Data Mining for Cryptographers
T-79.514, September 17th 2003

Ella Bingham
HUT Lab of Computer and Information Science
ella@iki.fi

About this presentation

We will discuss

- what is data mining (DM) in general
- some examples of application areas

We will not actually discuss

- specific relationships between DM and cryptography — hopefully revealed later
- specific methods and algorithms in DM —hopefully studied in context of their applications into privacy preserving DM; the choice of methods/algorithms is largely problem dependent

Data mining is . . .

(Material borrowed from H. Mannila's course notes)

"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 analyst"

In short, the goal is to **obtain useful knowledge from large masses of data**

At the intersection of machine learning, statistics and databases

An umbrella covering several types of methods, algorithms, and practices. No clear borders enforced.

An applied area, driven by successful applications

General characteristics of DM

Warnings

- terminology is overlapping: DM/KDD? algorithm? etc.
- hype
- heuristics that work in practice and are thus useful
- no universal methods
- no correct answers

Examples of discovered knowledge:

- association rules: “80% of customers who buy beer and sausage buy also mustard”
- rules: “if Age < 40 then Income < 10”
- models: $Y = aX + b$
- clusterings: similar data points grouped together

Data presentation

- as a matrix:
  - rows present different **observations**: customers in a transaction database, text documents, web pages etc.
  - columns present different **attributes** (also called **variables**) of the above observations: products, words, linked pages
- as a graph (can be written as a matrix)
- as a sequence/string such as telecom alarm log or DNA where events take place in time

Aims of DM

- classification or clustering
- regression
- summarization
- dependency modeling

In each DM algorithm/method one must decide

- what to look for in the data (see the above list)
- how to measure the goodness of the estimated pattern or model
- what is the optimization method to maximise the goodness
- data handling; is the data in main memory during computation, or where

Dichotomies

DM is about finding

- **patterns**: interesting (small) local features that
  - occur frequently, or
  - are outliers: rare and different.
  - patterns may be groups of attributes or observations
  - do not describe the whole data
  - fast algorithms exist
- **models**: global descriptions of the whole data
  - rules, clusters, joint density of attributes, prediction. . .
  - algorithms often slower

The algorithms are typically

- combinatorial (fast; for discrete data)
- probabilistic (gives a continuous description)

DM versus KDD

Knowledge discovery in databases (KDD) is a more general term that includes DM as a subprocess: KDD is the total process of
data selection → data preprocessing/transformation → data mining → interpretation/evaluation.

Several steps of the above may be of interest to privacy preservation

For our purposes, we can use the terms DM and KDD interchangeably (this happens elsewhere, too)

Current methodological challenges

- formulating the problem and hence choosing the method → algorithmic details not important
- large databases (number of observations and attributes) → computational efficiency is a must
- did we find anything useful? does our description generalize to other data sets or did we overfit?
- nonstationary data: possible tasks include
  - updating the description
  - detecting what is new in the data
- breaking the dichotomies presented earlier

Hot applications (for PPDM, too?)

- link analysis of Web pages
  - hubs and authorities
  - spectral clustering of the link matrix
- personalization
- recommendation systems, collaborative filtering: e.g. movie ratings given by a large set of people are used to give recommendations
- fraud/anomality detection
  - in telecommunication logs: is a client planning to avoid the phone bill?

  - industrial etc. alarm logs: problems not voluntarily revealed in public
- bioinformatics

Privacy preserving DM?

Privacy not an issue in several applications.
On the other hand, methods developed for one such application could be used in another where privacy is an issue