7/15= 0,997

b)

$info_{Tipo}(S)$= 6/15*(-4/6*$\log_2$ 4/6-2/6*$\log_2$ 2/6) + 5/15*(-3/5*$\log_2$ 3/5-2/5*$\log_2$ 2/5)+4/15*(-1/4*$\log_2$ 1/4-3/4*$\log_2$ 3/4) = 0,4 *0,918+0,333 *0,971 +0,266 *0,811=0,906

gain(Tipo) =0,997 -0,906=0,091

splitinfo(Tipo)= - 6/15*$\log_2$(6/15) - 5/15*$\log_2$(5/15) - 4/15*$\log_2$(4/15)= 1,566

gainratio(Tipo) =0,091/1,566=0,058

Per calcolare il guadagno dell'attributo Mobilità non si usa l'entropia calcolata su tutto il training set ma solo sugli esempi che hanno Mobilità noto (insieme F):

info(F)=-5/12*$\log_2$ 5/12– 7/12*$\log_2$ 7/12= 0,980

$info_{Mobilità}(S)$= 5/12*(-4/5*$\log_2$ 4/5-1/5*$\log_2$ 1/5) + 7/12*(-1/7*$\log_2$ 1/7-6/7*$\log_2$ 6/7) = 0,426*0,722+0,583*0,592=0,652

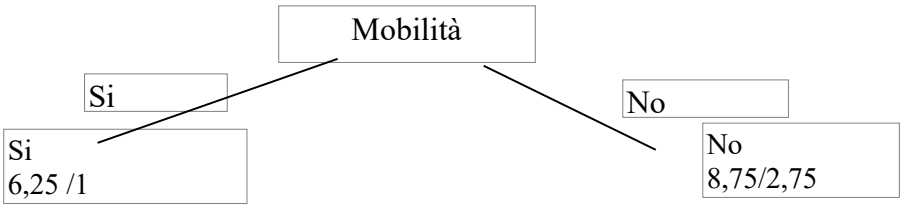
gain(Mobilità) =12/15*(0,980 -0,652)= 0,262

splitinfo(Mobilità)= - 5/15*$\log_2$(5/15) -7/15*$\log_2$(7/15) -3/15*$\log_2$(3/15)= 1,506

gainratio(Mobilità) =0,262/ 1,506=0,174

c) L'attributo scelto per la radice dell'albero è Mobilità

.



d) l'istanza viene divisa in due parti, di peso rispettivamente 6,25/15=0,416 e 8,75/15=0,583. La prima parte viene mandata lungo il ramo Si e classificata come Si con probabilità 5,25/6,25=84% e come No con probabilità 1/6,25=16%. La seconda parte viene mandata lungo il ramo No e classificata come No con probabilità 6 /8,75=68,6% e come Si con probabilità 2,75/8,75=31,4%. Quindi in totale la classificazione dell'istanza è

P(Si) =0,416*84%+0,583*31,4%=53,3%

P(No) = 0,416*16%+0,583*68,6%=46,7 %

Esercizio 3

conteggi

Itemset	Supporto
1	4
2	4
3	4
4	5
5	5
6	3

C2=

1,2
1,3
1,4
1,5
1,6
2,3
2,4
2,5
2,6
3,4
3,5
3,6
4,5
4,6
5,6

Conteggi

Itemset	Supporto
~~1,2~~	2
~~1,3~~	2
1,4	3
1,5	3
~~1,6~~	2
2,3	4
2,4	3
2,5	3
~~2,6~~	1
3,4	3
3,5	3
~~3,6~~	1
4,5	4
4,6	3
5,6	3

C3=

1,4,5
2,3,4
2,3,5
2,4,5
3,4,5
4,5,6

Conteggi

~~1,4,5~~	2
2,3,4	3
2,3,5	3
~~2,4,5~~	2
~~3,4,5~~	2
4,5,6	3

C4=

~~2,3,4,5~~

Esercizio 3

P(~T2 |T1, ~T3,T4, T5)= P(~T2 |T1, ~T3,T4)= P(T1, ~T2, ~T3,T4) / P(T1, ~T3,T4)

P(T1, ~T3,T4)= P(T1, ~T2, ~T3, T4)+ P(T1,T2, ~T3, T4)

P(T1,T2, ~T3, T4) = P(T1)P(T2)P(~T3)P(T4|T1,T2, ~T3)
=0,8*0,5*0,4*0,9=0,144

P(T1, ~T2, ~T3, T4) = P(T1)P(~T2)P(~T3)P(T4|T1, ~T2, ~T3)
=0,8*0,5*0,4*0,3=0,048

P(T1, ~T3, T4) = 0,144+0,048=0,192

P(~T2|T1, ~T3,T4,T5) = 0,048/0,192=0,25

oppure
P(~T2 |T1, ~T3,T4, T5)= P(T1, ~T2, ~T3,T4,T5) / P(T1, ~T3,T4,T5)

P(T1, ~T3,T4,T5)= P(T1, ~T2, ~T3, T4,T5)+ P(T1,T2, ~T3, T4,T5)

P(T1,T2, ~T3, T4,T5) = P(T1)P(T2)P(~T3)P(T4|T1,T2, ~T3)P(T5|~T3)
=0,8*0,5*0,4*0,9*0,7=0,1008

P(T1, ~T2, ~T3, T4,T5) = P(T1)P(~T2)P(~T3)P(T4|T1, ~T2, ~T3) P(T5|~T3)
=0,8*0,5*0,4*0,3*0,7=0,0336

P(T1, ~T3, T4,T5) = 0,1008+0,0336=0,1344

P(~T2|T1, ~T3,T4,T5) = 0,0336/0,1344 =0,25

Esercizio 3

Grounding

```
a : 0.3 ; b : 0.7:- c(1).
a : 0.3 ; b : 0.7:- c(2).

c(1).
c(2).

d:0.3.

c:- d,a.
c:- \+ d,b.
```

Mondi possibili

`a:-c(1).`
`a:-c(2).`
`d.`
	`a:-c(1).`
	`b:-c(2).`
	`d.`

P(w1)=0,027	P(w2)=0,063
b:-c(1). a:-c(2). d. P(w3)= 0,063	b:-c(1). b:-c(2). d. P(w4)=0,147
a:-c(1). a:-c(2). P(w5)=0,063	a:-c(1). b:-c(2). P(w6)=0,147
b:-c(1). a:-c(2). P(w7)=0,147	b:-c(1). b:-c(2). P(w8)=0,343

P(c)=P(w1)+P(w2)+P(w3)+P(w6)+P(w7)+P(w8) = 0.027+0.063+0.063+0.147+0.147+0.343 =0.79