COMPITO DI DATA MINING & ANALYTICS

29 gennaio 2018 (Punteggio 17; Tempo 2h)

Esercizio 1 (punti 4)

Dato il seguente training set S:

Sesso	**CAP**	**Classe**
f	1	Si
m	2	Si
m	1	No
f	?	No
m	1	No
f	3	Si
f	1	No
m	3	Si
m	2	No
f	3	Si
m	2	No
m	?	Si
f	2	No
f	3	Si

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

m	?

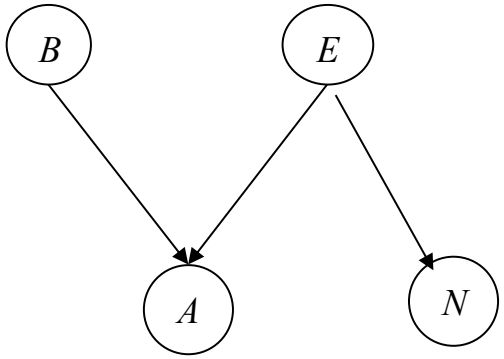
Esercizio 2 (punti 5)

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

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

Esercizio 3 (punti 4)

Sia data la seguente rete bayesiana:



dove tutte le variabili assumono i valori yes e no.
Le tabelle di probabilità condizionata sono

P(B)	
B=yes	0.1
B=no	0.9

P(E)	
E=yes	0.05
E=no	0.95

P(A|BE)	no,no	no,yes	yes,no	yes,yes
A=yes	0.1	0.85	0.9	0.99
A=no	0.9	0.15	0.1	0.01

P(N|E)	E=no	E=yes
N=yes	0.1	0.95
N=no	0.9	0.05

Si calcoli la probabilità P(~E|~N,A)

Esercizio 4 (punti 4)

Dato il seguente LPAD

```
earthquake(strong) : 0.3 ; earthquake(moderate) : 0.5 :-
  fault_rupture,volcanic_eruption.

fault_rupture.
volcanic_eruption:0.3:-
  volcano(X).

volcano(stromboli).
volcano(eyjafjallajkull).
```

Si calcoli la probabilità di `earthquake(strong)`

SOLUZIONE

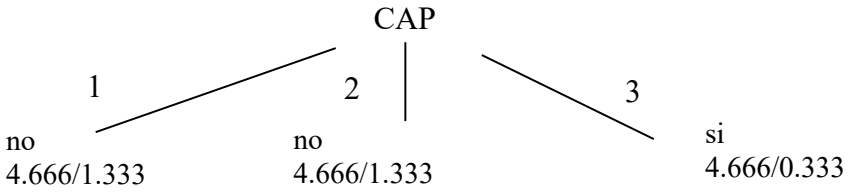
Esercizio 1

a) info(S)=-7/14*$\log_2$ 7/14-7/14*$\log_2$ 7/14= 1.0
b)

$info_{Sesso}$(S)= 7/14*(-4/7*$\log_2$4/7-3/7*$\log_2$ 3/7) + 7/14*(-3/7*$\log_2$3/7-4/7*$\log_2$ 4/7) = 0.5 *0.985+0.5 *0.985 =0.985
gain(Sesso) =1 -0.985=0.015
splitinfo(Sesso)= - 7/14*$\log_2$(7/14) - 7/14*$\log_2$(7/14)= 1
gainratio(Sesso) =0.015/ 1=0.015

Per calcolare il guadagno dell'attributo CAP non si usa l'entropia calcolata su tutto il training set ma solo sugli esempi che hanno CAP noto (insieme F):
info(F)=-6/12*$\log_2$ 6/12– 6/12*$\log_2$ 6/12= 1.0
$info_{CAP}$(F)=4/12*(-1/4*$\log_2$1/4-3/4*$\log_2$ 3/4)+4/12(-1/4*$\log_2$ 1/4-3/4*$\log_2$ 3/4)+4/12(-4/4*$\log_2$ 4/4-0/4*$\log_2$ 0/4) =0.333 *0.811+0.333 *0.811 +0.333 *0=0.540
gain(CAP) =12/14*(1-0.540) =0.394
splitinfo(CAP) =-4/14*$\log_2$(4/14)-4/14*$\log_2$(4/14)-4/14*$\log_2$(4/14)-2/14*$\log_2$(2/14)= 1.950
gainratio(CAP) =0.394/1.950=0.202

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



d) l'istanza viene divisa in tre parti, di peso rispettivamente 4.666/14=0.333, 4.666/14=0.333 e 4.666/14=0.333. La prima parte viene mandata lungo il ramo 1 e classificata come no con probabilità 3.333/4.666=71.4% e come Si con probabilità 1.333/4.666=28.6%. La seconda parte viene mandata lungo il ramo 2 e classificata come come no con probabilità 3.333/4.666=71.4% e come Si con probabilità 1.333/4.666=28.6%. La terza parte viene mandata lungo il ramo 3 e classificata come come si con probabilità 4.333/4.666=92.9% e come no con probabilità 0.333/4.666=7.1%. Quindi in totale la classificazione dell'istanza è
P(Si) =0.333*28.6%+0.333*28.6%+0.333*92.9%=0.5
P(No) = 0.333*71.4%+0.333*71.4%+0.333*7.1%=0.5

Esercizio 3

conteggi

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

C3=

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

Conteggi

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

C4={}

Esercizio 3

Si calcoli la probabilità P(~E|~N,A)

P(~E|~N,A)=P(~E,~N,A)/P(~N,A)

P(~E,~N,A)= P(B,A,~E,~N)+ P(~B,A,~E,~N)

P(B,A,~E,~N)=P(B)P(~E)P(A|B,~E)P(~N|~E)=0.1*0.95*0.9*0.9=0.07695

P(~B,A,~E,~N)=P(~B)P(~E)P(A|~B,~E)P(~N|~E)=0.9*0.95*0.1*0.9=0.07695

P(~N,A)=P(~E,~N,A)+P(B,A,E,~N)+ P(~B,A,E,~N)

P(B,A,E,~N)=P(B)P(E)P(A|B,E)P(~N|E)=0.1*0.05*0.99*0.05=0.0002475

P(~B,A,E,~N)=P(~B)P(E)P(A|~B,E)P(~N|E)=0.9*0.05*0.85*0.05=0.0019125

P(~E,A,~N)= 0.07695+0.07695=0.1539

P(~N,A)=0.1539+0.0002475+0.0019125=0.15606

P(N|~A,B)= 0.1539/0.15606=0.9861591696

Esercizio 3

Grounding

```
earthquake(strong) : 0.3 ; earthquake(moderate) : 0.5 :-
  fault_rupture,volcanic_eruption.

fault_rupture.

volcanic_eruption:0.3:-
  volcano(stromboli).
volcanic_eruption:0.3:-
  volcano(eyjafjallajkull).

volcano(stromboli).
volcano(eyjafjallajkull).



s= earthquake(strong) m= earthquake(moderate)
v= volcanic_eruption
vs= volcano(stromboli) ve= volcanic_eruption(eyjafjallajkull)
f=fault_rupture
n=null

s:0.3;m:0.5;n:0.2:-f,v.
v:0.3:-vs.
v:0.3:-ve.
f.
vs.
```

ve.

Mondi possibili

s:-f,v. v:-vs. v:-ve. P(w1)=0,027	m:-f,v. v:-vs. v:-ve. P(w2)=0,045	n:-f,v. v:-vs. v:-ve. P(w3)=0,018
s:-f,v. n:-vs. v:-ve. P(w4)=0,063	m:-f,v. n:-vs. v:-ve. P(w5)=0,105	n:-f,v. n:-vs. v:-ve. P(w6)=0,042
s:-f,v. v:-vs. n:-ve. P(w7)=0,063	m:-f,v. v:-vs. n:-ve. P(w8)=0,105	n:-f,v. v:-vs. n:-ve. P(w9)=0,042
s:-f,v. n:-vs. n:-ve. P(w10)=0,147	m:-f,v. n:-vs. n:-ve. P(w11)=0,245	n:-f,v. n:-vs. n:-ve. P(w12)=0,098

P(s)=0,027+0,063+0,063=0,153