DATA MINING - THEMES

Theoretical Foundations of Data Mining

The theoretical foundations of data mining includes the following concepts −

- **Data Reduction** − The basic idea of this theory is to reduce the data representation which trades accuracy for speed in response to the need to obtain quick approximate answers to queries on very large databases. Some of the data reduction techniques are as follows −
    - Singular value Decomposition
    - Wavelets
    - Regression
    - Log-linear models
    - Histograms
    - Clustering
    - Sampling
    - Construction of Index Trees
- **Data Compression** − The basic idea of this theory is to compress the given data by encoding in terms of the following −
    - Bits
    - Association Rules
    - Decision Trees
    - Clusters
- **Pattern Discovery** − The basic idea of this theory is to discover patterns occurring in a database. Following are the areas that contribute to this theory −
    - Machine Learning
    - Neural Network
    - Association Mining
    - Sequential Pattern Matching
    - Clustering
- **Probability Theory** − This theory is based on statistical theory. The basic idea behind this theory is to discover joint probability distributions of random variables.
- **Probability Theory** − According to this theory, data mining finds the patterns that are interesting only to the extent that they can be used in the decision-making process of some enterprise.
- **Microeconomic View** − As per this theory, a database schema consists of data and patterns that are stored in a database. Therefore, data mining is the task of performing induction on databases.
- **Inductive databases** − Apart from the database-oriented techniques, there are statistical techniques available for data analysis. These techniques can be applied to scientific data and data from economic and social sciences as well.

Statistical Data Mining

Some of the Statistical Data Mining Techniques are as follows −

- **Regression** − Regression methods are used to predict the value of the response variable from one or more predictor variables where the variables are numeric. Listed below are the forms of Regression −
    - Linear
    - Multiple
    - Weighted
    - Polynomial
    - Nonparametric
    - Robust
- **Generalized Linear Models** - Generalized Linear Model includes −
    - Logistic Regression
    - Poisson Regression

  The model's generalization allows a categorical response variable to be related to a set of predictor variables in a manner similar to the modelling of numeric response variable using linear regression.
- **Analysis of Variance** − This technique analyzes −
    - Experimental data for two or more populations described by a numeric response variable.
    - One or more categorical variables *factors*.
- **Mixed-effect Models** − These models are used for analyzing grouped data. These models describe the relationship between a response variable and some co-variates in the data grouped according to one or more factors.
- **Factor Analysis** − Factor analysis is used to predict a categorical response variable. This method assumes that independent variables follow a multivariate normal distribution.
- **Time Series Analysis** − Following are the methods for analyzing time-series data −
    - Auto-regression Methods.
    - Univariate ARIMA *AutoRegressiveIntegratedMovingAverage* Modeling.
    - Long-memory time-series modeling.

Visual Data Mining

Visual Data Mining uses data and/or knowledge visualization techniques to discover implicit knowledge from large data sets. Visual data mining can be viewed as an integration of the following disciplines −

- Data Visualization
- Data Mining

Visual data mining is closely related to the following −

- Computer Graphics
- Multimedia Systems
- Human Computer Interaction
- Pattern Recognition
- High-performance Computing

Generally data visualization and data mining can be integrated in the following ways −

- **Data Visualization** − The data in a database or a data warehouse can be viewed in several visual forms that are listed below −
    - Boxplots
    - 3-D Cubes
    - Data distribution charts
    - Curves
    - Surfaces
    - Link graphs etc.
- **Data Mining Result Visualization** − Data Mining Result Visualization is the presentation of the results of data mining in visual forms. These visual forms could be scattered plots, boxplots, etc.
- **Data Mining Process Visualization** − Data Mining Process Visualization presents the several processes of data mining. It allows the users to see how the data is extracted. It also allows the users to see from which database or data warehouse the data is cleaned, integrated, preprocessed, and mined.

Audio Data Mining

Audio data mining makes use of audio signals to indicate the patterns of data or the features of data mining results. By transforming patterns into sound and musing, we can listen to pitches and tunes, instead of watching pictures, in order to identify anything interesting.

Data Mining and Collaborative Filtering

Consumers today come across a variety of goods and services while shopping. During live customer transactions, a Recommender System helps the consumer by making product recommendations. The Collaborative Filtering Approach is generally used for recommending products to customers. These recommendations are based on the opinions of other customers.