Università di Pisa	A.A. 2006- 2007

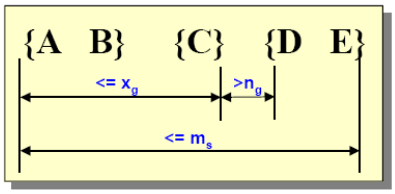
Data Mining - Corso di Laurea Specialistica in Informatica per l'economia e l'Azienda

Verifica 5 giugno 2007

Esercizio 1 - Sequential Patterns (8 punti)

1) Per ognuna delle sequenze w_i qua sotto, determinare se sono una sottosequenza oppure no della sequenza <{1,2,3}{2,4}{2,4,5}{3,5}{6}>.
   - w_1 = <{1}{2}{3}>
   - w_2 = <{1,2,3,4}{5,6}>
   - w_3 = <{2,4}{2,4}{6}>
   - w_4 = <{1}{2,4}{6}>
   - w_5 = <{1,2}{3,4}{5,6}>
2) Ripetere l'esercizio del punto 1) con il vincolo mingap = 0
3) Ripetere l'esercizio del punto 1) con il vincolo maxgap = 3
4) Ripetere l'esercizio del punto 1) con il vincolo maxspan = 5
5) Ripetere l'esercizio del punto 1) con i 3 vincoli (punti 2),3) e 4)) tutti insieme.

Si riassume graficamente il significato di min-gap, max-gap, max-span:



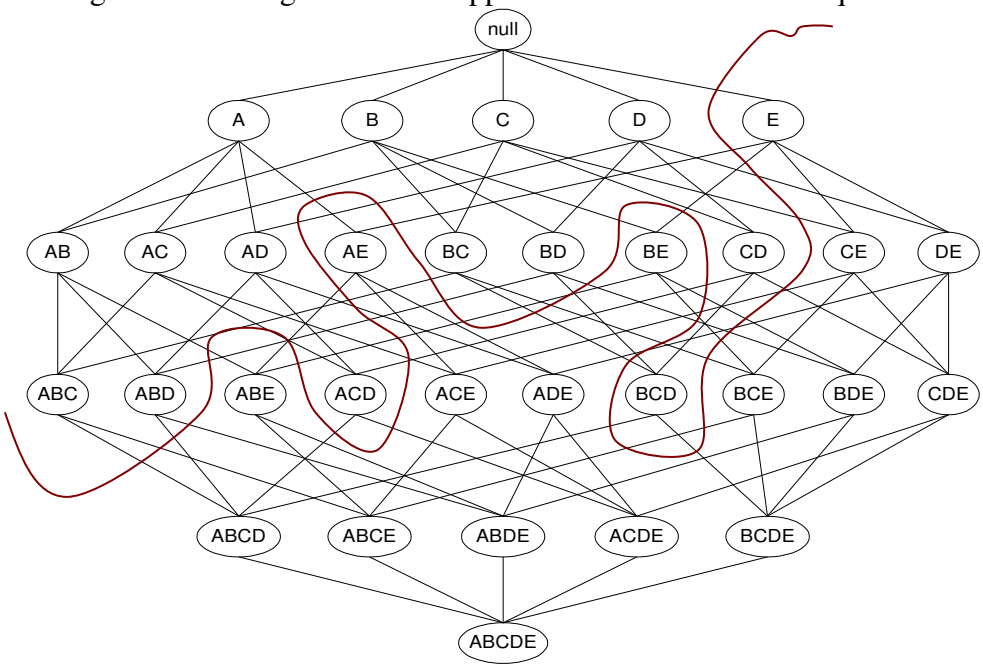
x_g: max-gap

n_g: min-gap

m_s: maximum span

Esercizio 2 – Frequent Itemsets (3 punti)

E' dato il lattice degli itemset in figura in cui è rappresentato il bordo della frequenza.



- Secondo l'algoritmo Apriori, quali sono gli itemsets **candidati** di dimensione 2? E quali di questi sono **frequenti**?
- Secondo l'algoritmo Apriori, quali sono gli itemsets **candidati** di dimensione 4? E quali di questi sono **frequenti**?

Esercizio 3 - Classificazione Aspetti Pratici (5 punti)

E' dato un problema di classificazione binaria in cui la classe di maggioranza rappresenta il 95% del training set. Se costruiamo un albero di decisione su tale training set, cosa possiamo aspettarci che succeda? Quali sono le possibili soluzioni a tale problema?

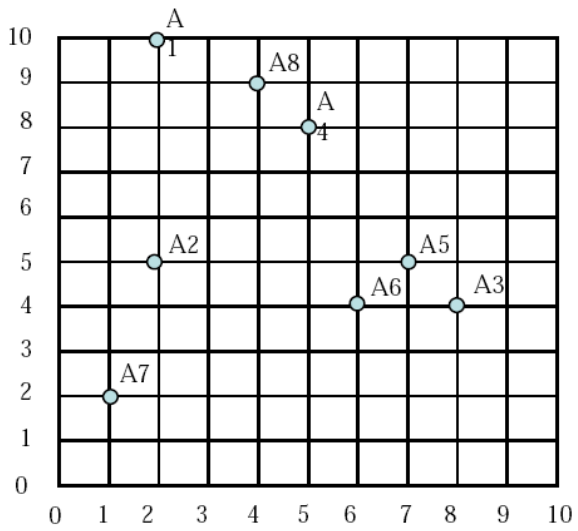
Esercizio 4 - Classificazione (7 punti)

Si costruisca un albero di decisione in riferimento al seguente training set, indicandone l'accuratezza (in riferimento al training set):

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

Esercizio 5 - Clustering (8 punti)

Si consideri il seguente dataset formato da 8 punti:



1. Determinare i cluster trovati da k-means partendo dai centri A1, A4 e A7, mostrando anche i cluster ottenuti nei passi intermedi.
2. Determinare il dendogramma ottenuto con un algoritmo gerarchico agglomerativo max-link (=la distanza tra due cluster è pari alla massima delle distanza tra coppie di punti). *Tagliando* il dendogramma per ottenere 3 cluster, confrontare i risultati ottenuti al punto precedente .

Esercizio 6 - Data preparation (4 punti)

1. Nel preparare i dati su cui applicare un algoritmo di clustering basato su distanza euclidea, di solito è sconsigliato avere attributi ridondanti o espressi in unità di misura diverse. Perchè?
2. Nel seguente database di esempio, quali problemi di questo genere si presentano, e come potremmo risolverli, in preparazione del passo di clustering?

Età	**Altezza**	**Reddito netto**	**Reddito lordo**
30	175	21.000	21.100
45	180	32.500	32.700
34	183	28.300	28.350
56	179	41.200	41.300