Soluzione Esercizi del Corso di Sistemi per l'Elaborazione dell'Informazione

Terza Parte

Esercizio 1

b) dato il seguente insieme di apprendimento

Esempio	*Target*	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_7
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 GainRatio(S,A)) costruisce l'albero di decisione corrispondente. Per ogni attributo calcolare $GainRatio(S, A)$ e giustificare la scelta dell'attributo utilizzato ad ogni nodo interno.

Risposta:

Entropia: 1.000000

Attributo A_1

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

valore 5:[+,-] 1 1

Gain: 0.400000, Split: 2.321928, Ratio: 0.172271

Attributo A_2

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

Gain: 0.400000, Split: 1.570951, Ratio: 0.254623

Attributo A_3

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

Gain: 0.039036, Split: 1.485475, Ratio: 0.026278

Attributo A_4

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

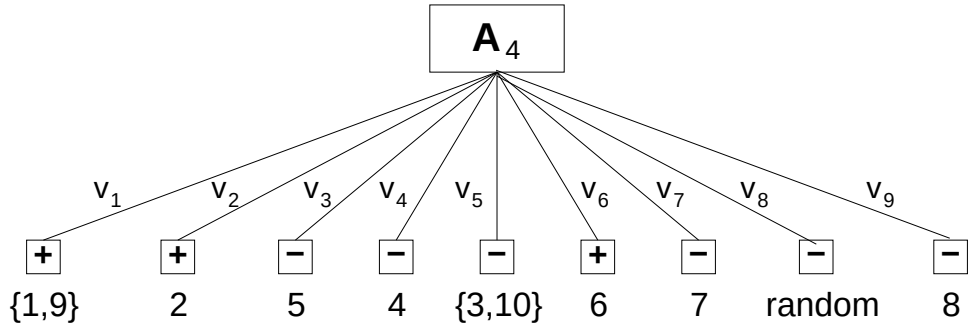
valore 5:[+,-] 0 2, valore 6:[+,-] 1 0, valore 7:[+,-] 1 0, valore 8:[+,-] 0 0, valore 9:[+,-] 0 1

Gain: 1.000000, Split: 2.921928, Ratio: 0.342240

Quindi si sceglie A_4

$S_1 = \{1[+],9[+]\}$, $S_2 = \{2[+]\}$, $S_3 = \{5[-]\}$, $S_4 = \{4[-]\}$, $S_5 = \{3[-],10[-]\}$, $S_6 = \{6[+]\}$, $S_7 = \{7[-]\}$, $S_8 =$ vuoto, $S_9 = \{8[-]\}$

e l'albero finale è:



c) Mostrare cosa succede se si aggiunge l'esempio *(4,3,2,1)* con target **-**

Risposta:

Entropia: 0.994030

Attributo A_1

valore 1:[+,-] 2 0, valore 2:[+,-] 0 2, valore 3:[+,-] 1 1, valore 4:[+,-] 1 2, valore 5:[+,-] 1 1

Gain: 0.379950 Split: 2.299896 Ratio: 0.165203

Attributo A_2

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

Gain: 0.335384 Split: 1.572624 Ratio: 0.213264

Attributo A_3

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

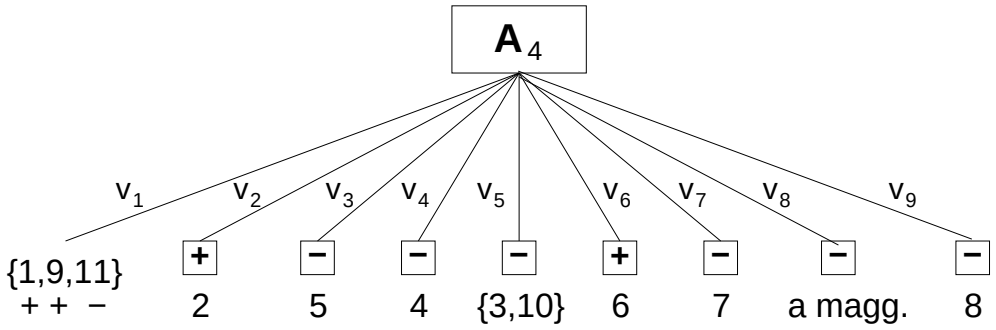
Gain: 0.007234 Split: 1.494919 Ratio: 0.004839

Attributo A_4

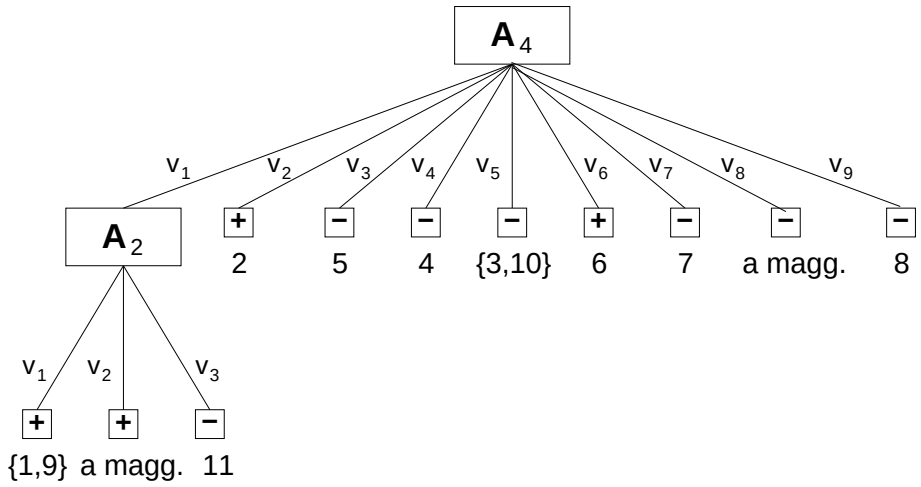
valore 1:[+,-] 2 1, valore 2:[+,-] 1 0, valore 3:[+,-] 0 1, valore 4:[+,-] 0 1, valore 5:[+,-] 0 2, valore 6:[+,-] 1 0, valore 7:[+,-] 1 0, valore 8:[+,-] 0 0, valore 9:[+,-] 0 1

Gain: 0.743586 Split: 2.845351 Ratio: 0.261334

Quindi si sceglie sempre A_4, però l'albero non è completo:



In particolare si deve ancora elaborare l'insieme $S_1 = \{1[+],9[+],11[-]\}$, che risulta essere classificato correttamente da A_1, o A_2, o A_3. Supponendo di scegliere l'attributo che assume il numero minimo di valori distinti, cioé A_2, otteniamo l'albero finale:



Esercizio 2

b) dato il seguente insieme di apprendimento

Esempio	*Target*	A_1 (5 *val.*)	A_2 (3 *val.*)	A_3 (4 *val.*)	A_4 (4 *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_1
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_2
7	+	v_4	v_3	v_1	v_1
8	−	v_4	v_2	v_4	v_3
9	+	v_5	v_1	v_4	v_1
10	+	v_5	v_3	v_2	v_2

mostrare come ID3 (con GainRatio(S,A)) costruisce l'albero di decisione corrispondente. Per ogni attributo calcolare $GainRatio(S, A)$ e giustificare la scelta dell'attributo utilizzato ad ogni nodo interno.

Risposta:

Entropia: 0.970951

Attributo A_1

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

valore 5:[+,-] 2 0

Gain: 0.570951 Split: 2.321928 Ratio: 0.245895

Attributo A_2

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

Gain: 0.646439 Split: 1.570951 Ratio: 0.411496

Attributo A_3

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

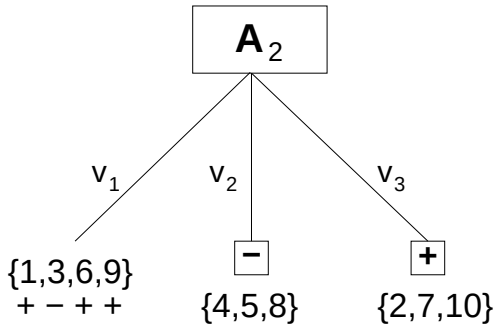
Gain: 0.285475 Split: 1.485475 Ratio: 0.192178

Attributo A_4

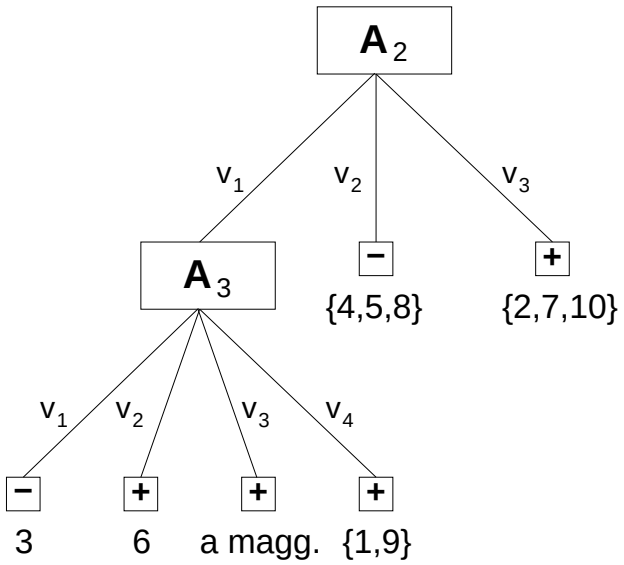
valore 1:[+,-] 3 1, valore 2:[+,-] 3 0, valore 3:[+,-] 0 2, valore 4:[+,-] 0 1

Gain: 0.646439 Split: 1.846439 Ratio: 0.350101

Quindi si sceglie A_2 per la radice:
$S_1 = \{1[+],3[-],6[+],9[+]\}$, $S_2 = \{4[-],5[-],8[-]\}$, $S_3 = \{2[+],7[+],10[+]\}$
e l'albero parziale è:



In particolare si deve ancora elaborare l'insieme $S_1 = \{1[+],3[-],6[+],9[+]\}$, che risulta essere classificato correttamente da A_1, o A_3. Supponendo di scegliere l'attributo che assume il numero minimo di valori distinti, cioé A_3, otteniamo l'albero finale:



c) Mostrare cosa succede se si aggiunge l'esempio *(4,2,2,1)* con target **-**

Risposta:

Entropia: 0.994030

Attributo A_1

valore 1:[+,-] 2 0, valore 2:[+,-] 0 2, valore 3:[+,-] 1 1, valore 4:[+,-] 1 2,
valore 5:[+,-] 2 0

Gain: 0.561768 Split: 2.299896 Ratio: 0.244258

Attributo A_2

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

Gain: 0.699020 Split: 1.572624 Ratio: 0.444493

Attributo A_3

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

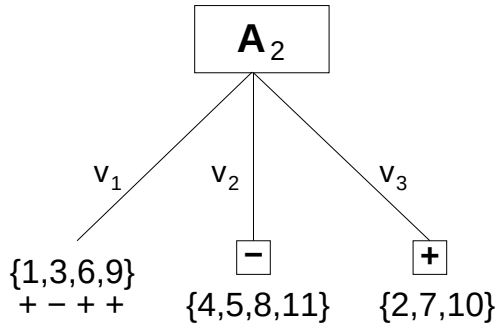
Gain: 0.075861 Split: 1.494919 Ratio: 0.050746

Attributo A_4

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

Gain: 0.552689 Split: 1.789929 Ratio: 0.308777

Quindi si sceglie di nuovo A_2 per la radice:
$S_1 = \{1[+],3[-],6[+],9[+]\}$, $S_2 = \{4[-],5[-],8[-],11[-]\}$, $S_3 = \{2[+],7[+],10[+]\}$
e l'albero parziale è:



e poiché non cambia l'insieme S_1, si ottiene di nuovo l'albero finale precedente.

Esercizio 3

b) dato il seguente insieme di apprendimento

Esempio	*Target*	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 GainRatio(S,A)) costruisce l'albero di decisione corrispondente. Per ogni attributo calcolare $GainRatio(S, A)$ e giustificare la scelta dell'attributo utilizzato ad ogni nodo interno.

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

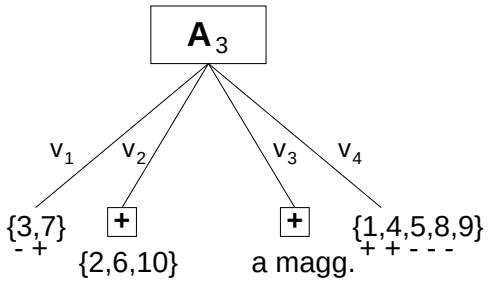
valore 1:[+,-] 1 1, valore 2:[+,-] 1 0, valore 3:[+,-] 0 1, valore 4:[+,-] 1 0,

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:

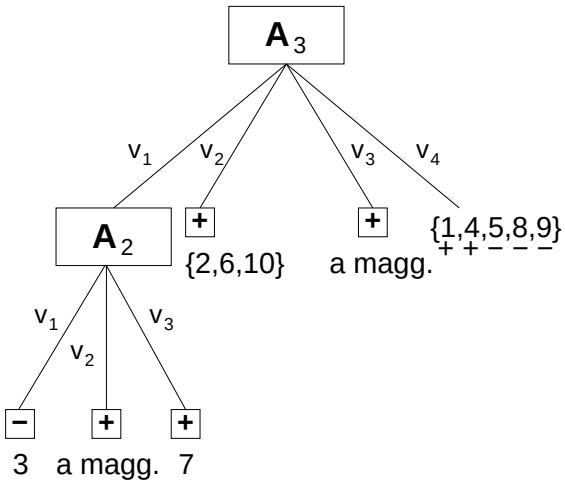


e rimangono i seguenti sottoinsiemi di esempi da elaborare:

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

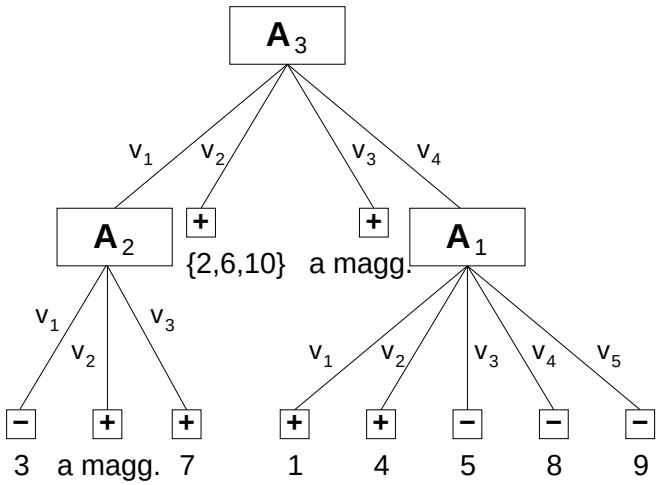
poiché S_2 origina una foglia con etichetta +, mentre S_3, essendo vuoto, origina, a maggioranza, una foglia con etichetta +.

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.

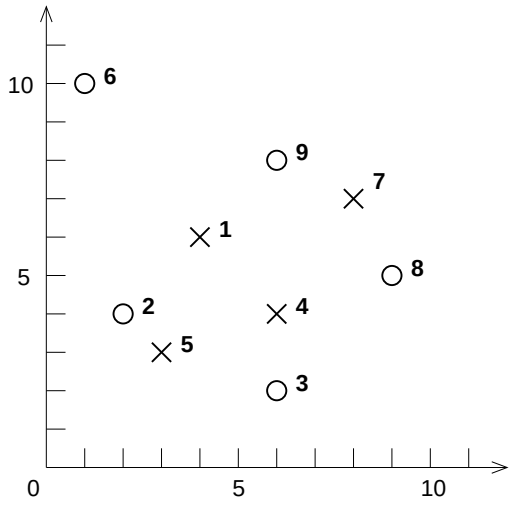
Infine, S_4 è chiaramente classificato correttamente da A_1:



Esercizio 4

b) sia dato uno spazio delle istanze costituito da punti sul piano a coordinate intere nell'intervallo [1,10], e uno spazio delle ipotesi dato da rettangoli con lati paralleli agli assi e vertici a coordinate intere (sempre nell'intervallo [1,10]), come visto a lezione.

Dato il seguente insieme di esempi

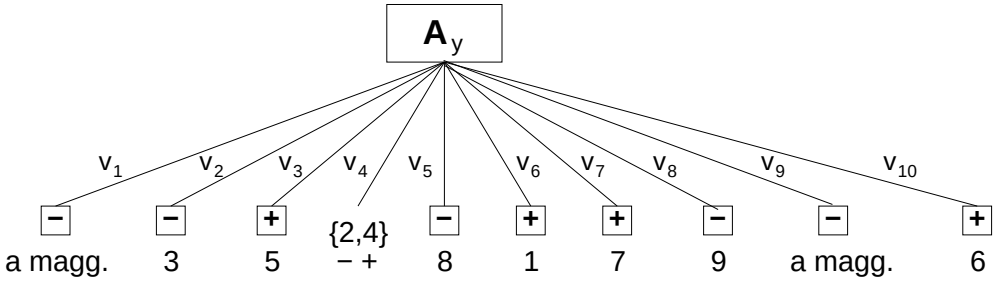


Calcolare i valori di *Ratio(S,A)* e *GainRatio(S,A)* per i due attributi (cioè la coordinata x e la coordinata y). Quindi usare *Ratio(S,A)* per mostrare come ID3 costruisce l'albero di decisione.

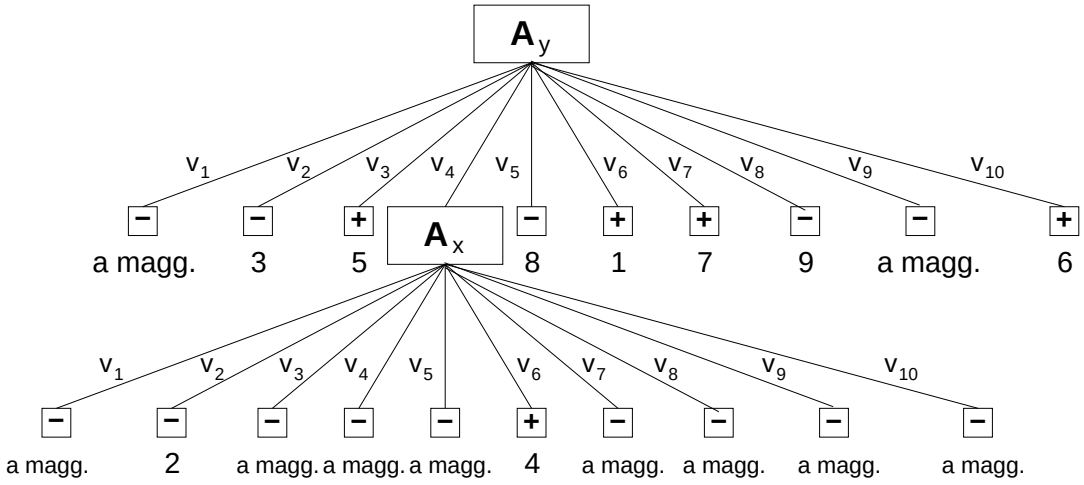
Risposta:

Entropia: 0.991076
Attributo A_x
valore 1:[+,-] 0 1, valore 2:[+,-] 0 1, valore 3:[+,-] 1 0, valore 4:[+,-] 1 0,
valore 5:[+,-] 0 0, valore 6:[+,-] 1 2, valore 7:[+,-] 0 0, valore 8:[+,-] 1 0,
valore 9:[+,-] 0 1, valore 10:[+,-] 0 0
Gain: 0.684977 Split: 2.641604 Ratio: 0.259304

Attributo A_y
valore 1:[+,-] 0 0, valore 2:[+,-] 0 1, valore 3:[+,-] 1 0, valore 4:[+,-] 1 1,
valore 5:[+,-] 0 1, valore 6:[+,-] 1 0, valore 7:[+,-] 1 0, valore 8:[+,-] 0 1,
valore 9:[+,-] 0 0, valore 10:[+,-] 0 1
Gain: 0.768854 Split: 2.947703 Ratio: 0.260832



Quindi si sceglie A_y per la radice:
e $S_4 = \{2[\text{-}],4[+]\}$ è infine classificato correttamente da A_x:



dove le decisioni a maggioranza sono prese risalendo alla radice a causa della parità riscontrata al livello del nodo.

Esercizio 5

b) dato il seguente insieme di apprendimento

Esempio	*Target*	A_1 (5 *val.*)	A_2 (3 *val.*)	A_3 (4 *val.*)	A_4 (4 *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_1
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_2
7	+	v_4	v_3	v_1	v_1
8	−	v_4	v_2	v_4	v_3
9	+	v_5	v_1	v_4	v_1
10	+	v_5	v_3	v_2	v_2

mostrare come ID3 (con GainRatio(S,A)) costruisce l'albero di decisione corrispondente. Per ogni attributo calcolare $GainRatio(S, A)$ e giustificare la scelta dell'attributo utilizzato ad ogni nodo interno.

Risposta:

Entropia: 0.970951

Attributo A_1

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

valore 5:[+,-] 2 0

Gain: 0.570951 Split: 2.321928 Ratio: 0.245895

Attributo A_2

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

Gain: 0.646439 Split: 1.570951 Ratio: 0.411496

Attributo A_3

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

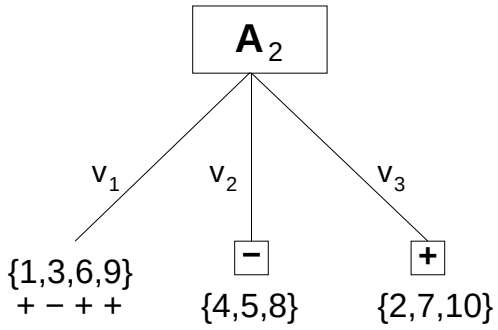
Gain: 0.285475 Split: 1.485475 Ratio: 0.192178

Attributo A_4

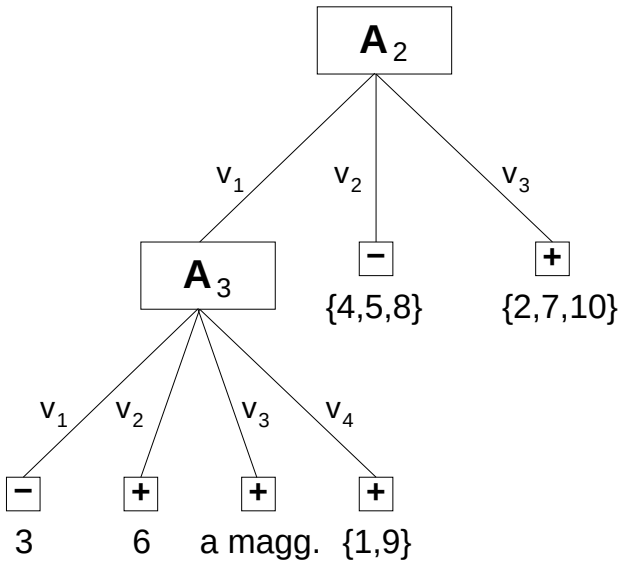
valore 1:[+,-] 3 1, valore 2:[+,-] 3 0, valore 3:[+,-] 0 2, valore 4:[+,-] 0 1

Gain: 0.646439 Split: 1.846439 Ratio: 0.350101

Quindi si sceglie A_2 per la radice:
$S_1 = \{1[+],3[-],6[+],9[+]\}$, $S_2 = \{4[-],5[-],8[-]\}$, $S_3 = \{2[+],7[+],10[+]\}$
e l'albero parziale è:



In particolare si deve ancora elaborare l'insieme $S_1 = \{1[+],3[-],6[+],9[+]\}$, che risulta essere classificato correttamente da A_1, o A_3. Supponendo di scegliere l'attributo che assume il numero minimo di valori distinti, cioé A_3, otteniamo l'albero finale:



c) Mostrare cosa succede se si aggiunge l'esempio *(4,2,2,1)* con target **-**

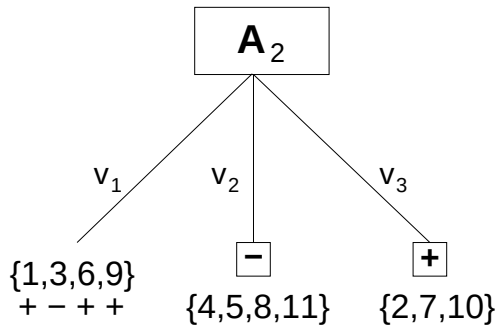
Risposta:
Entropia: 0.994030
Attributo A_1
valore 1:[+,-] 2 0, valore 2:[+,-] 0 2, valore 3:[+,-] 1 1, valore 4:[+,-] 1 2,
valore 5:[+,-] 2 0
Gain: 0.561768 Split: 2.299896 Ratio: 0.244258

Attributo A_2
valore 1:[+,-] 3 1, valore 2:[+,-] 0 4, valore 3:[+,-] 3 0
Gain: 0.699020 Split: 1.572624 Ratio: 0.444493

Attributo A_3
valore 1:[+,-] 1 1, valore 2:[+,-] 3 1, valore 3:[+,-] 0 0, valore 4:[+,-] 2 3
Gain: 0.075861 Split: 1.494919 Ratio: 0.050746

Attributo A_4
valore 1:[+,-] 3 2, valore 2:[+,-] 3 0, valore 3:[+,-] 0 2, valore 4:[+,-] 0 1
Gain: 0.552689 Split: 1.789929 Ratio: 0.308777

Quindi si sceglie di nuovo A_2 per la radice:
$S_1 = \{1[+],3[-],6[+],9[+]\}$, $S_2 = \{4[-],5[-],8[-],11[-]\}$, $S_3 = \{2[+],7[+],10[+]\}$
e l'albero parziale è:



e poiché non cambia l'insieme S_1, si ottiene di nuovo l'albero finale precedente.

Esercizio 7

b) dato il seguente insieme di apprendimento

Esempio	*Target*	*Out*	*Temp*	*Hum*	*Wind*
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 con stesso target.

Risposta:

Di seguito viene riportato il numero di occorrenza dei valori dei vari attributi per esempi con stesso target dell'esempio che presenta il valore mancante:

Out (+)	*Temp* (+)	*Hum* (-)	*Wind* (+)
$\#\text{“}S'' = 2$	$\#\text{“}H'' = 1$	$\#\text{“}H'' = 3$	$\#\text{“}W'' = 5$
$\#\text{“}O'' = 3$	$\#\text{“}M'' = 4$	$\#\text{“}N'' = 1$	$\#\text{“}S'' = 3$
$\#\text{“}R'' = 2$	$\#\text{“}C'' = 3$		

Scegliendo il valore con il maggiore numero di occorrenze per ogni attributo,

l'insieme di apprendimento diventa:

$Esempio$	$Target$	Out	$Temp$	Hum	$Wind$
$1'$	$-$	S	H	H	W
$2'$	$-$	S	H	H	S
$3'$	$+$	O	M	H	W
$4'$	$+$	R	M	H	W
$5'$	$+$	O	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
$12'$	$+$	O	M	H	S
$13'$	$+$	O	H	N	W
$14'$	$-$	R	M	H	S

Entropia: 0.940286

Attributo Out

valore S:[+,-] 2 3, valore O:[+,-] 5 0, valore R:[+,-] 2 2

Gain: 0.307804 ← **valore massimo**

Attributo $Temp$

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

Gain: 0.080154

Attributo Hum

valore H:[+,-] 3 4, valore N:[+,-] 6 1

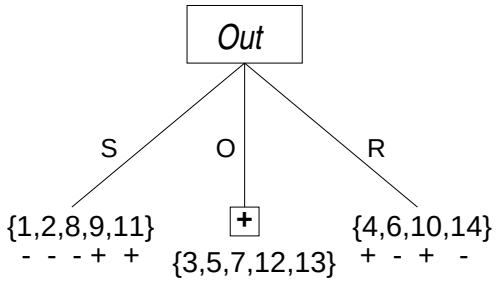
Gain: 0.151835

Attributo $Wind$

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

Gain: 0.048127

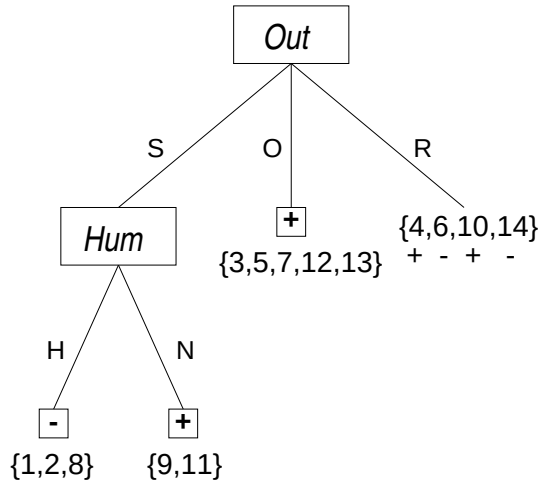
Quindi si sceglie Out come attributo per la radice:



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

$Esempio$	$Target$	Out	$Temp$	Hum	$Wind$
1	$-$	S	H	H	W
2	$-$	S	H	H	S
8	$-$	S	M	H	W
9	$+$	S	C	N	W
11	$+$	S	M	N	S

Si vede subito che (solo) Hum riesce a classificare correttamente S_S:



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

$Esempio$	$Target$	Out	$Temp$	Hum	$Wind$
4	$+$	R	M	H	W
6	$-$	R	C	N	S
10	$+$	R	M	N	W
14	$-$	R	M	_	S

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 esempi in S_R con target -:

$Wind$
$\#\text{“}H\text{”} = 0$
$\#\text{“}N\text{”} = 1$

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

$Esempio$	$Target$	Out	$Temp$	Hum	$Wind$
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: