COMPITO DI DATA MINING & ANALYTICS

29 settembre 2017 (Punteggio 17; Tempo 2h)

Esercizio 1 (punti 4)

Dato il seguente training set S:

Seme	**Valore**	**Classe**
Cuori	Numero	Pos
Quadri	Figura	Neg
Fiori	Numero	Neg
Picche	Figura	Pos
Fiori	Figura	Neg
Fiori	Figura	Pos
Quadri	Numero	Neg
Cuori	Figura	Pos
Picche	Figura	Pos
Fiori	Numero	Neg
Quadri	Figura	Pos
Cuori	Figura	Neg
Picche	Numero	Neg
Quadri	?	Pos
Cuori	?	Neg

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:

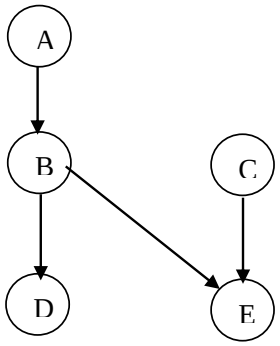
Quadri	?

Esercizio 2 (punti 5)

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

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

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
	0.2	0.8

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	0.1	0.9
Vero	Falso	0.4	0.6
Vero	Vero	0.3	0.7

Si calcoli la probabilità P(D|A,~C,~E).

Esercizio 4 (punti 4)

Dato il seguente LPAD

```
popular(X) :- friends(X,Y), popular(Y).

friends(john,david).
friends(john,robert):0.2.
popular(david):0.3.
popular(robert):0.6.
```

Si calcoli la probabilità di `popular(john).`

SOLUZIONE

Esercizio 1

a) info(S)=-7/15*$\log_2$ 5/15-8/15*$\log_2$ 8/15=0.997

b)

info$_{Seme}$(S)= 4/15*(-2/4*$\log_2$2/4-2/4*$\log_2$ 2/4) +4/15*(-2/4*$\log_2$2/4-2/4*$\log_2$ 2/4)+ 4/15*(-1/4*$\log_2$1/4-3/4*$\log_2$ 3/4)+ 3/15*(-2/3*$\log_2$2/3-1/3*$\log_2$ 1/3) =

=0.267*1+0.267*1+0.267*0.811+0.2*0.918=0.934

gain(Seme) =0.997-0.934=0.063

splitinfo(Seme)= - 4/15*$\log_2$(4/15) - 4/15*$\log_2$(4/15) -4/15*$\log_2$(4/15)- 3/15*$\log_2$(3/15)= 1.990

gainratio(Seme) =0.063/1.990=0.032

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

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

info$_{Valore}$ (F)=5/13*(-1/5*$\log_2$1/5-4/5*$\log_2$ 4/5)+8/13(-5/8*$\log_2$ 5/8-3/8*$\log_2$ 3/8) =

=0.385*0.722+0.615*0.954=0.865

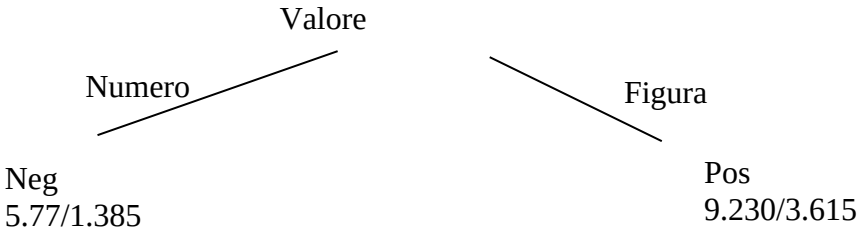
gain(Valore) =13/15*(0.996-0.865) =0.114

splitinfo(Valore) =-5/15*$\log_2$(5/15)-8/15*$\log_2$(8/15)-2/15*$\log_2$(2/15) =1.400

gainratio(Valore) =0.114/1.400=0.08143

c) L’attributo scelto per la radice dell’albero è Valore

.



d) l'istanza viene divisa in due parti, di peso rispettivamente 0.385 e 0.615. La prima parte viene mandata lungo il ramo Numero e classificata come Neg con probabilità =4.385/5.77=76% e come Pos con probabilità =1.385/5.77=24%. La seconda parte viene mandata lungo il ramo Figura e classificata come Pos con probabilità =5.615/9.23=60.8% e come Neg con probabilità =3.615/9.23=39.2%. Quindi in totale la classificazione dell'istanza è

P(Pos) =0.385*24%+0.615*60.8%=46.7%

P(Neg) = 0.385*76%+0.615*39.2%=53.4%

Esercizio 3

conteggi

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

C3=

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

Conteggi

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

C4=

2,3,4,5

Conteggi

2,3,4,5	2

Esercizio 3

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

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

P(A,~C,~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,B,~C,D,~E)=P(A)P(B|A)P(~C)P(D|B)P(~E|B,~C)=0.8*0.3*0.2*0.3*0.4=0.00576

P(A,~B,~C,D,~E)=P(A)P(~B|A)P(~C)P(D|~B)P(~E|~B,~C)=0.8*0.7*0.2*0.7*0.5=0.0392

P(A,B,~C,~D,~E)=P(A)P(B|A)P(~C)P(~D|B)P(~E|B,~C)=0.8*0.3*0.2*0.7*0.4=0.01344

P(A,~B,~C,~D,~E)=P(A)P(~B|A)P(~C)P(~D|~B)P(~E|~B,~C)=0.8*0.7*0.2*0.3*0.5=0.0168

P(A,~C,D,~E)= 0.00576+0.0392=0.04496

P(A,~C,~E)= 0.04496+0.01344+0.0168=0.0752

P(D|A,~C,~E)= 0.04496/0.0752=0.59787234042

Esercizio 3

```
popular(X) :- friends(X,Y), popular(Y).

friends(john,david).
friends(john,robert):0.2.
popular(david):0.3.
popular(robert):0.6.
```

Mondi possibili

```
f= friends(john,robert), d= popular(david) r= popular(robert)
```

f. d. r. P(w1)=0,036	f. d. P(w2)=0,024
f. r. P(w3)=0,084	f. P(w4)=0,056
d. r. P(w5)=0,144	d. P(w6)=0,096
r. P(w7)=0,336	P(w8)=0,224

P(e)=0,036+0,024+0,084+0,144+0,096=0,384