Introduction to Data Mining and Business Intelligence

What Are Data Mining and Business Intelligence?

- **Data mining** is the process of discovering hidden patterns in data, where
  - **Patterns** refer to inherent relationships and/or dependencies in the data, and
  - **Data** are typically stored in a database environment and are large in scale.
- Data mining is also referred to as *knowledge discovery in database* (KDD).
- **Business intelligence** is the transformation of raw data into knowledge and insight for making better business decisions.

Related Fields

- Data mining/analytics is closely related to the fields of database, artificial intelligence, statistics, and information retrieval. But there are considerable differences between data mining and these fields.
- Database: Focuses on data storage and access technology, while data mining focuses on data analysis and knowledge discovery.
- Artificial Intelligence (AI): There are overlaps between data mining and AI (including machine learning) techniques. However, AI techniques are not necessarily data-oriented (e.g., expert systems).
- Statistics:
  - Statistical science assumes data are scarce; it focuses on numeric data and parametric approach (e.g., assume data follows normal distribution).
  - Data mining assumes data are abundant; it deals with various data types and focuses on efficient algorithms for large-scale data.
- Information retrieval (IR) concerns finding materials (e.g., documents) of an unstructured nature (e.g., text) that satisfy an information need; it is closely related to text and web mining. A typical example of IR techniques is search engine.

Terminology

- **Attribute** (also called *variable* or *field*).
    - **Numeric** attribute (also called *continuous* or *real* attribute): Mathematical operations (e.g., addition, multiplication) can be applied to the values of this type of attribute.
    - **Categorical** attribute (also called *nominal* attribute): Mathematical operations cannot be applied to the values of such attribute, even if the values appear in a numeric format (e.g., social security number, credit card number).
- **Record** (also called *observation* or *instance*).
- Dataset (relation or relational database table): A set of data with attributes in columns and records in rows.

Business Applications

- Database marketing
- Credit evaluation
- Fraud detection
- Market basket analysis
- Market segmentation
- Web usage mining and personalization

Data Mining Tasks

- **Supervised learning** (where there is a predefined attribute whose values are to be predicted)
    - Classification
    - Prediction (of numeric values)
- **Unsupervised learning** (where there is no predefined attribute for prediction)
    - Association (or Association Rules Mining)
    - Clustering (or Cluster Analysis)

Classification

- **Classification** is the process of assigning data records into one of several predefined groups, called *classes*. Classification involves building a model (called *classifier*), which can be a mathematical function, a set of rules, or other representations.
- Examples of classification:
  - Fraud detection (true or false)
  - Security trading decision (buy, sell, or hold)
  - Medical diagnosis (presence or absence of a disease)

Prediction

- **Prediction** discovers the relationship between a set of variables, called *predictor* (or *input* or *independent*) variables, and another set of variables, called *target* (or *output* or *dependent*) variables in data, so that the past or current values of predictor variables can be used to predict the future values of target variables.
- Prediction vs. classification:
  - Prediction: the values of the attribute to be predicted (target variable) is numeric.
  - Classification: the values of the attribute to be predicted (class attribute) is categorical.
- Examples of prediction:
  - Sales volume or revenue prediction
  - Stock price prediction

Association

- **Association** refers to the presence of one set of items in a group of records implies the existence of another set of items in the same group of records.
- Examples of association:
  - Market basket analysis (i.e., what items were normally bought together in a customer's visit to the store?)
  - Recommender system (Amazon: Customers who bought this item also bought…)
  - Web usage mining (Clickstream analysis. See http://en.wikipedia.org/wiki/Clickstream)

Clustering

- **Clustering** is the process of grouping data records into a number of groups, called *clusters*, such that records within the same cluster are more similar than those between different clusters.
- Clustering differs from classification in that groups are formed as a result of the analysis instead of predefined.
- Examples of clustering:
  - Market segmentation
  - Grouping of library books by field

Data Mining Process

1. Problem identification
   - Purpose of the data-mining project and nature of the problem (classification, prediction, clustering, or association?)
2. Data preparation
   - Data collection: retrieving, merging and/or dividing data
   - Data cleaning: correcting errors, handling missing data, resolving inconsistencies
   - Data reduction: sampling (in rows), feature (attribute) selection (in columns)
   - Data transformation: standardizing data, reformatting data, conversion between numeric and categorical data
3. Model formulation and pattern exploration
   - Select appropriate data-mining techniques and tools and use the selected techniques and tools to build models and explore the patterns/relationships hidden in data.
4. Verification and modification
   - Test if the models built are valid; modify the models if necessary.
   - Compare different candidate models.
5. Interpretation and implementation (deployment)
   - Interpret the results of data mining in an intuitive manner.
   - Implement/deploy the model into related applications.

Big Data

- Big data can be characterized by 3Vs: volume, velocity, and variety.
  - Volume: amount and scale of data.
  - Velocity: speed of data in and out.
  - Variety: range and complexity of data types, structures and sources.
- Examples:
  - Google, Twitter, GPS data, Facebook, YouTube.
  - Financial market: 7 billion shares change hands every day on U.S. markets (MSC p.7).

Small Data vs. Bid Data

- Sampling vs. "N = All"
- Causality vs. Correlation
- Structured (SQL) vs. Unstructured (NoSQL, Not only SQL)
  - *MapReduce* is a framework for processing large-scale data using parallel and distributed computing technologies with a large number of computers. Apache *Hadoop* is an open-source implementation of MapReduce.

Missing Data Replacement

- Listwise deletion: disregard a record if it has any missing attribute values.
- Mean/Mode substitution:
  - For a missing value of a numeric attribute, replace it with the mean of the non-missing values of that attribute.
  - For a missing value of a categorical attribute, replace it with the mode (most frequent value) of the non-missing values of that attribute.
- Task-specific missing value replacement methods (we will discuss some of them later in class)

Normalizing Numeric Data

- Transform a series of numeric data into values within range [0, 1], as below:

$$\text{normalized value} = \frac{\text{original value} - \text{min value}}{\text{max value} - \text{min value}}$$

- Example:

  Original values: −1, 0, 2, and 4.
  Normalized values: 0, 0.2, 0.6, and 1.

Overfitting and Data Partitioning

- **Training** set is the portion of data used to build data-mining models.
- **Validation** set is the portion of data used to validate or adjust the models, and to prevent overfitting problems.
- **Test** set is the data used to evaluate the performance of the models. The test set serves as unseen future data.
- Overfitting occurs when a model fits the training data very well or even perfectly, but performs poorly when it is applied to unseen future data.