Data Warehouse
Technologies

cetinerg@itu.edu.tr Assoc.Prof.Dr.B.G.Çetiner ©2002

Data Mining

cetinerg@itu.edu.tr Assoc.Prof.Dr.B.G.Çetiner ©2002

Data Mining: Introduction

- More and more data is gathered due to progress in computers and databases
- Manual analysis, charts etc. are no longer feasible
- Scientifc progress stimulates needs and offers solutions
- Problem: Find *information nuggets* in vast amounts of data
  Solution: Knowledge Discovery in Databases (KDD)

Data Mining: Introduction (contd.)

- KDD: process of finding knowledge in data by "high level" application of data mining methods
- Data mining is only one step of KDD
- Blind application of data mining methods can be dangerous as invalid patterns might be detected
- Data mining is not a single method – "data mining" refers to various methods

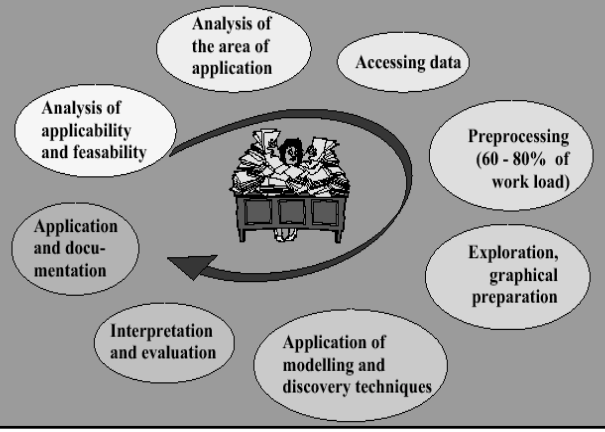
KDD: Knowledge Discovery in Databases

KDD is the non-trivial process of identifying

- valid,
- novel,
- potentially useful,
- ultimately understandable,

patterns in data (Fayyad, Piatetsky-Shapiro & Smyth, 1996).

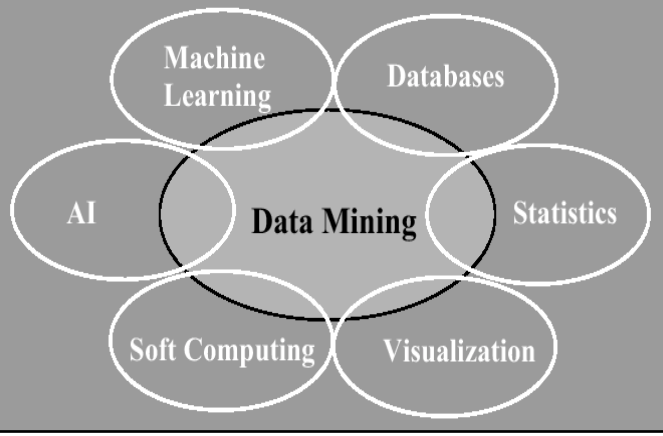
KDD: Knowledge Discovery in Databases



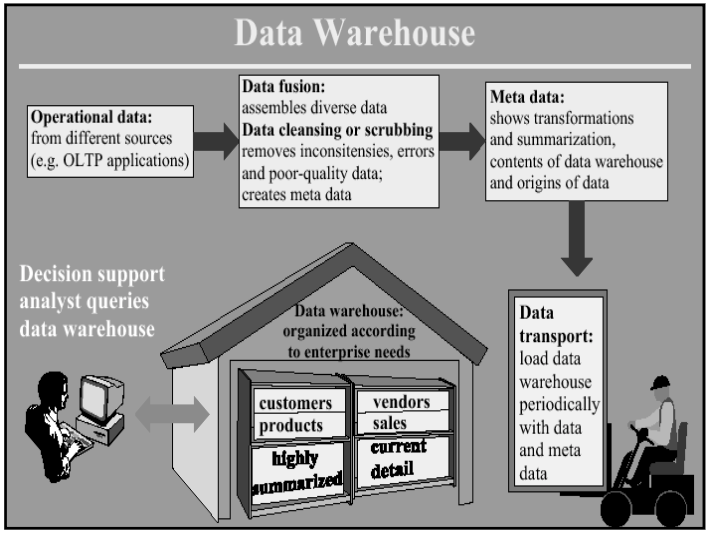
Data Mining

Data Mining is:	The application of various methods of analysis and learning to discover knowledge in data
or:	Torture your data until they confess.
Data Mining is not:	Ad-hoc queries, reports, data warehousing, OLAP software agents, XPS, alerting,...

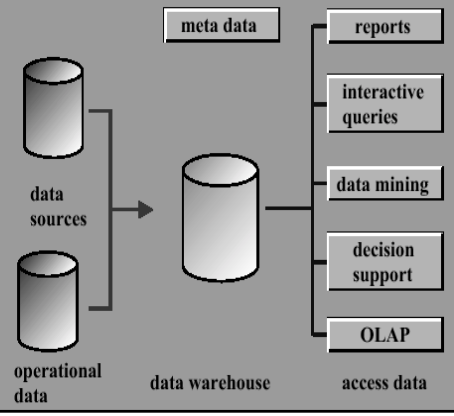
Data Mining is Interdisciplinary



Data Warehouse



Data Mining as Query Tool



Data Mining

- A lot of hype out there: Data Mining is a buzzword (yesterday C++ and statistics, today Java and data mining)
- There is a trade-off between usability and accuracy
- Most of the software aims at special applications. There are a lot of tools (over 50 in 1997)
- Severe errors occur if complex methods are not used correctly or are not explained to the end user
- The number and variety of data have more of an effect on the accuracy than the selected mining method

Data Mining Tasks

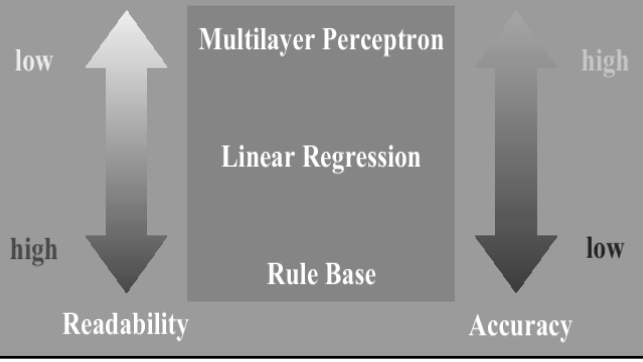
- Classification
  Is this a good customer ?
- Concept Description
  What makes a good customer ? (age, income, ...)
- Segmentation (Clustering)
  What kind of customers do I have ?

Data Mining Tasks

- **Prediction**
  What will be the demand for my product ?
- **Dependency Analysis**
  80% of customers who buy diapers buy beer, too
- **Deviation Analysis**
  Why do we sell less insurances in Cleveland ?

Model Selection

How readable / accurate is the selected model?



Uncertain Information

I am 90% certain that Peter is married

My belief that Peter is married is 0.9

Usually modeled by (subjective) probabilities
e.g. p(Peter is married) = 0.9

It can become worse:

I am 90% certain that Peter is tall
I am very certain that Peter is tall

Scenario A

You are a marketing manager for an insurance company

- The company sells liability insurance, personal effects insurance, life/accident insurance and car insurance.
- There are many cancelations / new contracts in car insurance each year.
- There are 1 million customers
- Goals: Prevent cancellations, cross-selling.

Solution A1

Prevent cancelations:

- Use historic data to predict which customers are likely to cancel their contract within the next 3 months.
- Contact these customers (send sales representative, offer better rates or benefits, ...).
- How can we predict future behavior?

Solution A2

Cross-selling:

- Use historical data:

  - find car insurance customers who bought life or accident insurance, create a classifier.

  - classify new customers according to your findings and send them an offer (mailing).

  - same method as used for solution to prevent cancellations

Solution A3

Cross-selling:

- If there is no historical data for this task:

  - find groups of customers, analyse and label them *(young parent, parsimonious, pensioners, ...)*

  - select groups that may be responsive to mailings *(pensioners don't buy life insurance, young parents do)*

  - How to find groups, how many are there?

Scenario B - Solution

Find the real end of the process

There is a database of (noisy) process data

Use this data to compute criteria to detect *c*.

Signal that *c* was reached by observing process data.

How can we find *c* in the process data?

Classical Statistical Approaches

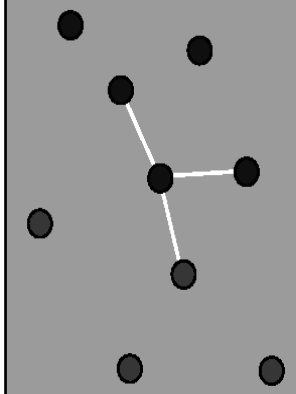
- Discriminant Analysis
- Regression Analysis
- Cluster Analysis
- Bayesian Learning

Problem: Often only linear models are applied, because non-linear models are not understood or cannot be handled by the user.

Classification

- Storing Known Cases: K-Nearest Neighbor
- Statistics: Discriminant Analysis, Logistic Regression
- Induction of Decision Trees
- Neural Networks: MLP or RBFN
- Fuzzy Classifier, Neuro-Fuzzy Classifier

Classification: K-Nearest Neighbor



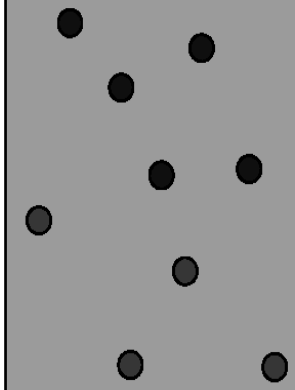
What color should the new ball have - red or blue?

Let the 3 nearest neighbors vote for it.

2 blue balls and 1 red ball: the new ball should be blue.

3-nearest neighbor classifier

Classification: K-Nearest Neighbor



Add the new case to the code book.

The code book can become very large (when to stop?)

Classifier simple to create, but classification takes long.

Possible: Store prototypes of clusters or mean values.

Classification: Induction of Decision Trees

Machine Learning (ML) Approach

A decision tree is a tree-like classifier that can be interpreted by rules

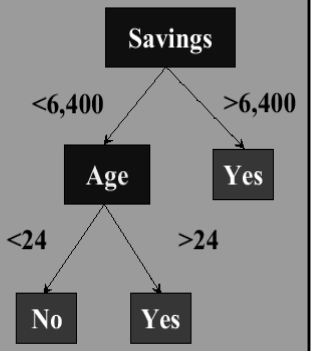
Inner nodes of tree: testing attributes

Leaf nodes: class labels

Idea: use an information theoretic measure to select "best" attributes first.

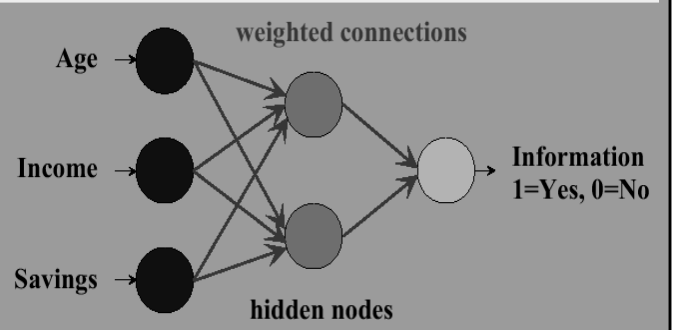
Classification: Decision Trees

Age	Income	Savings	Response
45	3,000	10,000	Yes
20	2,500	2,800	No
52	5,700	150,000	Yes
27	2,800	800	Yes
.	.	.	.
.	.	.	.



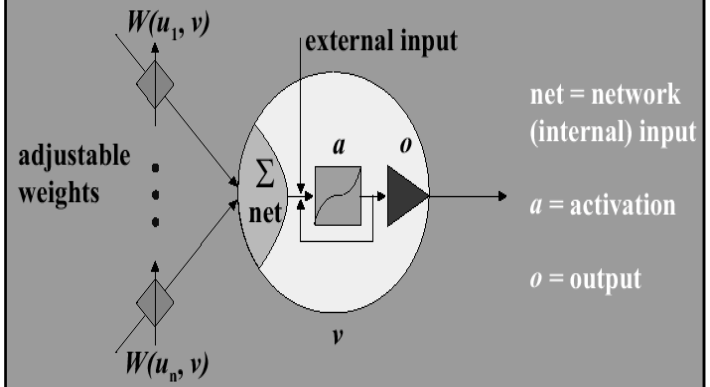
If Savings > 6,400 then send information
If Savings < 6,400 and Age > 24 then send information

Classification: Neural Networks



Non-linear model, universal function approximator, connection weights found by *"learning"*.

Neural Networks: Artificial Neuron



Classification: Neural Networks

Multilayer feedforward network with hidden layers

Nodes receive weighted sum as input

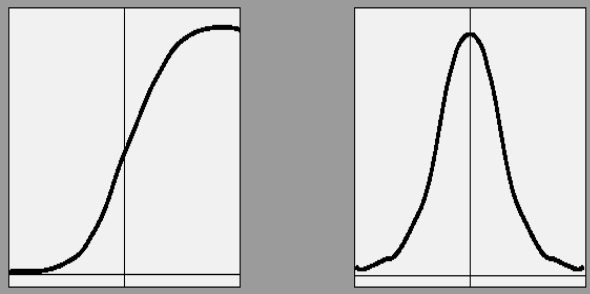
Hidden (and output) nodes use non-linear transfer functions

Multilayer Perceptron (MLP):
s-shaped (sigmoid) function

Radial Basis Function Network (RBFN):
bell-shaped (Gaussian) function

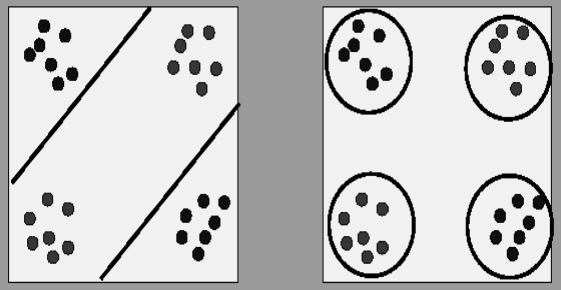
Classification: Neural Networks

Common activation functions in neural networks



MLP: sigmoid RBFN: Gaussian

Classification: Neural Networks



MLP:
global classification
using hyperplanes

RBFN:
local classification
using hyperellipsoids

Classification: Neural Networks

NN are called *model-free estimators*
(actually they represent a very general model, but there is no interpretation of model parameters)

NN do not rely on any special distribution of the data

NN cannot be interpreted (the ultimate black box)

Parameters of the NN are hard to determine without proper experience (e.g. how many hidden units?)

NN can outperform other methods,
but there is no guarantee

Association Rules

Bar-code technology makes it possible to store huge amounts of sales data.

Find rules in *basket data* for
- cross-marketing
- mailings
- catalog design
- store layout
- customer segmentation

Segmentation

Goal: Detect groups of cases that are similar and belong together

Problem: We don't know how many groups there are and how they should look like

Approach: Cluster Analysis

Preprocessing

Data Cleansing (Reduce Size, Improve Data)

Missing Values
- *delete cases with missing values*
- *estimate missing values by statistical methods*
- *do nothing, if your data mining method can handle them*

Remove Noise
- *filter the data to remove high frequency noise (mainly for function approximation and time series prediction)*

Preprocessing

Know Your Data Like Yourself

Compute basic descriptive statistics (mean, variance, ...).

Try simple linear models to see how they perform.

Visualize the data
- plot bar charts, 2D and 3D projections, ...

Ask "experts", i.e. persons who work with the data and collected it.

Validation

Always validate the model that is created during data mining!

N-fold Cross Validation:

Divide the data in *n* equal parts (same size and distribution).

Use *n*-1 parts to create a model and test on the remaining part.

Repeat *n* times, and compute the mean error.

Create a final model from the whole data set.

The mean error is an estimate for the error on unseen data.

Postprocessing

Interpret the result
Is is usable, efficient, easy to understand and to maintain?

Report all steps of the data mining process
It is essential that the result can be reproduced

Visualize the result
It is important that other persons can understand the result

Update the result, if your data changes
Specify when the result may be out of date

Evaluation Criteria

- Scalability
- Integretation with data warehouse
- Completeness
  - Is it an algorithm or a solution (application)?
- Usability
  - Does it solve a marketing problem?
  - Who is going to use it?
  - How is it going to be used?
  - How much does it cost?

Explaining the Results

- Depending on the selected model the results can be quite complex
- The results may influence strategical decisions
- Words are often better than numbers
- Interaction with users:
  - users must "get a feeling" for the result
  - let users identify their customers
  - reveal the data on several levels of detail, from a broad overview to the fine structure

Philosophies of Tools

Ground Level:
- add more sophisticated approaches to existing tools
- very flexible, but require a lot of expertise

One Step Up:
- data mining toolboxes
- problematic: often aim at users with insufficient expertise to consider tradeoffs

High Level Tools:
- end user applications, integrated into data warehouse
- interactive graphical tools: aimed at non-experts
- ease of use more important then accuracy

Some Tools

Statistics
- SAS (also data warehousing, statistics, NN, decision trees)
- SPSS (standalone statistics, add-ons for NN, CHAID)

Neural Networks
- SNNS (Stuttgart Neural Network Simulator, free)
- ECANSE, SENN (Siemens)

Data Mining
- Clementine (decision trees, NN)
- Data Engine (MIT GmbH, fuzzy, NN, plug-in extensions)
- IBM Data Mining Tool (statistics, NN, decision trees)
- Kepler (multi-relational data, logical rules, dec. trees)

Resources

Book:
Fayyad U.M. et al.:
Advances in Knowledge Discovery and Data Mining
MIT Press, Cambridge, MA 1996

WWW:
Knowledge Discovery Nuggets (with links to software)
http://www.kdnuggets.com

Journal on Data Mining and Knowledge Discovery
http://www.research.microsoft.com/research/datamine/

Conclusions

- There is not a single best method for data mining
- There are many methods, some are interchangeable.
- Thoroughly preprocess your data (get to know them).
- Know your objectives: interpretability or accuracy?
- At first, try methods and tools your are familiar with.
- Thoroughly validate and evaluate your results.