

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

# Risposta

---

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

- $\text{splitinfo}(A1)=1$
- $\text{gainratio}(A1)=(1-0.918)/1=0.082$

# Risposta

---

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

- Test su A2:  $\text{info}_{A_2}(T) = 4/6 * (-2/4 * \log_2 2/4 - 2/4 * \log_2 2/4) + 2/6 * (-1/2 * \log_2 1/2 - 1/2 * \log_2 1/2)$   
 $= 0.667 * 1 + 0.333 * 1 = 1$
- $\text{splitinfo}(A_2) = -4/6 * \log_2 4/6 - 2/6 * \log_2 2/6 = 0.918$
- $\text{gainratio}(A_2) = (1 - 1) / 0.918 = 0$
- Viene preferito A1

# Risposta

---

•  $T_{A1=T} = \{$

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

}

$T_{A1=F} = \{$

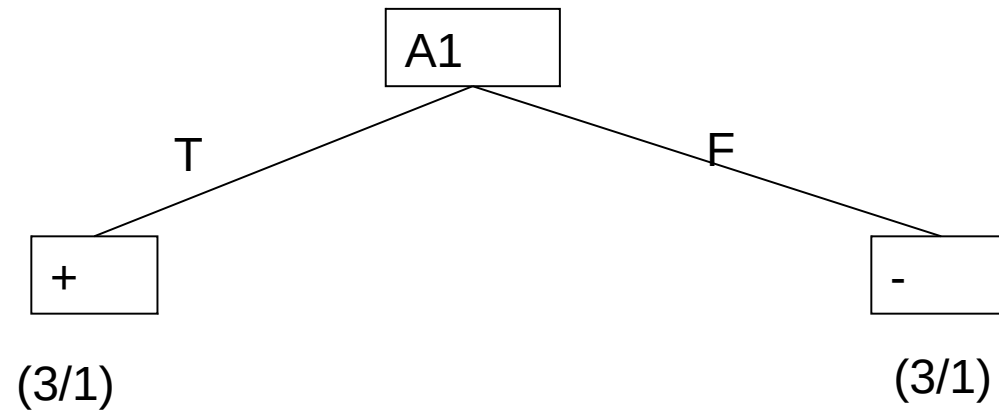
4	F	F	+
5	F	T	-
6	F	T	-

}

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

# Risposta

---



# Esercizio

---

- Dato il seguente training set S:

<b>A1</b>	<b>A2</b>	<b>Classe</b>
?	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)  $\text{info}(S) = -6/10 \cdot \log_2 6/10 - 4/10 \cdot \log_2 4/10 = 0.971$

b) Per calcolare 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):

$\text{info}(F) = -4/9 \cdot \log_2 4/9 - 5/9 \cdot \log_2 5/9 = 0,991$

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

$$\text{info}_{A_1}(F) = 4/9 * (-1/4 * \log_2 1/4 - 3/4 * \log_2 3/4) + 5/9 * (-4/5 * \log_2 4/5 - 1/5 * \log_2 1/5) =$$

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

$$\text{gain}(A_1) = 9/10 * (0.991 - 0.762) = 0.206$$

$$\text{splitinfo}(A_1) = -4/10 * \log_2(4/10) - 5/10 * \log_2(5/10) - 1/10 * \log_2(1/10) = 1.361$$

$$\text{gainratio}(A_1) = 0.206 / 1.361 = 0,151$$

# Risposta

---

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

$$\text{info}_{A_2}(S) = 3/10 * (-2/3 * \log_2 2/3 - 1/3 * \log_2 1/3) + 3/10 * (-1/3 * \log_2 1/3 - 2/3 * \log_2 2/3) + 4/10 * (-2/4 * \log_2 2/4 - 2/4 * \log_2 2/4) = 0.3 * 0.918 + 0.3 * 0.918 + 0.4 * 1 = 0.951$$

$$\text{gain}(A_2) = 0.971 - 0.951 = 0.020$$

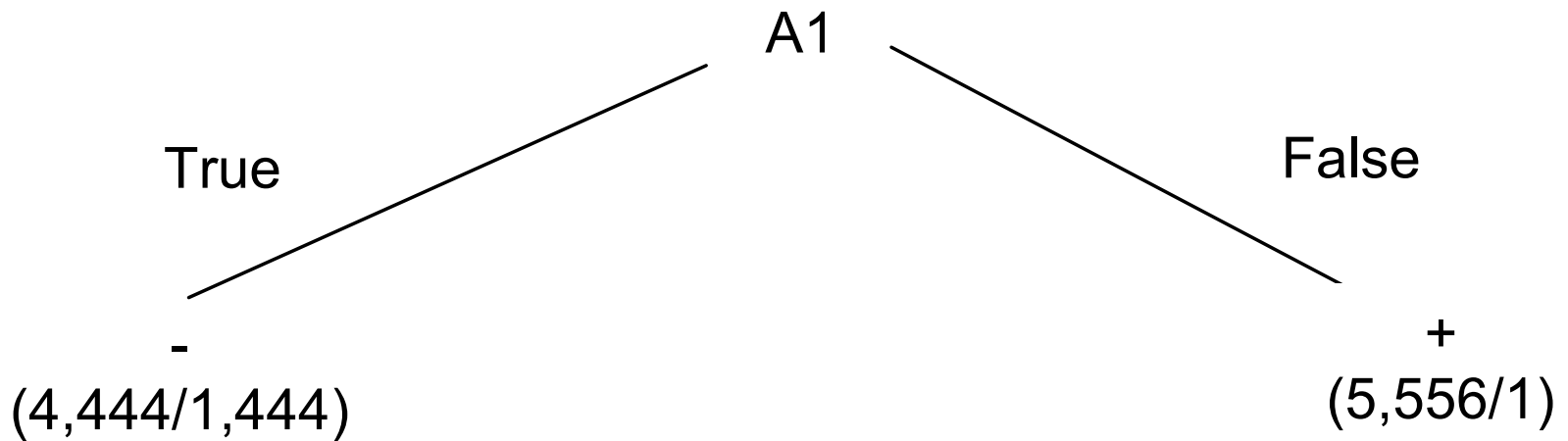
$$\text{splitinfo}(A_2) = -3/10 * \log_2(3/10) - 3/10 * \log_2(3/10) - 4/10 * \log_2(4/10) = 1.571$$

$$\text{gainratio}(A_2) = 0.020 / 1.571 = 0.013$$

# Risposta

---

- c)



# Risposta

---

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

# Esercizio

---

- Si consideri il training set

<b>Sky</b>	<b>Air Temp</b>	<b>Humid</b>	<b>Wind</b>	<b>EnjoySport</b>
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

<b>Sky</b>	<b>Air Temp</b>	<b>Humid</b>	<b>Wind</b>
?	Cold	Normal	Weak



# Risposta

---

a)  $\text{info}(T) = -2/6 \cdot \log_2 2/6 - 4/6 \cdot \log_2 4/6 = 0.918$

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

$\text{info}(F) = -2/5 \cdot \log_2 2/5 - 3/5 \cdot \log_2 3/5 = 0.971$

Per gli altri attributi invece si utilizza  $\text{info}(T)$

# Risposta

---

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

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

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

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

$$\text{Gainratio}(\text{Sky}) = 0.809 / 1.459 = 0.555$$

# Risposta

---

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

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

$$0.918 - 0.5 * 0.918 - 0.5 * 0.918 = 0$$

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

$$\text{Gainratio}(\text{Air Temp}) = 0/1 = 0$$

# Risposta

---

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

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

$$0.918 - 0.5 * 0.918 - 0.5 * 0.918 = 0$$

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

$$\text{Gainratio(Humid)} = 0/1 = 0$$

# Risposta

---

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

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

$$0.918 - 0.666 * 1 - 0.333 * 0 = 0.252$$

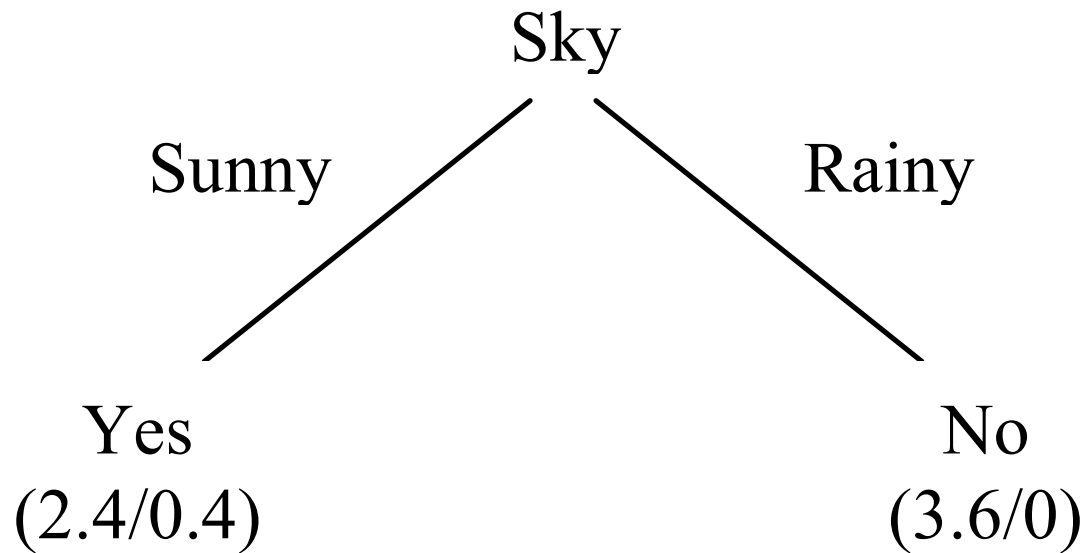
$$\text{Splitinfo(Humid)} = -4/6 * \log_2 4/6 - 2/6 * \log_2 2/6 = 0.918$$

$$\text{Gainratio(Humid)} = 0.252 / 0.918 = 0.274$$

# 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'  $2.4/6 * 0.4 / 2.4 + 3.6/6 * 3.6 / 3.6 = 0.4 * 0.166 + 0.6 * 1 = 0.667 = 66.7\%$ .