Esercizi sugli alberi di decisione

Esercizio

- Si consideri il dataset

Istanza	a1	a2	classe
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-

- Si costruisca a mano l'albero seguendo l'algoritmo di c4.5 nell'ipotesi che il minimo numero di esempi in almeno due sottoinsiemi sia pari a 2

Risposta

A1	+	-	Totale
T	2	1	3
F	1	2	3
Totale	3	3	6

- Nodo radice: info(T)=1
- Test su A1: $info_{A1}(T)=3/6*(-2/3*\log_2 2/3-1/3*\log_2 1/3)+3/6*(-1/3*\log_2 1/3 -2/3*\log_2 2/3)=$
- Ricorda: $\log_a x=\log_{10} x/\log_{10} a=\ln x/\ln a$
- $info_{A1}(T)=0.5*0.918+ 0.5*0.918=0.918$

Risposta

A1	+	-	Totale
T	2	1	3
F	1	2	3
Totale	3	3	6

- splitinfo(A1)=1
- gainratio(A1)=(1-0.918)/1=0.082

Risposta

A2	+	-	Totale
T	2	2	4
F	1	1	2
Totale	3	3	6

- Test su A2: $info_{A2}(T)=4/6*(-2/4*\log_2 2/4 -2/4*\log_2 2/4)+2/6(-1/2*\log_2 1/2 -1/2*\log_2 1/2)$ $=0.667*1+0.333*1=1$
- $splitinfo(A2)=-4/6*\log_2 4/6 -2/6*\log_2 2/6=0.918$
- $gainratio(A2)=(1-1)/0.918=0$
- Viene preferito A1

Risposta

- $T_{A1=T}$={

1	T	T	+
2	T	T	+
3	T	F	-

}

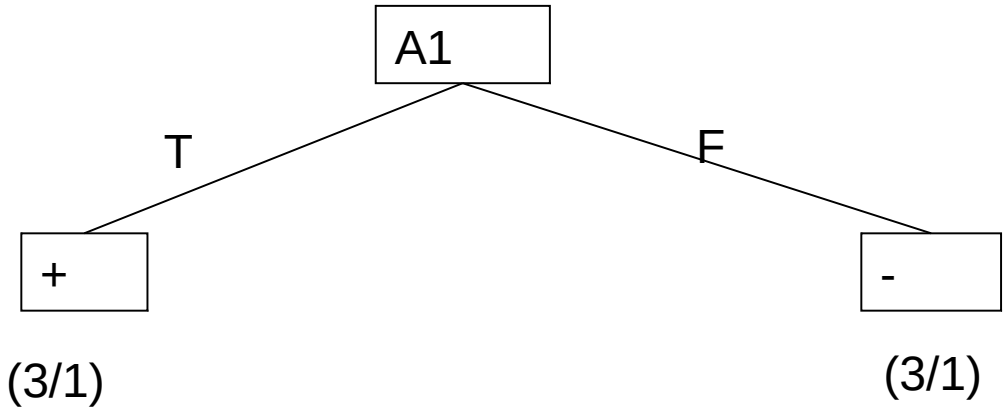
$T_{A1=F}$={

4	F	F	+
5	F	T	-
6	F	T	-

}

- c4.5 si ferma in quanto $T_{A1=T}$ e $T_{A1=F}$ non possono essere suddivisi in modo che almeno due sottoinsiemi abbiamo almeno due elementi

Risposta



Esercizio

- Dato il seguente training set S:

A1	A2	Classe
?	A	+
True	A	-
False	C	+
False	A	+
True	C	-
False	C	-
False	B	+
True	B	+
True	B	-
False	C	+

Esercizio

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

b) Si calcoli il gain ratio 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

True A

Risposta

a) info(S)=-6/10*$\log_2$ 6/10-4/10*$\log_2$ 4/10=0.971

b)Per caclolare il guadagno dell'attributo A1 non si usa l'entropia calcolata su tutto il training set ma solo sugli esempi che hanno A1 noto (insieme F):

info(F)=-4/9*$\log_2$ 4/9 – 5/9*$\log_2$ 5/9=0,991

A1	+	-	Totale
True	1	3	4
False	4	1	5
Totale	5	4	9

Risposta

A1	+	-	Totale
True	1	3	4
False	4	1	5
Totale	5	4	9

$info_{A1}(F)=4/9*(-1/4*\log_2 1/4-3/4*\log_2 3/4)+5/9(-4/5*\log_2 4/5-1/5*\log_2 1/5) =$

$=0.444*0.811+0.556*0.722=0.762$

$gain(A1)=9/10*(0.991-0.762)=0.206$

$splitinfo(A1)=-4/10*\log_2(4/10)-5/10*\log_2(5/10)-1/10*\log_2(1/10)=1.361$

$gainratio(A1)=0.206/1.361=0,151$

Risposta

A2	+	-	Totale
A	2	1	3
B	2	1	3
C	2	2	4
Totale	6	4	10

$info_{A2}(S)=3/10*(-2/3*\log_2 2/3-1/3*\log_2 1/3)+3/10*(-1/3*\log_2 1/3-2/3*\log_2 2/3)+4/10*(-2/4*\log_2 2/4-2/4*\log_2 2/4)=0.3*0.918+0.3*0.918+0.4*1=0.951$

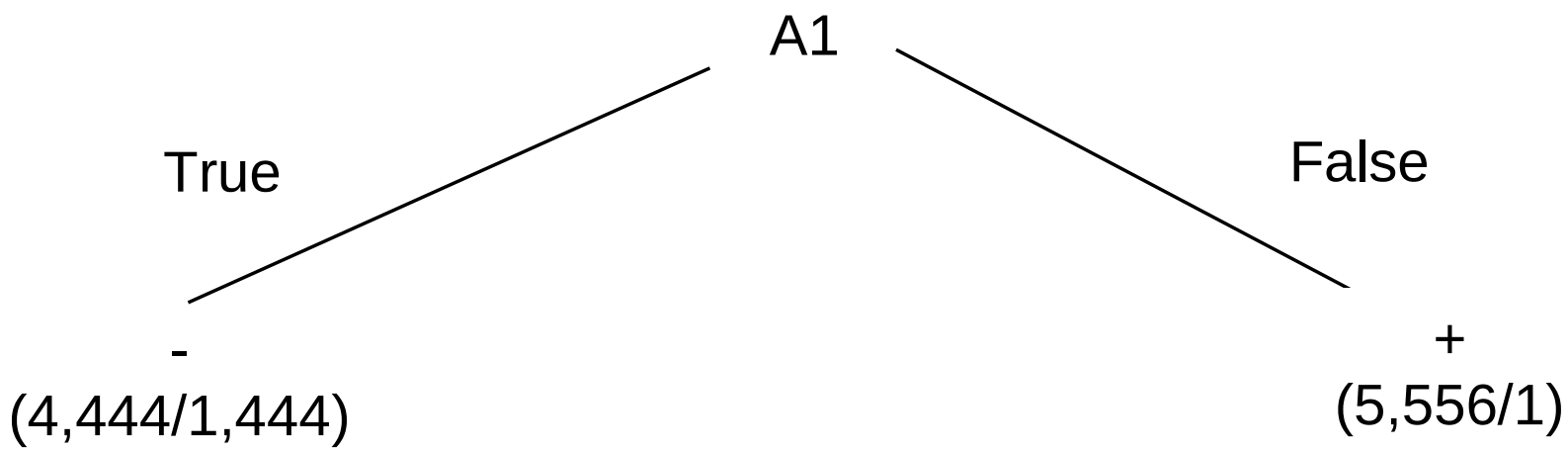
$gain(A2)=0.971-0.951=0.020$

$splitinfo(A2)=-3/10*\log_2(3/10)-3/10*\log_2(3/10)-4/10*\log_2(4/10)=1.571$

$gainratio(A2)=0.020/1.571=0.013$

Risposta

- c)



Risposta

d) l'istanza viene classificata nella foglia di sinistra, quindi appartiene alla classe – con probabilita' 3/4,444=0,675 e alla classe + con probabilità 1,444/4,444=0,325

Esercizio

- Si consideri il training set

Sky	Air Temp	Humid	Wind	EnjoySport
Sunny	Warm	Normal	Strong	Yes
Sunny	Cold	High	Strong	Yes
Rainy	Cold	High	Strong	No
Rainy	Warm	High	Strong	No
Rainy	Warm	Normal	Weak	No
?	Cold	Normal	Weak	No

a) qual è l'entropia del training set rispetto all'attributo EnjoySport?

Esercizio

b) si calcoli il rapporto di guadagno per i quattro attributi Sky, Air Temp, Humid e Wind.

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) dato l'albero costruito, si classifichi l'istanza

Sky	**Air Temp**	**Humid**	**Wind**
?	Cold	Normal	Weak

Risposta

a) info(T)=-2/6*$\log_2$ 2/6 –4/6*$\log_2$ 4/6=0.918

b)Per calcolare il guadagno dell'attributo Sky non si usa l'entropia calcolata su tutto il training set ma solo sugli esempi che hanno sky noto:

Sky	Yes	No	Totale
Sunny	2	0	2
Rainy	0	3	3
Totale	2	3	5

info(F)=-2/5*$\log_2$ 2/5 –3/5*$\log_2$ 3/5=0.971

Per gli altri attibuti invece si utilizza info(T)

Risposta

Sky	Yes	No	Totale
Sunny	2	0	2
Rainy	0	3	3
Totale	2	3	5

Gain(Sky)=5/6*(0,971-2/5*(-2/2*$\log_2$ 2/2 –0/2*$\log_2$ 0/2)-3/5*(–3/3*$\log_2$ 3/3-0/3*$\log_2$ 0/3))=

0.833*(0.971-0.4*0-0.6*0)=0.809

Splitinfo(Sky)=-2/6*$\log_2$2/6-3/6*$\log_2$3/6-1/6*$\log_2$1/6=1.459

Gainratio(Sky)=0.809/1.459=0.555

Risposta

Air Temp	Yes	No	Totale
Warm	1	2	3
Cold	1	2	3
Totale	2	4	6

Gain(Air Temp)=0,918-3/6*(-1/3*$\log_2$ 1/3 –2/3*$\log_2$ 2/3)-3/6*(-1/3*$\log_2$ 1/3 –2/3*$\log_2$ 2/3)=

0.918-0.5*0.918-0.5*0.918=0

Splitinfo(Air Temp)=-3/6*$\log_2$3/6-3/6*$\log_2$3/6=1

Gainratio(Air Temp)=0/1=0

Risposta

Humid	Yes	No	Totale
Normal	1	2	3
High	1	2	3
Totale	2	4	6

Gain(Humid)=0,918-3/6*(-1/3*$\log_2$ 1/3 –2/3*$\log_2$ 2/3)-3/6*(-1/3*$\log_2$ 1/3 –2/3*$\log_2$ 2/3)=

0.918-0.5*0.918-0.5*0.918=0

Splitinfo(Humid)=-3/6*$\log_2$3/6-3/6*$\log_2$3/6=1

Gainratio(Humid)=0/1=0

Risposta

Wind	Yes	No	Totale
Strong	2	2	4
Weak	0	2	2
Totale	2	4	6

Guadagno(Wind)=0,918-4/6*(-2/4*$\log_2$ 2/4 –2/4*$\log_2$ 2/4)-2/6*(-2/2*$\log_2$ 2/2 –0/2*$\log_2$ 0/2)=

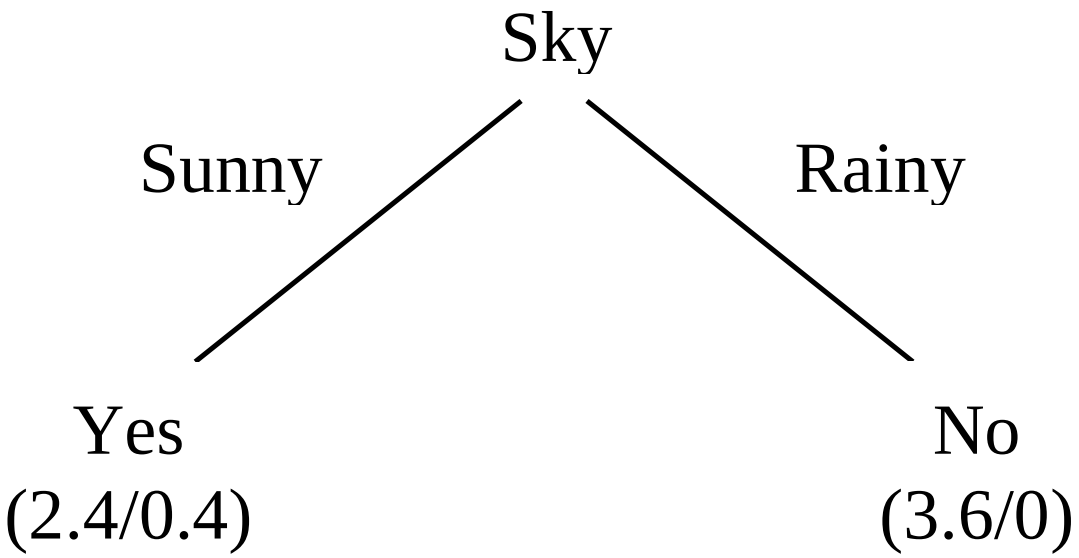
0.918-0.666*1-0.333*0=0.252

Splitinfo(Humid)=-4/6*$\log_2$4/6-2/6*$\log_2$2/6=0.918

Gainratio(Humid)=0.252/0.918=0.274

Risposta

c)



Risposta

- d) l'istanza ha l'attributo Sky sconosciuto, quindi viene suddivisa in due parti: una parte, con peso 2.4/(2.4+3.6)=2.4/6, va nella foglia Yes e una parte, con peso 3.6/6 va nella foglia No. Quindi la classificazione risultante e' Yes con probabilita' 2.4/6*2/2.4+3.6/6*0/3.6=0.4*0.833+0=0.333=33.3% e No con probabilita' 2.4/6*0.4/2.4+3.6/6*3.6/3.6=0.4*0.166+0.6*1=0.667=66.7%.