Università di Pisa	A.A. 2006- 2007

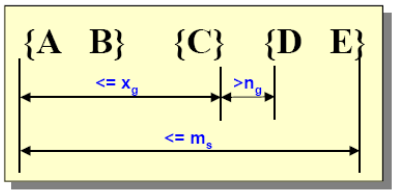
Data Mining - Corso di Laurea Specialistica in Informatica per l'economia e l'Azienda

Verifica 26 giugno 2007

Esercizio 1 - Sequential Patterns (7 punti)

1) Per ognuna delle sequenze w_i qua sotto, determinare se sono una sottosequenza oppure no della sequenza <{1,2,3,7}{2,4}{2,4,5,8}{3,5,7}{6}>.
   - w_1 = <{1,2,3,4}{5,6}>
   - 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}>
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 – Regole associative multilivello (9 punti)

Siano dati il dataset della pagina seguente e una semplice gerarchia su 4 attributi:

status: {freshman, sophomore, junior, senior} ∈ undergraduate

{M.Sc., M.A., Ph.D.} ∈ graduate

major: {physics, chemistry, math} ∈ science

{cs, engineering} ∈ appl_science

{French, philosophy} ∈ arts

age: {16..20, 21..25} ∈ young

{26..30, over_30} ∈ old

nationality: {Asia, Europe, Latin_America} ∈ foreign

{Canada, USA} ∈ North_America

Determinare quali sono le regole associative al livello più basso della gerarchia, e quali al livello più alto, con un supporto minimo fissato a 20% ed una confidenza del 60%.

major	status	age	nationality	gpa
French	M.A.	over_30	Canada	2.8-3.2
cs	junior	16..20	Europe	3.2-3.6
physics	M.Sc.	26..30	Latin_America	3.2-3.6
engineering	Ph.D.	26..30	Asia	3.6-4.0
philosophy	Ph.D.	26..30	Europe	3.2-3.6
French	senior	16..20	Canada	3.2-3.6
chemistry	junior	21..25	USA	3.6-4.0
cs	senior	16..20	Canada	3.2-3.6
philosophy	M.Sc.	over_30	Canada	3.6-4.0
French	junior	16..20	USA	2.8-3.2
philosophy	junior	26..30	Canada	2.8-3.2
philosophy	M.Sc.	26..30	Asia	3.2-3.6
French	junior	16..20	Canada	3.2-3.6
math	senior	16..20	USA	3.6-4.0
cs	junior	16..20	Canada	3.2-3.6
philosophy	Ph.D.	26..30	Canada	3.6-4.0
philosophy	senior	26..30	Canada	2.8-3.2
French	Ph.D.	over_30	Canada	2.8-3.2
engineering	junior	21..25	Europe	3.2-3.6
math	Ph.D.	26..30	Latin_America	3.2-3.6
chemistry	junior	16..20	USA	3.6-4.0
engineering	junior	21..25	Canada	3.2-3.6
French	M.Sc.	over_30	Latin_America	3.2-3.6
philosophy	junior	21..25	USA	2.8-3.2
math	junior	16..20	Canada	3.6-4.0

Esercizio 3 - Classificazione / Lift chart (**5** punti)

E' dato un problema di classificazione binaria su un dataset di 1000 elementi, in cui la classe *positiva* conta 100 elementi. Disegnare i lift chart rispettivamente de:

- il classificatore casuale
- il classificatore ottimo
- il classificatore pessimo
- un buon classificatore

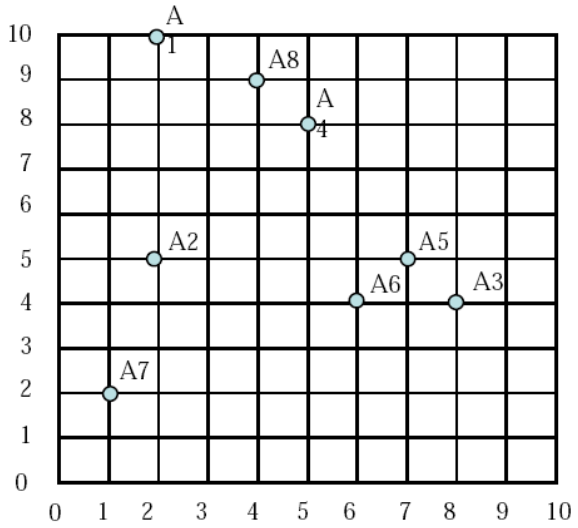
Esercizio 4 - Classificazione (7 punti)

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	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 (7 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 sono rumore, e quali cluster l'algoritmo DBSCAN individua.