BUSINESS INTELLIGENCE LABORATORY

Association Rules

Computer Science Department, University of Pisa

Business Informatics Degree

Items, transactions, transaction db

- Let $I = \{ a_1, \ldots, a_n \}$ be a finite set
  - $a_i \in \mathcal{I}$ is called an **item**
- A **itemset** I is a subset of $\mathcal{I}$
  - $I \subseteq \mathcal{I}$
- A **transaction** t is an itemset with an identifier
  - $t = (i, I)$ with $I \subseteq \mathcal{I}$ also written $t_i \subseteq \mathcal{I}$
- A **transaction database** is a finite set of transactions
  - $D = \{ t_i \mid i = 1 \ldots d, t_i \subseteq \mathcal{I} \}$

Format of transaction db

- Transactional
  - A row for each transaction (id not necessary)
  - List of items in the transaction

milk, sugar, water
beer, diapers
…

- Not available in Weka

Format of transaction db

- Tabular
  - Two columns
    - transaction ID
    - item

tID	item
1	milk
1	sugar
1	water
2	beer
…	…

    - Filter available in Weka: denormalize
      - Use the GUI Chooser -> Tools -> Package Manager to install it

Format of transaction db

- Binary
  - A column for each item
  - A row for each transaction (id not necessary)
  - Cell value
    - true (false) if the item is (not) in the transaction

milk	sugar	…	diapers
true	true		false
false	false		true
…	…	…	…

- In Weka use '?' for false

Format of transaction db

- Relational
  - Item are of the form **att=value**
    - **att** is an attribute, **value** is a value in its domain
  - A row for each transaction (id not necessary)
  - Cell value
    - **value** if the item **att=value** is in the transaction

income	status	…	age
high	married		20-30
medium	single		40-50
…	…	…	…

- Available in Weka

Support & Confidence

- Association rule $\mathbf{X} \rightarrow \mathbf{Y}$
  - X, Y itemsets and $X \cap Y = \varnothing$
- Classification rule $\mathbf{X} \rightarrow \mathbf{C}$
  - X itemset, C class item, $C \notin X$
  - Common in the relational format
- Support of an itemset
  - $supp(I) = |\{t \in D \mid I \subseteq t\}|$
  - relative support: $supp(I)/|D|$

Support & Confidence

- 4-fold contingency table

$X \rightarrow Y$

	Y	$\neg Y$
X	a	b
$\neg X$	c	d

- $\text{supp}(X \rightarrow Y) = a = \text{supp}(X,Y)$
- $\text{conf}(X \rightarrow Y) = a/(a+b) = \text{supp}(X,Y)/\text{supp}(X)$
- $\text{coverage}(X \rightarrow Y) = a+b = \text{supp}(X)$
- $\text{lift}(X \rightarrow Y) = \text{conf}(X \rightarrow Y)/\text{conf}(\text{true} \rightarrow Y) = \text{supp}(X,Y)/(\text{supp}(X)*\text{supp}(Y))$

Software for AR mining

- Weka
  - Binary and relational format
- Frida
  - http://www.borgelt.net/frida.html
  - Transactional format
- SQL Server Analysis Services
  - Relational format
- A lot of research and commercial systems
  - http://fimi.cs.helsinki.fi
  - http://www.kdnuggets.com/software/associations.html

Demo and practice

- Demo on the supermarket.arff dataset
- Practice on the credit-g.arff dataset
  - Objective:
    - Find conditions of past bad credit
  - Method
    - Find classification rules with **class=bad**
    - Rank them wrt *which measure*?