DME Introduction

Data Mining and Exploration

Nigel Goddard

School of Informatics, University of Edinburgh

Data Mining and Exploration: Introduction

`http://www.inf.ed.ac.uk/teaching/courses/dme/`

Please sign up on nb.mit.edu by following the link in your email.

These lecture slides are based extensively on previous versions of the course written by Chris Williams.

Data Mining and Exploration

Course Introduction

- Welcome
- Administration
  - Books (Hand Mannila and Smyth)
  - Mini Project
  - Paper presentations
  - Lab classes

Overview

- Relationships between courses
- What is data mining?
- Example applications
- Data mining and KDD (Knowledge Discovery in Databases)
- Models and patterns
- Data mining tasks
- Components of data mining algorithms
- Issues in data mining

Relationships between courses

PMR Probabilistic modelling and reasoning. Learning and inference for probabilistic models.

IAML Introductory Applied Machine Learning. Basic introductory course on supervised and unsupervised learning.

MLPR Machine Learning and Pattern Recognition. More detailed course on Bayesian Machine Learning.

RL Reinforcement Learning. Apologies - this course is not running this year.

DME Develops ideas from MLPR, IAML, PMR to deal with real-world data sets. Also data visualization and new techniques.

This course.

Beginning to End of the machine learning and data mining process.

What is data mining?

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. **Hand, Mannila, Smyth**

We are drowning in information, but starving for knowledge! **Naisbett**

[Data mining is the] extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases. **Han**

Data mining: pejorative sense

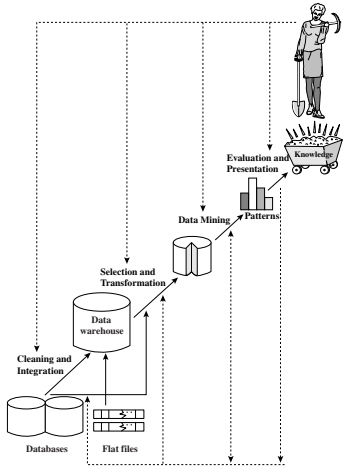
- Historically data mining was used in a pejorative sense by statisticians for the idea that, if you search long enough, you can always find some model to fit your data arbitrarily well.
- Example: David Rhine, a "parapsychologist" at Duke in the 1950's tested students for "extrasensory perception", by asking them to guess 10 playing cards—red or black.
  - Hypothesis was: all correct suggests student has ESP.
  - If he tested 10,000 students, and none had ESP, how many would guess correctly?
  - When he tested the 10 again, what would you expect him to find?
  - What conclusion would you draw?

Example applications

- **Scientific** SKICAT (Sky Image Cataloging and Analysis Tool) developed at JPL and Caltech. See `http://www-aig.jpl.nasa.gov/public/mls/skicat/skicat_home.html`. Predict if object is a star or galaxy.
- **Commercial** Decision trees constructed from bank-loan histories to decide whether or not to grant a loan
- **Marketing** "Diapers and beer". Observation that customers who buy diapers are more likely to buy beer than average allowed supermarkets to place beer and diapers nearby, knowing that many customers would walk between them. Placing potato chips between increased sales of all three items
- **Financial** Predict price movements in order to make more lucrative investments

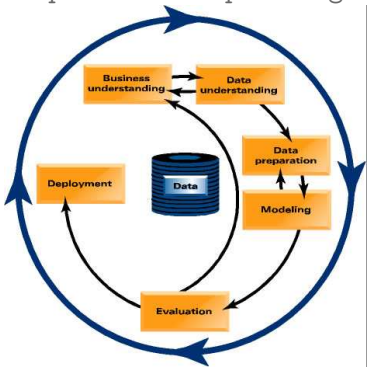
Datamining and KDD

Knowledge Discovery in Databases. Figure from Han and Kamber.



CRISP-DM methodology

Cross Industry Standard Process for Data Mining, `http://www.crisp-dm.org/`



Six Phases

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

Data Mining: History

- 1989 IJCAI workshop on KDD (Piatetsky-Shapiro)
- 1991-1994 workshops on KDD
- 1996 Advances in Knowledge Discovery and Data Mining (eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy)
- 1995 onwards: International Conferences
- 2000s: becomes known as Data Analytics
- 2010s: subsumed into new discipline of Data Science

Data Mining: Relationships to Other Fields

- Statistics
- Machine Learning
- Database technology
- Visualization
- ...

Relationship of Machine Learning to Data Mining

- Machine Learning is concerned with making computers that learn things for themselves.
- Data mining is more concerned with enabling humans to learn from data

Data Science

- In fact this course is really about Data Science
- Data Science is about integrating data driven enterprise into a whole process of doing things.
- Data Science is about the skills of a practitioner. It recognises the need for people in the process.
- However Data Science is also about automation (systems)- do the *right* things efficiently.

Models and Patterns

- A model structure is a *global* summary of the data set. Example: linear regression, makes a prediction for all input values. $Y = aX + c$
- Pattern structures make statements only about restricted regions of the space spanned by the variables.
  Example:
  `if` $X > x_1$ `then` $\text{prob}(Y > y1) = p1$
  [Equivalently $\text{prob}(Y > y1 | X > x_1) = p1$]
  Example: detection of outliers
- In both cases, we are interested in estimating the parameters

Data Mining Tasks

- Exploratory Data Analysis
- Descriptive Modelling
  - Density estimation
  - Cluster analysis/segmentation
- Predictive Modelling: Classification and Regression
- Discovering Patterns and Rules
  - Association rules
  - Outlier detection
- Mining Complex Types of Data
  - Retrieval by Content (RBC) for text, images
  - Time series and sequence data
  - Spatial data
  - Text mining
  - Mining the WWW (content, structure, usage)

Components of Data Mining Algorithms

Headings	Example: Neural Network
• Task	Regression
• Structure of model or pattern	Neural network function
• Score function	Squared error
• Optimization and search method	Gradient descent
• Data Management Strategy	unspecified

Ref: HMS chapter 1

Some Issues in Data Mining

(based on list by Han)

- Mining methodology and user interaction
  - e.g. Incorporation of background knowledge
  - e.g. Handling noise and incomplete data
- Performance and scalability
- Diversity of data types
  - Handling relational and complex types of data
  - Mining information from heterogeneous databases and WWW
- Applications, social impacts, ethics, ...

Tentative Lecture Outline

- Visualizing and Exploring Data
- Descriptive Data Modelling
  - Including hierarchical clustering
- Data Preprocessing
  - Data cleaning
  - Data integration and transformation
  - Data reduction
- Predictive Modelling
  - Overview of regression and classification
  - Decision trees
  - Support Vector machines
  - Performance evaluation
  - Dealing with unbalanced classes

Tentative Lecture Outline

- Patterns
  - A priori algorithm
- Mining Complex Data
  - Web mining: Page Rank (google)
  - Retrieval by Content
  - Text, time series, images
- Paper presentations.