

COMPITO DI DATA MINING & ANALYTICS

14 luglio 2017 (Punteggio 17; Tempo 2h)

Esercizio 1 (punti 4)

Dato il seguente training set S:

Età	Mobilità	Classe
A	Si	Si
B	No	Si
B	No	No
C	No	No
C	?	No
A	Si	Si
B	No	No
C	Si	No
C	?	Si
A	No	No
B	Si	Si
C	?	Si
A	No	No
A	Si	Si
B	?	Si

- Si calcoli l'entropia del training set rispetto all'attributo Classe
- Si calcoli il guadagno dei due attributi rispetto a questi esempi di training
- 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).
- si classifichi l'istanza:

A	?
---	---

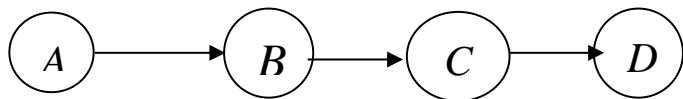
Esercizio 2 (punti 4)

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

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

Esercizio 3 (punti 5)

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(B A)	A=no	A=yes
B=yes	0.2	0.8
B=no	0.8	0.2

P(C B)	B=no	B=yes
C=yes	0.2	0.8
C=no	0.8	0.2

P(D C)	C=no	C=yes
D=yes	0.2	0.8
D=no	0.8	0.2

Si calcoli la probabilità $P(A|C,D)$

Esercizio 4 (punti 4)

Dato il seguente LPAD

```
nature_lover(X):- has_animal(X,Y),pet(Y).
pet(X):- cat(X).
has_animal(kevin, tom).
has_animal(kevin, fluffy):0.6.
cat(tom):0.3.
cat(fluffy):0.4.
```

Si calcoli la probabilità di $\text{nature_lover(kevin)}$.

SOLUZIONE

Esercizio 1

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

b)

$$\text{info}_{\text{Età}}(S) = 5/15 * (-3/5 * \log_2 3/5 - 2/5 * \log_2 2/5) + 5/15 * (-3/5 * \log_2 3/5 - 2/5 * \log_2 2/5) + 5/15 * (-2/5 * \log_2 2/5 - 3/5 * \log_2 3/5) = 0.333 * 0.971 + 0.333 * 0.971 + 0.333 * 0.971 = 0.971$$

$$\text{gain}(\text{Età}) = 0.997 - 0.971 = 0.026$$

$$\text{splitinfo}(\text{Età}) = -5/15 * \log_2(5/15) - 5/15 * \log_2(5/15) - 5/15 * \log_2(5/15) = 1.585$$

$$\text{gainratio}(\text{Età}) = 0.026 / 1.585 = 0.016$$

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

$$\text{info}(F) = -5/11 * \log_2 5/11 - 6/11 * \log_2 6/11 = 0.994$$

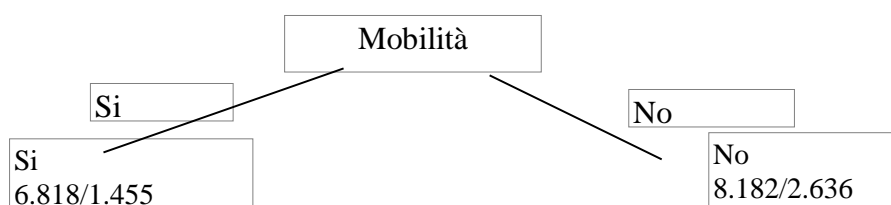
$$\text{info}_{\text{Mobilità}}(S) = 5/11 * (-4/5 * \log_2 4/5 - 1/5 * \log_2 1/5) + 6/11 * (-1/6 * \log_2 1/6 - 5/6 * \log_2 5/6) = 0.455 * 0.722 + 0.545 * 0.650 = 0.683$$

$$\text{gain}(\text{Mobilità}) = 11/15 * (0.994 - 0.683) = 0.228$$

$$\text{splitinfo}(\text{Mobilità}) = -5/15 * \log_2(5/15) - 6/15 * \log_2(6/15) - 4/15 * \log_2(4/15) = 1.566$$

$$\text{gainratio}(\text{Mobilità}) = 0.228 / 1.566 = 0.146$$

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



d) l'istanza viene divisa in due parti, di peso rispettivamente $6.818/15=0.455$ e $8.182/15=0.545$. La prima parte viene mandata lungo il ramo Si e classificata come Si con probabilità $5.363/6.818=78.7\%$ e come No con probabilità $1.455/6.818=21.3\%$. La seconda parte viene mandata lungo il ramo No e classificata come No con probabilità $5.546/8.182=67.8\%$ e come Si con probabilità $2.636/8.182=32.2\%$. Quindi in totale la classificazione dell'istanza è

$$P(\text{Si}) = 0.455 * 78.7\% + 0.545 * 32.2\% = 53.4\%$$

$$P(\text{No}) = 0.455 * 21.3\% + 0.545 * 67.8\% = 46.6\%$$

Esercizio 3

Database:

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

conteggi

Itemset	Supporto
---------	----------

1	3
2	5
3	3
4	3
5	5

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

C3=

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

Conteggi

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

C4={}

Esercizio 3

Si calcoli la probabilità $P(A|C,D)$

$$P(A|C,D)=P(A,C)/P(C)$$

$$P(A,C)=P(A,B,C)+P(A,\sim B,C)$$

