

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

Verifica 24 luglio 2007

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

---

Dato il seguente dataset di transazioni:

1. A B C D E
2. A C D E
3. C D
4. A E
5. A C D F
6. A D
7. B D E
8. D E F

determinare:

- Gli itemset frequenti
- Gli itemset frequenti *closed* (I è *closed* se nessun suo superinsieme ha lo stesso supporto)
- Gli itemset frequenti massimali

utilizzando un *supporto minimo* = 30%.Esercizio 2 - Classificazione (7 punti)

---

Sia dato un problema di classificazione in cui la variabile target assume 4 valori diversi: {1,2,3,4}. L'algoritmo C5.0, di norma, non tiene conto del fatto le classi sono confrontabili numericamente, ovvero che 2 è più simile ad 1 che a 4, che 4 è più simile a 3 che a 1, ecc. In pratica, per l'algoritmo classificare come "1" un oggetto in realtà di classe "4" *costa* quanto classificarlo come "3" (mentre la seconda è chiaramente da preferire).

Definire una **matrice dei costi** per C5.0 che induca l'algoritmo a tener conto di queste relazioni.Esercizio 3 - Classificazione (8 punti)

---

A. Si costruisca un albero di decisione in riferimento al seguente training set, 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:

## # Outlook Wind Play

- 1 rainy FALSE yes
- 2 rainy TRUE no
- 3 rainy TRUE yes
- 4 sunny FALSE no

- 5 sunny FALSE yes
- 6 rainy FALSE yes
- 7 sunny TRUE yes
- 8 sunny TRUE yes
- 9 sunny FALSE yes
- 10 rainy TRUE no

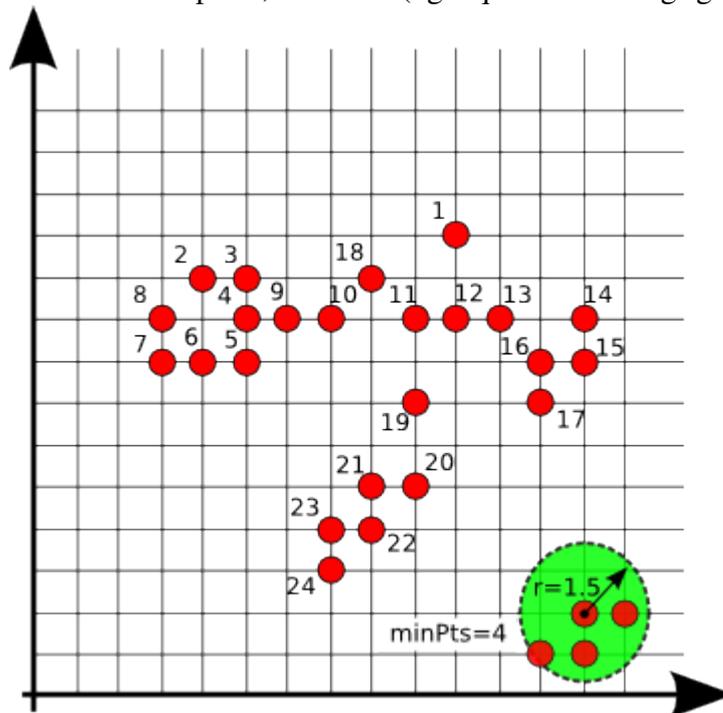
B. Calcolare la precisione dell'albero ottenuto sul seguente test set:

# Outlook Wind Play

- 1 sunny FALSE yes
- 2 rainy TRUE no
- 3 sunny TRUE yes
- 4 sunny FALSE no
- 5 rainy FALSE yes
- 6 rainy FALSE yes

Esercizio 5 - Clustering (9 punti)

Si consideri il seguente dataset di 24 punti, da 1 a 24 (ogni quadrato della griglia ha dimensione 1x1):



Determinare quali cluster vengono trovati dai seguenti algoritmi:

- Gerarchico Agglomerativo Single-Link (=Min-Link), tagliando il dendrogramma 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  $\leq 1.1$ .
- 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)