Università di Pisa	A.A. 2007- 2008

Tecniche di Data Mining - Corsi di Laurea Specialistica in Informatica e Tecnologie Informatiche

Verifica 3 aprile 2008

Esercizio 1 - Classificazione / Alberi di decisione (7 punti)

Considerando C come attributo di classe ed A e B come variabili numeriche continue, costruire due alberi di decisione:
(a) Discretizzando A e B.
(b) Assumendo A e B come attributi numerici.
Motivare il metodo di discretizzazione scelto e discutere le differenze fra i due alberi generati.

A	B	C
3	1	X
4	2	X
4	1	X
3	2	X
12	1	Y
13	2	Y
13	3	Y
18	7	X
16	8	X
18	9	X
23	7	Y
23	8	Y
24	9	Y

Esercizio 2 – Classificazione / overfitting (7 punti)

Descrivere il problema dell'overfitting e illustrare le tecniche per alleviarne l'effetto.

Esercizio 3 - Clustering / Single link (7 punti)

Si trovi il dendogramma dei clusters utilizzando l'algoritmo gerarchico Single-Link nell'ipotesi in cui si utilizzi la distanza di Jaccard.

	A	*B*	*C*	*D*	*E*
1	1	0	1	1	0
2	1	1	0	1	1
3	1	0	1	1	0
4	0	1	0	1	0
5	1	0	1	0	1
6	0	1	1	1	0

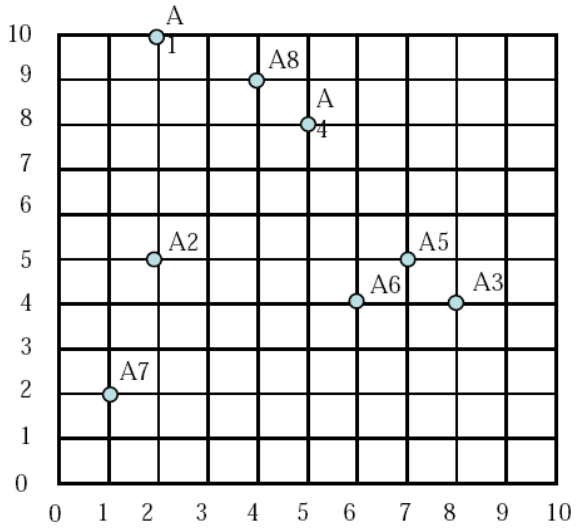
Esercizio 4 - Classificazione / Alberi di decisione (7 punti)

Si costruisca un albero di decisione di accuratezza 100% in riferimento al seguente training se:

Capacità (Mb)	Durata batterie	Prezzo ($)	Soddisfatto? (TARGET)
> 4	Lunga	<= 150	SI
<= 4	Lunga	> 150	SI
<= 4	Bassa	<= 150	NO
> 4	Media	<= 150	NO
<= 4	Media	<= 150	NO
> 4	Media	<= 150	NO
> 4	Lunga	> 150	SI
> 4	Lunga	> 150	SI
> 4	Bassa	> 150	SI
<= 4	Bassa	> 150	NO
> 4	Bassa	<= 150	SI
<= 4	Media	> 150	SI
> 4	Media	> 150	SI
<= 4	Bassa	> 150	NO
<= 4	Lunga	<= 150	SI

Esercizio 5 - Clustering (5 punti)

Si consideri il seguente dataset formato da 8 punti:



Dati i seguenti parametri: epsilon=2, min-points=2, determinare quali punti sono core-objects, quali border-objects, quali rumore, e quali cluster l'algoritmo DBSCAN individua.