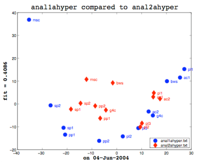
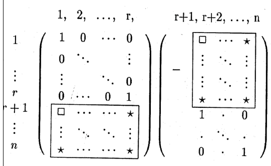
Data Mining
and
Information retrieval

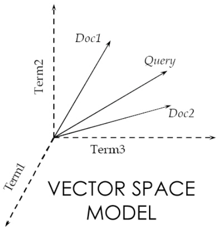




Pedro Contreras
pedro@cs.rhul.ac.uk

Department of Computer Science
Royal Holloway, University of London

20 February 2008



VECTOR SPACE MODEL

Overview, Lecture I

Data Mining

- What's Data?
  - Record data, numerical data, data matrix, document data, graph data, chemical data, etc.
- What's Data Mining?
- Why Data Mining?
  - Commercial viewpoint
  - Scientific viewpoint

Overview, Lecture I

Mining Data sets

- Association Rules
- Classification
- Clustering
- Forecasting

Challenges of Data Mining

References

Overview, Lecture II

What's Information Retrieval?

Parsing

- Part of speech
- Stop words removal
- Stemming
- Entity detection

Indexing

- Index construction
- Index compression

Querying

- Exploiting data

Lecture I

Data Mining

20 February 2008

Introduction

Data mining, what's all about?

It is all about:

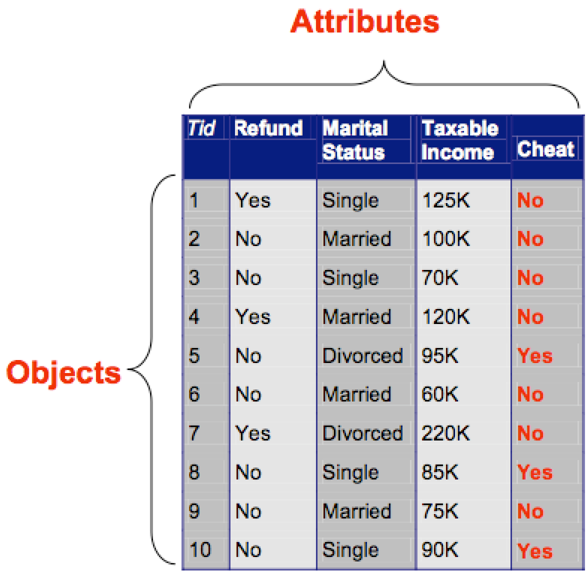
- Taking data and transforming it to valuable information, which in turns help in the decision making process
- Use information for competitive advantage

Data → Information → knowledge

What’s data

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  – Examples: eye colour of a person, temperature, etc.
  – Attribute is also known as:
  variable, field,
  characteristic, or feature
- A collection of attributes describe an object
  – Object is also known as:
  record, point,
  case, sample,
  entity, or instance

What’s data: an example


Attributes
Objects

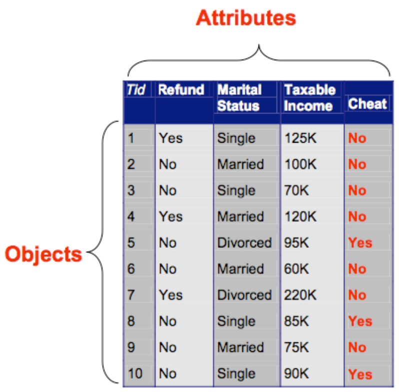
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Type of data sets

- Record
  - – Data Matrix
  - – Document Data
  - – Transaction Data
- Graph
  - – World Wide Web
  - – Molecular Structures
- Ordered
  - – Spatial Data
  - – Temporal Data
  - – Genetic Sequence Data

Type of data sets: record

Record data: Data that consists of a collection of records, each of which consists of a fixed set of attributes.


Attributes
Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Type of data sets: data matrix

If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute.

Such data set can be represented by an ***m*** by ***n*** matrix, where there are ***m*** rows, one for each object, and ***n*** columns, one for each attribute

	C_1	**C_2**	**C_3**	**-**	**C_n**
Obs 1	1	1	0	.	1
Obs 2	0	0	1	-	0
.	-	-	-	-	-
Obs n	1	0	1	-	0

Type of data sets: document data

- Each document becomes a "term" vector,
  – each term is a component (attribute) of the vector,
  – the value of each component is the number of times the corresponding term occurs in the document.

	book	tittle	search	price	sales	time
Document 1	5	0	1	3	0	1
Document 2	1	3	5	0	0	3
Document 3	3	1	1	0	0	1
Document 4	3	1	1	3	2	2

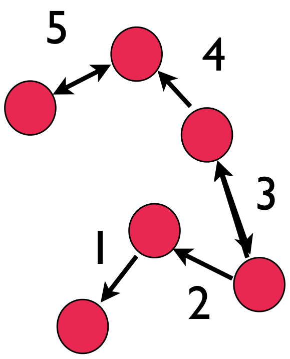
Type of data sets: transaction data

A special type of record data, where
– each record (transaction) involves a set of items.
– For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the item

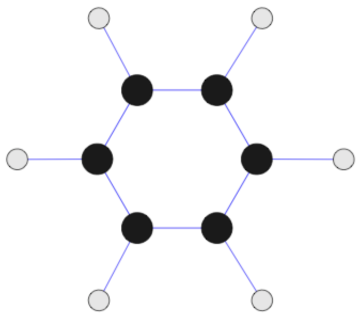
ID	Item
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Type of data sets: graph

– World Wide Web: Hyperlink



– Molecular Structures: Benzene Molecule C_6H_6

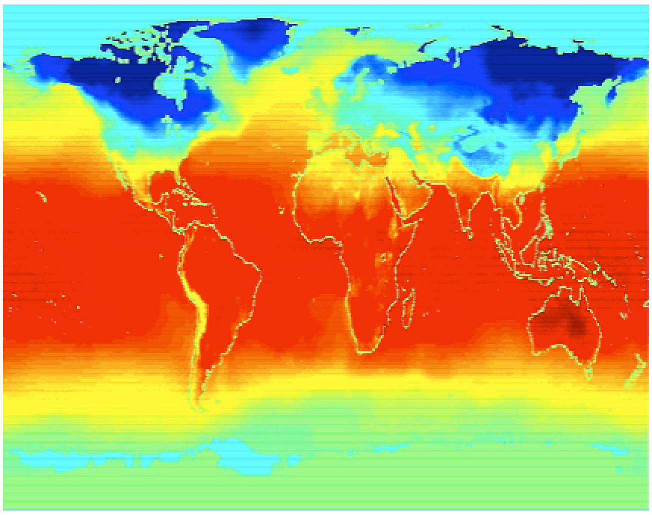


Type of data sets

Ordered
– Spatial Data
– Temporal Data
– Genetic Sequence Data

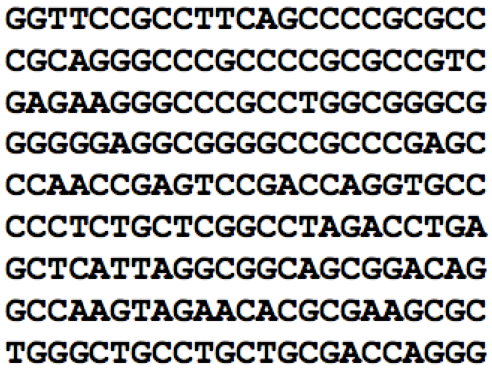
Type of data sets: spatio-temporal

Average monthly temperature of land and ocean



Type of data sets: ordered

Genomic sequence data



```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Other considerations about data

Data quality problems:

- Noise: e.g. errors when recording, in general any distortion in data is called noise.
- Missing values: information isn't collected, attributes not applicable for every case.
- Duplicate data: e.g. when merging data set for different sources

Why mine data, commercial viewpoint

- Lots of data is being collected and warehoused
  - – Web data, e-commerce
  - – Purchases at department stores
  - – Bank/Credit card transactions

- Computers have become cheaper and more powerful

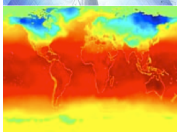
- Competitive pressure is strong
  - – Provide better, customised services

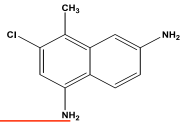
Why mine data, scientific viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - – remote sensors on a satellite
  - – telescopes scanning the skies
  - – micro-arrays generating gene expression data
  - – scientific simulations generating terabytes of data

- Traditional techniques infeasible for raw data

- Data mining may help scientists
  - – in classifying and segmenting data
  - – in hypothesis formation





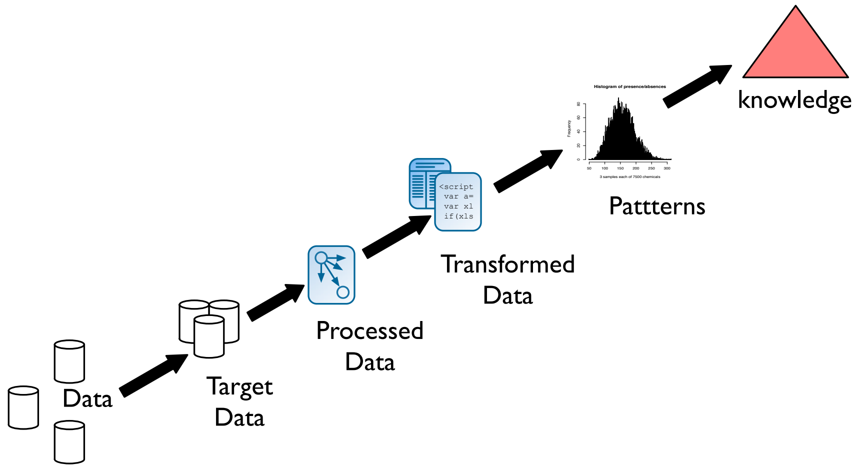
What’s data mining?

Many Definitions

– Non-trivial extraction of implicit, previously unknown and potentially useful information from data

– Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

What’s data mining?



Data

Target Data

Processed Data

Transformed Data

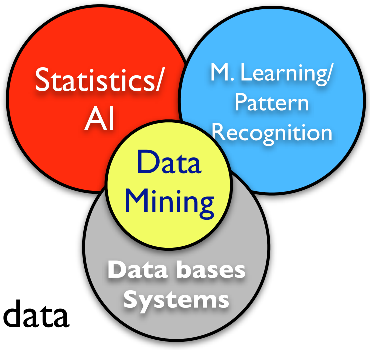
Pattterns

knowledge

Data → Information → knowledge

Origins of data mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to:
  – Enormity of data
  – High dimensionality of data
  – Heterogeneous, distributed nature of data



What data mining can do

Data mining is primarily used today by companies with a strong **consumer focus** - **retail**, **financial**, **communication**, and **marketing** organisations.

It enables these companies to determine **relationships** among "internal" factors such as **price**, **product positioning**, or **staff skills**, and "**external**" factors such as **economic indicators**, **competition**, and **customer demographics**.

And, it enables them to **determine** the **impact** on **sales**, **customer satisfaction**, and **profits.**

Data mining examples:

- Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers.
- Amazon mines customer's profile to offer new books/ products.
- And so on with Banks, Supermarkets, etc.

Data mining tasks

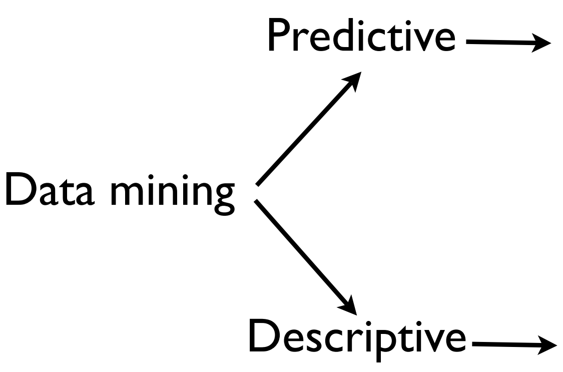
- Prediction Methods
  – Use some variables to predict unknown or future values of other variables.

- Description Methods
  – Find human-interpretable patterns that describe the data.

Data mining tasks

- Association Rule Discovery [Descriptive]
- Classification [Predictive]
- Clustering [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Data mining tasks



Classification: maps data into predefined classes
Regression: maps data into functions
Prediction: predicts future data states
Time series analysis: analyse data over time

(Supervised learning)

Clustering: find similar groups
Assoc. rules: find relationships between items
Sequence discovery: find sequential patterns

(Unsupervised learning)

Association Rules discovery

Association Rules are used to identify uncover hidden patterns in data sets, that is to find relationships or correlations in the data.

For example, let us consider the Amazon case:

- People that have bought book "A" also have bought book "B" and "C".
- People that have queried for product "A" also often query for products "C" and "D".

Association Rules

Coming back to our example

ID	Item
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

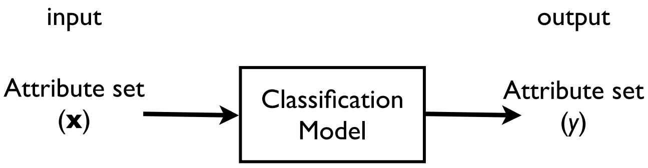
$\{\text{Milk}\} \longrightarrow \{\text{Coke}\}$

$\{\text{Diaper, Milk}\} \longrightarrow \{\text{Beer}\}$

Classification

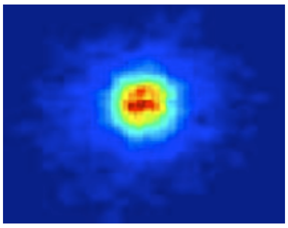
Given a set of record assign a class (or label) according to the record's attributes.

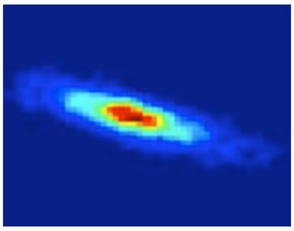
Remark: the "labels" are known

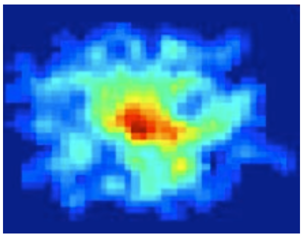


Classification example I

Sky survey catalogue: e.g. galaxy morphological classification, that's elliptic, spiral, irregular, etc.







Classification example II

Fraud Detection, predict fraudulent cases in credit card transactions.

Approach:

Use credit card transactions and the information on its account-holder as attributes.

– When does a customer buy, what does he buy, how often he pays on time, etc

Label past transactions as fraud or fair transactions. This forms the class attribute.
Learn a model for the class of the transactions.
Use this model to detect fraud by observing credit card transactions on an account

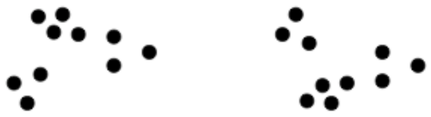
Clustering

Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that:

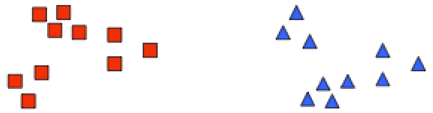
– Data points in one cluster are more similar to one another.
– Data points in separate clusters are less similar to one another.

Similarity Measures:
–Euclidean distance if attributes are continuous.
–Other Problem-specific Measures.

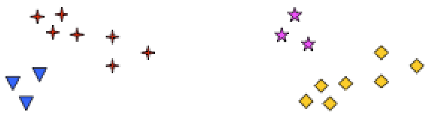
Notion of clusters can be ambiguous



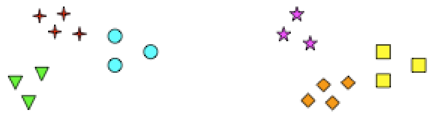
How many clusters?



Two Clusters



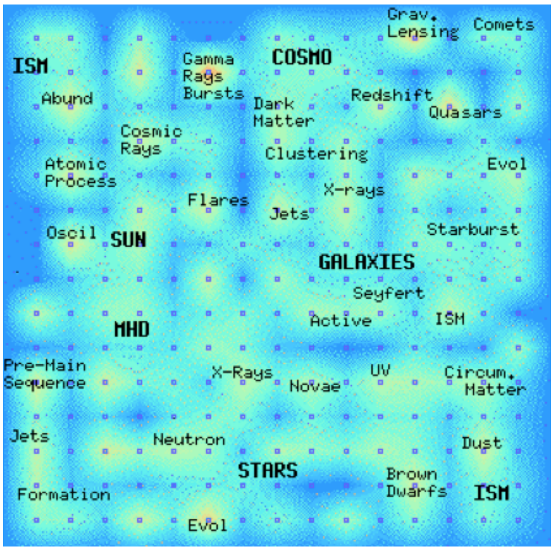
Four Clusters



Six Clusters

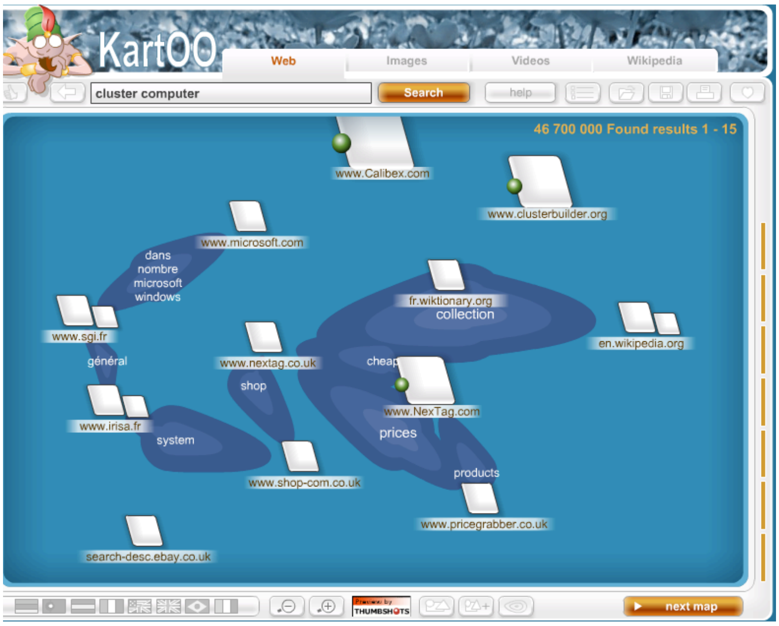
Clustering example I

Document clustering using a metric to determine "closeness" between documents.



Clustering example II

www.kartoo.com
Searching terms
cluster based

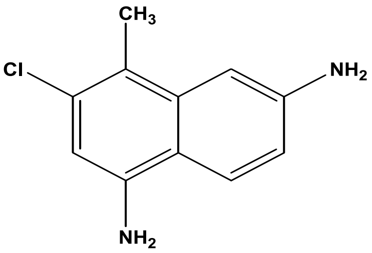


cluster computer

46 700 000 Found results 1 - 15

Clustering example III

Binary Fingerprints: Chemical compounds



Encode

1 0 0 0 1 0 0 0 1 ...1

Clustering of compounds based on chemical descriptors or chemical representations, in the pharmaceutical industry

Clustering example III

Baire, or longest common prefix

Case of vectors x and y, with 1 attribute. Precision: digits 1, 2, …, |K|

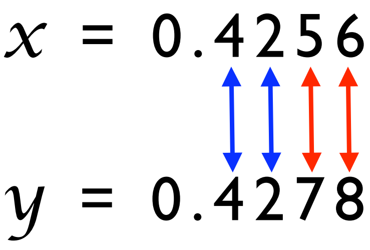
$$d_B(x_K, y_K) = \begin{cases} 1 & \text{if } x_1 \neq y_1 \\ \inf 2^{-n} & x_n = y_n \quad 1 \leq n \leq |K| \end{cases}$$

- Each coordinate is normalised, so is a floating point value.
- Then: we will define $d_B(x,y)$ based on sharing common prefix in all coordinates.

Clustering example III

Baire, or longest common prefix

An example of Baire distance for two numbers (x and y) using a precision of 4



x = 0.4256

y = 0.4278

Baire distance between x and y:

$d_B (x_4, y_4) = 2^{-3} = |K| = 3$

That is:
k=1 -> $x_k = y_k$ -> 4
k=2 -> $x_k = y_k$ -> 2
k=3 -> $x_k \neq y_k$ -> 5≠7

Clustering example III

Simple clustering hierarchy

Random projection schematically

A =

0	1	1	0
0	0	1	0
0	0	1	0
1	1	0	1

matrix Normalised by column sums

N =

0	0.5	0.33	0
0	0	0.33	0
0	0	0.33	0
1	0.5	0	1

Random vector; k = 2

R =

0.13	0.45	0.76	0.49

N x R = Random projected vector

0.47
0.25
0.25
0.84

Sorting ....

0.25	} 1st cluster
0.25	
0.47	} 2nd cluster
0.84	} 3rd cluster

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  – Marketing: predicting sales amounts of new product based on advertising expenditure.
  – Weather: predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  – Economy and finances: time series prediction of stock market indices.

Final remarks

Data mining

Data
Information
knowledge

But....

- We need to use the right method according the problem.
- And data not always is clean (noise)

References

[1] Pang-NingTan, Michael Steinbach and Vipin Kumar. Introduction to Data Mining. Addison Wesley. 2005.

[2] Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer, 2007.

[3] Soumen Chakrabarti. Mining the Web, Discovering Knowledge from Hypertext Data. M. Kaufmann. 2003.

[4] Oded Maimon and Lior Rokach, Data Mining and Knowledge Discovery Handbook. Springer. 2005.

[5] Ian Witten, Alistair Moffat and Timothy Bell. Managing Gigabytes. Compressing and Indexing Documents and Images. 2nd. Ed. M. Kaufman.1999.

[6] Ian Witten and Frank Eibe. Data Mining: Practical Machine Learning Tools and Techniques, 2nd Ed., M. Kaufmann, 2005.

[7] Margaret Dunham. Data Mining: Introductory and Advanced Topics, Prentice Hall, 2003.