

Esercizio 1

a) dato il seguente insieme di apprendimento

<i>Esempio</i>	<i>Target</i>	A_1 (5 val.)	A_2 (3 val.)	A_3 (4 val.)	A_4 (9 val.)
1	+	v_1	v_1	v_4	v_1
2	+	v_1	v_3	v_2	v_2
3	-	v_2	v_1	v_1	v_5
4	+	v_2	v_2	v_4	v_4
5	-	v_3	v_2	v_4	v_3
6	+	v_3	v_1	v_2	v_6
7	+	v_4	v_3	v_1	v_9
8	-	v_4	v_2	v_4	v_9
9	-	v_5	v_1	v_4	v_1
10	+	v_5	v_3	v_2	v_5

mostrare come ID3 (con $\text{GainRatio}(S,A)$) costruisce l'albero di decisione corrispondente. Per ogni attributo calcolare $\text{GainRatio}(S, A)$ e giustificare la scelta dell'attributo utilizzato ad ogni nodo interno. (Ricordare che $\log(\frac{a}{b}) = \log(a) - \log(b)$ e $\log(a \cdot b) = \log(a) + \log(b)$; usare $\log_2(3) = 1.585$ e $\log_2(5) = 2.32$;))

Risposta:

Entropia: 0.970951

Attributo A_1

valore 1:[+,-] 2 0, valore 2:[+,-] 1 1, valore 3:[+,-] 1 1, valore 4:[+,-] 1 1,

valore 5:[+,-] 1 1

Gain: 0.170951 Split: 2.321928 Ratio: 0.073624

Attributo A_2

valore 1:[+,-] 2 2, valore 2:[+,-] 1 2, valore 3:[+,-] 3 0

Gain: 0.295462 Split: 1.570951 Ratio: 0.188078

Attributo A_3

valore 1:[+,-] 1 1, valore 2:[+,-] 3 0, valore 3:[+,-] 0 0, valore 4:[+,-] 2 3

Gain: 0.285475 Split: 1.485475 Ratio: 0.192178

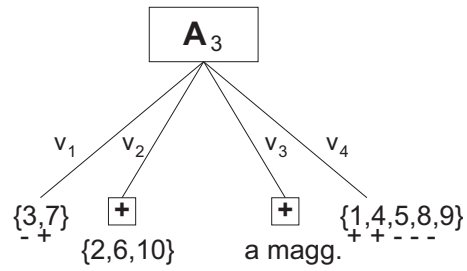
Attributo A_4

valore 1:[+,-] 1 1, valore 2:[+,-] 1 0, valore 3:[+,-] 0 1, valore 4:[+,-] 1 0,

valore 5:[+,-] 1 1, valore 6:[+,-] 1 0, valore 7:[+,-] 0 0, valore 8:[+,-] 0 0,

valore 9:[+,-] 1 1

Gain: 0.370951 Split: 2.721928 Ratio: 0.136282



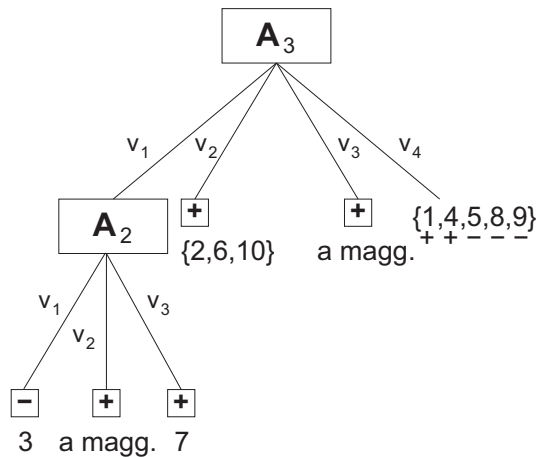
Quindi si sceglie l'attributo A_3 per la radice:

e rimangono i seguenti sottoinsiemi di esempi da elaborare:

$$S_1 = \{3[-], 7[+]\}, S_4 = \{1[+], 4[+], 5[-], 8[-], 9[-]\}$$

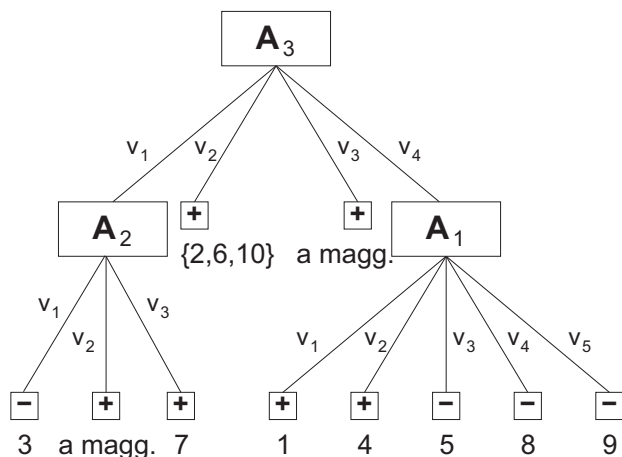
poiché S_2 origina una foglia con etichetta +, mentre S_3 , essendo vuoto, origina, a maggioranza, una foglia con etichetta +.

Si nota facilmente che S_1 è classificato correttamente da A_1 o A_2 . Supponendo di scegliere l'attributo che assume il numero minimo di valori distinti, cioè A_2 , si ottiene il seguente albero parziale:



dove l'etichetta della foglia nel mezzo è stata decisa risalendo alla radice, visto che gli esempi associati al nodo non mostrano il prevalere di una etichetta sull'altra.

Infine, S_4 è chiaramente classificato correttamente da A_1 :



Esercizio 2

b) dato il seguente insieme di apprendimento

<i>Esempio</i>	<i>Target</i>	<i>Out</i>	<i>Temp</i>	<i>Hum</i>	<i>Wind</i>
1	-	S	H	H	W
2	-	S	H	H	S
3	+	O	-	H	W
4	+	R	M	H	W
5	+	-	C	N	-
6	-	R	C	N	S
7	+	O	C	N	S
8	-	S	M	H	W
9	+	S	C	N	W
10	+	R	M	N	W
11	+	S	M	N	S
12	+	-	M	H	S
13	+	O	H	N	W
14	-	R	M	-	S

dove “-” indica un dato mancante. Mostrare come ID3 (con $Gain(S,A)$) costruisce l’albero di decisione corrispondente. Per ogni attributo calcolare $Gain(S,A)$ e giustificare la scelta dell’attributo utilizzato ad ogni nodo interno. Per il trattamento dei dati mancanti utilizzare l’approccio del valore più frequente.

Risposta:

Di seguito viene riportato il numero di occorrenza dei valori dei vari attributi:

<i>Out</i>	<i>Temp</i>	<i>Hum</i>	<i>Wind</i>
#“S” = 5	#“H” = 3	#“H” = 6	#“W” = 7
#“O” = 3	#“M” = 6	#“N” = 7	#“S” = 6
#“R” = 4	#“C” = 4		

Scegliendo il valore con il maggiore numero di occorrenze per ogni attributo, l'insieme di apprendimento diventa:

<i>Esempio</i>	<i>Target</i>	<i>Out</i>	<i>Temp</i>	<i>Hum</i>	<i>Wind</i>
1'	-	S	H	H	W
2'	-	S	H	H	S
3'	+	O	M	H	W
4'	+	R	M	H	W
5'	+	S	C	N	W
6'	-	R	C	N	S
7'	+	O	C	N	S
8'	-	S	M	H	W
9'	+	S	C	N	W
10'	+	R	M	N	W
11'	+	S	M	N	S
12'	+	S	M	H	S
13'	+	O	H	N	W
14'	-	R	M	N	S

Entropia: 0.940286

Attributo *Out*

valore S:[+,-] 4 3, valore O:[+,-] 3 0, valore R:[+,-] 2 2

Gain: 0.161958 ← **valore massimo**

Attributo *Temp*

valore H:[+,-] 1 2, valore M:[+,-] 5 2, valore C:[+,-] 3 1

Gain: 0.080154

Attributo *Hum*

valore H:[+,-] 3 3, valore N:[+,-] 6 2

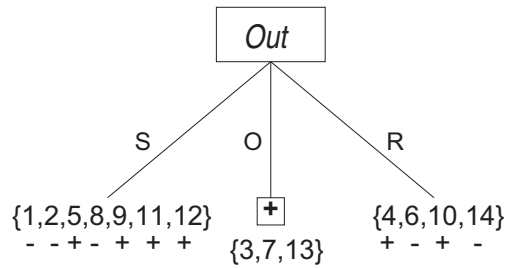
Gain: 0.048127

Attributo *Wind*

valore W:[+,-] 6 2, valore S:[+,-] 3 3

Gain: 0.048127

Quindi si sceglie *Out* come attributo per la radice:



Consideriamo adesso $S_S = \{1[-],2[-],5[+],8[-],9[+],11[+],12[+]\}$:

<i>Esempio</i>	<i>Target</i>	<i>Out</i>	<i>Temp</i>	<i>Hum</i>	<i>Wind</i>
1	-	S	H	H	W
2	-	S	H	H	S
5	+	-	C	N	-
8	-	S	M	H	W
9	+	S	C	N	W
11	+	S	M	N	S
12	+	-	M	H	S

Ricordando che non possiamo più utilizzare l'attributo *Out*, l'unico attributo che presenta un valore mancante è *Wind* (esempio 5). Quindi calcoliamo il numero di occorrenze per i valori di *Wind* relativamente ad S_S :

<i>Wind</i>
"W" = 3
"S" = 3

Poiché si presenta un caso di parità, scegliamo il valore più frequente a livello superiore (radice), cioè "W":

Entropia: 0.985228

Attributo *Temp*

valore H:[+,-] 0 2, valore M:[+,-] 2 1, valore C:[+,-] 2 0

Gain: 0.591673 ← **valore massimo**

Attributo *Hum*

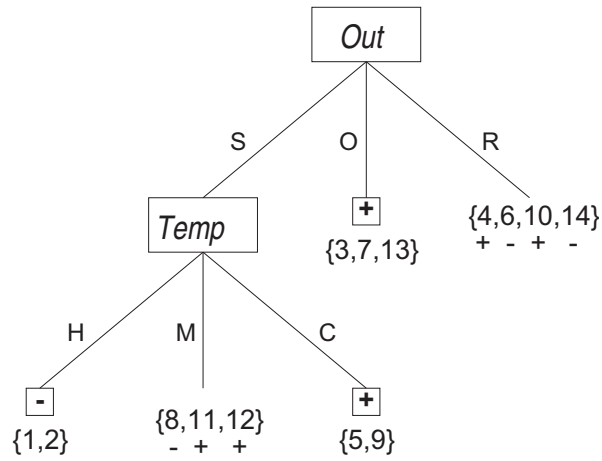
valore H:[+,-] 1 3, valore N:[+,-] 3 0

Gain: 0.521641

Attributo *Wind*

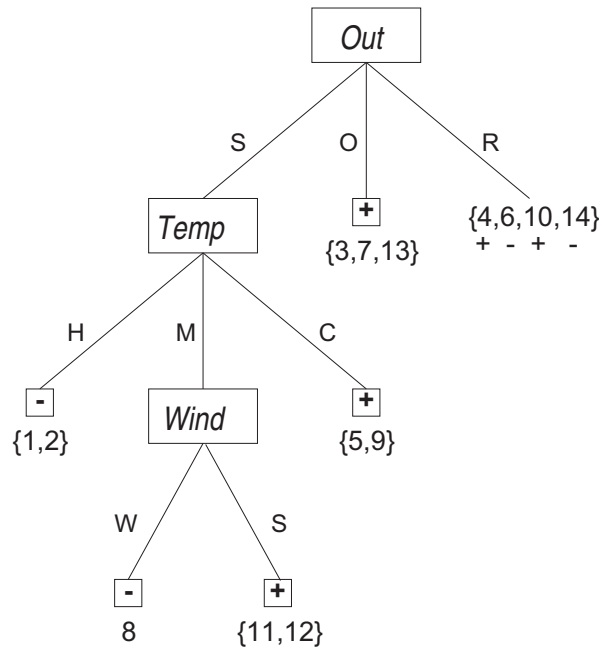
valore W:[+,-] 2 2, valore S:[+,-] 2 1

Gain: 0.020244



Quindi si sceglie *Temp* come attributo per la radice:

ed infine $S_M = \{8[-],11[+],12[+]\}$ è classificato correttamente (e solo) dall'attributo *Wind*:



Infine consideriamo $S_R = \{4[+],6[-],10[+],14[-]\}$:

<i>Esempio</i>	<i>Target</i>	<i>Out</i>	<i>Temp</i>	<i>Hum</i>	<i>Wind</i>
4	+	<i>R</i>	<i>M</i>	<i>H</i>	<i>W</i>
6	-	<i>R</i>	<i>C</i>	<i>N</i>	<i>S</i>
10	+	<i>R</i>	<i>M</i>	<i>N</i>	<i>W</i>
14	-	<i>R</i>	<i>M</i>	-	<i>S</i>

L'unico attributo che presenta un valore mancante è *Hum* (esempio 14). Quindi calcoliamo il numero di occorrenze per i valori di *Hum* relativamente ad S_R :

<i>Hum</i>
#“H” = 1
#“N” = 2

e scegliamo il valore più frequente, cioè “N”:

<i>Esempio</i>	<i>Target</i>	<i>Out</i>	<i>Temp</i>	<i>Hum</i>	<i>Wind</i>
4	+	R	M	H	W
6	-	R	C	N	S
10	+	R	M	N	W
14	-	R	M	N	S

Si vede subito che *Wind* classifica correttamente S_R :

