Practices of Business Intelligence

Text and Web Mining

Min-Yuh Day
Assistant Professor

Dept. of Information Management, Tamkang University

http://mail.tku.edu.tw/myday/
2014-04-30
週次 (Week)  日期 (Date)  內容 (Subject/Topics)
1  103/02/19  商業智慧導論 (Introduction to Business Intelligence)
2  103/02/26  管理決策支援系統與商業智慧 (Management Decision Support System and Business Intelligence)
3  103/03/05  企業績效管理 (Business Performance Management)
4  103/03/12  資料倉儲 (Data Warehousing)
5  103/03/19  商業智慧的資料探勘 (Data Mining for Business Intelligence)
6  103/03/26  商業智慧的資料探勘 (Data Mining for Business Intelligence)
7  103/04/02  教學行政觀摩日 (Off-campus study)
8  103/04/09  資料科學與巨量資料分析 (Data Science and Big Data Analytics)
<table>
<thead>
<tr>
<th>週次</th>
<th>日期</th>
<th>內容（Subject/Topics）</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>103/04/16</td>
<td>期中報告 (Midterm Project Presentation)</td>
</tr>
<tr>
<td>10</td>
<td>103/04/23</td>
<td>期中考試週 (Midterm Exam)</td>
</tr>
<tr>
<td>11</td>
<td>103/04/30</td>
<td>文字探勘與網路探勘 (Text and Web Mining)</td>
</tr>
<tr>
<td>12</td>
<td>103/05/07</td>
<td>意見探勘與情感分析 (Opinion Mining and Sentiment Analysis)</td>
</tr>
<tr>
<td>13</td>
<td>103/05/14</td>
<td>社會網路分析 (Social Network Analysis)</td>
</tr>
<tr>
<td>14</td>
<td>103/05/21</td>
<td>期末報告 (Final Project Presentation)</td>
</tr>
<tr>
<td>15</td>
<td>103/05/28</td>
<td>畢業考試週 (Final Exam)</td>
</tr>
</tbody>
</table>
Learning Objectives

• Describe text mining and understand the need for text mining
• Differentiate between text mining, Web mining and data mining
• Understand the different application areas for text mining
• Know the process of carrying out a text mining project
• Understand the different methods to introduce structure to text-based data

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

• Describe Web mining, its objectives, and its benefits
• Understand the three different branches of Web mining
  – Web content mining
  – Web structure mining
  – Web usage mining
• Understand the applications of these three mining paradigms

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

• Text Mining: Applications and Theory
• Web Mining and Social Networking
• Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites
• Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data
• Search Engines – Information Retrieval in Practice
Text Mining

Web Mining and Social Networking

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

http://www.amazon.com/Web-Data-Mining-Data-Centric-Applications/dp/3540378812
Search Engines: Information Retrieval in Practice

http://www.amazon.com/Search-Engines-Information-Retrieval-Practice/dp/0136072240
Text Mining

• Text mining (text data mining)
  – the process of deriving high-quality information from text
• Typical text mining tasks
  – text categorization
  – text clustering
  – concept/entity extraction
  – production of granular taxonomies
  – sentiment analysis
  – document summarization
  – entity relation modeling
    • i.e., learning relations between named entities.

http://en.wikipedia.org/wiki/Text_mining
Web Mining

• Web mining
  – discover useful information or knowledge from the **Web hyperlink structure, page content, and usage data.**

• Three types of web mining tasks
  – Web structure mining
  – Web content mining
  – Web usage mining

Mining Text For Security...

Cluster 1
- (L) Kampala
- (L) Uganda
- (P) Yoweri Museveni
- (L) Sudan
- (L) Khartoum
- (L) Southern Sudan

Cluster 2
- (P) Timothy McVeigh
- (P) Oklahoma City
- (P) Terry Nichols

Cluster 3
- (E) election
- (P) Norodom Ranariddh
- (P) Norodom Sihanouk
- (L) Bangkok
- (L) Cambodia
- (L) Phnom Penh
- (L) Thailand
- (P) Hun Sen
- (O) Khmer Rouge
- (P) Pol Pot

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

- 85-90 percent of all corporate data is in some kind of unstructured form (e.g., text)
- Unstructured corporate data is doubling in size every 18 months
- Tapping into these information sources is not an option, but a need to stay competitive
- Answer: **text mining**
  - A semi-automated process of extracting knowledge from unstructured data sources
  - a.k.a. **text data mining or knowledge discovery in textual databases**

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

• Both seek for novel and useful patterns
• Both are semi-automated processes
• Difference is the nature of the data:
  – Structured versus unstructured data
  – **Structured data**: in databases
  – **Unstructured data**: Word documents, PDF files, text excerpts, XML files, and so on
• Text mining – first, impose structure to the data, then mine the structured data

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

• Benefits of text mining are obvious especially in text-rich data environments
  – e.g., law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), marketing (customer comments), etc.

• Electronic communication records (e.g., Email)
  – Spam filtering
  – Email prioritization and categorization
  – Automatic response generation

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

• Information extraction
• Topic tracking
• Summarization
• Categorization
• Clustering
• Concept linking
• Question answering

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

• Unstructured or semistructured data
• Corpus (and corpora)
• Terms
• Concepts
• Stemming
• Stop words (and include words)
• Synonyms (and polysemes)
• Tokenizing

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

• Term dictionary
• Word frequency
• Part-of-speech tagging (POS)
• Morphology
• Term-by-document matrix (TDM)
  – Occurrence matrix
• Singular Value Decomposition (SVD)
  – Latent Semantic Indexing (LSI)

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

• What is a patent?
  – “exclusive rights granted by a country to an inventor for a limited period of time in exchange for a disclosure of an invention”

• How do we do patent analysis (PA)?

• Why do we need to do PA?
  – What are the benefits?
  – What are the challenges?

• How does text mining help in PA?

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

• Structuring a collection of text
  – Old approach: bag-of-words
  – New approach: natural language processing

• NLP is ...
  – a very important concept in text mining
  – a subfield of artificial intelligence and computational linguistics
  – the studies of "understanding" the natural human language

• Syntax versus semantics based text mining

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

• What is “Understanding”?  
  – Human understands, what about computers?  
  – Natural language is vague, context driven  
  – True understanding requires extensive knowledge of a topic  
  
  – Can/will computers ever understand natural language the same/accurate way we do?

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

• Challenges in NLP
  – Part-of-speech tagging
  – Text segmentation
  – Word sense disambiguation
  – Syntax ambiguity
  – Imperfect or irregular input
  – Speech acts

• Dream of AI community
  – to have algorithms that are capable of automatically reading and obtaining knowledge from text

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

• WordNet
  – A laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets
  – A major resource for NLP
  – Need automation to be completed

• Sentiment Analysis
  – A technique used to detect favorable and unfavorable opinions toward specific products and services
  – CRM application

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

- Information retrieval (IR)
- Information extraction (IE)
- Named-entity recognition (NER)
- Question answering (QA)
- Automatic summarization
- Natural language generation and understanding (NLU)
- Machine translation (ML)
- Foreign language reading and writing
- Speech recognition
- Text proofing
- Optical character recognition (OCR)

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

- Marketing applications
  - Enables better CRM

- Security applications
  - ECHELON, OASIS
  - Deception detection (…)

- Medicine and biology
  - Literature-based gene identification (…)

- Academic applications
  - Research stream analysis

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

• Application Case: Mining for Lies
• Deception detection
  – A difficult problem
  – If detection is limited to only text, then the problem is even more difficult
• The study
  – analyzed text based testimonies of person of interests at military bases
  – used only text-based features (cues)

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

• Application Case: Mining for Lies

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

- **Application Case: Mining for Lies**

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Cues</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quantity</strong></td>
<td>Verb count, noun-phrase count, ...</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td>Avg. no of clauses, sentence length, ...</td>
</tr>
<tr>
<td><strong>Uncertainty</strong></td>
<td>Modifiers, modal verbs, ...</td>
</tr>
<tr>
<td><strong>Nonimmediacy</strong></td>
<td>Passive voice, objectification, ...</td>
</tr>
<tr>
<td><strong>Expressivity</strong></td>
<td>Emotiveness</td>
</tr>
<tr>
<td><strong>Diversity</strong></td>
<td>Lexical diversity, redundancy, ...</td>
</tr>
<tr>
<td><strong>Informality</strong></td>
<td>Typographical error ratio</td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td>Spatiotemporal, perceptual information ...</td>
</tr>
<tr>
<td><strong>Affect</strong></td>
<td>Positive affect, negative affect, etc.</td>
</tr>
</tbody>
</table>

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

• Application Case: Mining for Lies
  – 371 usable statements are generated
  – 31 features are used
  – Different feature selection methods used
  – 10-fold cross validation is used
  – Results (overall % accuracy)
    • Logistic regression 67.28
    • Decision trees 71.60
    • Neural networks 73.46

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Applications
(gene/protein interaction identification)

..expression of Bcl-2 is correlated with insufficient white blood cell death and activation of p53.

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

Context diagram for the text mining process

- Unstructured data (text)
- Structured data (databases)

Extract knowledge from available data sources

Context-specific knowledge

- Software/hardware limitations
- Privacy issues
- Linguistic limitations
- Domain expertise
- Tools and techniques

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

The three-step text mining process:

**Task 1: Establish the Corpus:**
Collect & Organize the Domain Specific Unstructured Data

- The inputs to the process include a variety of relevant unstructured (and semi-structured) data sources such as text, XML, HTML, etc.

**Task 2: Create the Term-Document Matrix:**
Introduce Structure to the Corpus

- The output of Task 1 is a collection of documents in some digitized format for computer processing.

- The output of Task 2 is a flat file called term-document matrix where the cells are populated with the term frequencies.

**Task 3: Extract Knowledge:**
Discover Novel Patterns from the T-D Matrix

- The output of Task 3 is a number of problem specific classification, association, clustering models and visualizations.

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

• **Step 1:** Establish the corpus
  – Collect all relevant unstructured data
    (e.g., textual documents, XML files, emails, Web pages, short notes, voice recordings...)
  – Digitize, standardize the collection
    (e.g., all in ASCII text files)
  – Place the collection in a common place
    (e.g., in a flat file, or in a directory as separate files)

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

**Step 2: Create the Term–by–Document Matrix**

<table>
<thead>
<tr>
<th>Terms</th>
<th>investment risk</th>
<th>project management</th>
<th>software engineering</th>
<th>development</th>
<th>SAP</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 2</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 3</td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 4</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 5</td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 6</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

• **Step 2:** Create the Term–by–Document Matrix (TDM), cont.
  – Should all terms be included?
    • Stop words, include words
    • Synonyms, homonyms
    • Stemming
  – What is the best representation of the indices (values in cells)?
    • Row counts; binary frequencies; log frequencies;
    • Inverse document frequency

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

• **Step 2:** Create the Term–by–Document Matrix (TDM), cont.
  
  – TDM is a sparse matrix. How can we reduce the dimensionality of the TDM?
    
    • Manual - a domain expert goes through it
    • Eliminate terms with very few occurrences in very few documents (?)
    • Transform the matrix using singular value decomposition (SVD)
    • SVD is similar to principle component analysis

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

- **Step 3:** Extract patterns/knowledge
  - Classification (text categorization)
  - Clustering (natural groupings of text)
    - Improve search recall
    - Improve search precision
    - Scatter/gather
    - Query-specific clustering
  - Association
  - Trend Analysis (...)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Application
(research trend identification in literature)

• Mining the published IS literature
  – MIS Quarterly (MISQ)
  – Journal of MIS (JMIS)
  – Information Systems Research (ISR)
  – Covers 12-year period (1994-2005)
  – 901 papers are included in the study
  – Only the paper abstracts are used
  – 9 clusters are generated for further analysis

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
# Text Mining Application
(research trend identification in literature)

<table>
<thead>
<tr>
<th>Journal</th>
<th>Year</th>
<th>Author(s)</th>
<th>Title</th>
<th>Vol/No</th>
<th>Pages</th>
<th>Keywords</th>
<th>Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>MISQ</td>
<td>2005</td>
<td>A. Malhotra, S. Gosain and O. A. El Sawy</td>
<td>Absorptive capacity configurations in supply chains: Gearing for partner-enabled market knowledge creation</td>
<td>29/1</td>
<td>145-187</td>
<td>knowledge management supply chain absorptive capacity interorganizational information systems configuration approaches</td>
<td>The need for continual value innovation is driving supply chains to evolve from a pure transactional focus to leveraging interorganizational partner ships for sharing</td>
</tr>
<tr>
<td>ISR</td>
<td>1999</td>
<td>D. Robey and M. C. Boudreau</td>
<td>Accounting for the contradictory organizational consequences of information technology: Theoretical directions and methodological implications</td>
<td>2-Oct</td>
<td>167-185</td>
<td>organizational transformation impacts of technology organization theory research methodology intraorganizational power electronic communication mis implementation culture systems</td>
<td>Although much contemporary thought considers advanced information technologies as either determinants or enablers of radical organizational change, empirical studies have revealed inconsistent findings to support the deterministic logic implicit in such arguments. This paper reviews the contradictory</td>
</tr>
<tr>
<td>JMIS</td>
<td>2001</td>
<td>R. Aron and E. K. Clemons</td>
<td>Achieving the optimal balance between investment in quality and investment in self-promotion for information products</td>
<td>18/2</td>
<td>65-88</td>
<td>information products internet advertising product positioning signaling signaling games</td>
<td>When producers of goods (or services) are confronted by a situation in which their offerings no longer perfectly match consumer preferences, they must determine the extent to which the advertised features of</td>
</tr>
</tbody>
</table>

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Application
(research trend identification in literature)

Source: Turban et al. (2011), Decision Support and Business Intelligence Systems
Text Mining Application
(research trend identification in literature)

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

• Commercial Software Tools
  – SPSS PASW Text Miner
  – SAS Enterprise Miner
  – Statistica Data Miner
  – ClearForest, ...

• Free Software Tools
  – RapidMiner
  – GATE
  – Spy-EM, ...

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

https://www.youtube.com/watch?v=l1rYdrRCZJ4
Web Mining Overview

• Web is the largest repository of data
• Data is in HTML, XML, text format
• Challenges (of processing Web data)
  – The Web is too big for effective data mining
  – The Web is too complex
  – The Web is too dynamic
  – The Web is not specific to a domain
  – The Web has everything

• Opportunities and challenges are great!

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

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

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

User / Customer

Website

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

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
Web Mining Success Stories

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

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

<table>
<thead>
<tr>
<th>Product Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angoss Knowledge WebMiner</td>
<td>angoss.com</td>
</tr>
<tr>
<td>ClickTracks</td>
<td>clicktracks.com</td>
</tr>
<tr>
<td>LiveStats from DeepMetrix</td>
<td>deepmetrix.com</td>
</tr>
<tr>
<td>Megaputer WebAnalyst</td>
<td>megaputer.com</td>
</tr>
<tr>
<td>MicroStrategy Web Traffic Analysis</td>
<td>microstrategy.com</td>
</tr>
<tr>
<td>SAS Web Analytics</td>
<td>sas.com</td>
</tr>
<tr>
<td>SPSS Web Mining for Clementine</td>
<td>spss.com</td>
</tr>
<tr>
<td>WebTrends</td>
<td>webtrends.com</td>
</tr>
<tr>
<td>XML Miner</td>
<td>scientio.com</td>
</tr>
</tbody>
</table>

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

• Evaluation of Information Retrieval
• Evaluation of Classification Model (Prediction)
  – Accuracy
  – Precision
  – Recall
  – F-score
Accuracy

Precision

Validity

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

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

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%), and 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
### True Class (actual value)

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

**Total**

- P (Positive)
- N (Negative)

### 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

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

### Graph

This graph illustrates the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate against the False Positive Rate (1 - Specificity) for different values of a diagnostic test. The area under the ROC curve (AUC) ranges from 0.5 to 1.0, where 1.0 represents perfect classification.

True Positive Rate (Sensitivity) = \( \frac{TP}{TP + FN} \)

**Sensitivity**

= True Positive Rate

= Recall

= Hit rate

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

\[ \text{Recall} = \frac{TP}{TP + FN} \]
Specificity
= True Negative Rate
= \frac{TN}{N}
= \frac{TN}{(TP + TN)}

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

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

Source:  http://en.wikipedia.org/wiki/Receiver_operating_characteristic
Precision

= Positive Predictive Value (PPV)

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

Recall

= True Positive Rate (TPR)

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

Recall

= Sensitivity

= Hit Rate

F1 score (F-score)(F-measure)

is the harmonic mean of precision and recall

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

A

\[
\begin{array}{c|c|c}
& TP & FP \\
\hline
TP & 63 & 28 \\
FN & 37 & 72 \\
\hline
91 & 109 & 200 \\
\end{array}
\]

TPR = 0.63
FPR = 0.28
PPV = 0.69
F1 = 0.66
ACC = 0.68

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

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

Specificity
= True Negative Rate
= \frac{TN}{N}
= \frac{TN}{(TP + TN)}

\[
Specificity = \frac{TN}{TN + FP}
\]

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

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

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

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

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

### A

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

**TPR** = 0.63

**FPR** = 0.28

**PPV** = 0.69

= \frac{63}{63+28}

= \frac{63}{91}

**F1** = 0.66

= \frac{2 \times (0.63 \times 0.69)}{0.63 + 0.69}

= \frac{2 \times 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} = 67.5

### B

<table>
<thead>
<tr>
<th></th>
<th>TP (77)</th>
<th>FP (77)</th>
<th>FN (23)</th>
<th>TN (23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td>200</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TPR** = 0.77

**FPR** = 0.77

**PPV** = 0.50

= \frac{77}{77+23}

= \frac{77}{100}

**F1** = 0.61

= \frac{2 \times (0.77 \times 0.50)}{0.77 + 0.50}

= \frac{2 \times 77}{154}

= \frac{(0.77 + 0.50)}{2} = \frac{1.27}{2} = 0.635

**ACC** = 0.50

= \frac{77 + 23}{200}

= \frac{100}{200} = 50

----

**Recall**

= True Positive Rate (TPR)

= Sensitivity

= Hit Rate

**Precision**

= Positive Predictive Value (PPV)

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

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

---

Recall

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

Precision

= Positive Predictive Value (PPV)

Summary

• Text Mining
• Web Mining
References

• Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006, Elsevier
• Michael W. Berry and Jacob Kogan, Text Mining: Applications and Theory, 2010, Wiley
• Guandong Xu, Yanchun Zhang, Lin Li, Web Mining and Social Networking: Techniques and Applications, 2011, Springer
• Matthew A. Russell, Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites, 2011, O'Reilly Media
• Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2009, Springer