Esempio di Compitino del Corso di Apprendimento Automatico, Parte I

Anno Accademico 99/00

Novembre, 1999

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 1

Definire in modo completo

a) il *Version Space*;

b) la *VC-dimension*;

c) l' *Optimal Mistake Bound*;

d) il Teorema sul Version Space ϵ-esausto, includendo la dimostrazione;

Esercizio 2

Considerare il problema di apprendere il concetto *EnjoySport* visto a lezione:

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

$$\begin{array}{lccccccc} S:\{< & ? & Warm & Normal & Strong & ? & Same & >\} \\ G:\{< & ? & ? & Normal & ? & ? & ? & >\} \end{array}$$

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

c) scegliere una delle ipotesi all'interno del Version Space come concetto target e dire di quante query c'è bisogno per apprenderlo usando la strategia ottima vista a lezione. Dare una possibile sequenza di tali query;

d) si supponga che a partire dal Version Space definito in *a)* si presentino nuovi esempi $(x, c(x))$ generati secondo una distribuzione sconosciuta D. Ci interessa sapere di quanti esempi abbiamo bisogno per assicurare che con probabilità 0.95 tutti i membri del Version Space risultante abbiano errore esatto (true error) al più 0.1 (rispetto al concetto target e D).

Esercizio 3

a) definire *Entropy(S)* e *Gain(S,A)*;

b) dato il seguente insieme di apprendimento

Esempio	*Target*	A_1 (5 *val.*)	A_2 (3 *val.*)	A_3 (4 *val.*)
1	$+$	v_1	v_1	v_4
2	$+$	v_1	v_3	v_2
3	$-$	v_2	v_1	v_1
4	$-$	v_2	v_2	v_4
5	$-$	v_3	v_2	v_4
6	$+$	v_3	v_1	v_2
7	$+$	v_4	v_3	v_1
8	$-$	v_4	v_2	v_4
9	$+$	v_5	v_1	v_4
10	$-$	v_5	v_3	v_2

mostrare come ID3 costruisce l'albero di decisione corrispondente. Per ogni attributo calcolare $Gain(S, A)$ e giustificare la scelta dell'attributo utilizzato ad ogni nodo interno. (Ricordare che $\log(\frac{a}{b}) = \log(a) - \log(b)$; usare $\log_2(3) = 1.585$ e $\log_2(5) = 2.32$; consiglio: usare il valore numerico solo per il calcolo finale, cioè alla fine delle possibili trasformazioni algebriche).

c) Se si utilizza $GainRatio(S, A)$ invece di $Gain(S, A)$ cambia l'albero di decisione ? Se si, mostrare come.

c) Supponete che l'esempio 9 sia sostituito dal seguente esempio

$$9' \quad + \quad v_5 \quad v_1 \quad \text{-}$$

dove il valore per l'ultimo attributo è mancante. Mostrate come ID3 costruirebbe l'albero avendo scelto una delle strategie viste a lezione.