COMPITO DI DATA MINING & ANALYTICS

15 febbraio 2018 (Punteggio 17; Tempo 2h)

Esercizio 1 (punti 4)

Dato il seguente training set S:

Credit	**Loan**	**Classe**
1	2000	Sì
1	2000	No
1	2000	No
2	1000	Sì
1	2000	No
?	1000	Si
2	2000	No
1	1000	Sì
1	2000	No
2	1000	Si
2	1000	Sì
1	2000	No
2	1000	Sì

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:

1	?

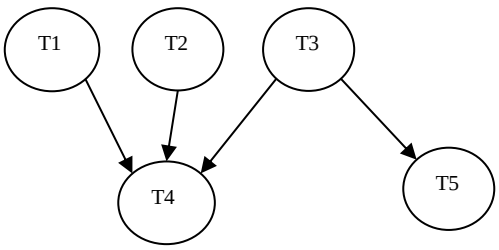
Esercizio 2 (punti 5)

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

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

```
a1 : 0.3; a2 : 0.2; a3 : 0.1.

b:0.3.

c:-b,a1.
C:-b,a3.
```

Si calcoli la probabilità di `c`

SOLUZIONE

Esercizio 1

a) info(S)=-7/13*$\log_2$ 7/13-6/13*$\log_2$ 6/13= 0.996

b)

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

info(F)=-6/12*$\log_2$ 6/12– 6/12*$\log_2$ 6/12= 1

$info_{Credit}$(F)=7/12*(-2/7*$\log_2$2/7-5/7*$\log_2$ 5/7)+5/12(-4/5*$\log_2$ 4/5-1/5*$\log_2$ 1/5) =0.583 *0.863 +0.417 *0.722=0.804

gain(Credit) =12/13*(1 -0.804) =0.180

splitinfo(Credit) =-6/13*$\log_2$(7/13)-6/13*$\log_2$(6/13)-1/13*$\log_2$(1/13)=1.314

gainratio(Credit) =0.180/1.314=0.137

$info_{Loan}$(S)= 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

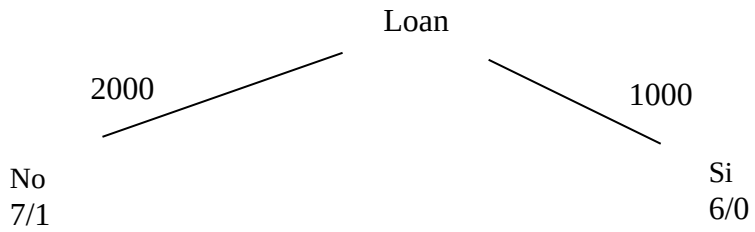
gain(Loan) =0.996 -0.318=0.678

splitinfo(Loan)= - 7/13*$\log_2$(7/13) - 6/13*$\log_2$(6/13) = 0.996

gainratio(Loan) =0.678 / 0.996=0.681

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

.



d) l'istanza viene divisa in due parti, di peso rispettivamente 7/13=0.538 e 6/13=0.462. La prima parte viene mandata lungo il ramo 2000 e classificata come No con probabilità 6/7=85.7.5% e come si con probabilità 1/7=14.3%. La seconda parte viene mandata lungo il ramo 1000 e classificata come Si con probabilità 100%. Quindi in totale la classificazione dell'istanza è

P(Si) =0.538*14.3%+0.462*100%=53.9%

P(No) = 0.538*85.7%+0.462*0%=46.1%

Esercizio 3

conteggi

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

C3=

1,4,5
2,3,4
~~4.5.6~~

Conteggi

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

C4={}

Esercizio 3

P(T2 |T1,T3,T4, T5)= 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*0.6*0.8=0.192

P(T1, ~T2,T3, T4) = P(T1)P(~T2)P(T3)P(T4|T1, ~T2,T3)
=0.8*0.5*0.6*0.1=0.024

P(T1,T3, T4) = 0.192+0.024=0.216

P(T2|T1,T3,T4,T5) = 0.192/0.216=0.8888888

Esercizio 3

Grounding

```
a1 : 0.3; a2 : 0.2; a3 : 0.1.

b:0.3.

c:-b,a1.
C:-b,a3.
```

Mondi possibili

a1. b. P(w1)=0,09	a1. P(w2)=0,21
a2. b. P(w3)=0,06	a2. P(w4)=0,14
a3. b. P(w5)=0,03	a3. P(w6)=0,07
 b. P(w7)=0,12	 P(w8)=0,28

P(c)=P(w1)+P(w5)=0,09+0,03=0,12