Data mining for prediction

Prof. Gianluca Bontempi
Département d’Informatique
Faculté de Sciences
ULB Université Libre de Bruxelles
email: gbonte@ulb.ac.be
Outline

• Extracting knowledge from observations.
• Introduction to Data Mining (What, Why, and How?)
• Our techniques for data mining.
• Prediction from data: applications.
• What can we do for you?
Knowledge from observations

Extracting knowledge from observations is a major issue in applied sciences.

Knowledge can be extracted for different purposes:

- qualitative description of phenomena,
- theoretical understanding of laws underlying nature,
- prediction of future observations.
Knowledge Discovery in Databases (KDD) is a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.
What is data-mining?

- Search for valuable information in large volumes of data.
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns & rules.
Why Mine Data? (for businesspersons)

• Lots of data has been and is being collected and warehoused.
• computing has become affordable.
• competitive pressure is strong
  – Provide better, customized services for an edge.
  – Information is becoming product in its own right.
• « convert data into business intelligence »
• « make management decision making based on facts not intuition »
• « get closer to the customers »
• « gain competitive advantage ». 
Business Cycle of Data Mining

- Employ Data Mining to Gain Knowledge
- Act on the Knowledge
- Measure Effectiveness of your Efforts (Dollar Value)
- Identify Business Opportunities Where Data Can Provide Value

From [Berry & Linoff] Data Mining Techniques, 1997

(Q 1999, Univ. of Minnesota)
Why mine data?  
(for scientists)

- Data collected and stored at enormous speeds (Gbyte/hour)
  - remote sensor on a satellite
  - telescope scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Large dimensionality (thousands of variables).
- Data mining for data reduction.
  - cataloging, classifying, segmenting data
- Helps scientists in hypothesis formation.
What does data-mining?

- **Description models**
  - Find human-interpretable patterns that describe the data.

- **Prediction models**
  - Use some variables to predict unknown or future values of other variables.
  - Examples: classification, regression, clustering, time series prediction.

From [Fayyad, et.al.] *Advances in Knowledge Discovery and Data Mining*, 1996
Example of descriptive measures

- Think to the difference between these two types of information:
  1. the whole historical record of your bank transactions
  2. the evaluation if you are a good or bad client (in your bank’s perspective)
• Finite amount of noisy observations.
• No a priori knowledge of the phenomenon.
Some theory..
Prediction modeling issues

• What are the relevant factors (inputs & outputs) in the data?
• Which kind of correlation (or model) exists between input and outputs? (linear, nonlinear,…)
• Is this correlation stationary or time varying?
• Which level of accuracy in our prediction can we guarantee?
Model discovery

- Model Generation
- Parametric Identification
- Model Validation
- Model Delivery
Nonlinear relationship

Target function

output

input
Global modeling
Prediction with global models
Prediction with local models
Our competences in modeling

• Conventional statistic techniques
• Neural networks techniques
• Regression trees
• Local modeling techniques
• Neuro-fuzzy techniques
In practice…
Financial prediction

daily stock market index

Task: predict the future trends of the financial series.

Goal: automatic trading system to anticipate the fluctuations of the market.

Industrial partner: Masterfood, Belgium.
**Task:** predict how many cars will be matriculated next year.

**Goal:** support the marketing campaign of a car dealer.

**Industrial partner:** D’ieteren, Belgium.
Task: predict the flow stress of the steel plate as a function of the chemical and physical properties of the material.

Goal: cope with different types of metals, reduce the production time and improve final quality.

Industrial partner: FaFer Usinor, Belgium.
Control

Task: model the dynamics of the plant on the basis of accessible information.

Goal: control the level of water pollutants.

Industrial partner: Honeywell Technology Center, US.
Environmental problems

**Task:** predicting the biological state (e.g. density of algae communities) as a function of chemicals.

**Goal:** make automatic the analysis of the state of the river by monitoring chemical concentrations.
Performance modeling of embedded systems

Task: model the performance of software tasks on different hardware platforms on the basis of profiling information.

Goal: select the hardware architecture that better meets the product requirements.

Industrial partner: Philips, The Netherlands.
Multimedia quality of service

Task: model the time/energy requirements of a multimedia application

Goal: find the best trade-off quality/performance

Industrial partner: IMEC, Belgium.
**Task**: predict the continuation of the series for next 100 steps.
Our prediction

Our methods are able to predict the abrupt change at $t = 1060$ !
Awards in international competitions

• Data analysis competition: awarded as a runner-up among 21 participants at the 1999 CoIL International Competition on *Protecting rivers and streams by monitoring chemical concentrations and algae communities*.

• Time series competition: ranked second among 17 participants to the *International Competition on Time Series* organized by the *International Workshop on Advanced Black-box techniques for nonlinear modeling* in Leuven, Belgium.
What can we do for you?

- **Take and analyze** your most valuable data (in electronic form).
- **Understand** if some relevant information is hidden within them.
- Choose, using our techniques, **the best model for your data**.
- **Make** (or providing you tools for doing) predictions on the basis of your historical information.