

Data Mining : extraction of relevant and useful hidden patterns from large databases

STATS24x7.com© 2010 ADI-NV, INC.

Knowledge Discovery in Databases (KDD)

- Formalized in 1989
- Process of identifying a valid, useful, and comprehensive structure in data
- Process involves:
 - Sampling data from data warehouse
 - Data cleansing
 - Reducing/transforming data
 - Applying an appropriate DM method to derive structure in data

STATS24x7.com© 2010 ADI-NV, INC.

KDD

- Selection
- Reprocessing
- Transformation

DM

- Extracts patterns
- Patterns interpreted/evaluated
- Data visualization can help in finding patterns/displaying DM results

STATS24x7.com© 2010 ADI-NV, INC.

DM draws from:

- Statistics
 - One of the key disciplines on which DM is built
- Machine Learning
 - Automated learning process which constructs rules from observation.
 - Inductive Learning (system infers knowledge itself from observing its environment)
 - Supervised Learning: learn from training sets
 - Unsupervised Learning: learn from observations and discovery
- Mathematical Programming (e.g. Linear Programming, Quadratic Programming, etc.)
 - Most of DM tasks can be formulated as a math programming problem. Support Vector Machines (SVM) falls in this category.

STATS24x7.com© 2010 ADI-NV, INC.

Discovery Driven DM Tasks

- Discovery of
 - Association Rules
 - Classification Rules
 - Clustering
 - Frequent episodes (temporal)
 - Deviation detection- outliers
 - Neural Networks
 - Genetic Algorithms
 - Support Vector Machines
 - Sequence Mining (Temporal DM)
 - Discovery of temporal sequences of events
 - Trend discovery
 - Web Mining
 - Use of DM to automatically discover and extract information from the web
 - Text Mining
 - Finding patterns in text databases
 - Spatial DM
 - DM on spatial data

STATS24x7.com© 2010 ADI-NV, INC.

Common Data Mining Techniques

- Predictive: predicts unknown or future values of interest
Examples: MLR using OLS, Weighted LS, Ridge Regression, Binary Logistic Regression, Poisson Regression, Multinomial Logistic Regression, Neural network, CART (Classification and Regression Trees), CHAID (Chi Square Automatic Interaction Detection), Time Series Methods
- Descriptive: finds patterns in data
EDA (Exploratory Data Analysis), Cluster analysis, discriminant analysis (logistic regression can be used)

STATS24x7.com© 2010 ADI-NV, INC.