

Data Mining Fundamentals

Length: 2 Days

Audience: Application Developers, Database Administrators, Business Analysts, Systems Analysts, Data Modelers, and Managers

Prerequisites: Experience working with a relational database system including a basic understanding of SQL

Overview: Data mining is an interdisciplinary field, employing techniques from areas such as statistics, machine learning, pattern recognition, artificial intelligence, data visualization, and database systems. With advances in technology, it has become increasingly easy to accumulate massive amounts of data. The challenge posed to the knowledge worker is how to make sense of all the data, i.e., how to identify interesting patterns to drive business decisions.

This course provides a survey of the main topics in data mining and knowledge discovery in databases (KDD), including classification, clustering, association rules, and sequence mining. Discussions will include algorithmic and system issues in KDD, as well as applications such as Web mining and multimedia mining.

This course is product-independent, providing a theoretical or conceptual overview of data mining mechanisms applicable to relational database systems. It is intended to prepare students for the challenge of exploring data repositories and performing analysis of discovered patterns.

Topics discussed include:

- Motivation for Data Mining
 - Concepts and Terminology
 - Applications of Data Mining
 - Distinguish Between Query/Reporting, OLAP, and Data Mining
 - Supervised and Unsupervised Learning
 - Interestingness
 - Efficiency and Scalability
- Measurement and Data
 - Types of Measurement
 - Errors of Measurement
 - Summary Statistics
- Security and Privacy
 - Disclosure Problem
 - Inference Problem
 - Randomization
 - Quasi-Identifiers
 - k-anonymity
 - Generalizing Data

- Suppression
- Data Preparation and Quality
 - Dealing with Missing Values and Noisy Data
 - Data Integration
 - Data Selection
 - Data Transformation
- Mining Association Rules
 - Market Basket Analysis
 - Apriori Principle
 - Mining Sequential Patterns
 - Inter-transaction Association Rules
- Classification and Regression
 - Using a Classification Model
 - Decision Tree Induction
 - Hunt's Algorithm
 - Tree Pruning
 - Extracting Classification Rules
 - Predictive Data Mining
- Clustering
 - Types of Data
 - Types of Clusterings
 - k-Means Algorithm
 - Similarity Measure
 - k-Medoids Method