Università di Pisa	A.A. 2006- 2007

Data Mining - Corso di Laurea Specialistica in
Informatica per l'economia e l'Azienda

Verifica 24 luglio 2007

Soluzioni

Esercizio 1 - Frequent / closed / maximal itemsets (**9** punti)

- Gli itemset frequenti:
  - C (50.0)
  - A (62.5)
  - E (62.5)
  - D (87.5)
  - C A (37.5)
  - C D (50.0)
  - A E (37.5)
  - A D (50.0)
  - E D (50.0)
  - C A D (37.5)

- Gli itemset frequenti *closed* (I è *closed* se nessun suo superinsieme ha lo stesso supporto):
  - A (62.5)
  - E (62.5)
  - D (87.5)
  - C D (50.0)
  - A E (37.5)
  - A D (50.0)
  - E D (50.0)
  - C A D (37.5)

- Gli itemset frequenti massimali:
  - A E (37.5)
  - E D (50.0)
  - C A D (37.5)

Esercizio 2 - Classificazione (**7** punti)

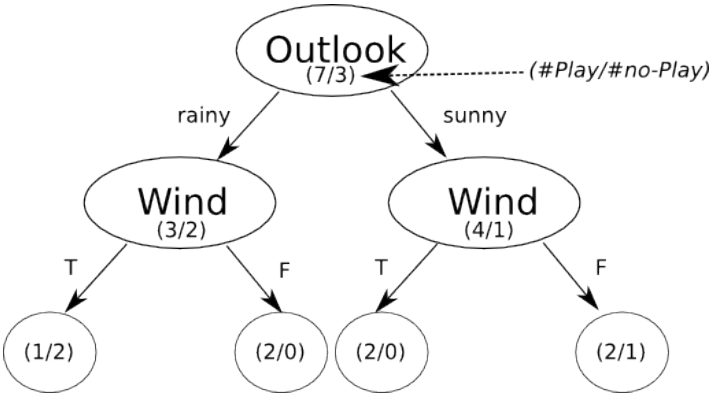
La **matrice dei costi** per C5.0 che induce l'algoritmo a tener conto delle relazioni date non è unica. Una possibile soluzione è la seguente:

1	2	3	4	<-- classe predetta
0	1	2	3	
1	0	1	2	
2	1	0	1	
3	2	1	0	

Esercizio 3 - Classificazione (**8** punti)

A. Si costruisca un albero di decisione usando il Gini Index per determinare gli attributi di splitting ad ogni nodo dell'albero. Terminare la costruzione solo quando ogni nodo ha un 100% di precisione.

NOTA: non è possibile costruire un albero con 100% di precisione. Gli alberi possibili sono due, a seconda di quale sia il primo attributo usato per lo split. Esempio iniziando da Outlook:

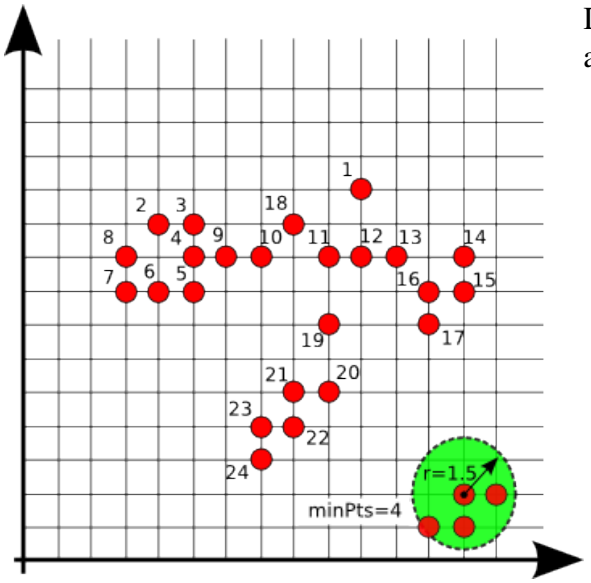


B. Calcolare la precisione dell'albero ottenuto sul test set:

#	Outlook	Wind	Play	Predizione -> confronto
1	sunny	FALSE	yes	yes ->OK
2	rainy	TRUE	no	no ->OK
3	sunny	TRUE	yes	yes ->OK
4	sunny	FALSE	no	yes ->**NO**
5	rainy	FALSE	yes	yes ->OK
6	rainy	FALSE	yes	yes ->OK

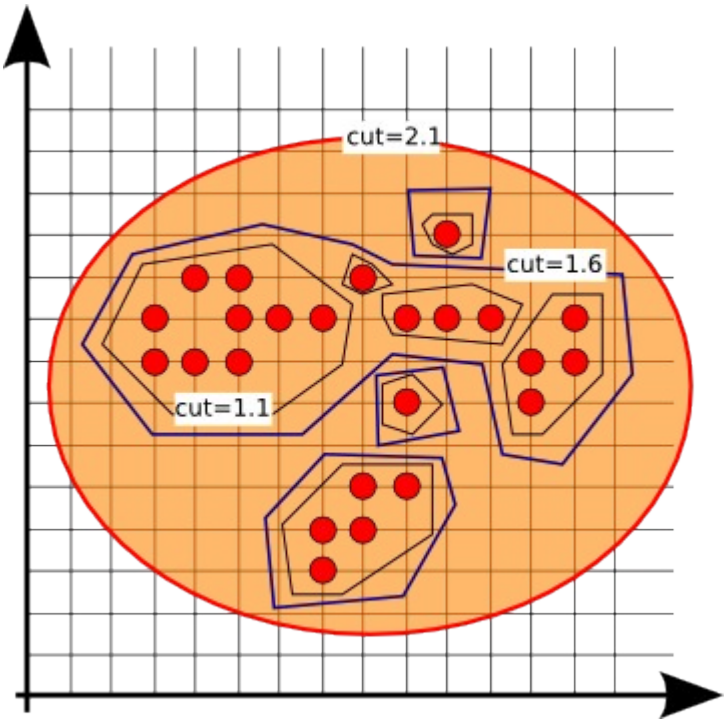
1 errore su 6 => precisione dell'83%

Esercizio 4 - Clustering (**9** punti)

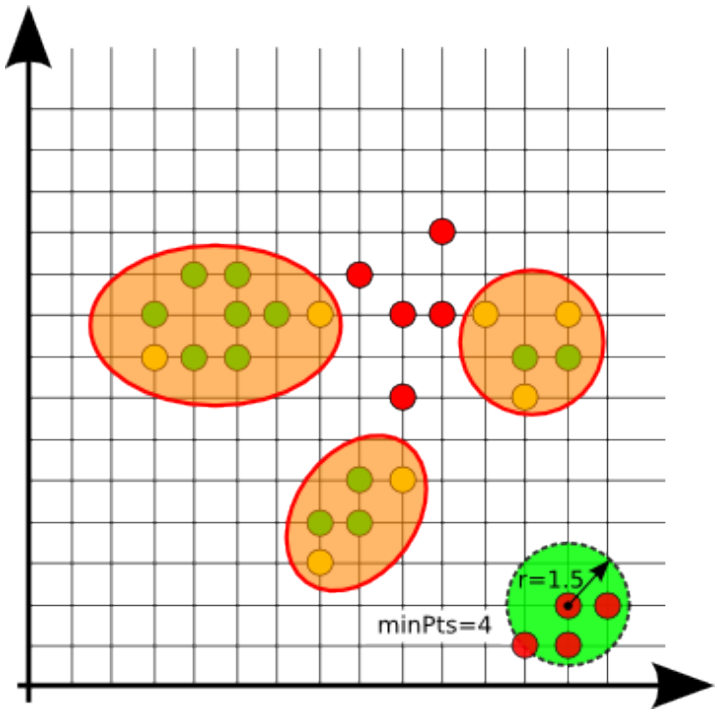


Determinare quali cluster vengono trovati dai seguenti algoritmi:

- Gerarchico Agglomerativo Single-Link (=Min-Link), *tagliando* il dendogramma in corrispondenza di una distanza pari a *cut = 1.1.* Suggerimento: ciò equivale a trovare le componenti connesse del grafo ottenuto connettendo le coppie di punti aventi distanza <= 1.1.
- Idem, con *cut = 1.6*
- Idem, con *cut = 2.1*
- DBSCAN, con *epsilon=1.5* e *minPts=4* (incluso il punto al centro dell'intorno)



Single Linkage con 3 soglie



DBSCAN
(verde=core, giallo=border,rosso=rumore)