大數據分析
Big Data Analysis (IM EMBA, TKU)
鄭啟斌 教授
資料探勘介紹
(Introduction to Data Mining)

Time: 2015/10/12 (19:20-22:10)
Place: D325

Min-Yuh Day
戴敏育
Assistant Professor
Dept. of Information Management, Tamkang University

http://mail.tku.edu.tw/myday/
2015-10-12
Outline

• Data Mining and Big Data Analytics
• Data Mining Process
• Data Mining Tasks
• Data Mining Evaluation
• Social Network Analysis

Source: Han & Kamber (2006)
Data Mining
and
Big Data Analytics
Stephan Kudyba (2014), *Big Data, Mining, and Analytics: Components of Strategic Decision Making*, Auerbach Publications

Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Data Mining

Web Mining and Social Networking

Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites

Source: http://www.amazon.com/Mining-Social-Web-Analyzing-Facebook/dp/1449388345
ENTERPRISE ANALYTICS

Optimize Performance, Process, and Decisions through Big Data

EDITED BY THOMAS DAVENPORT

Big Data: The Management Revolution

Exploiting vast new flows of information can radically improve your company’s performance. But first you’ll have to change your decision-making culture.

by Andrew McAfee and Erik Brynjolfsson

Architecture of Big Data Analytics

Big Data Sources
- Internal
- External
- Multiple formats
- Multiple locations
- Multiple applications

Big Data Transformation
- Middleware
- Extract Transform Load
- Data Warehouse
- Traditional Format CSV, Tables

Transformed Data
- Hadoop
- MapReduce
- Pig
- Hive
- Jaql
- Zookeeper
- Hbase
- Cassandra
- Oozie
- Avro
- Mahout
- Others

Big Data Platforms & Tools

Big Data Analytics Applications
- Queries
- Reports
- OLAP
- Data Mining

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Architecture of Big Data Analytics

Data Mining
Big Data Analytics Applications

Big Data Sources
- Internal
- External
- Multiple formats
- Multiple locations
- Multiple applications

Big Data Transformation

Big Data Platforms & Tools
- Queries
- Reports
- OLAP
- Data Mining

Big Data Analytics Applications

Source: Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
Architecture for Social Big Data Mining
(Hiroshi Ishikawa, 2015)

Enabling Technologies
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distrusted processing

Conceptual Layer
- Integrated analysis
- Multivariate analysis
- Application specific task

Logical Layer
- Data Mining
- Integrated analysis model
- Natural Language Processing
- Information Extraction
- Anomaly Detection
- Discovery of relationships among heterogeneous data
- Large-scale visualization
- Parallel distrusted processing

Analysts
- Model Construction
- Explanation by Model
- Construction and confirmation of individual hypothesis
- Description and execution of application-specific task

Software

Hardware

Social Data

Physical Layer

Source: Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press
Business Intelligence (BI) Infrastructure

Operational Data
Historical Data
Machine Data
Web Data
Audio/Video Data
External Data

Extract, transform, load

Data Mart

Data Warehouse

Hadoop Cluster

Analytic Platform

Casual users
- Queries
- Reports
- Dashboards

Power users
- Queries
- Reports
- OLAP
- Data mining

Data Warehouse

Data Mining and Business Intelligence

Increasing potential to support business decisions

Decision Making

Data Presentation

Visualization Techniques

Data Mining

Information Discovery

Data Exploration

Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources

Paper, Files, Web documents, Scientific experiments, Database Systems

End User

Business Analyst

Data Analyst

DBA

Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier
The Evolution of BI Capabilities

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Business Intelligence and Analytics

• Business Intelligence 2.0 (BI 2.0)
  – Web Intelligence
  – Web Analytics
  – Web 2.0
  – Social Networking and Microblogging sites

• Data Trends
  – Big Data

• Platform Technology Trends
  – Cloud computing platform

Business Intelligence and Analytics: Research Directions

1. Big Data Analytics
   - Data analytics using Hadoop / MapReduce framework

2. Text Analytics
   - From Information Extraction to Question Answering
   - From Sentiment Analysis to Opinion Mining

3. Network Analysis
   - Link mining
   - Community Detection
   - Social Recommendation

Big Data, Big Analytics:
Emerging Business Intelligence and Analytic Trends for Today's Businesses
Big Data, Prediction vs. Explanation

Big Data: The Management Revolution

Business Intelligence and Enterprise Analytics

• Predictive analytics
• Data mining
• Business analytics
• Web analytics
• Big-data analytics

Three Types of Business Analytics

• Prescriptive Analytics
• Predictive Analytics
• Descriptive Analytics

Three Types of Business Analytics

<table>
<thead>
<tr>
<th>Optimization</th>
<th>“What’s the best that can happen?”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized Testing</td>
<td>“What if we try this?”</td>
</tr>
<tr>
<td>Predictive Modeling / Forecasting</td>
<td>“What will happen next?”</td>
</tr>
<tr>
<td>Statistical Modeling</td>
<td>“Why is this happening?”</td>
</tr>
<tr>
<td>Alerts</td>
<td>“What actions are needed?”</td>
</tr>
<tr>
<td>Query / Drill Down</td>
<td>“What exactly is the problem?”</td>
</tr>
<tr>
<td>Ad hoc Reports / Scorecards</td>
<td>“How many, how often, where?”</td>
</tr>
<tr>
<td>Standard Report</td>
<td>“What happened?”</td>
</tr>
</tbody>
</table>

Big-Data Analysis

• Too Big, too Unstructured, too many different source to be manageable through traditional databases
Big Data with Hadoop Architecture

LOGICAL ARCHITECTURE

Processing: MapReduce

Job Tracker
Task Tracker → Mapper
Task Tracker → Mapper
Task Tracker → Mapper

Shuffle and Sort
Reducer → Reducer → Reducer

Storage: HDFS

NameNode
Data Node → BLOCK
Data Node → BLOCK
Data Node → BLOCK

PROCESS FLOW

Input Data Set

Split 0

Map 0 → Reduce 0

Split 1

Map 1

Split n

Map n → Reduce 0

PHYSICAL ARCHITECTURE

Hadoop Cluster

Master
Slave
Slave
Slave
Slave
Slave
Slave
Slave
Slave
Slave

Big Data with Hadoop Architecture

Logical Architecture

Processing: MapReduce

Big Data with Hadoop Architecture

Logical Architecture

Storage: HDFS

Big Data with Hadoop Architecture
Process Flow

Big Data with Hadoop Architecture

Hadoop Cluster

Traditional ETL Architecture

Offload ETL with Hadoop (Big Data Architecture)

Big Data Solution

HDP
A Complete Enterprise Hadoop Data Platform

Source: http://hortonworks.com/hdp/
Data Mining

Advanced Data Analysis

Evolution of Database System Technology
Evolution of Database System Technology

Data Collection and Database Creation
(1960s and earlier)
- Primitive file processing

Database Management Systems
(1970s–early 1980s)
- Hierarchical and network database systems
- Relational database systems
- Query languages: SQL, etc.
- Transactions, concurrency control and recovery
- On-line transaction processing (OLTP)

Advanced Database Systems
(mid-1980s–present)
- Advanced data models: extended relational, object-relational, etc.
  - Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, knowledge-based
  - XML-based database systems
- Integration with information retrieval
- Data and information integration

Advanced Data Analysis:
(late 1980s–present)
- Data mining and knowledge discovery:
  generalization, classification, association, clustering
  - Advanced data mining applications:
    stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.
- Data mining applications
- Data mining and society

New Generation of Information Systems
(present–future)

Source: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier
Internet Evolution

Internet of People (IoP): Social Media
Internet of Things (IoT): Machine to Machine

Data Mining at the Intersection of Many Disciplines

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining Technologies

- Statistics
- Database Systems
- Data Warehouse
- Information Retrieval
- Machine Learning
- Applications
- Pattern Recognition
- Visualization
- Algorithms
- High-performance Computing

Source: Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier
Data Mining Process
Data Mining Process

• A manifestation of best practices
• A systematic way to conduct DM projects
• Different groups has different versions
• Most common standard processes:
  – CRISP-DM
    (Cross-Industry Standard Process for Data Mining)
  – SEMMA
    (Sample, Explore, Modify, Model, and Assess)
  – KDD
    (Knowledge Discovery in Databases)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining Process (SOP of DM)

What main methodology are you using for your analytics, data mining, or data science projects?

# Data Mining Process

<table>
<thead>
<tr>
<th>Methodology</th>
<th>2014 Poll</th>
<th>2007 Poll</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRISP-DM (86)</td>
<td>43%</td>
<td>42%</td>
</tr>
<tr>
<td>My own (55)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEMMA (17)</td>
<td>8.5%</td>
<td>13%</td>
</tr>
<tr>
<td>Other, not domain-specific (16)</td>
<td>8%</td>
<td>4%</td>
</tr>
<tr>
<td>KDD Process (15)</td>
<td>7.5%</td>
<td>7.3%</td>
</tr>
<tr>
<td>My organizations' (7)</td>
<td>3.5%</td>
<td>5.3%</td>
</tr>
<tr>
<td>A domain-specific methodology (4)</td>
<td>2%</td>
<td>4.7%</td>
</tr>
<tr>
<td>None (0)</td>
<td>0%</td>
<td>4.7%</td>
</tr>
</tbody>
</table>

Data Mining: Core Analytics Process

The KDD Process for Extracting Useful Knowledge from Volumes of Data

Data Mining

Knowledge Discovery in Databases (KDD) Process

(Fayyad et al., 1996)

Knowledge Discovery in Databases (KDD) Process

Data mining: core of knowledge discovery process

Data Warehouses

Cleaning and Integration

Selection and Transformation

Data Mining

Task-relevant Data

Evaluation and Presentation

Patterns

Knowledge

Source: Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Second Edition, Elsevier
Data Mining Process: CRISP-DM

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining Process: CRISP-DM

Step 1: Business Understanding
Step 2: Data Understanding
Step 3: Data Preparation (!)
Step 4: Model Building
Step 5: Testing and Evaluation
Step 6: Deployment

• The process is highly repetitive and experimental (DM: art versus science?)

Accounts for ~85% of total project time

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Preparation – A Critical DM Task

Real-world Data

Data Consolidation
- Collect data
- Select data
- Integrate data

Data Cleaning
- Impute missing values
- Reduce noise in data
- Eliminate inconsistencies

Data Transformation
- Normalize data
- Discretize/aggregation data
- Construct new attributes

Data Reduction
- Reduce number of variables
- Reduce number of cases
- Balance skewed data

Well-formed Data

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining Process:

**SEMMA**

- **Sample**
  (Generate a representative sample of the data)

- **Modify**
  (Select variables, transform variable representations)

- **Explore**
  (Visualization and basic description of the data)

- **Assess**
  (Evaluate the accuracy and usefulness of the models)

- **Model**
  (Use variety of statistical and machine learning models)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining Processing Pipeline
(Charu Aggarwal, 2015)

Data Collection → Data Preprocessing
  - Feature Extraction
  - Cleaning and Integration
  → Analytical Processing
    - Building Block 1
    - Building Block 2
  → Output for Analyst

Feedback (Optional)

Source: Charu Aggarwal (2015), Data Mining: The Textbook Hardcover, Springer
Using Databases to Improve Business Performance and Decision Making

• **Big data**
  – Massive sets of unstructured/semi-structured data from Web traffic, social media, sensors, and so on
  – Petabytes, exabytes of data
    • Volumes too great for typical DBMS
  – Can reveal more patterns and anomalies
Using Databases to Improve Business Performance and Decision Making

• **Business intelligence infrastructure**
  – Today includes an array of tools for separate systems, and big data

• **Contemporary tools:**
  – Data warehouses
  – Data marts
  – Hadoop
  – In-memory computing
  – Analytical platforms

Data Warehouse vs. Data Marts

• **Data warehouse:**
  – Stores current and historical data from many core operational transaction systems
  – Consolidates and standardizes information for use across enterprise, but data cannot be altered
  – Provides analysis and reporting tools

• **Data marts:**
  – Subset of data warehouse
  – Summarized or focused portion of data for use by specific population of users
  – Typically focuses on single subject or line of business

Hadoop

- Enables distributed parallel processing of big data across inexpensive computers
- Key services
  - Hadoop Distributed File System (HDFS): data storage
  - MapReduce: breaks data into clusters for work
  - Hbase: NoSQL database
- Used by Facebook, Yahoo, NextBio

In-memory computing

- Used in big data analysis
- Use computers main memory (RAM) for data storage to avoid delays in retrieving data from disk storage
- Can reduce hours/days of processing to seconds
- Requires optimized hardware

Analytic platforms

• High-speed platforms using both relational and non-relational tools optimized for large datasets

• Examples:
  – IBM Netezza
  – Oracle Exadata

Analytical tools:
Relationships, patterns, trends

• Business Intelligence Analytics and Applications
• Tools for consolidating, analyzing, and providing access to vast amounts of data to help users make better business decisions
  – Multidimensional data analysis (OLAP)
  – Data mining
  – Text mining
  – Web mining

Online analytical processing (OLAP)

• Supports multidimensional data analysis
  – Viewing data using multiple dimensions
  – Each aspect of information (product, pricing, cost, region, time period) is different dimension
  – Example: How many washers sold in East in June compared with other regions?

• OLAP enables rapid, online answers to ad hoc queries

MULTIDIMENSIONAL DATA MODEL

Data mining

• Finds hidden patterns, relationships in datasets
  – Example: customer buying patterns

• Infers rules to predict future behavior
  – Data mining provides insights into data that cannot be discovered through OLAP, by inferring rules from patterns in data.

Types of Information Obtained from Data Mining

- **Associations**: Occurrences linked to single event
- **Sequences**: Events linked over time
- **Classification**: Recognizes patterns that describe group to which item belongs
- **Clustering**: Similar to classification when no groups have been defined; finds groupings within data
- **Forecasting**: Uses series of existing values to forecast what other values will be

Text mining

• Extracts key elements from large unstructured data sets
  – Stored e-mails
  – Call center transcripts
  – Legal cases
  – Patent descriptions
  – Service reports, and so on

• Sentiment analysis software
  – Mines e-mails, blogs, social media to detect opinions

Web mining

• Discovery and analysis of useful patterns and information from Web
  – Understand customer behavior
  – Evaluate effectiveness of Web site, and so on

• 3 Tasks of Web Mining
  – Web content mining
    • Mines content of Web pages
  – Web structure mining
    • Analyzes links to and from Web page
  – Web usage mining
    • Mines user interaction data recorded by Web server

Web Mining

- Web mining (or Web data mining) is the process of discovering intrinsic relationships from Web data (textual, linkage, or usage)

Web Mining

- **Web Content Mining**
  Source: unstructured textual content of the Web pages (usually in HTML format)

- **Web Structure Mining**
  Source: the unified resource locator (URL) links contained in the Web pages

- **Web Usage Mining**
  Source: the detailed description of a Web site’s visits (sequence of clicks by sessions)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Databases and the Web

• Many companies use Web to make some internal databases available to customers or partners

• Typical configuration includes:
  – Web server
  – Application server/middleware/CGI scripts
  – Database server (hosting DBMS)

• Advantages of using Web for database access:
  – Ease of use of browser software
  – Web interface requires few or no changes to database
  – Inexpensive to add Web interface to system

Web Content/Structure Mining

• Mining of the textual content on the Web
• Data collection via Web crawlers

• Web pages include hyperlinks
  – Authoritative pages
  – Hubs
  – hyperlink-induced topic search (HITS) alg

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Usage Mining

• Extraction of information from data generated through Web page visits and transactions...
  – data stored in server access logs, referrer logs, agent logs, and client-side cookies
  – user characteristics and usage profiles
  – metadata, such as page attributes, content attributes, and usage data

• Clickstream data

• Clickstream analysis

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Usage Mining

• Web usage mining applications
  – Determine the lifetime value of clients
  – Design cross-marketing strategies across products.
  – Evaluate promotional campaigns
  – Target electronic ads and coupons at user groups based on user access patterns
  – Predict user behavior based on previously learned rules and users' profiles
  – Present dynamic information to users based on their interests and profiles...

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Web Usage Mining
(clickstream analysis)

Pre-Process Data
- Collecting
- Merging
- Cleaning
- Structuring
  - Identify users
  - Identify sessions
  - Identify page views
  - Identify visits

Extract Knowledge
- Usage patterns
- User profiles
- Page profiles
- Visit profiles
- Customer value

User / Customer

How to better the data
How to improve the Web site
How to increase the customer value

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems

77
Web Mining Success Stories

- Amazon.com, Ask.com, Scholastic.com, ...
- Website Optimization Ecosystem

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining Tasks
A Taxonomy for Data Mining Tasks

<table>
<thead>
<tr>
<th>Data Mining</th>
<th>Learning Method</th>
<th>Popular Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>Supervised</td>
<td>Classification and Regression Trees, ANN, SVM, Genetic Algorithms</td>
</tr>
<tr>
<td></td>
<td>Supervised</td>
<td>Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms</td>
</tr>
<tr>
<td></td>
<td>Supervised</td>
<td>Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM</td>
</tr>
<tr>
<td>Association</td>
<td>Unsupervised</td>
<td>Apriory, OneR, ZeroR, Eclat</td>
</tr>
<tr>
<td></td>
<td>Unsupervised</td>
<td>Expectation Maximization, Apriory Algorithm, Graph-based Matching</td>
</tr>
<tr>
<td></td>
<td>Unsupervised</td>
<td>Apriory Algorithm, FP-Growth technique</td>
</tr>
<tr>
<td>Clustering</td>
<td>Unsupervised</td>
<td>K-means, ANN/SOM</td>
</tr>
<tr>
<td></td>
<td>Unsupervised</td>
<td>K-means, Expectation Maximization (EM)</td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Why Data Mining?

• More intense competition at the global scale
• Recognition of the value in data sources
• Availability of quality data on customers, vendors, transactions, Web, etc.
• Consolidation and integration of data repositories into data warehouses
• The exponential increase in data processing and storage capabilities; and decrease in cost
• Movement toward conversion of information resources into nonphysical form

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Definition of Data Mining

• The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases.
  - Fayyad et al., (1996)

• **Keywords in this definition:** Process, nontrivial, valid, novel, potentially useful, understandable.

• Data mining: a misnomer?

• **Other names:**
  – knowledge extraction, pattern analysis, knowledge discovery, information harvesting, pattern searching, data dredging,...

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining
Characteristics/Objectives

• Source of data for DM is often a consolidated data warehouse (not always!)
• DM environment is usually a client-server or a Web-based information systems architecture
• Data is the most critical ingredient for DM which may include soft/unstructured data
• The miner is often an end user
• Striking it rich requires creative thinking
• Data mining tools’ capabilities and ease of use are essential (Web, Parallel processing, etc.)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data in Data Mining

• Data: a collection of facts usually obtained as the result of experiences, observations, or experiments
• Data may consist of numbers, words, images, ...
• Data: lowest level of abstraction (from which information and knowledge are derived)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
What Does DM Do?

• DM extract patterns from data
  – Pattern?
    A mathematical (numeric and/or symbolic) relationship among data items

• Types of patterns
  – Association
  – Prediction
  – Cluster (segmentation)
  – Sequential (or time series) relationships

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining Applications

• Customer Relationship Management
  – Maximize return on marketing campaigns
  – Improve customer retention (churn analysis)
  – Maximize customer value (cross-, up-selling)
  – Identify and treat most valued customers

• Banking and Other Financial
  – Automate the loan application process
  – Detecting fraudulent transactions
  – Optimizing cash reserves with forecasting

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining Applications (cont.)

• Retailing and Logistics
  – Optimize inventory levels at different locations
  – Improve the store layout and sales promotions
  – Optimize logistics by predicting seasonal effects
  – Minimize losses due to limited shelf life

• Manufacturing and Maintenance
  – Predict/prevent machinery failures
  – Identify anomalies in production systems to optimize the use manufacturing capacity
  – Discover novel patterns to improve product quality

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining Applications (cont.)

• Brokerage and Securities Trading
  – Predict changes on certain bond prices
  – Forecast the direction of stock fluctuations
  – Assess the effect of events on market movements
  – Identify and prevent fraudulent activities in trading

• Insurance
  – Forecast claim costs for better business planning
  – Determine optimal rate plans
  – Optimize marketing to specific customers
  – Identify and prevent fraudulent claim activities

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Data Mining Applications (cont.)

- Computer hardware and software
- Science and engineering
- Government and defense
- Homeland security and law enforcement
- Travel industry
- Healthcare
- Medicine
- Entertainment industry
- Sports
- Etc.

Highly popular application areas for data mining

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
## A Taxonomy for Data Mining Tasks

<table>
<thead>
<tr>
<th>Learning Method</th>
<th>Popular Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>Classification and Regression Trees, ANN, SVM, Genetic Algorithms</td>
</tr>
<tr>
<td>Supervised</td>
<td>Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms</td>
</tr>
<tr>
<td>Supervised</td>
<td>Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Apriory, OneR, ZeroR, Eclat</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Expectation Maximization, Apriory Algorithm, Graph-based Matching</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Apriory Algorithm, FP-Growth technique</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>K-means, ANN/SOM</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>K-means, Expectation Maximization (EM)</td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Association Analysis: Mining Frequent Patterns, Association and Correlations

- Association Analysis
- Mining Frequent Patterns
- Association and Correlations
- Apriori Algorithm

Source: Han & Kamber (2006)
Market Basket Analysis

Which items are frequently purchased together by my customers?

Shopping Baskets

Customer 1: milk, bread, cereal
Customer 2: milk, bread, sugar, eggs
Customer 3: milk, bread, butter
Customer n: sugar, eggs

Source: Han & Kamber (2006)
Association Rule Mining

• Apriori Algorithm

<table>
<thead>
<tr>
<th>Raw Transaction Data</th>
<th>One-item Itemsets</th>
<th>Two-item Itemsets</th>
<th>Three-item Itemsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transaction No</td>
<td>Itemset (SKUs)</td>
<td>Itemset (SKUs)</td>
<td>Itemset (SKUs)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1, 2</td>
<td>1, 2, 4</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1, 3</td>
<td>2, 3, 4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1, 4</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>2, 3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1, 2, 4</td>
<td>2, 4</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2, 3, 4</td>
<td>3, 4</td>
<td></td>
</tr>
</tbody>
</table>

Support

1. 3
2. 6
3. 4
4. 5
5. 3
6. 2
7. 3
8. 4
9. 5
10. 3

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Association Rule Mining

• A very popular DM method in business
• Finds interesting relationships (affinities) between variables (items or events)
• Part of machine learning family
• Employs unsupervised learning
• There is no output variable
• Also known as market basket analysis
• Often used as an example to describe DM to ordinary people, such as the famous “relationship between diapers and beers!”

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Association Rule Mining

• **Input:** the simple point-of-sale transaction data
• **Output:** Most frequent affinities among items
• **Example:** according to the transaction data...
  "Customer who bought a laptop computer and a virus protection software, also bought extended service plan 70 percent of the time."

• **How do you use such a pattern/knowledge?**
  – Put the items next to each other for ease of finding
  – Promote the items as a package (do not put one on sale if the other(s) are on sale)
  – Place items far apart from each other so that the customer has to walk the aisles to search for it, and by doing so potentially seeing and buying other items

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Association Rule Mining

• A representative applications of association rule mining include
  – **In business:** cross-marketing, cross-selling, store design, catalog design, e-commerce site design, optimization of online advertising, product pricing, and sales/promotion configuration
  – **In medicine:** relationships between symptoms and illnesses; diagnosis and patient characteristics and treatments (to be used in medical DSS); and genes and their functions (to be used in genomics projects)...
Association Rule Mining

• Are all association rules interesting and useful?

A Generic Rule:  \( X \Rightarrow Y \) \([S\%, C\%]\)

\( X, Y \): products and/or services

\( X \): Left-hand-side (LHS)

\( Y \): Right-hand-side (RHS)

\( S \): **Support**: how often \( X \) and \( Y \) go together

\( C \): **Confidence**: how often \( Y \) go together with the \( X \)

**Example**: \{Laptop Computer, Antivirus Software\} \( \Rightarrow \)

\{Extended Service Plan\} \([30\%, 70\%]\)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Association Rule Mining

• Algorithms are available for generating association rules
  – Apriori
  – Eclat
  – FP-Growth
  – + Derivatives and hybrids of the three

• The algorithms help identify the frequent item sets, which are, then converted to association rules

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Association Rule Mining

• Apriori Algorithm
  – Finds subsets that are common to at least a minimum number of the itemsets
  – uses a bottom-up approach
    • frequent subsets are extended one item at a time (the size of frequent subsets increases from one-item subsets to two-item subsets, then three-item subsets, and so on), and
    • groups of candidates at each level are tested against the data for minimum

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Basic Concepts: Frequent Patterns and Association Rules

- Itemset $X = \{x_1, ..., x_k\}$
- Find all the rules $X \rightarrow Y$ with minimum support and confidence
  - support, $s$, probability that a transaction contains $X \cup Y$
  - confidence, $c$, conditional probability that a transaction having $X$ also contains $Y$

Let $sup_{min} = 50\%$, $conf_{min} = 50\%$
Freq. Pat.: $\{A:3, B:3, D:4, E:3, AD:3\}$
Association rules:
  - $A \rightarrow D$ (60%, 100%)
  - $D \rightarrow A$ (60%, 75%)

$A \rightarrow D$ (support $= 3/5 = 60\%$, confidence $= 3/3 = 100\%$)
$D \rightarrow A$ (support $= 3/5 = 60\%$, confidence $= 3/4 = 75\%$

Source: Han & Kamber (2006)
Market basket analysis

• Example
  – Which groups or sets of items are customers likely to purchase on a given trip to the store?

• Association Rule
  – Computer $\rightarrow$ antivirus_software
    [support = 2%; confidence = 60%]
    • A support of 2% means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together.
    • A confidence of 60% means that 60% of the customers who purchased a computer also bought the software.

Source: Han & Kamber (2006)
Association rules

• Association rules are considered interesting if they satisfy both
  – a minimum support threshold and
  – a minimum confidence threshold.

Source: Han & Kamber (2006)
Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of items. Let $D$, the task-relevant data, be a set of database transactions where each transaction $T$ is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let $A$ be a set of items. A transaction $T$ is said to contain $A$ if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subseteq I$, $B \subseteq I$, and $A \cap B = \emptyset$. The rule $A \Rightarrow B$ holds in the transaction set $D$ with support $s$, where $s$ is the percentage of transactions in $D$ that contain $A \cup B$ (i.e., the union of sets $A$ and $B$, or say, both $A$ and $B$). This is taken to be the probability, $P(A \cup B)$.

The rule $A \Rightarrow B$ has confidence $c$ in the transaction set $D$, where $c$ is the percentage of transactions in $D$ containing $A$ that also contain $B$. This is taken to be the conditional probability, $P(B|A)$. That is,

\[
\text{Support} (A \Rightarrow B) = P(A \cup B) \\
\text{Confidence} (A \Rightarrow B) = P(B|A)
\]
Support \((A \rightarrow B) = P(A \cup B)\)

Confidence \((A \rightarrow B) = P(B | A)\)

- The notation \(P(A \cup B)\) indicates the probability that a transaction contains the union of set \(A\) and set \(B\)
  - \((i.e.,\ it\ contains\ every\ item\ in\ A\ and\ in\ B)\).
- This should not be confused with \(P(A\ or\ B)\), which indicates the probability that a transaction contains either \(A\ or\ B\).
Does diaper purchase predict beer purchase?

- Contingency tables

<table>
<thead>
<tr>
<th>No diapers</th>
<th>Beer Yes</th>
<th>Beer No</th>
</tr>
</thead>
<tbody>
<tr>
<td>No diapers</td>
<td>6</td>
<td>94</td>
</tr>
<tr>
<td>diapers</td>
<td>40</td>
<td>60</td>
</tr>
</tbody>
</table>

DEPENDENT (yes)


<table>
<thead>
<tr>
<th>INDEPENDENT (no predictability)</th>
<th>Beer Yes</th>
<th>Beer No</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>77</td>
<td></td>
</tr>
</tbody>
</table>
Support \((A \rightarrow B) = P(A \cup B)\)

Confidence \((A \rightarrow B) = P(B | A)\)

\[\text{Conf} (A \rightarrow B) = \frac{\text{Supp} (A \cup B)}{\text{Supp} (A)}\]

\[\text{Lift} (A \rightarrow B) = \frac{\text{Supp} (A \cup B)}{(\text{Supp} (A) \times \text{Supp} (B))}\]

\[\text{Lift (Correlation)}\]

\[\text{Lift} (A \rightarrow B) = \frac{\text{Confidence} (A \rightarrow B)}{\text{Support}(B)}\]

Lift

Lift = Confidence / Expected Confidence if Independent

<table>
<thead>
<tr>
<th>Checking Saving</th>
<th>No (1500)</th>
<th>Yes (8500)</th>
<th>(10000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>500</td>
<td>3500</td>
<td>4000</td>
</tr>
<tr>
<td>Yes</td>
<td>1000</td>
<td>5000</td>
<td>6000</td>
</tr>
</tbody>
</table>

SVG=>CHKG Expect \( \frac{8500}{10000} = 85\% \) if independent
Observed Confidence is \( \frac{5000}{6000} = 83\% \)
Lift = \( \frac{83}{85} < 1 \).
Savings account holders actually LESS likely than others to have checking account !!!

• Rules that satisfy both a minimum support threshold \((\text{min\_sup})\) and a minimum confidence threshold \((\text{min\_conf})\) are called strong.

• By convention, we write support and confidence values so as to occur between 0% and 100%, rather than 0 to 1.0.

Source: Han & Kamber (2006)
• itemset
  – A set of items is referred to as an itemset.

• K-itemset
  – An itemset that contains \(k\) items is a \(k\)-itemset.

• Example:
  – The set \(\{computer, antivirus software\}\) is a 2-itemset.

Source: Han & Kamber (2006)
Absolute Support and Relative Support

• Absolute Support
  – The occurrence frequency of an itemset is the number of transactions that contain the itemset
    • frequency, support count, or count of the itemset
    – Ex: 3

• Relative support
  – Ex: 60%

Source: Han & Kamber (2006)
• If the **relative support** of an itemset $I$ satisfies a prespecified **minimum support threshold**, then $I$ is a **frequent itemset**.
  
  – i.e., the **absolute** support of $I$ satisfies the corresponding **minimum support count threshold**

• The set of **frequent $k$-itemsets** is commonly denoted by $L_k$

Source: Han & Kamber (2006)
the confidence of rule $A \to B$ can be easily derived from the support counts of $A$ and $A \cup B$.

once the support counts of $A$, $B$, and $A \cup B$ are found, it is straightforward to derive the corresponding association rules $A \to B$ and $B \to A$ and check whether they are strong.

Thus the problem of mining association rules can be reduced to that of mining frequent itemsets.

Source: Han & Kamber (2006)
Association rule mining: Two-step process

1. Find all frequent itemsets
   – By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, \( \text{min}_\text{sup} \).

2. Generate strong association rules from the frequent itemsets
   – By definition, these rules must satisfy minimum support and minimum confidence.

Source: Han & Kamber (2006)
Efficient and Scalable Frequent Itemset Mining Methods

• The Apriori Algorithm
  – Finding Frequent Itemsets Using Candidate Generation

Source: Han & Kamber (2006)
Apriori Algorithm

• **Apriori** is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules.

• The name of the algorithm is based on the fact that the algorithm uses *prior knowledge of frequent itemset properties, as we shall see following.*

Source: Han & Kamber (2006)
Apriori Algorithm

• Apriori employs an iterative approach known as a **level-wise search**, where **k-itemsets are used to explore (k+1)-itemsets**.

• *First, the set of frequent 1-itemsets is found* by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted \( L_1 \).

• *Next, \( L_1 \) is used to find \( L_2 \), the set of frequent 2-itemsets, which is used to find \( L_3 \), and so on, until no more frequent \( k \)-itemsets can be found.*

• *The finding of each \( L_k \) requires one full scan of the database.*

Source: Han & Kamber (2006)
Apriori Algorithm

• To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property.

• Apriori property
  – *All nonempty subsets of a frequent itemset must also be frequent.*

Source: Han & Kamber (2006)
Apriori algorithm
(1) Frequent Itemsets
(2) Association Rules
## Transaction Database

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Items bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>T01</td>
<td>A, B, D</td>
</tr>
<tr>
<td>T02</td>
<td>A, C, D</td>
</tr>
<tr>
<td>T03</td>
<td>B, C, D, E</td>
</tr>
<tr>
<td>T04</td>
<td>A, B, D</td>
</tr>
<tr>
<td>T05</td>
<td>A, B, C, E</td>
</tr>
<tr>
<td>T06</td>
<td>A, C</td>
</tr>
<tr>
<td>T07</td>
<td>B, C, D</td>
</tr>
<tr>
<td>T08</td>
<td>B, D</td>
</tr>
<tr>
<td>T09</td>
<td>A, C, E</td>
</tr>
<tr>
<td>T10</td>
<td>B, D</td>
</tr>
</tbody>
</table>
Table 1 shows a database with 10 transactions. Let *minimum support = 20%* and *minimum confidence = 80%*. Please use **Apriori algorithm** for generating **association rules** from frequent itemsets.

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Items bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>T01</td>
<td>A, B, D</td>
</tr>
<tr>
<td>T02</td>
<td>A, C, D</td>
</tr>
<tr>
<td>T03</td>
<td>B, C, D, E</td>
</tr>
<tr>
<td>T04</td>
<td>A, B, D</td>
</tr>
<tr>
<td>T05</td>
<td>A, B, C, E</td>
</tr>
<tr>
<td>T06</td>
<td>A, C</td>
</tr>
<tr>
<td>T07</td>
<td>B, C, D</td>
</tr>
<tr>
<td>T08</td>
<td>B, D</td>
</tr>
<tr>
<td>T09</td>
<td>A, C, E</td>
</tr>
<tr>
<td>T10</td>
<td>B, D</td>
</tr>
</tbody>
</table>
### Apriori Algorithm

C₁ \rightarrow L₁

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Items bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>T01</td>
<td>A, B, D</td>
</tr>
<tr>
<td>T02</td>
<td>A, C, D</td>
</tr>
<tr>
<td>T03</td>
<td>B, C, D, E</td>
</tr>
<tr>
<td>T04</td>
<td>A, B, D</td>
</tr>
<tr>
<td>T05</td>
<td>A, B, C, E</td>
</tr>
<tr>
<td>T06</td>
<td>A, C</td>
</tr>
<tr>
<td>T07</td>
<td>B, C, D</td>
</tr>
<tr>
<td>T08</td>
<td>B, D</td>
</tr>
<tr>
<td>T09</td>
<td>A, C, E</td>
</tr>
<tr>
<td>T10</td>
<td>B, D</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>7</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
</tr>
</tbody>
</table>

minimum support = 20%  
= 2 / 10  
Min. Support Count = 2

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>7</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
</tr>
</tbody>
</table>
## Apriori Algorithm

### Step 1-2

### $C_2 \rightarrow L_2$

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B</td>
<td>3</td>
</tr>
<tr>
<td>A, C</td>
<td>4</td>
</tr>
<tr>
<td>A, D</td>
<td>3</td>
</tr>
<tr>
<td>A, E</td>
<td>2</td>
</tr>
<tr>
<td>B, C</td>
<td>3</td>
</tr>
<tr>
<td>B, D</td>
<td>6</td>
</tr>
<tr>
<td>B, E</td>
<td>2</td>
</tr>
<tr>
<td>C, D</td>
<td>3</td>
</tr>
<tr>
<td>C, E</td>
<td>3</td>
</tr>
<tr>
<td>D, E</td>
<td>1</td>
</tr>
</tbody>
</table>

### $L_2$

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B</td>
<td>3</td>
</tr>
<tr>
<td>A, C</td>
<td>4</td>
</tr>
<tr>
<td>A, D</td>
<td>3</td>
</tr>
<tr>
<td>A, E</td>
<td>2</td>
</tr>
<tr>
<td>B, C</td>
<td>3</td>
</tr>
<tr>
<td>B, D</td>
<td>6</td>
</tr>
<tr>
<td>B, E</td>
<td>2</td>
</tr>
<tr>
<td>C, D</td>
<td>3</td>
</tr>
<tr>
<td>C, E</td>
<td>3</td>
</tr>
</tbody>
</table>

**minimum support = 20%**

$= 2 / 10$

Min. Support Count = 2
Apriori Algorithm

$C_3 \rightarrow L_3$

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Items bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>T01</td>
<td>A, B, D</td>
</tr>
<tr>
<td>T02</td>
<td>A, C, D</td>
</tr>
<tr>
<td>T03</td>
<td>B, C, D, E</td>
</tr>
<tr>
<td>T04</td>
<td>A, B, D</td>
</tr>
<tr>
<td>T05</td>
<td>A, B, C, E</td>
</tr>
<tr>
<td>T06</td>
<td>A, C</td>
</tr>
<tr>
<td>T07</td>
<td>B, C, D</td>
</tr>
<tr>
<td>T08</td>
<td>B, D</td>
</tr>
<tr>
<td>T09</td>
<td>A, C, E</td>
</tr>
<tr>
<td>T10</td>
<td>B, D</td>
</tr>
</tbody>
</table>

**C$_3$**

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B, C</td>
<td>1</td>
</tr>
<tr>
<td>A, B, D</td>
<td>2</td>
</tr>
<tr>
<td>A, B, E</td>
<td>1</td>
</tr>
<tr>
<td>A, C, D</td>
<td>1</td>
</tr>
<tr>
<td>A, C, E</td>
<td>2</td>
</tr>
</tbody>
</table>

Itemset $minimun support = 20% = 2 / 10$
Min. Support Count = 2

**L$_3$**

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B, D</td>
<td>2</td>
</tr>
<tr>
<td>A, C, E</td>
<td>2</td>
</tr>
<tr>
<td>B, C, D</td>
<td>2</td>
</tr>
<tr>
<td>B, C, E</td>
<td>2</td>
</tr>
</tbody>
</table>

Step 1-3
Generating Association Rules

minimum confidence = 80%

Association Rules Generated from \( L_2 \)

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B</td>
<td>3</td>
</tr>
<tr>
<td>A, C</td>
<td>4</td>
</tr>
<tr>
<td>A, D</td>
<td>3</td>
</tr>
<tr>
<td>A, E</td>
<td>2</td>
</tr>
<tr>
<td>B, C</td>
<td>3</td>
</tr>
<tr>
<td>B, D</td>
<td>6</td>
</tr>
<tr>
<td>B, E</td>
<td>2</td>
</tr>
<tr>
<td>C, D</td>
<td>3</td>
</tr>
<tr>
<td>C, E</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>7</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
</tr>
</tbody>
</table>

Transaction ID | Items bought
T01 | A, B, D
T02 | A, C, D
T03 | B, C, D, E
T04 | A, B, D
T05 | A, B, C, E
T06 | A, C
T07 | B, C, D
T08 | B, D
T09 | A, C, E
T10 | B, D

Step 2-1
Generating Association Rules

minimum confidence = 80%

Association Rules Generated from $L_3$

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>7</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
</tr>
</tbody>
</table>

$L_1$ | $L_2$ | $L_3$

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B</td>
<td>3</td>
</tr>
<tr>
<td>A, C</td>
<td>4</td>
</tr>
<tr>
<td>A, D</td>
<td>3</td>
</tr>
<tr>
<td>A, E</td>
<td>2</td>
</tr>
<tr>
<td>B, C</td>
<td>3</td>
</tr>
<tr>
<td>B, D</td>
<td>6</td>
</tr>
<tr>
<td>B, E</td>
<td>2</td>
</tr>
<tr>
<td>C, D</td>
<td>3</td>
</tr>
<tr>
<td>C, E</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B, D</td>
<td>2</td>
</tr>
<tr>
<td>A, C, E</td>
<td>2</td>
</tr>
<tr>
<td>B, C, D</td>
<td>2</td>
</tr>
<tr>
<td>B, C, E</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A→BD: 2/6</td>
<td></td>
</tr>
<tr>
<td>B→AD: 2/7</td>
<td></td>
</tr>
<tr>
<td>D→AB: 2/7</td>
<td></td>
</tr>
<tr>
<td>AB→D: 2/3</td>
<td></td>
</tr>
<tr>
<td>AD→B: 2/3</td>
<td></td>
</tr>
<tr>
<td>BD→A: 2/6</td>
<td></td>
</tr>
<tr>
<td>A→CE: 2/6</td>
<td></td>
</tr>
<tr>
<td>C→AE: 2/6</td>
<td></td>
</tr>
<tr>
<td>E→AC: 2/3</td>
<td></td>
</tr>
<tr>
<td>AC→E: 2/4</td>
<td></td>
</tr>
<tr>
<td>AE→C: 2/2=100%*</td>
<td></td>
</tr>
<tr>
<td>BE→C: 2/2=100%*</td>
<td></td>
</tr>
<tr>
<td>CE→A: 2/3</td>
<td></td>
</tr>
<tr>
<td>CE→B: 2/3</td>
<td></td>
</tr>
</tbody>
</table>

$B$→$CD$: 2/7

$D$→$BC$: 2/7

$BC$→$D$: 2/3

$BD$→$C$: 2/6

$CD$→$B$: 2/3

$B$→$CE$: 2/7

$C$→$BE$: 2/6

$E$→$BC$: 2/3

$BC$→$E$: 2/3
Frequent Itemsets and Association Rules

Transaction ID | Items bought
--- | ---
T01 | A, B, D
T02 | A, C, D
T03 | B, C, D, E
T04 | A, B, D
T05 | A, B, C, E
T06 | A, C
T07 | B, C, D
T08 | B, D
T09 | A, C, E
T10 | B, D

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>6</td>
</tr>
<tr>
<td>D</td>
<td>7</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B</td>
<td>3</td>
</tr>
<tr>
<td>A, C</td>
<td>4</td>
</tr>
<tr>
<td>A, D</td>
<td>3</td>
</tr>
<tr>
<td>A, E</td>
<td>2</td>
</tr>
<tr>
<td>B, C</td>
<td>3</td>
</tr>
<tr>
<td>B, D</td>
<td>6</td>
</tr>
<tr>
<td>B, E</td>
<td>2</td>
</tr>
<tr>
<td>C, D</td>
<td>3</td>
</tr>
<tr>
<td>C, E</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Support Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, B, D</td>
<td>2</td>
</tr>
<tr>
<td>A, C, E</td>
<td>2</td>
</tr>
<tr>
<td>B, C, D</td>
<td>2</td>
</tr>
<tr>
<td>B, C, E</td>
<td>2</td>
</tr>
</tbody>
</table>

minimum support = 20%
minimum confidence = 80%

Association Rules:

B → D (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
D → B (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
E → C (30%, 100%) (Sup.: 3/10, Conf.: 3/3)
AE → C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)
BE → C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)
Table 1 shows a database with 10 transactions. Let *minimum support = 20%* and *minimum confidence = 80%*. Please use **Apriori algorithm** for generating **association rules** from frequent itemsets.

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Items bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>T01</td>
<td>A, B, D</td>
</tr>
<tr>
<td>T02</td>
<td>A, C, D</td>
</tr>
<tr>
<td>T03</td>
<td>B, C, D, E</td>
</tr>
<tr>
<td>T04</td>
<td>A, B, D</td>
</tr>
<tr>
<td>T05</td>
<td>A, B, C, E</td>
</tr>
<tr>
<td>T06</td>
<td>A, C</td>
</tr>
<tr>
<td>T07</td>
<td>B, C, D</td>
</tr>
<tr>
<td>T08</td>
<td>B, D</td>
</tr>
<tr>
<td>T09</td>
<td>A, C, E</td>
</tr>
<tr>
<td>T10</td>
<td>B, D</td>
</tr>
</tbody>
</table>

**Association Rules:**

B→D (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
D→B (60%, 85.7%) (Sup.: 6/10, Conf.: 6/7)
E→C (30%, 100%) (Sup.: 3/10, Conf.: 3/3)
AE→C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)
BE→C (20%, 100%) (Sup.: 2/10, Conf.: 2/2)
A Taxonomy for Data Mining Tasks

Data Mining

- Prediction
  - Classification
  - Regression

- Association
  - Link analysis
  - Sequence analysis

- Clustering
  - Outlier analysis

<table>
<thead>
<tr>
<th>Learning Method</th>
<th>Popular Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>Classification and Regression Trees, ANN, SVM, Genetic Algorithms</td>
</tr>
<tr>
<td>Supervised</td>
<td>Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms</td>
</tr>
<tr>
<td>Supervised</td>
<td>Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Apriory, OneR, ZeroR, Eclat</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Expectation Maximization, Apriory Algorithm, Graph-based Matching</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>Apriory Algorithm, FP-Growth technique</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>K-means, ANN/SOM</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>K-means, Expectation Maximization (EM)</td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Classification vs. Prediction

• **Classification**
  – predicts *categorical class* labels (discrete or nominal)
  – classifies data (constructs a model) based on the training set and the values (*class labels*) in a classifying attribute and uses it in classifying new data

• **Prediction**
  – models *continuous-valued* functions
    • i.e., predicts unknown or missing values

• **Typical applications**
  – Credit approval
  – Target marketing
  – Medical diagnosis
  – Fraud detection

Source: Han & Kamber (2006)
Data Mining Methods: Classification

- Most frequently used DM method
- Part of the machine-learning family
- Employ supervised learning
- Learn from past data, classify new data
- The output variable is categorical (nominal or ordinal) in nature

- Classification versus regression?
- Classification versus clustering?

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Classification Techniques

• Decision tree analysis
• Statistical analysis
• Neural networks
• Support vector machines
• Case-based reasoning
• Bayesian classifiers
• Genetic algorithms
• Rough sets

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Example of Classification

• Loan Application Data
  – Which loan applicants are “safe” and which are “risky” for the bank?
  – “Safe” or “risky” for load application data

• Marketing Data
  – Whether a customer with a given profile will buy a new computer?
  – “yes” or “no” for marketing data

• Classification
  – Data analysis task
  – A model or Classifier is constructed to predict categorical labels
    • Labels: “safe” or “risky”; “yes” or “no”; “treatment A”, “treatment B”, “treatment C”

Source: Han & Kamber (2006)
What Is Prediction?

• (Numerical) prediction is similar to classification
  – construct a model
  – use model to predict continuous or ordered value for a given input
• Prediction is different from classification
  – Classification refers to predict categorical class label
  – Prediction models continuous-valued functions
• Major method for prediction: regression
  – model the relationship between one or more independent or predictor variables and a dependent or response variable
• Regression analysis
  – Linear and multiple regression
  – Non-linear regression
  – Other regression methods: generalized linear model, Poisson regression, log-linear models, regression trees

Source: Han & Kamber (2006)
Prediction Methods

- Linear Regression
- Nonlinear Regression
- Other Regression Methods

Salaries data:

<table>
<thead>
<tr>
<th>years experience</th>
<th>salary (in $1000s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>57</td>
</tr>
<tr>
<td>9</td>
<td>64</td>
</tr>
<tr>
<td>13</td>
<td>72</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td>6</td>
<td>43</td>
</tr>
<tr>
<td>11</td>
<td>59</td>
</tr>
<tr>
<td>21</td>
<td>90</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>83</td>
</tr>
</tbody>
</table>

Source: Han & Kamber (2006)
Classification and Prediction

- **Classification** and **prediction** are two forms of data analysis that can be used to extract **models** describing important data classes or to predict future data trends.

- **Classification**
  - Effective and scalable methods have been developed for **decision trees induction**, **Naive Bayesian classification**, **Bayesian belief network**, **rule-based classifier**, **Backpropagation**, **Support Vector Machine (SVM)**, **associative classification**, **nearest neighbor classifiers**, and **case-based reasoning**, and other classification methods such as **genetic algorithms**, **rough set and fuzzy set** approaches.

- **Prediction**
  - Linear, nonlinear, and **generalized linear models of regression** can be used for **prediction**. Many nonlinear problems can be converted to linear problems by performing transformations on the predictor variables. **Regression trees** and **model trees** are also used for prediction.

Source: Han & Kamber (2006)
Classification—A Two-Step Process

1. Model construction: describing a set of predetermined classes
   - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
   - The set of tuples used for model construction is training set
   - The model is represented as classification rules, decision trees, or mathematical formulae

2. Model usage: for classifying future or unknown objects
   - Estimate accuracy of the model
     • The known label of test sample is compared with the classified result from the model
     • Accuracy rate is the percentage of test set samples that are correctly classified by the model
     • Test set is independent of training set, otherwise over-fitting will occur
   - If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known

Source: Han & Kamber (2006)
Supervised vs. Unsupervised Learning

• **Supervised learning (classification)**
  – Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  – New data is classified based on the training set

• **Unsupervised learning (clustering)**
  – The class labels of training data is unknown
  – Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Source: Han & Kamber (2006)
Issues Regarding Classification and Prediction: Data Preparation

• Data cleaning
  – Preprocess data in order to reduce noise and handle missing values
• Relevance analysis (feature selection)
  – Remove the irrelevant or redundant attributes
  – Attribute subset selection
    • Feature Selection in machine learning
• Data transformation
  – Generalize and/or normalize data
  – Example
    • Income: low, medium, high

Source: Han & Kamber (2006)
Issues:
Evaluating Classification and Prediction Methods

• Accuracy
  – classifier accuracy: predicting class label
  – predictor accuracy: guessing value of predicted attributes
  – estimation techniques: cross-validation and bootstrapping

• Speed
  – time to construct the model (training time)
  – time to use the model (classification/prediction time)

• Robustness
  – handling noise and missing values

• Scalability
  – ability to construct the classifier or predictor efficiently given large amounts of data

• Interpretability
  – understanding and insight provided by the model

Source: Han & Kamber (2006)
Data Classification Process 1: **Learning (Training) Step**

(a) **Learning**: Training data are analyzed by classification algorithm

\[ y = f(X) \]

(source: Han & Kamber (2006))
Data Classification Process 2

(b) **Classification**: Test data are used to estimate the accuracy of the classification rules.

Source: Han & Kamber (2006)
### Process (1): Model Construction

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>YEARS</th>
<th>TENURED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mike</td>
<td>Assistant Prof</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>Mary</td>
<td>Assistant Prof</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Bill</td>
<td>Professor</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>Jim</td>
<td>Associate Prof</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Dave</td>
<td>Assistant Prof</td>
<td>6</td>
<td>no</td>
</tr>
<tr>
<td>Anne</td>
<td>Associate Prof</td>
<td>3</td>
<td>no</td>
</tr>
</tbody>
</table>

**Training Data**

**Classification Algorithms**

IF rank = ‘professor’ OR years > 6 THEN tenured = ‘yes’

Source: Han & Kamber (2006)
Process (2): Using the Model in Prediction

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>YEARS</th>
<th>TENURED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom</td>
<td>Assistant Prof</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>Merlisa</td>
<td>Associate Prof</td>
<td>7</td>
<td>no</td>
</tr>
<tr>
<td>George</td>
<td>Professor</td>
<td>5</td>
<td>yes</td>
</tr>
<tr>
<td>Joseph</td>
<td>Assistant Prof</td>
<td>7</td>
<td>yes</td>
</tr>
</tbody>
</table>

Unseen Data: (Jeff, Professor, 4)

Tenured? Yes

Source: Han & Kamber (2006)
Decision Trees
Decision Trees

A general algorithm for decision tree building

• Employs the divide and conquer method
• Recursively divides a training set until each division consists of examples from one class
  1. Create a root node and assign all of the training data to it
  2. Select the best splitting attribute
  3. Add a branch to the root node for each value of the split. Split the data into mutually exclusive subsets along the lines of the specific split
  4. Repeat the steps 2 and 3 for each and every leaf node until the stopping criteria is reached

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Decision Trees

• DT algorithms mainly differ on
  – Splitting criteria
    • Which variable to split first?
    • What values to use to split?
    • How many splits to form for each node?
  – Stopping criteria
    • When to stop building the tree
  – Pruning (generalization method)
    • Pre-pruning versus post-pruning

• Most popular DT algorithms include
  – ID3, C4.5, C5; CART; CHAID; M5

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Decision Trees

• Alternative splitting criteria
  – **Gini index** determines the purity of a specific class as a result of a decision to branch along a particular attribute/value
    • Used in CART
  – **Information gain** uses entropy to measure the extent of uncertainty or randomness of a particular attribute/value split
    • Used in ID3, C4.5, C5
  – **Chi-square statistics** (used in CHAID)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
### Classification by Decision Tree Induction

#### Training Dataset

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>student</th>
<th>credit_rating</th>
<th>buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31…40</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31…40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>31…40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>31…40</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
</tbody>
</table>

This follows an example of Quinlan’s ID3 (Playing Tennis)

Source: Han & Kamber (2006)
Classification by Decision Tree Induction

Output: A Decision Tree for “buys_computer”

```
buys_computer=“yes” or buys_computer=“no”
```

Source: Han & Kamber (2006)
Three possibilities for partitioning tuples based on the splitting Criterion

Source: Han & Kamber (2006)
Algorithm for Decision Tree Induction

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a top-down recursive divide-and-conquer manner
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)

- Conditions for stopping partitioning
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning — majority voting is employed for classifying the leaf
  - There are no samples left

Source: Han & Kamber (2006)
Attribute Selection Measure

- Notation: Let $D$, the data partition, be a training set of class-labeled tuples.
  Suppose the class label attribute has $m$ distinct values defining $m$ distinct classes, $C_i$ (for $i = 1, \ldots, m$).
  Let $C_{i,D}$ be the set of tuples of class $C_i$ in $D$.
  Let $|D|$ and $|C_{i,D}|$ denote the number of tuples in $D$ and $C_{i,D}$, respectively.

- Example:
  - Class: $\text{buys\_computer}= \text{“yes”}$ or \text{“no”}
  - Two distinct classes ($m=2$)
    - Class $C_i$ ($i=1,2$):
      - $C_1$ = \text{“yes”}
      - $C_2$ = \text{“no”}

Source: Han & Kamber (2006)
Attribute Selection Measure: Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- Let $p_i$ be the probability that an arbitrary tuple in $D$ belongs to class $C_i$, estimated by $|C_{i,D}|/|D|$
- **Expected information** (entropy) needed to classify a tuple in $D$: 
  $$Info(D) = - \sum_{i=1}^{m} p_i \log_2(p_i)$$
- Information needed (after using $A$ to split $D$ into $v$ partitions) to classify $D$: 
  $$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$
- Information gained by branching on attribute $A$
  $$Gain(A) = Info(D) - Info_A(D)$$

Source: Han & Kamber (2006)
Decision Tree
Information Gain
<table>
<thead>
<tr>
<th>ID</th>
<th>age</th>
<th>income</th>
<th>student</th>
<th>credit_rating</th>
<th>Class: buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>youth</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>middle_aged</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>youth</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>senior</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>senior</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>senior</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>middle_aged</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>youth</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>youth</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>senior</td>
<td>medium</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
</tbody>
</table>
Table 2 shows the class-labeled training tuples from customer database. Please calculate and illustrate the final decision tree returned by decision tree induction using information gain.

(1) What is the Information Gain of “age”?
(2) What is the Information Gain of “income”?
(3) What is the Information Gain of “student”?
(4) What is the Information Gain of “credit_rating”?
(5) What is the class (buys_computer = “yes” or buys_computer = “no”) for a customer (age=youth, income=low, student=yes, credit= fair ) based on the classification result by decision tree induction?

<table>
<thead>
<tr>
<th>ID</th>
<th>age</th>
<th>income</th>
<th>student</th>
<th>credit_rating</th>
<th>Class: buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>youth</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>middle_aged</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>youth</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>senior</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>senior</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>senior</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>middle_aged</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>youth</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>youth</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>senior</td>
<td>medium</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
</tbody>
</table>
Attribute Selection Measure: Information Gain (ID3/C4.5)

- Select the attribute with the highest information gain
- Let $p_i$ be the probability that an arbitrary tuple in $D$ belongs to class $C_i$, estimated by $|C_i, D|/|D|$
- **Expected information** (entropy) needed to classify a tuple in $D$:

  \[
  \text{Info}(D) = - \sum_{i=1}^{m} p_i \log_2(p_i)
  \]

- **Information** needed (after using $A$ to split $D$ into $v$ partitions) to classify $D$:

  \[
  \text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)
  \]

- **Information gained** by branching on attribute $A$

  \[
  \text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)
  \]

Source: Han & Kamber (2006)
Step 1: Expected information

\[
Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)
\]

\[
Info(D) = I(6,4) = -\frac{6}{10} \log_2\left(\frac{6}{10}\right) + \left(-\frac{4}{10} \log_2\left(\frac{4}{10}\right)\right)
\]

\[
= -0.6 \times \log_2(0.6) - 0.4 \times \log_2(0.4)
\]

\[
= -0.6 \times (-0.737) - 0.4 \times (-1.3219)
\]

\[
= 0.4422 + 0.5288
\]

\[
= 0.971
\]

\[\text{Class P (Positive): buys\_computer = "yes"} \]

\[\text{Class N (Negative): buys\_computer = "no"} \]

\[P(buys = yes) = P_{i=1} = p_1 = \frac{6}{10} = 0.6 \]

\[P(buys = no) = P_{i=2} = p_2 = \frac{4}{10} = 0.4 \]

\[
\begin{align*}
\log_2(0.1) & = -3.3219 \\
\log_2(0.2) & = -2.3219 \\
\log_2(0.3) & = -1.7370 \\
\log_2(0.4) & = -1.3219 \\
\log_2(0.5) & = -1 \\
\log_2(0.6) & = -0.7370 \\
\log_2(0.7) & = -0.5146 \\
\log_2(0.8) & = -0.3219 \\
\log_2(0.9) & = -0.1520 \\
\log_2(1) & = 0
\end{align*}
\]

\[\text{Class P (Positive): buys\_computer = "yes"} \]

\[\text{Class N (Negative): buys\_computer = "no"} \]

\[P(buys = yes) = P_{i=1} = p_1 = \frac{6}{10} = 0.6 \]

\[P(buys = no) = P_{i=2} = p_2 = \frac{4}{10} = 0.4 \]

\[
\begin{align*}
\log_2(0.1) & = -3.3219 \\
\log_2(0.2) & = -2.3219 \\
\log_2(0.3) & = -1.7370 \\
\log_2(0.4) & = -1.3219 \\
\log_2(0.5) & = -1 \\
\log_2(0.6) & = -0.7370 \\
\log_2(0.7) & = -0.5146 \\
\log_2(0.8) & = -0.3219 \\
\log_2(0.9) & = -0.1520 \\
\log_2(1) & = 0
\end{align*}
\]
<table>
<thead>
<tr>
<th>ID</th>
<th>age</th>
<th>income</th>
<th>student</th>
<th>credit_rating</th>
<th>Class: buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>youth</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>middle_aged</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>youth</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>senior</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>senior</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>senior</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>middle_aged</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>youth</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>youth</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>senior</td>
<td>medium</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>age</th>
<th>( p_i )</th>
<th>( n_i )</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>youth</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>middle_aged</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>senior</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>income</th>
<th>( p_i )</th>
<th>( n_i )</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>midium</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>low</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>student</th>
<th>( p_i )</th>
<th>( n_i )</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>no</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>credit_rating</th>
<th>( p_i )</th>
<th>( n_i )</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>excellent</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>fair</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>
### Information Gain Calculation

#### Step 2: Information

\[
I(1,3) = -\frac{1}{4} \log_2 \left( \frac{1}{4} \right) + (-\frac{3}{4} \log_2 \left( \frac{3}{4} \right))
\]
\[
= -0.25 \times [\log_2 1 - \log_2 4] + (-0.75 \times [\log_2 3 - \log_2 4])
\]
\[
= -0.25 \times [0 - 2] - 0.75 \times [1.585 - 2]
\]
\[
= -0.25 \times [-2] - 0.75 \times [-0.415]
\]
\[
= 0.5 + 0.3112 = 0.8112
\]

#### Step 3: Information Gain

\[
I(2,0) = -\frac{2}{2} \log_2 \left( \frac{2}{2} \right) + (-\frac{0}{2} \log_2 \left( \frac{0}{2} \right))
\]
\[
= -1 \times \log_2 1 + (-0 \times \log_2 0)
\]
\[
= -1 \times 0 + (-0 \times -\infty)
\]
\[
= 0 + 0 = 0
\]

\[
I(3,1) = -\frac{3}{4} \log_2 \left( \frac{3}{4} \right) + (-\frac{1}{4} \log_2 \left( \frac{1}{4} \right))
\]
\[
= -0.75 \times [\log_2 3 - \log_2 4] + (-0.25 \times [\log_2 1 - \log_2 4])
\]
\[
= -0.75 \times [1.585 - 2] - 0.25 \times [0 - 2]
\]
\[
= -0.75 \times [-0.415] - 0.25 \times [-2]
\]
\[
= 0.3112 + 0.5 = 0.8112
\]

#### Gain(A) = Info(D) - Info_A(D)

\[
Gain(age) = Info(D) - Info_{age}(D)
\]
\[
= 0.971 - 0.6488 = 0.3221
\]

**Gain(age) = 0.3221**
\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{income} & p_i & n_i & \text{total} & I(p_i, n_i) & I(p, n_i) \\
\hline
\text{high} & 2 & 2 & 4 & I(2,2) & 1 \\
\text{midium} & 2 & 1 & 3 & I(2,1) & 0.9182 \\
\text{low} & 2 & 1 & 3 & I(2,1) & 0.9182 \\
\hline
\end{array}
\]

\[
Info(D) = - \sum_{i=1}^{m} p_i \log_2 (p_i)
\]

\[
Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)
\]

\[
Info_{income}(D) = \frac{4}{10} I(2,2) + \frac{3}{10} I(2,1) + \frac{3}{10} I(2,1)
\]

\[
= \frac{4}{10} \times 1 + \frac{3}{10} \times 0.9182 + \frac{3}{10} \times 0.9182
\]

\[
= 0.4 + 0.2755 + 0.2755 = 0.951
\]

\[
Gain(A) = Info(D) - Info_A(D)
\]

\[
Gain(income) = Info(D) - Info_{income}(D)
\]

\[
= 0.971 - 0.951 = 0.02
\]
\[
\text{Info}(D) = - \sum_{i=1}^{m} p_i \log_2 (p_i)
\]

\[
\text{Info}(D) = I(6,4) = 0.971
\]

\[
\text{Info}_A(D) = \sum_{j=1}^{v} \left| \frac{D_j}{D} \right| \times I(D_j)
\]

\[
\text{Info}_{\text{student}}(D) = \frac{5}{10} I(4,1) + \frac{5}{10} I(2,3)
\]

\[
= 0.5 \times 0.7219 + 0.5 \times 0.971
\]

\[
= 0.36095 + 0.48545 = 0.8464
\]

\[
\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)
\]

\[
\text{Gain(student)} = \text{Info}(D) - \text{Info}_{\text{student}}(D)
\]

\[
= 0.971 - 0.8464 = 0.1245
\]
<table>
<thead>
<tr>
<th>credit</th>
<th>( p_i )</th>
<th>( n_i )</th>
<th>total</th>
<th>( I(p_i, n_i) )</th>
<th>( I(p_i, n_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>excellent</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>( I(2,2) )</td>
<td>1</td>
</tr>
<tr>
<td>fair</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>( I(4,2) )</td>
<td>0.9183</td>
</tr>
</tbody>
</table>

\[
\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)
\]

\[
\text{Info}(D) = I(6,4) = 0.971
\]

\[
\text{Info}_A(D) = \sum_{j=1}^{v} \left| \frac{D_j}{D} \right| \times I(D_j)
\]

\[
\text{Info}_{\text{credit}}(D) = \frac{4}{10} I(2,2) + \frac{6}{10} I(4,2)
\]

\[
= \frac{4}{10} \times 1 + \frac{6}{10} \times 0.9182
\]

\[
= 0.4 + 0.5509 = 0.9509
\]

\[
\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)
\]

\[
\text{Gain}(\text{credit}) = \text{Info}(D) - \text{Info}_{\text{credit}}(D)
\]

\[
= 0.971 - 0.9509 = 0.019
\]
(5) What is the class (buys_computer = “yes” or buys_computer = “no”) for a customer (age=youth, income=low, student =yes, credit= fair ) based on the classification result by decision tree induction?

(5) Yes =0.0889  (No=0.0167)

age (0.3221) > student (0.1245) > income (0.02) > credit (0.019)

buys_computer = “yes”

age:youth (1/4) x student:yes (4/5) x income:low (2/3) x credit:fair (4/6)

Yes: $\frac{1}{4} \times \frac{4}{5} \times \frac{2}{3} \times \frac{4}{6} = \frac{4}{45} = 0.0889$

buys_computer = “no”

age:youth (3/4) x student:yes (1/5) x income:low (1/3) x credit:fair (2/6)

No: $\frac{3}{4} \times \frac{1}{5} \times \frac{1}{3} \times \frac{2}{6} = 0.01667$
## A Taxonomy for Data Mining Tasks

<table>
<thead>
<tr>
<th>Data Mining</th>
<th>Learning Method</th>
<th>Popular Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction</td>
<td>Supervised</td>
<td>Classification and Regression Trees, ANN, SVM, Genetic Algorithms</td>
</tr>
<tr>
<td></td>
<td>Supervised</td>
<td>Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms</td>
</tr>
<tr>
<td></td>
<td>Supervised</td>
<td>Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM</td>
</tr>
<tr>
<td>Association</td>
<td>Unsupervised</td>
<td>Apriori, OneR, ZeroR, Eclat</td>
</tr>
<tr>
<td>Link analysis</td>
<td>Unsupervised</td>
<td>Expectation Maximization, Apriori Algorithm, Graph-based Matching</td>
</tr>
<tr>
<td>Sequence analysis</td>
<td>Unsupervised</td>
<td>Apriori Algorithm, FP-Growth technique</td>
</tr>
<tr>
<td>Clustering</td>
<td>Unsupervised</td>
<td>K-means, ANN/SOM</td>
</tr>
<tr>
<td>Outlier analysis</td>
<td>Unsupervised</td>
<td>K-means, Expectation Maximization (EM)</td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Cluster Analysis

• Used for automatic identification of natural groupings of things
• Part of the machine-learning family
• Employ unsupervised learning
• Learns the clusters of things from past data, then assigns new instances
• There is not an output variable
• Also known as segmentation

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Cluster Analysis

Clustering of a set of objects based on the *k*-means method. *(The mean of each cluster is marked by a “+”.)*

Source: Han & Kamber (2006)
Cluster Analysis

• Clustering results may be used to
  – Identify natural groupings of customers
  – Identify rules for assigning new cases to classes for targeting/diagnostic purposes
  – Provide characterization, definition, labeling of populations
  – Decrease the size and complexity of problems for other data mining methods
  – Identify outliers in a specific domain (e.g., rare-event detection)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Example of Cluster Analysis

<table>
<thead>
<tr>
<th>Point</th>
<th>P</th>
<th>P(x,y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p01</td>
<td>a</td>
<td>(3, 4)</td>
</tr>
<tr>
<td>p02</td>
<td>b</td>
<td>(3, 6)</td>
</tr>
<tr>
<td>p03</td>
<td>c</td>
<td>(3, 8)</td>
</tr>
<tr>
<td>p04</td>
<td>d</td>
<td>(4, 5)</td>
</tr>
<tr>
<td>p05</td>
<td>e</td>
<td>(4, 7)</td>
</tr>
<tr>
<td>p06</td>
<td>f</td>
<td>(5, 1)</td>
</tr>
<tr>
<td>p07</td>
<td>g</td>
<td>(5, 5)</td>
</tr>
<tr>
<td>p08</td>
<td>h</td>
<td>(7, 3)</td>
</tr>
<tr>
<td>p09</td>
<td>i</td>
<td>(7, 5)</td>
</tr>
<tr>
<td>p10</td>
<td>j</td>
<td>(8, 5)</td>
</tr>
</tbody>
</table>
Cluster Analysis for Data Mining

• Analysis methods
  – Statistical methods (including both hierarchical and nonhierarchical), such as *k*-means, *k*-modes, and so on
  – Neural networks (adaptive resonance theory [ART], self-organizing map [SOM])
  – Fuzzy logic (e.g., fuzzy c-means algorithm)
  – Genetic algorithms

• Divisive versus Agglomerative methods

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Cluster Analysis for Data Mining

- **How many clusters?**
  - There is not a “truly optimal” way to calculate it
  - Heuristics are often used
    1. Look at the sparseness of clusters
    2. Number of clusters = \((n/2)^{1/2}\) (n: no of data points)
    3. Use Akaike information criterion (AIC)
    4. Use Bayesian information criterion (BIC)

- **Most cluster analysis methods involve the use of a distance measure** to calculate the closeness between pairs of items
  - Euclidian versus Manhattan (rectilinear) distance

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
**k-Means Clustering Algorithm**

- **k**: pre-determined number of clusters
- **Algorithm** *(Step 0: determine value of k)*

**Step 1**: Randomly generate *k* random points as initial cluster centers

**Step 2**: Assign each point to the nearest cluster center

**Step 3**: Re-compute the new cluster centers

**Repetition step**: Repeat steps 2 and 3 until some convergence criterion is met (usually that the assignment of points to clusters becomes stable)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Cluster Analysis for Data Mining - \(k\)-Means Clustering Algorithm

Step 1

Step 2

Step 3

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Similarity and Dissimilarity Between Objects

- **Distances** are normally used to measure the **similarity** or **dissimilarity** between two data objects.

- Some popular ones include: **Minkowski distance**:

  \[ d(i, j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q)} \]

  where \( i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \) and \( j = (x_{j1}, x_{j2}, \ldots, x_{jp}) \) are two \( p \)-dimensional data objects, and \( q \) is a positive integer.

- If \( q = 1 \), \( d \) is **Manhattan distance**:

  \[ d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}| \]

Source: Han & Kamber (2006)
Similarity and Dissimilarity Between Objects (Cont.)

• *If* $q = 2$, $d$ is Euclidean distance:

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2)}$$

— Properties

• $d(i,j) \geq 0$

• $d(i,i) = 0$

• $d(i,j) = d(j,i)$

• $d(i,j) \leq d(i,k) + d(k,j)$

• Also, one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures

Source: Han & Kamber (2006)
Euclidean distance vs Manhattan distance

• Distance of two points \( x_1 = (1, 2) \) and \( x_2 = (3, 5) \)

**Euclidean distance:**
\[
= ((3-1)^2 + (5-2)^2)^{1/2}
= (2^2 + 3^2)^{1/2}
= (4 + 9)^{1/2}
= (13)^{1/2}
= 3.61
\]

**Manhattan distance:**
\[
= (3-1) + (5-2)
= 2 + 3
= 5
\]
The *K-Means* Clustering Method

- **Example**

1. **Assign each object to most similar center**
2. **Update the cluster means**
3. **Reassign objects**
4. **Update the cluster means**

*K=2*

Arbitrarily choose K object as initial cluster center

Source: Han & Kamber (2006)
**K-Means Clustering**

**Step by Step**

<table>
<thead>
<tr>
<th>Point</th>
<th>P</th>
<th>P(x, y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p01</td>
<td>a</td>
<td>(3, 4)</td>
</tr>
<tr>
<td>p02</td>
<td>b</td>
<td>(3, 6)</td>
</tr>
<tr>
<td>p03</td>
<td>c</td>
<td>(3, 8)</td>
</tr>
<tr>
<td>p04</td>
<td>d</td>
<td>(4, 5)</td>
</tr>
<tr>
<td>p05</td>
<td>e</td>
<td>(4, 7)</td>
</tr>
<tr>
<td>p06</td>
<td>f</td>
<td>(5, 1)</td>
</tr>
<tr>
<td>p07</td>
<td>g</td>
<td>(5, 5)</td>
</tr>
<tr>
<td>p08</td>
<td>h</td>
<td>(7, 3)</td>
</tr>
<tr>
<td>p09</td>
<td>i</td>
<td>(7, 5)</td>
</tr>
<tr>
<td>p10</td>
<td>j</td>
<td>(8, 5)</td>
</tr>
</tbody>
</table>
K-Means Clustering

Step 1: K=2, Arbitrarily choose K object as initial cluster center

Point P P(x, y)
p01 a (3, 4)
p02 b (3, 6)
p03 c (3, 8)
p04 d (4, 5)
p05 e (4, 7)
p06 f (5, 1)
p07 g (5, 5)
p08 h (7, 3)
p09 i (7, 5)
p10 j (8, 5)

Initial m1 (3, 4)
Initial m2 (8, 5)
Step 2: Compute seed points as the centroids of the clusters of the current partition

Step 3: Assign each objects to most similar center

<table>
<thead>
<tr>
<th>Point</th>
<th>P</th>
<th>P(x,y)</th>
<th>$m_1$ distance</th>
<th>$m_2$ distance</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>p01</td>
<td>a</td>
<td>(3, 4)</td>
<td>0.00</td>
<td>5.10</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p02</td>
<td>b</td>
<td>(3, 6)</td>
<td>2.00</td>
<td>5.10</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p03</td>
<td>c</td>
<td>(3, 8)</td>
<td>4.00</td>
<td>5.83</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p04</td>
<td>d</td>
<td>(4, 5)</td>
<td>1.41</td>
<td>4.00</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p05</td>
<td>e</td>
<td>(4, 7)</td>
<td>3.16</td>
<td>4.47</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p06</td>
<td>f</td>
<td>(5, 1)</td>
<td>3.61</td>
<td>5.00</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p07</td>
<td>g</td>
<td>(5, 5)</td>
<td>2.24</td>
<td>3.00</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p08</td>
<td>h</td>
<td>(7, 3)</td>
<td>4.12</td>
<td>2.24</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p09</td>
<td>i</td>
<td>(7, 5)</td>
<td>4.12</td>
<td>1.00</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p10</td>
<td>j</td>
<td>(8, 5)</td>
<td>5.10</td>
<td>0.00</td>
<td>Cluster2</td>
</tr>
</tbody>
</table>

$M_2 = (8, 5)$

$m_1 = (3, 4)$

Initial $m_1$ (3, 4)

Initial $m_2$ (8, 5)
Step 2: Compute seed points as the centroids of the clusters of the current partition

Step 3: Assign each objects to most similar center

Euclidean distance $b(3,6) \leftrightarrow m1(3,4)$

$= ((3-3)^2 + (4-6)^2)^{1/2}$

$= (0^2 + (-2)^2)^{1/2}$

$= (0 + 4)^{1/2}$

$= (4)^{1/2}$

$= 2.00$

Euclidean distance $b(3,6) \leftrightarrow m2(8,5)$

$= ((8-3)^2 + (5-6)^2)^{1/2}$

$= (5^2 + (-1)^2)^{1/2}$

$= (25 + 1)^{1/2}$

$= (26)^{1/2}$

$= 5.10$

<table>
<thead>
<tr>
<th>Point</th>
<th>P</th>
<th>P(x,y)</th>
<th>m1 distance</th>
<th>m2 distance</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>p01</td>
<td>a</td>
<td>(3, 4)</td>
<td>0.00</td>
<td>5.10</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p02</td>
<td>b</td>
<td>(3, 6)</td>
<td>2.00</td>
<td>5.10</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p03</td>
<td>c</td>
<td>(3, 8)</td>
<td>4.00</td>
<td>5.83</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p04</td>
<td>d</td>
<td>(4, 5)</td>
<td>1.41</td>
<td>4.00</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Step 4: Update the cluster means, Repeat Step 2, 3, stop when no more new assignment

<table>
<thead>
<tr>
<th>Point</th>
<th>P</th>
<th>P(x,y)</th>
<th>m1 distance</th>
<th>m2 distance</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>p01</td>
<td>a</td>
<td>(3, 4)</td>
<td>1.43</td>
<td>4.34</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p02</td>
<td>b</td>
<td>(3, 6)</td>
<td>1.22</td>
<td>4.64</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p03</td>
<td>c</td>
<td>(3, 8)</td>
<td>2.99</td>
<td>5.68</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p04</td>
<td>d</td>
<td>(4, 5)</td>
<td>0.20</td>
<td>3.40</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p05</td>
<td>e</td>
<td>(4, 7)</td>
<td>1.87</td>
<td>4.27</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p06</td>
<td>f</td>
<td>(5, 1)</td>
<td>4.29</td>
<td>4.06</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p07</td>
<td>g</td>
<td>(5, 5)</td>
<td>1.15</td>
<td>2.42</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p08</td>
<td>h</td>
<td>(7, 3)</td>
<td>3.80</td>
<td>1.37</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p09</td>
<td>i</td>
<td>(7, 5)</td>
<td>3.14</td>
<td>0.75</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p10</td>
<td>j</td>
<td>(8, 5)</td>
<td>4.14</td>
<td>0.95</td>
<td>Cluster2</td>
</tr>
</tbody>
</table>

$m_1 = (3.86, 5.14)$  
$M_2 = (7.33, 4.33)$

**K-Means Clustering**
Step 4: Update the cluster means,
Repeat Step 2, 3,
stop when no more new assignment

<table>
<thead>
<tr>
<th>Point</th>
<th>P</th>
<th>P(x,y)</th>
<th>m1</th>
<th>m2</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>p01</td>
<td>a</td>
<td>(3, 4)</td>
<td>1.95</td>
<td>3.78</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p02</td>
<td>b</td>
<td>(3, 6)</td>
<td>0.69</td>
<td>4.51</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p03</td>
<td>c</td>
<td>(3, 8)</td>
<td>2.27</td>
<td>5.86</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p04</td>
<td>d</td>
<td>(4, 5)</td>
<td>0.89</td>
<td>3.13</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p05</td>
<td>e</td>
<td>(4, 7)</td>
<td>1.22</td>
<td>4.45</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p06</td>
<td>f</td>
<td>(5, 1)</td>
<td>5.01</td>
<td>3.05</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p07</td>
<td>g</td>
<td>(5, 5)</td>
<td>1.57</td>
<td>2.30</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p08</td>
<td>h</td>
<td>(7, 3)</td>
<td>4.37</td>
<td>0.56</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p09</td>
<td>i</td>
<td>(7, 5)</td>
<td>3.43</td>
<td>1.52</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p10</td>
<td>j</td>
<td>(8, 5)</td>
<td>4.41</td>
<td>1.95</td>
<td>Cluster2</td>
</tr>
</tbody>
</table>

**K-Means Clustering**

$m_1 = (3.67, 5.83)$

$m_2 = (6.75, 3.50)$
stop when no more new assignment

<table>
<thead>
<tr>
<th>Point</th>
<th>P</th>
<th>P(x,y)</th>
<th>m1 distance</th>
<th>m2 distance</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>p01</td>
<td>a</td>
<td>(3, 4)</td>
<td>1.95</td>
<td>3.78</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p02</td>
<td>b</td>
<td>(3, 6)</td>
<td>0.69</td>
<td>4.51</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p03</td>
<td>c</td>
<td>(3, 8)</td>
<td>2.27</td>
<td>5.86</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p04</td>
<td>d</td>
<td>(4, 5)</td>
<td>0.89</td>
<td>3.13</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p05</td>
<td>e</td>
<td>(4, 7)</td>
<td>1.22</td>
<td>4.45</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p06</td>
<td>f</td>
<td>(5, 1)</td>
<td>5.01</td>
<td>3.05</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p07</td>
<td>g</td>
<td>(5, 5)</td>
<td>1.57</td>
<td>2.30</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p08</td>
<td>h</td>
<td>(7, 3)</td>
<td>4.37</td>
<td>0.56</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p09</td>
<td>i</td>
<td>(7, 5)</td>
<td>3.43</td>
<td>1.52</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p10</td>
<td>j</td>
<td>(8, 5)</td>
<td>4.41</td>
<td>1.95</td>
<td>Cluster2</td>
</tr>
</tbody>
</table>

*K-Means Clustering*

m1 (3.67, 5.83)
m2 (6.75, 3.50)
**K-Means Clustering** *(K=2, two clusters)*

**stop when no more new assignment**

<table>
<thead>
<tr>
<th>Point</th>
<th>P(x,y)</th>
<th>Distance m1</th>
<th>Distance m2</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>p01</td>
<td>a (3, 4)</td>
<td>1.95</td>
<td>3.78</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p02</td>
<td>b (3, 6)</td>
<td>0.69</td>
<td>4.51</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p03</td>
<td>c (3, 8)</td>
<td>2.27</td>
<td>5.86</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p04</td>
<td>d (4, 5)</td>
<td>0.89</td>
<td>3.13</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p05</td>
<td>e (4, 7)</td>
<td>1.22</td>
<td>4.45</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p06</td>
<td>f (5, 1)</td>
<td>5.01</td>
<td>3.05</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p07</td>
<td>g (5, 5)</td>
<td>1.57</td>
<td>2.30</td>
<td>Cluster1</td>
</tr>
<tr>
<td>p08</td>
<td>h (7, 3)</td>
<td>4.37</td>
<td>0.56</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p09</td>
<td>i (7, 5)</td>
<td>3.43</td>
<td>1.52</td>
<td>Cluster2</td>
</tr>
<tr>
<td>p10</td>
<td>j (8, 5)</td>
<td>4.41</td>
<td>1.95</td>
<td>Cluster2</td>
</tr>
</tbody>
</table>

**K-Means Clustering**

m1 (3.67, 5.83)  
m2 (6.75, 3.50)
Data Mining Evaluation
Evaluation
(Accuracy of Classification Model)
Assessment Methods for Classification

• Predictive accuracy
  – Hit rate
• Speed
  – Model building; predicting
• Robustness
• Scalability
• Interpretability
  – Transparency, explainability

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Accuracy vs. Precision

A: High Accuracy, High Precision
B: Low Accuracy, High Precision
C: High Accuracy, Low Precision
D: Low Accuracy, Low Precision
Accuracy vs. Precision

- **A**: High Accuracy, High Precision, High Validity, High Reliability
- **B**: Low Accuracy, High Precision, Low Validity, High Reliability
- **C**: High Accuracy, Low Precision, High Validity, Low Reliability
- **D**: Low Accuracy, Low Precision, Low Validity, Low Reliability
Accuracy vs. Precision

A

High Accuracy
High Precision
High Validity
High Reliability

B

Low Accuracy
High Precision
Low Validity
High Reliability

C

High Accuracy
Low Precision
High Validity
Low Reliability

D

Low Accuracy
Low Precision
Low Validity
Low Reliability
Accuracy of Classification Models

- In classification problems, the primary source for accuracy estimation is the confusion matrix

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>True Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>True Positive Count (TP)</td>
<td>True Positive Count (TP)</td>
</tr>
<tr>
<td>False Positive Count (FP)</td>
<td>False Positive Count (FP)</td>
</tr>
<tr>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>False Negative Count (FN)</td>
<td>False Negative Count (FN)</td>
</tr>
<tr>
<td>True Negative Count (TN)</td>
<td>True Negative Count (TN)</td>
</tr>
</tbody>
</table>

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
\text{True Positive Rate} = \frac{TP}{TP + FN}
\]

\[
\text{True Negative Rate} = \frac{TN}{TN + FP}
\]

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Estimation Methodologies for Classification

• **Simple split** (or holdout or test sample estimation)
  – Split the data into 2 mutually exclusive sets training (~70%) and testing (30%)
  – For ANN, the data is split into three sub-sets (training [~60%], validation [~20%], testing [~20%])

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Estimation Methodologies for Classification

• **k-Fold Cross Validation** (rotation estimation)
  – Split the data into $k$ mutually exclusive subsets
  – Use each subset as testing while using the rest of the subsets as training
  – Repeat the experimentation for $k$ times
  – Aggregate the test results for true estimation of prediction accuracy training

• Other estimation methodologies
  – Leave-one-out, bootstrapping, jackknifing
  – Area under the ROC curve

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Estimation Methodologies for Classification – ROC Curve

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Sensitivity = True Positive Rate

Specificity = True Negative Rate
**True Class (actual value)**

<table>
<thead>
<tr>
<th>Predictive Class (prediction outcome)</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td>Negative</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

**total**

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>N</th>
</tr>
</thead>
</table>

**Total**

**Accuracy**

\[ \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \]

**True Positive Rate**

\[ \text{True Positive Rate} = \frac{TP}{TP + FN} \]

**True Negative Rate**

\[ \text{True Negative Rate} = \frac{TN}{TN + FP} \]

**Precision**

\[ \text{Precision} = \frac{TP}{TP + FP} \]

**Recall**

\[ \text{Recall} = \frac{TP}{TP + FN} \]

**True Positive Rate (Sensitivity)**

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

= True Positive Rate

= Recall

= Hit rate

= \( TP / (TP + FN) \)
**Specificity**

\[
\text{Specificity} = \text{True Negative Rate} = \frac{TN}{N} = \frac{TN}{TN + FP}
\]

\[
\text{True Negative Rate (Specificity)} = \frac{TN}{TN + FP}
\]

\[
\text{False Positive Rate (1 - Specificity)} = \frac{FP}{FP + TN}
\]

Precision
= Positive Predictive Value (PPV)
\[ Precision = \frac{TP}{TP + FP} \]

Recall
= True Positive Rate (TPR)
= Sensitivity
= Hit Rate
\[ Recall = \frac{TP}{TP + FN} \]

F1 score (F-score)(F-measure)
is the harmonic mean of precision and recall
= \( \frac{2TP}{P + P'} \)
= \( \frac{2TP}{2TP + FP + FN} \)

\[ F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]

A

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FP + TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>63</td>
<td>28</td>
<td></td>
<td>91</td>
</tr>
<tr>
<td>FN</td>
<td>37</td>
<td>72</td>
<td></td>
<td>109</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

**Recall**

- True Positive Rate (TPR)
- Sensitivity
- Hit Rate
- TP / (TP + FN)

**Specificity**

- True Negative Rate
- TN / N
- TN / (TN + FP)

**TPR = 0.63**

**FPR = 0.28**

**PPV = 0.69**

- \( \text{Precision} = \frac{TP}{TP + FP} \)
- \( \text{PPV} = \frac{TP}{(TP + FP)} \)

**F1 = 0.66**

- \( F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \)
- \( F = 2 \times \frac{0.63 \times 0.69}{0.63 + 0.69} \)
- \( F = \frac{2 \times 63}{100 + 91} \)
- \( F = \frac{0.63 + 0.69}{2} = 1.32 / 2 = 0.66 \)

**ACC = 0.68**

- \( \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \)
- \( \text{ACC} = \frac{63 + 72}{200} \)
- \( \text{ACC} = \frac{135}{200} = 67.5 \)

A

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(TP)</td>
<td>63</td>
<td>28</td>
<td>72</td>
<td>200</td>
</tr>
<tr>
<td>(FN)</td>
<td>37</td>
<td>72</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>

\[
TPR = 0.63 = \frac{63}{63+28} = \frac{63}{91}
\]

\[
FPR = 0.28 = \frac{28}{72+28}
\]

\[
PPV = 0.69 = \frac{63}{63+72}
\]

\[
F1 = 0.66 = \frac{2 \cdot 0.63 \cdot 0.69}{0.63 + 0.69} = \frac{2 \cdot 63}{100 + 91} = \frac{0.63 + 0.69}{2} = \frac{1.32}{2} = 0.66
\]

\[
ACC = 0.68 = \frac{63 + 72}{200} = \frac{135}{200} = 0.675
\]

B

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(TP)</td>
<td>77</td>
<td>23</td>
<td>23</td>
<td>200</td>
</tr>
<tr>
<td>(FN)</td>
<td>23</td>
<td>77</td>
<td>77</td>
<td>200</td>
</tr>
</tbody>
</table>

\[
TPR = 0.77 = \frac{77}{77+23} = \frac{77}{100}
\]

\[
FPR = 0.77 = \frac{23}{77+23}
\]

\[
PPV = 0.50 = \frac{77}{77+23}
\]

\[
F1 = 0.61 = \frac{2 \cdot 0.77 \cdot 0.50}{0.77 + 0.50} = \frac{2 \cdot 77}{100 + 91} = \frac{0.77 + 0.50}{2} = \frac{1.27}{2} = 0.635
\]

\[
ACC = 0.50 = \frac{77 + 23}{200} = \frac{100}{200} = 0.50
\]

Recall

Recall = True Positive Rate (TPR)
= Sensitivity
= Hit Rate

\[
Recall = \frac{TP}{TP + FN}
\]

Precision

Precision = Positive Predictive Value (PPV)

\[
Precision = \frac{TP}{TP + FP}
\]

Source: http://en.wikipedia.org/wiki/Receiver_operating_characteristic
Recall
= True Positive Rate (TPR)
= Sensitivity
= Hit Rate

Precision
= Positive Predictive Value (PPV)

\[
Recall = \frac{TP}{TP + FN} \quad \text{Precision} = \frac{TP}{TP + FP}
\]
Social Network Analysis
Jennifer Golbeck (2013), Analyzing the Social Web, Morgan Kaufmann

Social Network Analysis (SNA)
Facebook TouchGraph
Social Network Analysis

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
Social Network Analysis

• A **social network** is a social structure of people, related (directly or indirectly) to each other through a common relation or interest

• **Social network analysis (SNA)** is the study of social networks to understand their structure and behavior

Source: (c) Jaideep Srivastava, srivasta@cs.umn.edu, Data Mining for Social Network Analysis
Using Social Network Analysis, you can get answers to questions like:

- How highly connected is an entity within a network?
- What is an entity's overall importance in a network?
- How central is an entity within a network?
- How does information flow within a network?
Social Network Analysis

• Social network is the study of social entities (people in an organization, called actors), and their interactions and relationships.

• The interactions and relationships can be represented with a network or graph,
  – each vertex (or node) represents an actor and
  – each link represents a relationship.

• From the network, we can study the properties of its structure, and the role, position and prestige of each social actor.

• We can also find various kinds of sub-graphs, e.g., communities formed by groups of actors.

Social Network and the Web

- Social network analysis is useful for the Web because the Web is essentially a virtual society, and thus a virtual social network,
  - Each page: a social actor and
  - Each hyperlink: a relationship.

- Many results from social network can be adapted and extended for use in the Web context.

- Two types of social network analysis,
  - Centrality
  - Prestige

  closely related to hyperlink analysis and search on the Web.

Social Network Analysis (SNA)

Centrality

Prestige
Degree

Source: https://www.youtube.com/watch?v=89mxOdwPfxA
Degree

A: 2
B: 4
C: 2
D: 1
E: 1

Source: https://www.youtube.com/watch?v=89mxOdwPfxA
Density

Source: https://www.youtube.com/watch?v=89mxOdwpfxA
Density

Edges (Links): 5
Total Possible Edges: 10
Density: \( \frac{5}{10} = 0.5 \)

Source: https://www.youtube.com/watch?v=89mxOdwPfxA
Nodes (n): 10
Edges (Links): 13
Total Possible Edges: \( \frac{n \times (n-1)}{2} = \frac{10 \times 9}{2} = 45 \)
Density: \( \frac{13}{45} = 0.29 \)
Which Node is Most Important?
Centrality

• Important or prominent actors are those that are linked or involved with other actors extensively.

• A person with extensive contacts (links) or communications with many other people in the organization is considered more important than a person with relatively fewer contacts.

• The links can also be called ties. A central actor is one involved in many ties.

Social Network Analysis (SNA)

• Degree Centrality
• Betweenness Centrality
• Closeness Centrality
Alice has the highest degree centrality, which means that she is quite active in the network. However, she is not necessarily the most powerful person because she is only directly connected within one degree to people in her clique—she has to go through Rafael to get to other cliques.

Social Network Analysis: Degree Centrality

• Degree centrality is simply the number of direct relationships that an entity has.

• An entity with high degree centrality:
  – Is generally an active player in the network.
  – Is often a connector or hub in the network.
  – Is not necessarily the most connected entity in the network (an entity may have a large number of relationships, the majority of which point to low-level entities).
  – May be in an advantaged position in the network.
  – May have alternative avenues to satisfy organizational needs, and consequently may be less dependent on other individuals.
  – Can often be identified as third parties or deal makers.

Social Network Analysis: Degree Centrality
Social Network Analysis: Degree Centrality

<table>
<thead>
<tr>
<th>Node</th>
<th>Score</th>
<th>Standardized Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>2/10 = 0.2</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>2/10 = 0.2</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>5/10 = 0.5</td>
</tr>
<tr>
<td>D</td>
<td>3</td>
<td>3/10 = 0.3</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>3/10 = 0.3</td>
</tr>
<tr>
<td>F</td>
<td>2</td>
<td>2/10 = 0.2</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>4/10 = 0.4</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>3/10 = 0.3</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>1/10 = 0.1</td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>1/10 = 0.1</td>
</tr>
</tbody>
</table>
Social Network Analysis: Betweenness Centrality

Rafael has the highest betweenness because he is between Alice and Aldo, who are between other entities. Alice and Aldo have a slightly lower betweenness because they are essentially only between their own cliques. Therefore, although Alice has a higher degree centrality, Rafael has more importance in the network in certain respects.

Social Network Analysis: Betweenness Centrality

- Betweenness centrality identifies an entity's position within a network in terms of its ability to make connections to other pairs or groups in a network.
- An entity with a high betweenness centrality generally:
  - Holds a favored or powerful position in the network.
  - Represents a single point of failure—take the single betweenness spanner out of a network and you sever ties between cliques.
  - Has a greater amount of influence over what happens in a network.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
Betweenness centrality: Connectivity

Number of shortest paths going through the actor
Betweenness Centrality

\[ C_B(i) = \sum_{j<k} \frac{g_{ik}(i)}{g_{jk}} \]

Where \( g_{jk} \) = the number of shortest paths connecting \( jk \)
\( g_{jk}(i) \) = the number that actor \( i \) is on.

Normalized Betweenness Centrality

\[ C'_B(i) = \frac{C_B(i)}{\left[ (n-1)(n-2)/2 \right]} \]

Number of pairs of vertices excluding the vertex itself

Source: https://www.youtube.com/watch?v=RXohUeNCJiU
A: Betweenness Centrality = 0

A → C: 0/1 = 0
A → D: 0/1 = 0
A → E: 0/1 = 0
B → C: 0/1 = 0
B → D: 0/1 = 0
B → E: 0/1 = 0
C → D: 0/1 = 0
C → E: 0/1 = 0
D → E: 0/1 = 0

Total: 0
Betweenness Centrality

B: Betweenness Centrality = 5

B:  
A→C: 0/1 = 0  
A→D: 1/1 = 1  
A→E: 1/1 = 1  
C→D: 1/1 = 1  
C→E: 1/1 = 1  
D→E: 1/1 = 1  

Total: 5
C: Betweenness Centrality = 0
Betweenness Centrality

A: 0
B: 5
C: 0
D: 0
E: 0
Which Node is Most Important?
Which Node is Most Important?
Betweenness Centrality

\[ C_B(i) = \sum_{j<k} \frac{g_{ik}(i)}{g_{jk}} \]
Betweenness Centrality

A: Betweenness Centrality = 0

A

B

C

D

E

B→C: 0/1 = 0
B→D: 0/1 = 0
B→E: 0/1 = 0
C→D: 0/1 = 0
C→E: 0/1 = 0
D→E: 0/1 = 0

Total: 0
Rafael has the highest closeness centrality because he can reach more entities through shorter paths. As such, Rafael's placement allows him to connect to entities in his own clique, and to entities that span cliques.
Social Network Analysis: Closeness Centrality

• Closeness centrality measures how quickly an entity can access more entities in a network.
• An entity with a high closeness centrality generally:
  – Has quick access to other entities in a network.
  – Has a short path to other entities.
  – Is close to other entities.
  – Has high visibility as to what is happening in the network.

Source: http://www.fmsasg.com/SocialNetworkAnalysis/
Social Network Analysis: Closeness Centrality

C: Closeness Centrality = 15/9 = 1.67

C → A: 1
C → B: 1
C → D: 1
C → E: 1
C → F: 2
C → G: 1
C → H: 2
C → I: 3
C → J: 3

Total = 15
Social Network Analysis: Closeness Centrality

G: Closeness Centrality = 14/9 = 1.56
Social Network Analysis: Closeness Centrality

H: Closeness Centrality = 17/9 = 1.89
Social Network Analysis: Closeness Centrality

G: Closeness Centrality = 14/9 = 1.56
C: Closeness Centrality = 15/9 = 1.67
H: Closeness Centrality = 17/9 = 1.89
Social Network Analysis: Closeness Centrality

Sum of the reciprocal distances

$$C_C(p_k) = \sum_{i=1}^{n} d(p_i, p_k)^{-1}$$

where $d(p_j, p_k)$ is the geodesic distance (shortest paths) linking $p_j, p_k$
Social Network Analysis: Betweenness Centrality

\[ C_B(p_k) = \sum_{i<j}^{n} \frac{g_{ij}(p_k)}{g_{ij}}; \quad i \neq j \neq k \]

where \( g_{ij} \) is the geodesic distance (shortest paths) linking \( p_i \) and \( p_j \) and \( g_{ij}(p_k) \) is the geodesic distance linking \( p_i \) and \( p_j \) that contains \( p_k \).
Social Network Analysis: Degree Centrality

\[ C_D(p_k) = \sum_{i=1}^{n} a(p_i, p_k) \]

where \( a(p_i, p_k) = 1 \) if and only if \( p_i \) and \( p_k \) are connected by a line
0 otherwise

\[ C'_D(p_k) = \frac{\sum_{i=1}^{n} a(p_i, p_k)}{n-1} \]
Alice and Rafael are closer to other highly close entities in the network. Bob and Frederica are also highly close, but to a lesser value.

Social Network Analysis: Eigenvalue

- Eigenvalue measures how close an entity is to other highly close entities within a network. In other words, Eigenvalue identifies the most central entities in terms of the global or overall makeup of the network.

- A high Eigenvalue generally:
  - Indicates an actor that is more central to the main pattern of distances among all entities.
  - Is a reasonable measure of one aspect of centrality in terms of positional advantage.

Eigenvector centrality:

Importance of a node depends on the importance of its neighbors
Hubs are entities that point to a relatively large number of authorities. They are essentially the mutually reinforcing analogues to authorities. Authorities point to high hubs. Hubs point to high authorities. You cannot have one without the other.
Social Network Analysis: Hub and Authority

• Entities that many other entities point to are called Authorities. In Sentinel Visualizer, relationships are directional—they point from one entity to another.

• If an entity has a high number of relationships pointing to it, it has a high authority value, and generally:
  – Is a knowledge or organizational authority within a domain.
  – Acts as definitive source of information.

Source: http://www.fmsasq.com/SocialNetworkAnalysis/
### Social Network Analysis

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Degree</th>
<th>Betweenness</th>
<th>Closeness</th>
<th>Eigenvector</th>
<th>Hub</th>
<th>Authority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Osama bin Laden</td>
<td>Person</td>
<td>44</td>
<td>0.92043209235823</td>
<td>1</td>
<td>0.0271</td>
<td>0</td>
<td>0.011</td>
</tr>
<tr>
<td>Abdallah Al-Halabi</td>
<td>Person</td>
<td>2</td>
<td>0</td>
<td>0.65486725663797</td>
<td>0.0001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Abu Mussab al-Zarqawi</td>
<td>Person</td>
<td>84</td>
<td>0.93488784732609</td>
<td>0.86945169712794</td>
<td>0.7028</td>
<td>0.6572</td>
<td>0.1076</td>
</tr>
<tr>
<td>Al Qaeda</td>
<td>Terrorist Organizational Entity</td>
<td>85</td>
<td>1</td>
<td>0.96242774566634</td>
<td>0.0416</td>
<td>0.3941</td>
<td>0.0156</td>
</tr>
<tr>
<td>Ayman Al-Zawahiri</td>
<td>Person</td>
<td>14</td>
<td>0.06579490878324</td>
<td>0.71612903225801</td>
<td>0</td>
<td>0</td>
<td>0.0173</td>
</tr>
<tr>
<td>Enaan Arnaout</td>
<td>Person</td>
<td>4</td>
<td>0.31189325814109</td>
<td>0.65680473372701</td>
<td>0.0001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Imad Eddin Barakat Yarbas</td>
<td>Person</td>
<td>11</td>
<td>0.06504958903841</td>
<td>0.70401691331900</td>
<td>0.0015</td>
<td>0</td>
<td>0.0025</td>
</tr>
<tr>
<td>Khalid Shaikh Mohammed</td>
<td>Person</td>
<td>32</td>
<td>0.33991646472406</td>
<td>0.86609817934502</td>
<td>0.002</td>
<td>0</td>
<td>0.1528</td>
</tr>
<tr>
<td>Mohamed Atta</td>
<td>Person</td>
<td>61</td>
<td>0.66626874007400</td>
<td>0.82019704433402</td>
<td>0.0015</td>
<td>0</td>
<td>0.6816</td>
</tr>
</tbody>
</table>
Social Network Analysis (SNA) Tools

- UCINet
- Pajek
Summary

• Data Mining and Big Data Analytics
• Data Mining Process
• Tasks of Data Mining
• Evaluation of Data Mining
• Social Network Analysis

Source: Han & Kamber (2006)
References

• Jiawei Han and Micheline Kamber (2011), Data Mining: Concepts and Techniques, Third Edition, Elsevier
• Jennifer Golbeck (2013), Analyzing the Social Web, Morgan Kaufmann
• Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
• Hiroshi Ishikawa (2015), Social Big Data Mining, CRC Press