COMPITO DI DATA MINING & ANALYTICS

27 luglio 2017 (Punteggio 17; Tempo 2h)

Esercizio 1 (punti 4)

Dato il seguente training set S:

InCorso	**VotoMaturità**	**Classe**
Si	Alto	Si
Si	Basso	No
No	Medio	Si
?	Medio	No
No	Alto	Si
No	Medio	Si
Si	Basso	No
Si	Alto	No
Si	Basso	No
No	Basso	Si
?	Alto	No
Si	Medio	No
No	Medio	Si
Si	Alto	No
No	Basso	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 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:

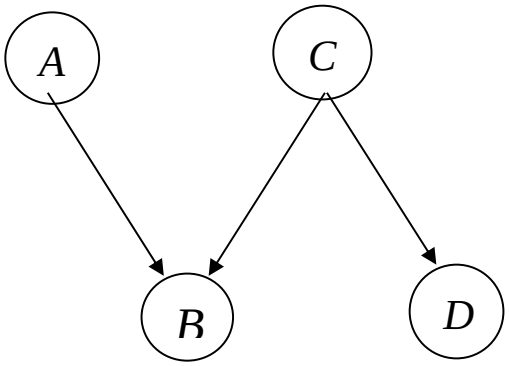
?	Alto

Esercizio 2 (punti 5)

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

ID transazione	Items acquistati
1	2,3,5
2	1,2,3,4
3	1,2,3,5
4	2,3,5
5	1,2,4
6	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(A)	
A=yes	0.2
A=no	0.8

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

P(D|C)	no	yes
D=yes	0.1	0.95
D=no	0.9	0.05

P(B|AC)	no,no	no,yes	yes,no	yes,yes
B=yes	0.1	0.85	0.9	0.99
B=no	0.9	0.15	0.1	0.01

Si calcoli la probabilità P(~D|~A,~B)

Esercizio 4 (punti 4)

Dato il seguente LPAD

```
epidemic :- flu(X), cold.

cold:0.3.
flu(david):0.4.
flu(robert):0.5.
```

Si calcoli la probabilità di `epidemic`.

SOLUZIONE

Esercizio 1

a) info(S)=-7/15*$\log_2$ 7/15-8/15*$\log_2$ 8/15= 0.997

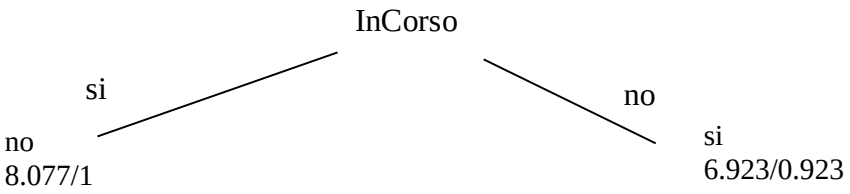
b)

Per calcolare il guadagno dell'attributo InCorso non si usa l'entropia calcolata su tutto il training set ma solo sugli esempi che hanno InCorso noto (insieme F):
info(F)=-7/13*$\log_2$ 7/13– 6/13*$\log_2$ 6/13= 0.996
$info_{InCorso}$(F)=7/13*(-1/7*$\log_2$1/7-6/7*$\log_2$ 6/7)+6/13(-6/6*$\log_2$ 6/6-0/6*$\log_2$ 0/6) =0.538 *0.592 +0.462 *0=0.318
gain(InCorso) =13/15*(0.996-0.318) =0.588
splitinfo(InCorso) =-7/15*$\log_2$(7/15)-6/15*$\log_2$(6/15)-2/15*$\log_2$(2/15)= 1.429
gainratio(InCorso) =0.588/1.429=0.411

$info_{VotoMaturità}$(S)= 5/15*(-2/5*$\log_2$2/5-3/5*$\log_2$ 3/5) + 5/15*(-3/5*$\log_2$3/5-2/5*$\log_2$ 2/5) + 5/15*(-2/5*$\log_2$2/5-3/5*$\log_2$ 3/5) = 0.333 *0.971+0.333 *0.971+0.333 *0.971=0.971
gain(VotoMaturità) =0.997 -0.971=0.026
splitinfo(VotoMaturità)= - 5/15*$\log_2$(5/15) - 5/15*$\log_2$(5/15)- 5/15*$\log_2$(5/15)= 1.585
gainratio(VotoMaturità) =0.026/ 1.585=0.016

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

.



d) l'istanza viene divisa in due parti, di peso rispettivamente 8.077/15=0.538 e 6.923/15=0.462. La prima parte viene mandata lungo il ramo si e classificata come no con probabilità 7.077/8.077=87.6% e come Si con probabilità 1/8.077=12.4%. La seconda parte viene mandata lungo il ramo no e classificata come si con probabilità 6/6.923=86.7% e come no con probabilità 0.923/6.923=13.3%. Quindi in totale la classificazione dell'istanza è
P(Si) =0.538*12.4%+0.462*86.7%=0.467
P(No) = 0.538*87.6%+0.462*13.3%=0.533

Esercizio 3

conteggi

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

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

C3=

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

Conteggi

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

C4=

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

Esercizio 3

Si calcoli la probabilità P(~D|~A,~B)

P(~D|~A,~B)=P(~D, ~A, ~B)/P(~A, ~B)

P(~D, ~A, ~B)=P(~D, ~A, ~B, ~C)+ P(~D, ~A, ~B,C)

P(~A, ~B)=P(~D, ~A, ~B)+P(D, ~A, ~B, ~C)+P(D, ~A, ~B,C)

P(~D, ~A, ~B, ~C)=P(~A)P(~C)P(~B|~A, ~C)P(~D|~C)=0.8*0.95*0.9*0.9=0.6156

P(~D, ~A, ~B,C)=P(~A)P(C)P(~B|~A,C)P(~D|C)=0.8*0.05*0.15*0.05=0.0003

P(D, ~A, ~B, ~C)=P(~A)P(~C)P(~B|~A, ~C)P(D|~C)=0.8*0.95*0.9*0.1=0.0684

P(D, ~A, ~B,C)=P(~A)P(C)P(~B|~A,C)P(D|C)=0.8*0.05*0.15*0.95=0.0057

P(~D, ~A, ~B)= 0.6156+0.0003=0.6159

P(~A, ~B)= 0.6159 +0.0684 +0.0057 =0.69

P(~D|~A, ~B)= 0.6159 /0.69 =0.892608696

Esercizio 3

```
epidemic :- flu(X), cold.

cold:0.3.
flu(david):0.4.
flu(robert):0.5.
```

Mondi possibili

```
c= cold, d=flu(david) r=flu(Robert)
```

c. d. r.	c. d.
P(w1)=0,06	P(w2)=0,06
c. r.	c.
P(w3)=0,09	P(w4)=0,09
d. r.	d.
P(w5)=0,14	P(w6)=0,14
r.	
P(w7)=0,21	P(w8)=0,21

P(e)=0,06+0,06+0,09=0,21