CSE 546

Data Mining

Instructor:

Pedro Domingos

Logistics

- **Instructor:** Pedro Domingos
  Email: pedrod@cs
  Office: 648 Allen Center
  Office hours: TBA
- **TA:** Matt Richardson
  Email: mattr@cs
  Office: TBA
  Office hours: TBA
- **Web:** www.cs.washington.edu/546
- **Mailing list:** cse546

Projects

- Individual or in groups of two
- From list or proposed by students
- Written report and oral presentation
- Class grade = Project grade

Source Materials

- Tom Mitchell, *Machine Learning*, McGraw-Hill, 1997 (required).
- David Hand, Heikki Mannila, and Padhraic Smyth, *Principles of Data Mining*, MIT Press, 2001 (recommended).
- Papers

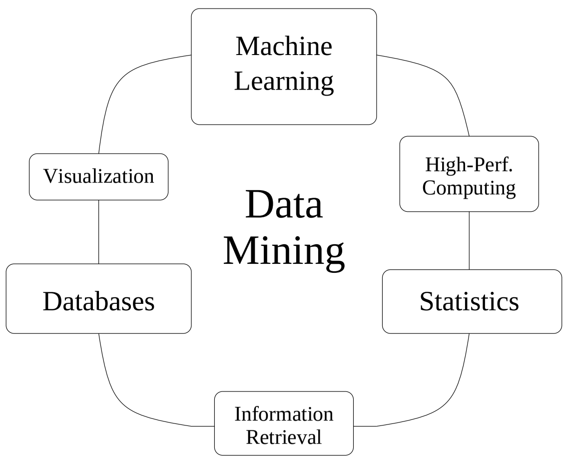
What Is Data Mining?

Data mining is the process of identifying valid, novel, useful and understandable patterns in data.

Also known as **KDD** (<u>K</u>nowledge <u>D</u>iscovery in <u>D</u>atabases).

"We're drowning in information, but starving for knowledge."

John Naisbett



Applications of Data Mining

- E-commerce, marketing, retail, CRM
- Finance
- Telecoms
- World Wide Web
- Drug design
- Process control
- Space and earth sensing
- Etc.

The Data Mining Process

- Understanding domain, prior knowledge, and goals
- Data integration and selection
- Data cleaning and preprocessing
- Modeling and searching for patterns
- Interpreting results
- Consolidating and deploying discovered knowledge
- Loop

Data Mining Tasks

- Classification
- Regression
- Probability estimation
- Clustering
- Association detection
- Summarization
- Trend and deviation detection
- Etc.

Inductive Learning

Inductive learning:

Given examples of a function $\langle X, F(X) \rangle$

Predict function $F(X)$ for new examples X

- Discrete $F(X)$: Classification
- Continuous $F(X)$: Regression
- $F(X) = \text{Probability}(X)$: Probability estimation

Widely-Used Approaches

- Decision trees
- Rule induction
- Instance-based learning (kNN)
- Neural networks
- Bayesian networks
- Etc.

Requirements for a Data Mining System

Data mining systems should be

- Computationally sound
- Statistically sound
- Ergonomically sound

Components of a Data Mining System

- Representation
- Evaluation
- Search
- Data management
- User interface

Topics for this Quarter (Slide 1 of 2)

- Decision trees
- Rule induction
- Instance-based learning
- Bayesian networks
- Neural networks
- Genetic algorithms

Topics for this Quarter (Slide 2 of 2)

- Model ensembles
- Learning theory
- Support vector machines
- Clustering
- Association rules

Projects: Factors to Consider

- How leading-edge?
- How well-defined?
- How open-ended?
- Potential impact?
- How much math & statistics?
- How much implementation work?
- Data available?
- Expertise available?

Project Ideas (Slide 1 of 3)

- Mining multi-relational databases
- Extracting knowledge bases from the World Wide Web
- Mining social networks from the Internet
- Incorporating prior knowledge into data mining algorithms
- Learning models of an adversary's behavior

Project Ideas (Slide 2 of 3)

- Mining sequences of program versions to automate debugging
- Merging and refining databases for data mining
- Predicting the evolution of scientific communities
- Simultaneously classifying sets of related entities
- Discovering unobserved relations from observable ones

Project Ideas (Slide 3 of 3)

- Learning across different subpopulations
- Learning given a query distribution
- Modeling the hidden Web through querying
- Modeling Web surfing behavior
- Scaling up learning algorithms via sampling and compression