COMPITO DI DATA MINING & ANALYTICS

22 luglio 2019 (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	?

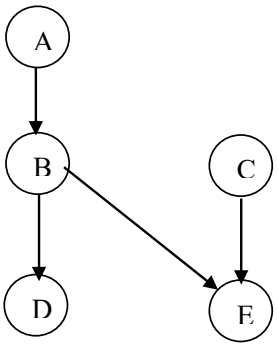
Esercizio 2 (punti 5)

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

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

	A=Falso	A=Vero
	0.3	0.7

per C:

	C=Falso	C=Vero
	0.2	0.8

per B:

A	B=falso	B =vero
Falso	0.3	0.7
Vero	0.7	0.3

per D:

B	D=falso	D =vero
Falso	0.3	0.7
Vero	0.7	0.3

per E:

B	C	E=falso	E=vero
Falso	Falso	0.5	0.5
Falso	Vero	0.2	0.8
Vero	Falso	0.4	0.6
Vero	Vero	0.6	0.4

Si calcoli la probabilità P(E|A,B,D).

Esercizio 4 (punti 4)

Dato il seguente LPAD

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

c(1).
c(2).

d(X):0.3.

c:- d(X),a.
c:- b.
```

Si calcoli la probabilità di `c`

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

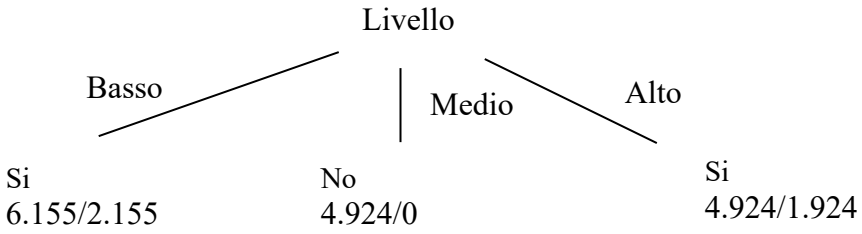
gain(Punteggio) =0.989 -0.977=0.012

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

conteggi

Itemset	Supporto
1	3
2	3
3	4
4	6
5	3
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~~	1
~~1,3~~	1
1,4	3
~~1,5~~	2
~~1,6~~	2
2,3	3
2,4	3
~~2,5~~	1
~~2,6~~	0
3,4	4
~~3,5~~	1
~~3,6~~	1
4,5	3
4,6	3
~~5,6~~	2

C3=

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

Conteggi

2,3,4	3

C4={}

Esercizio 3

P(E|A,B,D)=P(A,B,D,E)/P(A,B,D)
P(A,B,D,E)=P(A,B,C,D,E)+ P(A,B,~C,D,E)
P(A,B,D)=P(A,B,~C,D,~E)+ P(A,B,~C,D,E)+ P(A,B,C,D,~E)+ P(A,B,C,D,E)

P(A,B,~C,D, ~E)=P(A)P(B|A)P(~C)P(D|B)P(~E|B,~C)= 0.7*0.3*0.2*0.3*0.4=0. 00504

P(A,B,~C,D,E)=P(A)P(B|A)P(~C)P(D|B)P(E|B,~C)= 0.7*0.3*0.2*0.3*0.6=0.00756

P(A,B,C,D,~E)= P(A)P(B|A)P(C)P(D|B)P(~E|B,C)=0.7*0.3*0.8*0.3*0.6=0.03024

P(A,B,C,D,E)=P(A)P(B|A)P(C)P(D|B)P(E|B,C)=0.7*0.3*0.8*0.3*0.4=0.02016

P(A,B,D,E) =0.00756+0.02016=0.02772

P(A,B,D) = 0. 02772+0.03024+0.00504=0.063

P(E|A,B,D) =0.02772/0.063=0.44

Esercizio 3

Grounding

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

c(1).
c(2).

d(1):0.3.
d(2):0.3.

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

Mondi possibili

a:-c(1). d(1). d(2). P(w1)=0,027	a:-c(1). d(1). P(w2)=0,063
a:-c(1). d(2). P(w3)= 0,063	a:-c(1). P(w4)=0,147
b:-c(1). d(1). d(2). P(w5)=0,027	b:-c(1). d(1). P(w6)=0,063
b:-c(1). d(2). P(w7)=0,063	b:-c(1). P(w8)=0,147
n:-c(1). d(1). d(2). P(w9)=0,036	n:-c(1). d(1). P(w10)=0,084
n:-c(1). d(2). P(w11)=0,084	n:-c(1). P(w12)=0,196

P(c)=P(w1)+P(w2)+P(w3)+P(w5)+P(w6)+P(w7)+P(w8) = 0,027 + 0,063 + 0,063 + 0,027 + 0,063+ 0,063 + 0,147 =0,453