Data Mining Overview

Motivation

- We live in a world overwhelmed by data.
- Data accumulates faster than humans can interpret.
- Hidden patterns = valuable information.
- Machine learning & data mining automate pattern discovery.

Data Mining Defined

Data mining = discovering meaningful patterns in large datasets

- Must be automatic or semi-automatic
- Patterns must be useful
- Patterns must support prediction
- Data volume usually large

Black Box vs Transparent Box

- **Black box:** Makes predictions, but reasoning opaque.
- **Transparent box:** Uses structural patterns (rules, trees) humans can understand.

Concepts

- A **concept** is the object of learning.
- A **concept description** is the resulting model.
- Four learning styles:
  - Classification
  - Numeric prediction
  - Association
  - Clustering

Association Learning

- No class label
- Find any interesting relationships
- Often many rules -> filter by frequency and accuracy

Clustering

- No labels
- Group by similarity

Data Organization

- Instances
  - A row of data representing a single example
  - Values assigned to attributes
- Relations & Denormalization
  - Some problems involve relationships, not independent rows
  - Must flatten relational data before learning
- Multi-instance Learning
  - One training example contains multiple sub-instances
  - Example: drug molecules with multiple shapes

Attribute Types

- Numeric
- Nominal
- Ordinal
- Interval
- Ratio

Attribute Levels Table

Type	Example	Allowed Ops
Nominal	sunny/rainy	equality
Ordinal	cool $<$ mild $<$ hot	comparison
Interval	dates	addition/subtraction
Ratio	distance	full arithmetic

Machine Learning vs Statistics

- Statistics: hypothesis testing
- ML: search through model space

Generalization as Search

Model learning = search + bias + generalization.

1. All Possible Models
2. Apply Constraints
3. Search / Heuristics
4. Candidate Models
5. Prune Overfit Models
6. Final Model

Bias Types

- Language bias: the concept description language
- Search bias: the order in which the space is searched
- Overfitting avoidance: the way that overfitting to the training data is avoided

Fielded Applications

- Web mining
- Credit approval
- Oil-spill detection
- Power load forecasting
- Industrial diagnostics
- Marketing & churn analysis

Ethics in Data Mining

- Discrimination concerns
- Reidentification risks
  - Example: ZIP + birthdate + sex identifies 85% of Americans
- Responsible use of personal data

Summary

- Data mining extracts valuable patterns
- Structural models support understanding
- Bias enables learning
- ML widely applied across industries
- Ethics essential