An Introduction to Data Mining

Joseph Heng

Intel Beijing

wei.heng@intel.com

January 17, 2014

Outline

1. DW Overview
   - What is Data Mining
   - Notable Application of Data Mining
   - Conference, Software and Applications
   - Major Process in Data Mining
2. Major Tasks in Detail
   - Regression
   - Classification
   - Clustering
   - Summarization
3. Future Reading
   - Recent Evolution
   - Reference

What is Data Mining

Data Mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistic and database system.
The actual task is the automatic or semi-automatic analysis of big data to extract **unknown interesting patterns**, such as groups of data records(cluster analysis), unusual records(anomaly detection), dependencies(association rule mining).

1. Early methods of identifying patterns in data include **Bayes' theorem** (1700s) and **regression analysis** (1800s).
2. Aided by other discoveries in computer science, such as **neural networks**, **cluster analysis**, **genetic algorithms** (1950s), **decision trees** (1960s), and **support vector machines** (1990s).

Notable Application of Data Mining

- **Games** since the early 1960s, with the availability of oracles for certain combinatorial games, extraction of human-usable strategies from these oracles.
- **Business** analysis of historical business activities, to reveal hidden patterns,trends and unknown strategic business information.such as CRM, Customer Behavior, return on investment.
- **Science and Engineering** bioinformatics, genetics, medicine, education and electrical power engineering.
- **Others** Medical, Sensor, Music, Visual data Mining.

Conference and Research

- KDD Conference ACM SIGKDD Conference on Knowledge Discovery and Data Mining.
- MLDM Conference Machine Learning and Data Mining in Pattern Recognition.
- WSDM Conference ACM Conference on Web Search and Data Mining.

Software and Applications

Free open-source data mining software and applications

- NLTK (Natural Language Toolkit) A suite of libraries and programs for symbolic and statistical NLP for the Python.
- R A programming language and software environment for statistical computing, data mining and graphics.
- Weka A suite of machine learning software applications written in the Java.

Commercial data-mining software and applications

- IBM SPSS Modeler data mining software provided by IBM.
- Microsoft Analysis Services data mining software Microsoft.
- Oracle Data Mining data mining software by Oracle.
- Google Predict API Google's cloud-based machine learning tools.

Major Process in Data Mining

- Pre-processing Before data mining algorithms can be used, a target data set must be assembled.As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit.
- Data mining Data mining involves six common classes of tasks:Anomaly detection, Association rule learning, Clustering, Classification, Regression, Summarization.
- Results validation verify that the patterns produced by the data mining algorithms occur in the wider data set. such as overfitting and underfitting.

Regression Analysis

estimating the relationships among variables, it estimates the conditional expectation of the dependent variable given the independent variables.

common used for:

1. prediction and forecasting.
2. explore the forms of these relationships among the independent and dependent variables.

common techniques:

1. linear regression or ordinary least squares.
2. non-linear regression, such as polynomial regression.

Regression Analysis

illustration of linear regression:

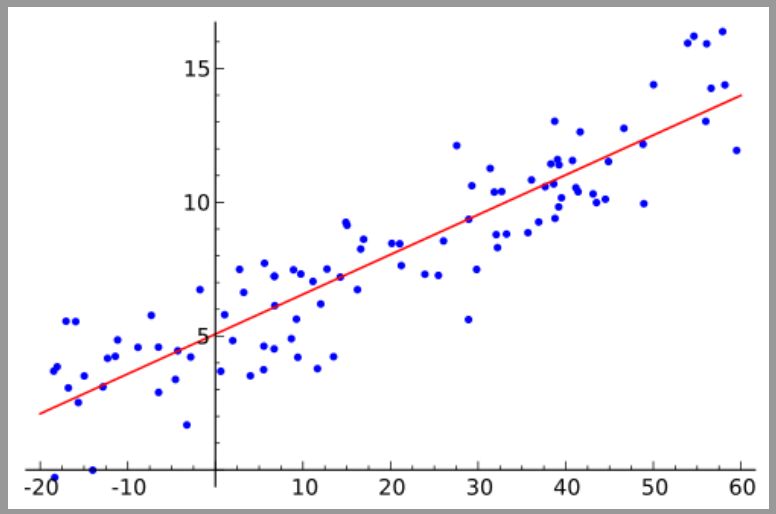

Figure : simple linear regression

Minimizing Errors

an error is the distance between the actual and model data.

- Actual Data: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Error: $\varepsilon_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
- Sum of Errors: $S = \sum \varepsilon_i^2$
- Best Fit Y-Intercept: $\hat{\beta}_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$

Figure : error function

Other Type of Regression

what if your data is not a straight line ?

- Logarithmic Regression
  $Y = a + b * (\ln(x))$
- Quadratic Regression
  $Y = a * x^2 + b * x + c$
- Power Regression
  $Y = a * x^b$
- Exponential Regression
  $Y = a * b^x$

Classification

classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

- Bayesian procedure and discrete procedure.
- Binary classification and Multiclass classification.

Common Algorithms

Classification Algorithms

- Linear classifiers
  - Logistic Regression
  - Naive Bayes Classifier
  - Perceptron
- Support Vector Machines
- Kernel estimation
  - k-nearest neighbor
- Boosting (meta-algorithm)
- Decision trees
  - ID3 algorithm
  - C4.5 algorithm
- Neural networks
- Bayesian networks
- Hidden Markov models

Clustering

Cluster Analysis
clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).
Cluster analysis itself is not one specific algorithm, but the general task to be solved.

Clustering Algorithm

Clustering Algorithm

- Connectivity models for example hierarchical clustering, hierarchical latent dirichlet process, hierarchical dirichlet process.
- Centroid models for example k-means algorithm represents each cluster by a single mean vector.
- Distribution models for example topic modeling based LDA, hLDA rCRP etc.
- Graph-based models commonly also known as Bayesian Network, or Probabilistic Graphic Model.

Summarization

the meaning of summarization is general, and mainly contains two aspects: visualization and report generation.

Common Visualization Tools

- Matlab
- R language
- Octave
- Dot language and graphize tools package.

Summarization

Here we lay attention on automatic summarization techniques.
Automatic summarization *is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document*
Generally, there are two approaches to automatic summarization: extraction and abstraction. And the most used is extraction based techniques, which scores every sentences in documents, and ranking them according to scors rank:

- Graph based
- Topic Modeling based
- linguistics rule based.

Recent Research Direction

- hierarchical topic modeling, such as HDP, rCRP, hLDA.
- multi-layer neural network based, such as Deep Learning.

Reference

John Smith (2012)
History of Data Mining
Association of Compuationall Linguistics 12(3), 45 – 678.

The Beginning