COMPITO DI DATA MINING & ANALYTICS

16 giugno 2017 (Punteggio 17; Tempo 2h)

Esercizio 1 (punti 4)

Dato il seguente training set S:

Livello	**Punteggio**	**Classe**
Basso	0-30	Sì
Medio	31-70	No
Medio	31-70	No
Basso	0-30	Sì
?	31-70	No
Medio	0-30	No
Alto	31-70	No
Basso	0-30	Sì
Medio	0-30	No
Basso	31-70	No
Basso	0-30	Sì
?	0-30	No
Alto	31-70	Sì
?	0-30	No
Alto	31-70	Sì
Alto	31-70	Sì

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:

Alto	?

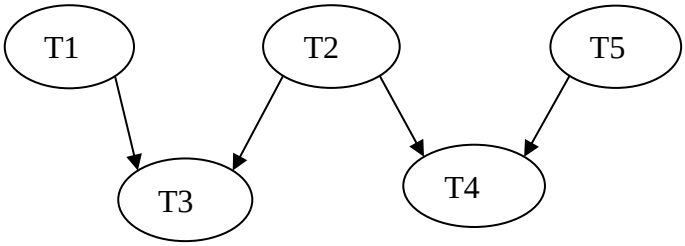
Esercizio 2 (punti 4)

Si trovino gli itemset con supporto maggiore o uguale al 33% dal database:

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

Esercizio 3 (punti 5)

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
	0.1	0.9

per T2:

	T2=Falso	T2=Vero
	0.4	0.6

per T3:

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

per T4:

T2	T5	T4=falso	T4=vero
Falso	Falso	0.5	0.5
Falso	Vero	0.1	0.9
Vero	Falso	0.4	0.6
Vero	Vero	0.3	0.7

per T5:

T5	T5=falso	T5=vero
	0.1	0.9

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

Esercizio 4 (punti 4)

Dato il seguente LPAD

```
epidemic : 0.6 ; pandemic : 0.3 :- flu(X), cold.
% if somebody has the flu and the climate is cold, there is the possibility
% that an epidemic arises with probability 0.6 and the possibility that a
% pandemic arises with probability 0.3

cold.
flu(david).
flu(robert).
```

Si calcoli la probabilità di `epidemic`.

SOLUZIONE

Esercizio 1

a) info(S)=-7/16*$\log_2$ 7/16-9/16*$\log_2$ 9/16= 0.989

b)

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

info(F)=-7/13*$\log_2$ 7/13– 6/13*$\log_2$ 6/13= 0.996

$info_{Livello}$(F)=5/13*(-1/5*$\log_2$1/5-4/5*$\log_2$ 4/5)+4/13(-4/4*$\log_2$ 4/4-0/4*$\log_2$ 0/4) +4/13(-1/4*$\log_2$ 1/4-3/4*$\log_2$ 3/4) =0.385*0.722 +0.308*0 +0.308*0.811=0.528

gain(Livello) =13/16*(0.996-0.528) =0.380

splitinfo(Livello) =-5/16*$\log_2$(5/16)-4/16*$\log_2$(4/16)-4/16*$\log_2$(4/16) -3/16*$\log_2$(3/16) =1.977

gainratio(Livello) =0.380/1.977=0.192

$info_{Punteggio}$(S)= 8/16*(-4/8*$\log_2$4/8-4/8*$\log_2$ 4/8) + 8/16*(-3/8*$\log_2$3/8-5/8*$\log_2$ 5/8) =
0.5*1+0.5*0.9544=0.977

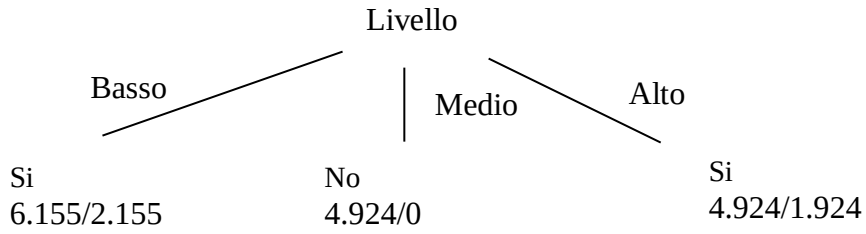
gain(Punteggio) =0.989 -0.977=0.012

splitinfo(Punteggio)= - 8/16*$\log_2$(8/16) - 8/16*$\log_2$(8/16) = 1

gainratio(Punteggio) =0.012/1=0.012

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

.



d) l'istanza viene mandata lungo il ramo Alto e classificata come Si con probabilità 3/4.924=60.9% e come No con probabilità 1-0.609=39.1%.

Esercizio 3

Database:

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

conteggi

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

C2=

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

Conteggi

Itemset	Supporto
~~1,2~~	~~1~~
1,3	2
~~1,4~~	~~1~~
1,5	3
2,3	2
2,4	2
2,5	3
~~3,4~~	~~0~~
3,5	3
4,5	2

C3=

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

Conteggi

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

C4={}

Esercizio 3

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

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

P(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,T2,T3,T4,T5)

P(~T1,~T2,T3,T4,T5)=P(~T1)P(~T2)P(T3|~T1,~T2)P(T5)P(T4|~T2,T5)
=0.1*0.4*0.2*0.9*0.9=0.00648

P(~T1,T2,T3,T4,T5)=P(~T1)P(T2)P(T3|~T1,T2)P(T5)P(T4|T2,T5) =0.1*0.6*0.4*0.9*0.7= 0.01512

P(T1,~T2,T3,T4,T5)=P(T1)P(~T2)P(T3|T1,~T2)P(T5)P(T4|~T2,T5) =0.9*0.4*0.9*0.9*0.9= 0.26244

P(T1,T2,T3,T4,T5)=P(T1)P(T2)P(T3|T1,T2)P(T5)P(T4|T2,T5) =0.9*0.6*0.7*0.9*0.7=0.23814

P(~T1,T3,T4,T5)= 0.00648+0.01512 =0.0216
P(T3,T4,T5)= 0.0216+0.26244+0.23814=0.52218

P(~T1|T3,T4,~T5) = 0.0216/0.52218=0.04136504653

Esercizio 3

Grounding

```
epidemic : 0.6 ; pandemic : 0.3 :- flu(david), cold.
epidemic : 0.6 ; pandemic : 0.3 :- flu(robert), cold.

cold.
flu(david).
flu(robert).
```

Mondi possibili

```
e=epidemic,  p=pandemic,  fd=flu(david),  fr=flu(robert),  c=cold,
n=null.
cold.
flu(david).
flu(robert).
```

`e:-fd.c  e:-fr,c.`	`e:-fd,c. p:-fr,c`	`e:-fd,c. n:-fr,c`
P(w1)=0,36	P(w2)=0,18	P(w3)=0,06
`p:-fd.c  e:-fr,c.`	`p:-fd,c. p:-fr,c`	`p:-fd,c. n:-fr,c`
P(w4)=0,18	P(w5)=0,09	P(w6)=0,03
`n:-fd.c  e:-fr,c.`	`n:-fd,c. p:-fr,c`	`n:-fd,c. n:-fr,c`
P(w7)=0,06	P(w8)=0,03	P(w9)=0,01

P(e)=0,36+0,18+0,06+0,18+0,06=0,84