Compitino del Corso di Apprendimento Automatico

Anno Accademico 1999/2000

22 Settembre 2000

Compitino del Corso di Apprendimento Automatico

Anno Accademico 99/00

22 Settembre 2000 - Prima Parte

Istruzioni

- Scrivere *Nome*, *Cognome* e *Matricola* su **ogni** foglio.
- Scrivere la risposta nello spazio bianco al di sotto della domanda; Non è possibile allegare fogli aggiuntivi, quindi cercate di essere chiari e non prolissi.
- In caso di errori indicate chiaramente quale parte della risposta deve essere considerata; annullate le parti non pertinenti.
- Assicurarsi che non manchi alcun foglio al momento della consegna.

Esercizio 2

Considerare il problema di apprendere il concetto *EnjoySport* visto a lezione, dove però si considerano i seguenti valori per gli attributi

$$\begin{aligned}
\text{Sky} &\rightarrow \{Sunny, Cloudy, Rainy\};\\
\text{AirTemp} &\rightarrow \{Warm, Cold\};\\
\text{Humidity} &\rightarrow \{Normal, High\};\\
\text{Wind} &\rightarrow \{Weak, Strong\};\\
\text{Water} &\rightarrow \{Warm, Cool\};\\
\text{Forecast} &\rightarrow \{Same, Change\};
\end{aligned}$$

a) definire un insieme minimo di esempi che, se presentato all'algoritmo *Candidate-Elimination*, produce il seguente Version Space:

$$\begin{array}{llllllll}
S: \{< & Sunny & ? & High & Weak & Cool & ? & >\} \\
G: \{< & Sunny & ? & ? & ? & ? & ? & >, \\
< & ? & ? & High & Weak & ? & ? & >\}
\end{array}$$

Risposta: Una possibile sequenza di esempi è la seguente:

$$\begin{array}{l}
1.\ (< S \quad W \quad H \quad W \quad C \quad C \quad >, \quad Yes) \\
2.\ (< S \quad C \quad H \quad W \quad C \quad S \quad >, \quad Yes) \\
3.\ (< C \quad C \quad N \quad W \quad C \quad S \quad >, \quad No) \\
4.\ (< C \quad C \quad H \quad S \quad C \quad S \quad >, \quad No)
\end{array}$$

b) dire quante ipotesi contiene il Version Space di sopra;

Risposta:

10

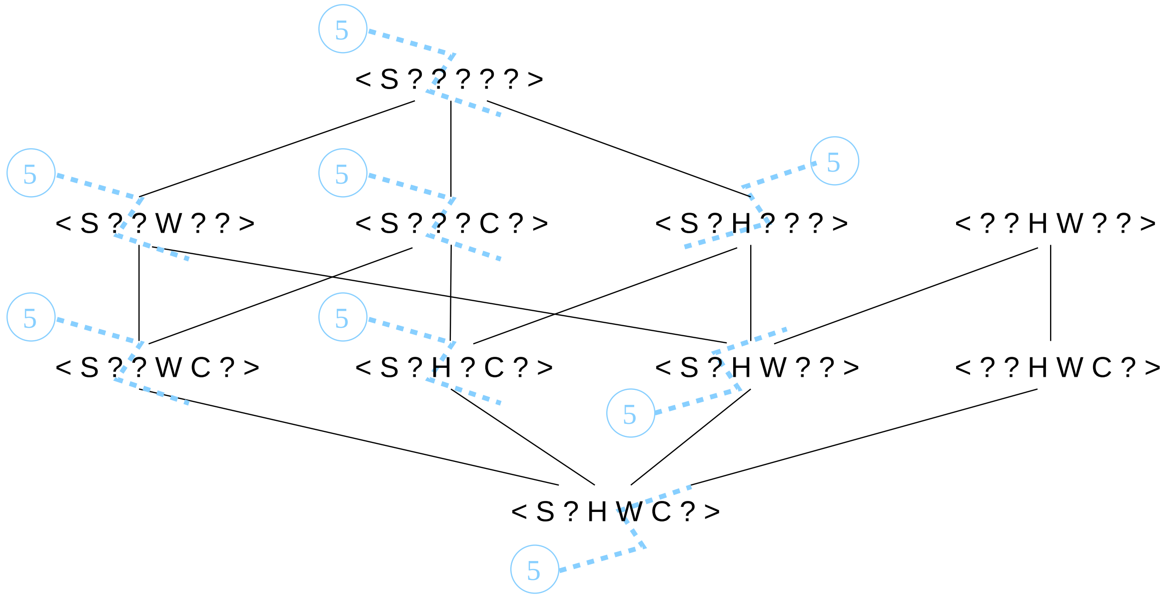
c) dare un insieme minimo di esempi addizionali per far convergere il Version Space di sopra al seguente Version Space

$$\begin{array}{llllllll}
S: \{< & ? & ? & High & Weak & Cool & ? & >\} \\
G: \{< & ? & ? & High & Weak & ? & ? & >\}
\end{array}$$

Risposta:

Basta un esempio:

$$5.\ (< C \quad C \quad H \quad W \quad C \quad S \quad >, \quad Yes)$$



Esercizio 3

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*

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 con valore mancante:

Out	$Temp$	$Wind$
$\#``S'' = 4$	$\#``H'' = 2$	$\#``W'' = 6$
$\#``O'' = 2$	$\#``M'' = 3$	$\#``S'' = 3$
$\#``R'' = 3$	$\#``C'' = 4$	

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	C	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

Entropia: 0.970951

Attributo Out

valore S:[+,-] 2 3, valore O:[+,-] 2 0, valore R:[+,-] 2 1

$\boxed{\text{Gain: } 0.209987}$

Attributo $Temp$

valore H:[+,-] 0 2, valore M:[+,-] 2 1, valore C:[+,-] 4 1

$\boxed{\text{Gain: } 0.334498}$ $\leftarrow$ **valore massimo**

Attributo Hum

valore H:[+,-] 2 3, valore N:[+,-] 4 1

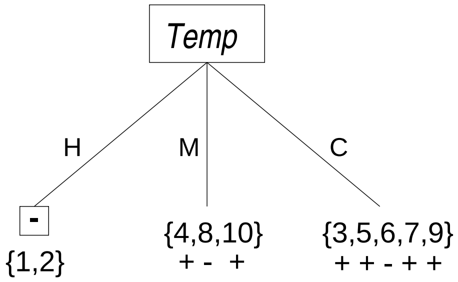
$\boxed{\text{Gain: } 0.124511}$

Attributo $Wind$

valore W:[+,-] 5 2, valore S:[+,-] 1 2

$\boxed{\text{Gain: } 0.091277}$

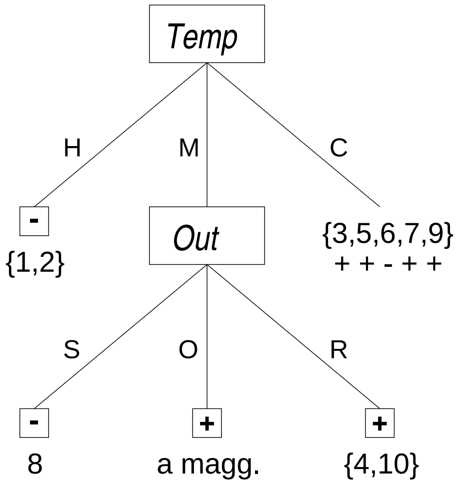
Quindi si sceglie $Temp$ come attributo per la radice:



Consideriamo adesso S_M = {4[+],8[-],10[+]}:

$Esempio$	$Target$	Out	$Temp$	Hum	$Wind$
4	$+$	R	M	H	W
8	$-$	S	M	H	W
10	$+$	R	M	N	W

Si vede subito che Out classifica correttamente S_M:



Consideriamo quindi S_C={3[+],5[+],6[-],7[+],9[+]}:

$Esempio$	$Target$	Out	$Temp$	Hum	$Wind$
3	$+$	O	_	H	W
5	$+$	_	C	N	_
6	$-$	R	C	N	S
7	$+$	O	C	N	S
9	$+$	S	C	N	W

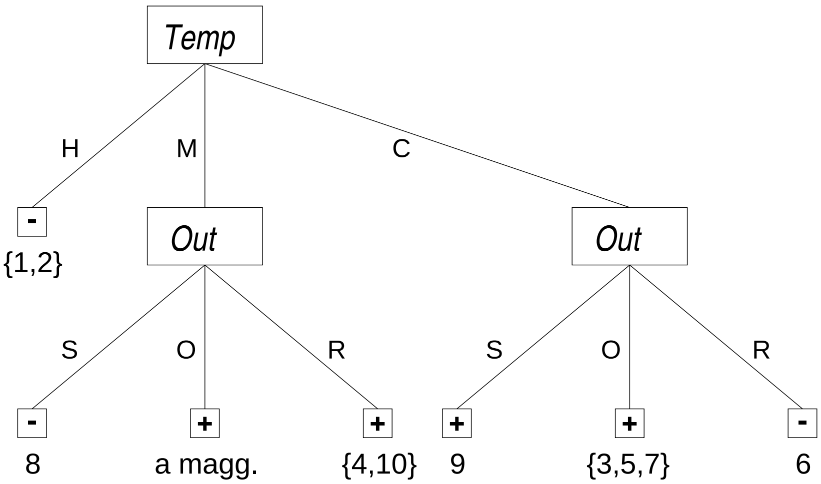
Di seguito viene riportato il numero di occorrenza dei valori dei vari attributi con valore mancante:

Out	*Wind*
$\#``S'' = 1$	$\#``W'' = 2$
$\#``O'' = 2$	$\#``S'' = 2$
$\#``R'' = 1$	

Per Out si sceglie quindi "O", mentre per $Wind$, essendoci parità, si sale di un livello (radice) e si sceglie "W":

$Esempio$	$Target$	Out	$Temp$	Hum	$Wind$
3	$+$	O	_	H	W
5	$+$	O	C	N	W
6	$-$	R	C	N	S
7	$+$	O	C	N	S
9	$+$	S	C	N	W

L'attributo Out è l'unico che classifica correttamente S_C:



c) Utilizzando le frequenze per i valori degli attributi calcolate durante la costruzione dell'albero per ogni nodo che lo richiedeva, mostrare come le seguenti istanze vengono classificati utilizzanzo la tecnica delle probabilità per il trattamento dei dati mancanti:

$Esempio$	Out	$Temp$	Hum	$Wind$
11	R	_	H	W
12	_	C	N	W

Risposta:

Le frequenze calcolate sono:

Radice	*Figlio destro*
$p_H = 2/9$	$p_S = 1/4$
$p_M = 3/9$	$p_O = 2/4$
$p_C = 4/9$	$p_R = 1/4$

Quindi per l'istanza 11:

$$- \quad \rightarrow \quad \underbrace{\frac{2}{9}}_{\text{foglia sin. radice}} + \underbrace{\frac{4}{9}}_{\text{foglia più a dest.}} = \frac{6}{9} \tag{1}$$

$$+ \quad \rightarrow \quad \underbrace{\frac{3}{9}}_{\text{foglia più a dest. figlio cent. radice}} \tag{2}$$

quindi viene classificata $-$. Per l'istanza 12:

$$+ \quad \rightarrow \quad \frac{1}{4} + \frac{2}{4} = \frac{3}{4} \tag{3}$$

$$- \quad \rightarrow \quad \frac{1}{4} \tag{4}$$

quindi viene classificata $+$.