



Principles of Data Mining by Hand&Mannila&Smyth

— Slides for Textbook —

©Ari Visa, , Institute of Signal Processing
Tampere University of Technology



Differences between Data Mining (Han&Kamber) and Principles of Data Mining (Hand&Mannila&Smith)

- 1. Introduction
- 2. Measurement and Data
- 3. Visualizing and Exploring Data
- 4. Data Analysis and Uncertainty
- 5. A Systematic Overview of Data Mining Algorithms
- 6. Models and Patterns

Differences between Data Mining (Han&Kamber) and Principles of Data Mining (Hand&Mannila&Smith)

- 7. Score Functions for Data Mining Algorithms
- 8. Search and Optimization Methods
- 9. Descriptive Modeling
- 10. Predictive Modeling for Classification
- 11. Predictive Modeling for Regression

Differences between Data Mining (Han&Kamber) and Principles of Data Mining (Hand&Mannila&Smith)

- 12. Data Organization and DataBases
- 13. Finding Patterns and Rules
- 14. Retrieval by Content



1. Introduction

- More statistical view
- Han & Kamber is more oriented towards computational and data-management issues

- A global model \sim a local pattern
- Exploratory Data Analysis: to explore the data without any clear ideas of what we are looking for. Typically, techniques are interactive and visual.



1. Introduction

- **Descriptive Modeling:** to describe all of the data. Density estimation, cluster analysis, segmentation, dependency modeling are examples.
- **Predictive Modeling:** Classification and Regression, to build a model that will permit the value of one variable to be predicted from known values of other variables.



1. Introduction

- Discovering Patterns and Rules: are concerned with pattern detection, association rules.
- Retrieval by Content: the user has a pattern of interest and wishes to find similar patterns in the data set.



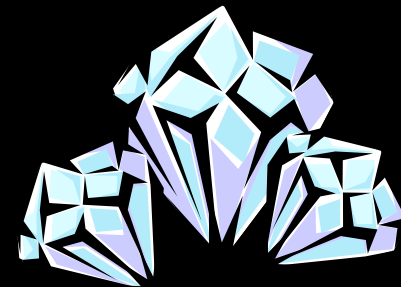
1. Introduction

- Components of Data Mining
 - Model/Pattern Structure: determining the underlying structure or functional forms that we seek from the data.
 - Score Function: judging the quality of a fitted model.
 - Optimization and Search Method: optimizing the score function and searching over different model and pattern structures.
 - Data Management Strategy: handling data access efficiently during the search/optimization



2. Measurement and Data

- Types of measurement
- Distance measures
- Transforming data
- The form of data
- Data quality for individual measurements
- Data quality for collections of data





3. Visualizing and Exploring Data

- Summarizing data
- Tools for displaying single variables
- Tools for displaying relationships between two variables
- Tools for displaying more than two variables
- Multidimensional scaling

4. Data Analysis and Uncertainty



- Dealing with uncertainty
- Samples and statistical inference
- Estimation
 - Maximum Likelihood Estimation
 - Bayesian Estimation
- Hypothesis testing
- Sampling methods



5. A Systematic Overview of Data Mining Algorithms

- The data mining task the algorithm is used to address
- The structure of the model/pattern we are fitting to the data
- The score function we are using to judge the quality of our fitted models/patterns based on observed data
- The search/optimization method we use to search over parameters and structures
- The data management technique to be used for storing, indexing, and retrieving data

5. A Systematic Overview of Data Mining Algorithms

	Classification and Regression trees CART	Backpropagation	A Prior
Task	Classification and Regression	Regression	Rule Pattern Discovery
Structure	Decision Tree	Neural Network	Association Rules
Score Function	Cross-validated Loss Function	Squared Error	Support/Accuracy
Search Method	Greedy Search over Structures	Gradient Descent on Parameters	Breadth-First with Pruning
Data Management Technique	Unspecified	Unspecified	Linear Scan

6. Models and Patterns



- Fundamentals of Modeling
- Model structures for prediction
 - regression models with linear structure
 - local piecewise model structures for regression
 - nonparametric “memory-based” local methods
 - stochastic components of model structures
 - predictive models for classification



6. Models and Patterns

- Models for probability distributions and density functions
 - parametric models: where a particular functional form is assumed.
 - nonparametric models: where the distribution or density estimate is data-driven and relatively few assumptions are made a priori about the functional form.
 - mixtures of parametric models
 - joint distributions for unordered categorical data



6. Models and Patterns

- The curse of dimensionality:
 - factorization and independence in high dimensions
 - variable selection for high-dimensional data
 - transformations for high-dimensional data
- Models for structured data
 - Markov models
- Pattern structures

7. Score Functions for Data Mining Algorithms



- Scoring patterns
- Predictive versus descriptive score functions
- Scoring models with different complexities
- Evaluation of models and patterns
- Robust methods

8. Search and Optimization Methods



- Searching for models and patterns
 - the state-space formulation for search in data mining
 - a simple greedy search algorithm
 - systematic search and search heuristics
 - branch-and-bound
- Parameter optimization methods
 - closed form and linear algebra methods
 - gradient-based methods for optimizing smooth functions
 - univariate parameter optimization
 - multivariate parameter optimization
 - constrained optimization



8. Search and Optimization Methods

- Optimization with missing data -> the Expectation-Maximization algorithm
- Online and single-scan algorithms
- Stochastic search and optimization techniques
 - genetic search
 - simulated annealing



9. Descriptive Modeling

- Describing data by probability distributions and densities
 - score functions for estimating probability distributions and densities
 - parametric density models
 - mixture distributions and densities
 - the EM algorithm for mixture models
 - nonparametric density estimation
 - joint distributions for categorical data
- Partition-based clustering algorithms
 - score functions for partition-based clustering
 - basic algorithms for partition-based clustering



9. Descriptive Modeling

- Hierarchical clustering
 - agglomerative methods
 - divisive methods
- Probabilistic model-based clustering using mixture models

10. Predictive Modeling for Classification



- Discriminative classification and decision boundaries
- Probabilistic models for classification
- The perceptron
- Tree models
- Nearest Neighbor Methods
- Logistic Discriminant Analysis
- The Naïve Bayes Model

11. Predictive Modeling for Regression



- Linear models and least squares fitting
- Generalized linear models
- Artificial neural networks
- Other highly parameterized models
 - generalized additive models
 - projection pursuit regression

12. Data Organization and Databases



- Memory hierarchy
- Index structures
 - B-trees
 - Hash indices
- Multidimensional indexing
 - R*-tree
- Relational databases



12. Data Organization and Databases

- Manipulating tables
 - Union
 - Intersection
 - Difference
 - Projection
 - Selection
 - Join
- The Structured Query Language (SQL)
- Query execution and optimization



12. Data Organization and Databases

- Data warehousing and online analytical processing (OLAP)
- Data structures for OLAP
- String Databases
- Massive Data Sets, Data Management, and Data Mining

force the data into main memory
scalable versions of data mining algorithms
special-purpose algorithms for disk access
pseudo data sets and sufficient statistics



13. Finding Patterns and Rules

- Rule presentations
- Frequent itemsets and association rules
- Generalizations
- Finding episodes from sequences
- Selective discovery of patterns and rules
- From local patterns to global models
- Predictive rule induction



14. Retrieval by Content

- Evaluation of retrieval systems
- Text retrieval
- Modeling individual preferences
- Image retrieval
- Time series and sequence retrieval



8009065 Course Schedule

- Chapter 1. Introduction {L2}
- Chapter 2. Data warehousing and OLAP technology for data mining {L3}
- Chapter 3. Data preprocessing {L3}
- Chapter 4. Data mining primitives, languages and system architectures {L4}
- Chapter 5. Concept description: Characterization and comparison {L5}
- Chapter 6. Mining association rules in large databases {L6}

Homework distribution (DecisionMaster)

- Chapter 7. Classification and prediction { L7}, Midterm
- Chapter 8. Clustering analysis {L8}
- Chapter 9. Mining complex types of data {L9}
- Chapter 10. Data mining applications and trends in data mining {L10}
- The other book {L11}
- The other book {L12}
- The other book {L13}
 - Course project due
- Repetition and feedback {L14}
- Final Exam