



CIS338

Data mining:

Course organization

Lecturer: Dr Daniel Stamate



Data Mining (DM)

- DM is at the confluence of fields as
 - Machine learning – AI
 - Mathematical Statistics
 - Databases
- Combines practical aspects with very technical ones



Course structure

- 2 lecture hours
 - See reading list at the end
- 1 lab hour
 - DM/Machine Learning Software: Weka
 - Coding DM algorithms in Java
 - Demos



Course content

- Introduction to Data Mining
- Presentation of Data Mining strategies + algorithms
 - Supervised learning
 - Decision Trees, Rules
 - 1R, ID3, C4.5, K-nearest neighbour algorithms
 - Back propagation algorithm for Neural Networks, etc
 - Association Analysis
 - Association Rules and the Apriori algorithm
 - Unsupervised Clustering
 - K-means algorithm, Self-organizing maps
 - Expectation Maximization, Conceptual Clustering, etc
 - Genetic learning



Course content

- Apriori Data Mining algorithm

Apriori (T, ε)

$L_1 \leftarrow \{ \text{large 1-itemsets} \}$

$k \leftarrow 2$

while $L_{k-1} \neq \emptyset$

$C_k \leftarrow \text{Generate}(L_{k-1})$

 for transactions $t \in T$

$C_t \leftarrow \text{Subset}(C_k^f)$

 for candidates $c \in C_t$

$\text{count}[c] \leftarrow \text{count}[c] + 1$

$L_k \leftarrow \{ c \in C_k \mid \text{count}[c] \geq \varepsilon \}$

$k \leftarrow k + 1$

return $\bigcup_k L_k$



Course content

Knowledge Discovery in Data

- Goal identification
- Creating a Target Dataset
- Data Pre-processing
- Data Transformation (normalisation, type conversion, attributes/instances selection)
- Data Mining
- Interpretation and Evaluation
- Action



Course content

- Probabilistic Statistical Techniques
 - Regression, classification, clustering techniques, performance evaluation

Logistic regression

$$p(y = 1 | x) = \frac{e^{\mathbf{ax} + c}}{1 + e^{\mathbf{ax} + c}}$$

Bayesian classifiers

$$P(H | E) = \frac{P(E | H) \times P(H)}{P(E)}$$

where $P(A | B)$ conditional probability of A given B

H is the hypothesis to be tested

E is the evidence associated with H



Course content

- Neural Networks
- Specialized Techniques
 - Time Series
 - Mining the Web
 - Mining Text Databases



Course content

- Data Warehousing
 - Data Warehouse design
 - the star, snowflake schemas
 - On-line Analytical Processing (OLAP)



Labs

Coding data mining algorithms in Java

Example:

K-Nearest Neighbour algorithm ⇒ Java code

Labs

Application of data mining techniques using available software

The screenshot displays the Weka Explorer application window. The interface is divided into several sections:

- Preprocess** (selected), **Classify**, **Cluster**, **Associate**, **Select attributes**, and **Visualize** tabs.
- Buttons for **Open file...**, **Open URL...**, **Open DB...**, **Undo**, **Edit...**, and **Save...**.
- Filter** section with a **Choose** button and a dropdown menu set to **None**, and an **Apply** button.
- Current relation** section showing **Relation: diagnoses_train**, **Instances: 11**, and **Attributes: 7**.
- Attributes** section with **All**, **None**, and **Invert** buttons, and a table of attributes:

No.	Name
1	<input checked="" type="checkbox"/> PatientID
2	<input type="checkbox"/> SoreThroat
3	<input type="checkbox"/> Fever
4	<input type="checkbox"/> SwollenGlands
5	<input type="checkbox"/> Congestion
6	<input type="checkbox"/> Headache
7	<input type="checkbox"/> Diagnosis

Below the attributes table is a **Remove** button.

- Selected attribute** section for **PatientID** (Type: Numeric):

Statistic	Value
Minimum	1
Maximum	11
Mean	6
StdDev	3.317

Missing: 0 (0%), Distinct: 11, Unique: 11 (100%)

- Class: Diagnosis (Nom)** dropdown menu and **Visualize All** button.
- Visualization** area showing a horizontal bar chart with segments in dark grey, cyan, red, and blue. The x-axis is labeled with 1, 6, and 11.
- Status** section at the bottom left showing **OK**.
- Log** button and a small icon with **x 0** at the bottom right.



Assessment

- All the material comprised in the course may be subject of assessment
- Coursework 20%
- Written exam 80%



Assessment

- Coursework profile
 - tackling exercises on machine learning techniques and algorithms
 - coding data mining algorithms in Java
 - building data models with Weka
- Written exam profile
 - Theoretical and practical questions (no coding in exam, but work with algorithms)
 - Based on the course material



Course material

- **Course website** <http://www.doc.gold.ac.uk/~mas01ds/cis338/>
- **Indicative reading list (main titles in bold):**
 - **Lecture**
 1. **Richard Roiger and Michael Geatz** "*Data Mining, a tutorial-based primer*", Addison Wesley, 2003
 2. **Jiawei Han and Micheline Kamber** "*Data Mining: Concepts and Techniques*", Morgan Kaufmann, 2000 or 2006
 - **Lab**
 3. **Ian Witten and Eibe Frank** "*Data Mining: Practical Machine Learning Tools and Techniques*", Morgan Kaufmann, 2005
 - **Additional**
 4. Margaret Dunham "*Data Mining: Introductory and Advanced Topics*", Prentice Hall, 2002
 5. Sheldon Ross "*Introductory Statistics*", Elsevier Academic Press, 2005

You can find the books in the list from one or more of the following sources:

Goldsmiths College Library + local bookshop

Amazon

pearsoned.co.uk