Data Management & Big Data

Big Data Methodologies

- Associated Rule Learning
- Sentiment Analysis
- Natural Language Processing
- Classification (Supervised Machine Learning)
- Clustering (Unsupervised Machine Learning)
- Statistical Analysis (A/B, MVT testing, Regression)
- Graph Analysis

Associated Rule Learning

Association rule learning is a method for discovering interesting correlations between variables in large databases. It was first used by major supermarket chains to discover interesting relations between products, using data from supermarket point-of-sale (POS) systems.

- Association rule learning is being used to help:
  - place products in better proximity to each other in order to increase sales
  - extract information about visitors to websites from web server logs
  - analyze biological data to uncover new relationships
  - monitor system logs to detect intruders and malicious activity
  - identify if people who buy milk and butter are more likely to buy diapers

Sentiment Analysis

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker.

- Sentiment analysis is being used to help:
  - improve service at a hotel chain by analyzing guest comments
  - customize incentives and services to address what customers are really asking for
  - determine what consumers really think based on opinions from social media

Natural Language Processing

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken. NLP is a component of artificial intelligence.

- The development of NLP applications is challenging because computers traditionally require humans to "speak" to them in a programming language that is precise, unambiguous and highly structured, or through a limited number of clearly enunciated voice commands. Human speech, however, is not always precise -- it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.

Classification (Supervised Machine Learning)

Supervised learning is a type of machine learning algorithm that uses a known dataset (called the training dataset) to make predictions. The training dataset includes input data and response values. From it, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new dataset. A test dataset is often used to validate the model. Using larger training datasets often yield models with higher predictive power that can generalize well for new datasets.

- Common classification algorithms include:
- Support vector machines (SVM)
- Neural networks
- Naïve Bayes classifier
- Decision trees
- Discriminant analysis
- Nearest neighbors (*k*NN)

Clustering (Unsupervised Machine Learning)

Cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise.

- A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

- K-means clustering method.

Statistical Analysis (A/B & MVT testing)

A/B testing is a method of comparing two versions of a webpage or app against each other to determine which one performs better. AB testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

Multivariate testing is a technique for testing a hypothesis in which multiple variables are modified. The goal of multivariate testing is to determine which combination of variations performs the best out of all of the possible combinations.

Graph Analytics

Graph analytics leverage graph structures to understand, codify, and visualize relationships that exist between people or devices in a network. Graph analytics, built on the mathematics of graph theory, is used to model pairwise relationships between people, objects, or nodes in a network

- Spotting fraud, which applies to fraudulent transactions and applications in banking, benefits fraud in government, applications and claims fraud in insurance and fraudulent activities in telecommunications
- Preventing crime and performing counterterrorism
- Applying influencer analysis in social network communities

Supervised vs. Unsupervised

- **Supervised**: All data is labeled and the algorithms learn to predict the output from the input data.
- **Unsupervised**: All data is unlabeled and the algorithms learn to inherent structure from the input data.
- **Semi-supervised**: Some data is labeled but most of it is unlabeled and a mixture of supervised and unsupervised techniques can be used.