

# Data Mining

---

CS57300

Purdue University

December 9, 2010

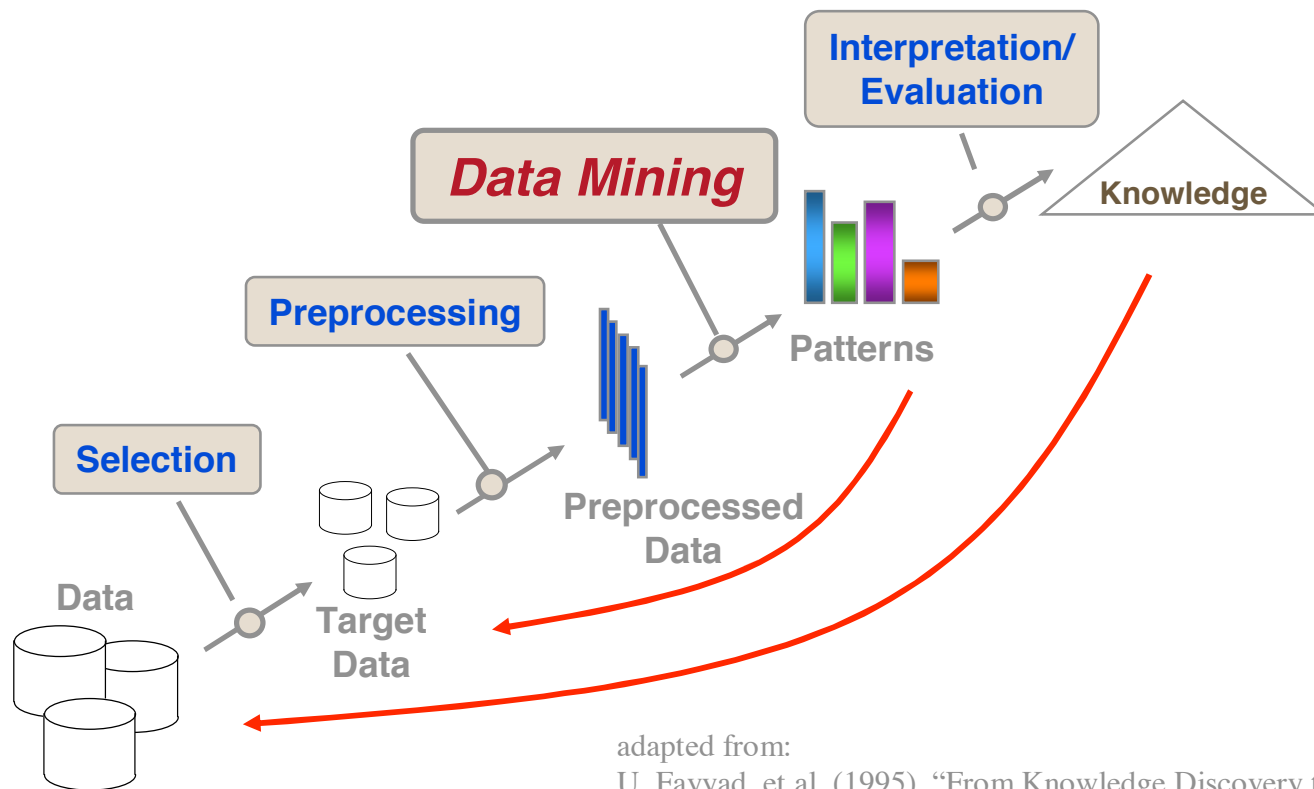
Final review

# Topics

---

- Elements of data mining algorithms
- Data preparation and exploration
- Statistical foundations
- Predictive modeling
- Descriptive modeling
- Pattern mining
- Anomaly detection
- Data mining in practice

# Elements of data mining



adapted from:  
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

# Elements of data mining

---

- Task specification
- Data representation
- Knowledge representation
  - Defines a set of possible models or patterns
- Learning technique
  - Search: Method for generating possible models/patterns and optimizing their score
  - Scoring: Associates a numerical score with each possible model/pattern
- Inference and/or interpretation

Basics

# Statistics

---

- Bayesian vs. Frequentist
- Random variables and common distributions
- Expectation, variance, independence, conditional independence
- Populations and samples
- Properties of estimators (bias and variance)
- Parameter estimation: sufficiency, MLE and MAP
- Parameter estimation vs. structure learning

# Data

---

- Measurement
- Noise and outliers
- Similarity and distance measures
- Exploratory data analysis and visualization
- Dimensionality reduction

# Hypothesis testing

---

- Decision making: biases and heuristics
- Sampling distributions
- Hypothesis testing
- Type I and type II error, statistical power
-

Predictive modeling

# Predictive models

---

- Predictive models predict the value of one variable of interest given known values of other variables
  - Focus on modeling conditional distribution  $P(Y | X)$  or decision boundary for  $Y$
- Data representation: training set of  $\mathbf{x}(i), y(i)$  pairs
- Task: estimate a function  $y=f(\mathbf{x};\theta)$  which maps observed  $x$  values to  $y$  value

# Topics

---

- Discriminative vs. generative
- Parametric vs. non-parametric
- Ensembles
- Search:
  - Set of states (defined by knowledge representation)
  - Set of search operators (actions to move in state space)
  - Search algorithm (input state, choice of actions, stopping criterion)
- Scoring function:
  - Internal: associate score with state for use in search
  - External: measure quality of pattern/model

# Examples

---

- Classification or decision trees
- Nearest neighbor classifiers
- Perceptron classifiers
- Naive Bayes classifiers
- Support vector machines

# External evaluation

---

- Score functions
  - Zero-one loss, squared loss, etc.
- Methodology
  - Disjoint training and test sets
  - Learning curves
  - Cross validation
  - Paired t-test to assess significance
- Other approaches
  - Cost-sensitive analysis, ROC analysis, bias-variance

# Pathologies

---

- Overfitting
- Oversearching
- Attribute selection errors
- Multiple comparison problems

Descriptive modeling

# Descriptive models

---

- Goal is to summarize the data
  - Global summary
  - Model main features of the data
- Data representation: training set of  $x^{(i)}$  instances
- Task—depends on approach
  - Clustering: partition the instances into groups of similar instances
  - Density estimation: determine a compact representation of the full joint distribution  $P(\mathbf{X})=P(X_1, X_2, \dots, X_p)$

# Algorithms

---

- K-means clustering
- Spectral clustering
- Nearest neighbor clustering
- Mixture models
  - Expectation maximization
- Graphical models (Bayes networks and Markov networks)
  - Structure learning
  - Inference (exact and approximate)

Pattern mining

# Pattern mining

---

- Models vs. patterns
  - Pattern characterize local aspects of data
- Task: find descriptive associations between variables
- Algorithms
  - Association rules
  - Graph mining

Anomaly detection

# Anomaly detection

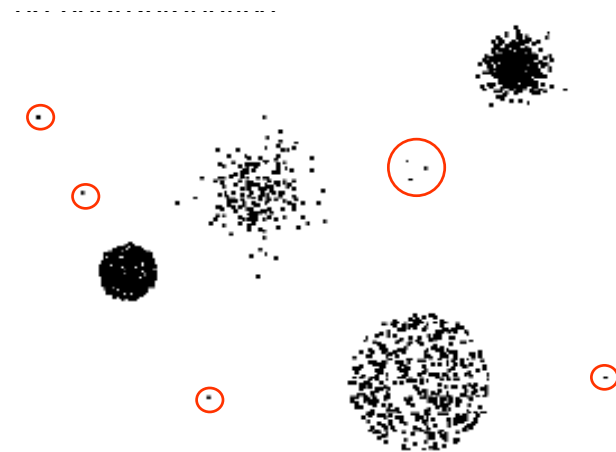
---

- Find data points that are considerably different from the remainder of the data
- Approaches
  - Supervised
    - Similar to classification with imbalanced classes
  - Semi-supervised
    - Labels available only for normal data
  - Unsupervised
    - No labels -- based on the assumption that anomalies are very rare compared to normal data

# Unsupervised (point) anomaly detection

---

- General method
  - Build a profile of “normal” behavior based on patterns or summary statistics for the overall population
  - Use deviations from “normal” to detect anomalies
- Types of methods
  - Visual and statistical-based
  - Distance-based
  - Model-based
  - Projection-based/spectral decomposition



# Data mining in practice

# Topics

---

- How to choose a data mining system
- Top ten data mining mistakes
- Myths and pitfalls

# Continuing with DM/ML

---

- Courses: Spring 2010
  - CS54701: Information Retrieval (*Luo Si*)
  - CS59000-SML: Statistical Machine Learning (Alan Qi)
  - STAT598A: Introduction to Machine Learning (*SVN Vishwanathan*)
- Machine Learning group
  - <https://learning.stat.purdue.edu/wiki/sml/start>
  - Join mlstudents mailing list
  - Attend weekly machine learning seminar

Thanks!