Data Mining
Chapter1:DM Overview

Associate Professor Yachai Limpiyakorn, Ph.D.

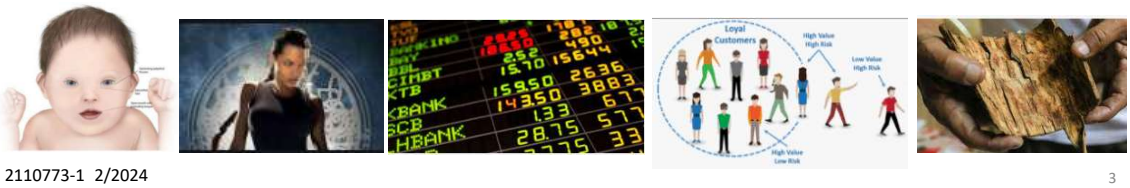
รศ.ดร. ญาใจ ลิ่มปิยะกรณ์

บทนำ

- การทำเหมืองข้อมูล (Data Mining) เป็นศาสตร์ที่เกิดขึ้นราวต้นทศวรรษ 1990s หรือที่รู้จักในราวทศวรรษ 1970s ว่า การค้นพบความรู้ในฐานข้อมูลขนาดใหญ่ (Knowledge Discovery in large Databases – KDD) หมายถึง กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหา แพตเทิร์น (patterns) และ ความสัมพันธ์ (associations) ที่ซ่อน (hidden) อยู่ในชุดข้อมูลนั้น กระบวนการดังกล่าวมีความเป็นอัตโนมัติ (automation) ไม่สามารถประมวลผลได้ด้วยมือ ต้องใช้คอมพิวเตอร์เข้าช่วยเหลือ เนื่องจากข้อมูลมีปริมาณมาก มีเกร็ดเกี่ยวกับการตั้งชื่อการทำเหมืองข้อมูลว่าเป็นการตั้งชื่อที่ผิดหรือไม่เหมาะสม (misnomer) เนื่องจากโดยทั่วไปแล้ว คำนามที่ตามหลังคำว่า “การทำเหมือง” มักเป็นสิ่งมีค่ามีราคา เช่น การทำเหมืองทอง เหมืองเพชร เหมืองถ่านหิน เป็นต้น ซึ่งสิ่งมีค่าเหล่านั้นมักจะมีปริมาณน้อยมากเมื่อเปรียบเทียบกับปริมาณมหาศาลของกรวด หิน ดิน ทราย ที่ปะปนกันอยู่ สำหรับชื่อ “การทำเหมืองข้อมูล”นั้น เป็นการตั้งชื่อที่ผิดแปลกไปจากปกติ ซึ่งจากหลักการดังกล่าวข้างต้น ควรตั้งชื่อว่า “การทำเหมืองความรู้” (knowledge mining) แต่ศัพท์ “การทำเหมืองข้อมูล” เป็นคำที่นิยมใช้แพร่หลายสืบต่อกันมา โดยมุ่งเน้นทำให้เห็นภาพที่ชัดเจนถึงกระบวนการค้นหาสิ่งมีค่า คือ ความรู้ (knowledge) จากกองวัตถุดิบปริมาณมหาศาล คือ ข้อมูล

บทนำ

- ผลลัพธ์จากการทำเหมืองข้อมูล คือ ความรู้ ซึ่งเป็นแพตเทิร์น และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลหนึ่งๆ
- Pattern– เหตุการณ์หรือสิ่งที่เกิดซ้ำแล้วซ้ำอีก (repeat) จนสามารถทำนายได้ (predictable) ตัวอย่าง:
  - คนที่เป็นโรคชนิดหนึ่ง มักจะมียีนส์ลักษณะแบบนี้ ซึ่งความรู้ดังกล่าวสามารถใช้ในการวินิจฉัยโรคทางการแพทย์ได้
- แพตเทิร์นหรือความสัมพันธ์ของลักษณะต่างๆที่พบในข้อมูลดิบ จึงมีศักยภาพพัฒนาไปสู่ความรู้ที่ช่วยให้มนุษย์นำไปใช้ประโยชน์ในด้านต่างๆได้ เช่น ในเชิงธุรกิจ การวินิจฉัยหรือรักษาโรคทางการแพทย์ การกีฬา ฯลฯ
- **Patterns (knowledge representation)**



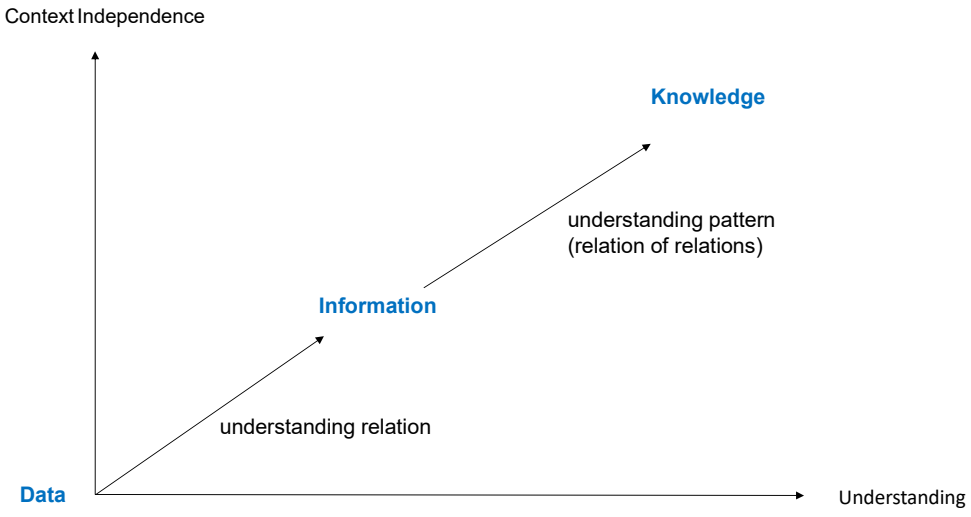
Data mining: what is it?

- Extraction of useful patterns from *data sources, e.g. databases, texts, web, images*
- Analysis of data for relationships that have *not previously been discovered or known*.
- A term coined for a new *discipline* lying at the interface of *database technology, machine learning, pattern recognition, statistics, and visualization*.
- Key element in much more elaborate process called "**Knowledge Discovery in Databases**".
- The efficient extraction of previously unknown patterns in very large data bases.

Motivations – data explosion problem

- Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories.
- More data is generated:
  - Bank, telecom, other business transactions ...
  - Scientific data: astronomy, biology, etc.
  - Web, text, and e-commerce
- We are drowning in data, but starving for knowledge! → data tomb
- Data Mining can help discover knowledge nuggets

ข้อมูล สารสนเทศ และความรู้



Data Mining vs. Query Processing

Data query can help us find answer to questions v ask about info stored in data repository.

DM gives us the ability to find answer to questions v never thought about asking

Machine Learning, Tom Mitchell, McGraw Hill, 1997.

Machine Learning is the study of computer algorithms that improve automatically through experience. Applications range from datamining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests.

This book provides a single source introduction to the field. It is written for advanced undergraduate and graduate students, and for developers and researchers in the field. No prior background in artificial intelligence or statistics is assumed.

Chapter Outline: (or see the detailed table of contents (postscript))

- 1. Introduction
- 2. Concept Learning and the General-to-Specific Ordering
- 3. Decision Tree Learning
- 4. Artificial Neural Networks
- 5. Evaluating Hypotheses
- 6. Bayesian Learning
- 7. Computational Learning Theory
- 8. Instance-Based Learning
- 9. Genetic Algorithms
- 10. Learning Sets of Rules
- 11. Analytical Learning
- 12. Combining Inductive and Analytical Learning
- 13. Reinforcement Learning

414 pages. ISBN 0070428077

Database Community

Machine Learning

- Concept
  - an abstract or generic idea generalized from particular instances. [Webster]
- Induction-based Learning
  - process of forming a general concept definition by observing specific examples of the concept 2b learned

Abstract idea denoting all **objects** in a given category/ class of entities, leaving differences in their extension

Cannot be visualized

Bearers of meaning, as opposed to Agents of meaning (a single concept can be expressed by any number of languages)

Acquisition of concepts is studied in ML: supervised/ unsupervised

Result from reason

Inductive Reasoning (1)

- การให้เหตุผลโดยอาศัยข้อสังเกตหรือผลการทดลองจากหลายๆตัวอย่างมาสรุปเป็นข้อตกลง หรือข้อคาดเดาทั่วไป หรือคำพยากรณ์ ซึ่งจะเห็นว่าการจะนำเอาข้อสังเกตหรือผลการทดลองจากบางหน่วยมาสนับสนุนให้ได้ข้อตกลง หรือ ข้อความทั่วไปซึ่งกินความถึงทุกหน่วย ย่อมไม่สมเหตุสมผล เพราะเป็นการอนุมานเกินสิ่งที่กำหนดให้ หมายความว่า การให้เหตุผลแบบอุปนัยจะต้องมีกฎของความสมเหตุสมผลเฉพาะของตนเอง คือจะต้องมีข้อสังเกต หรือผลการทดลอง หรือ มีประสบการณ์ที่มากมายพอที่จะปักใจเชื่อได้ แต่ก็ยังไม่สามารถแน่ใจในผลสรุปได้เต็มที่ เหมือนการให้เหตุผลแบบนิรนัย กล่าวได้ว่าการให้เหตุผลแบบนิรนัยจะให้ความแน่นอน (certain) แต่การให้เหตุผลแบบอุปนัย จะให้ความน่าจะเป็น (probable)

Inductive Reasoning (2)

- "ปลาทุกชนิดออกลูกเป็นไข่"
- ใช้สร้าง Axiom (สัจพจน์)
  - "เส้นตรงสองเส้นตัดกันเพียงจุด ๆ เดียวเท่านั้น"
  - "เส้นมัธยฐานของสามเหลี่ยมใดๆ พบกันที่จุดๆหนึ่งเสมอ"
  - ในการบวกเลข 2 จำนวน พบว่า 1+2 = 2+1; 2+3 = 3+2; .....
    จะได้ว่า a + b = b + a
- ข้อสังเกต
  - ข้อสรุปของการให้เหตุผลแบบอุปนัยอาจจะไม่จริงเสมอไป
  - ข้อสรุปที่ได้จากการให้เหตุผลแบบอุปนัยไม่จำเป็นต้องเหมือนกัน

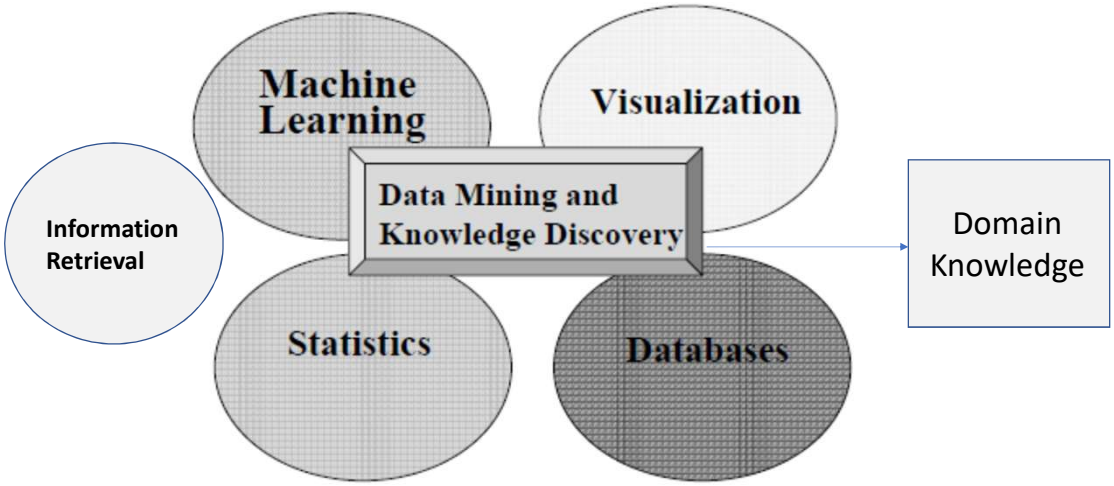
ตัวอย่าง กำหนด จำนวน 2, 4, 6 , a จงหา จำนวน a จะได้ a = ...

กำหนด จำนวน 2, 4, 6 , a จงหา จำนวน a จะได้ a = ...

- สัญชาน, การรับรู้
- กระบวนการที่มนุษย์ติดต่อสื่อสารกับสิ่งแวดล้อมรอบๆตัว โดยมนุษย์จะทำการตีความสิ่งแวดล้อมที่สัมผัสได้ แล้วตอบสนองกลับไปอย่างเหมาะสม แต่ละคนอาจจะตีความในสิ่งแวดล้อมที่เหมือนกันออกไปในทางต่างๆกัน ขึ้นอยู่กับพื้นฐานทางจิตใจและความคิดของแต่ละคน

Perception

Related Fields



Data Mining: On What Kind of Data?

- Attribute-value tables (standard form / data table)
- Relational data
- Structured data (graphs, workflows, ontologies, ...)
- Other more complex data repositories
  - Data warehouse
  - Time-series data and temporal data (time-related DB)
  - Text databases and multimedia databases
  - WWW
  - Spatial
  - Images: mining to discover the images using certain keywords alphanumeric data and patterns

Data Warehouse

- สถาปัตยกรรมแหล่งที่เก็บข้อมูลที่เกิดขึ้นมาสำหรับจัดเก็บข้อมูลจากหลายๆแหล่งที่มีโครงสร้างการจัดเก็บแตกต่างกัน (heterogeneous data sources) มารวมไว้ในที่เดียวกันด้วยโครงสร้างการจัดเก็บเดียวกัน เพื่ออำนวยความสะดวกในการบริหารการตัดสินใจ เทคโนโลยีคลังข้อมูลประกอบด้วย การทำความสะอาดข้อมูล การบูรณาการข้อมูล และ online analytical processing (OLAP) ซึ่งเป็นเทคนิคการวิเคราะห์ที่ทำหน้าที่การสรุป (summarization) การรวมเข้าเป็นหน่วยเดียว (consolidation) การรวมกลุ่ม (aggregation) อีกทั้งมีความสามารถในการแสดงสารสนเทศในมุมมองต่างๆได้ ถึงแม้ว่าเครื่องมือต่างๆของ OLAP สามารถสนับสนุนการวิเคราะห์แบบหลายมิติ (multidimensional) และการตัดสินใจ แต่ยังคงมีความต้องการเครื่องมืออื่นเพิ่มเติมสำหรับการวิเคราะห์เชิงลึก อาทิเช่น เครื่องมือในการทำเหมืองข้อมูลซึ่งมีความสามารถในการจำแนกประเภทข้อมูล (classification) การจัดกลุ่ม (clustering) การตรวจจับความแปลกแยก/ความผิดปกติ (outlier/anomaly detection) และการอธิบายลักษณะของการเปลี่ยนแปลงในชุดข้อมูลตามกาลเวลา

DB+SQL vs. DW+OLAP

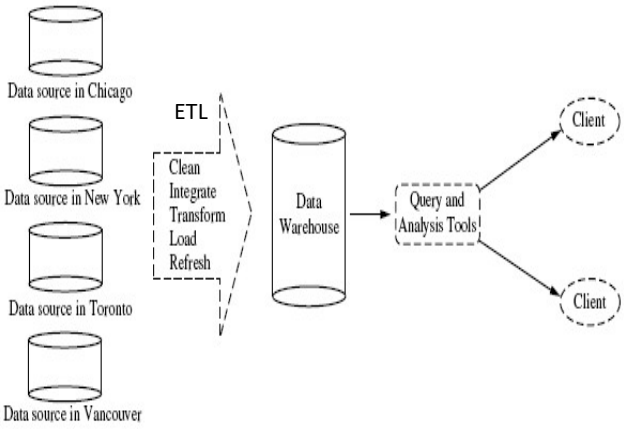
Database- A DB designed 4 processing the Day-to-day transaction of a company.

- Support daily operations
- Transaction-oriented
- Up-to-date data
- R/W
- Redundancy not allowed (Normalization)

Data Warehouse- A historical DB designed 4 decision support rather than transaction processing

- Decision support
- Subject-oriented
- Historical data
- Read only
- Redundancy allowed

Data Warehouse Architecture



Data Integration Challenges:

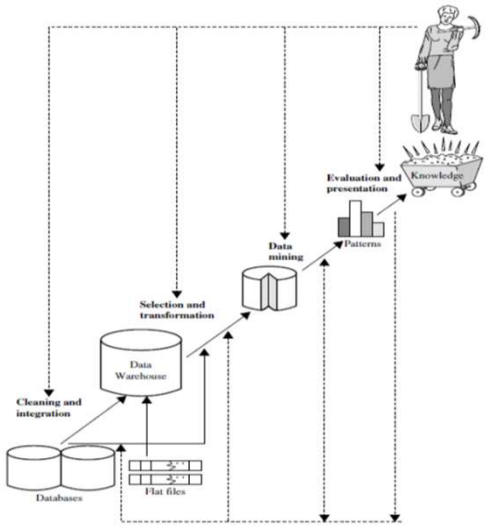
- ความแตกต่างทางโครงสร้างการจัดเก็บ (Schema Differences)
- ความแตกต่างจากการตั้งชื่อ (Naming Differences)
- ความแตกต่างจากประเภทข้อมูล (Data Type Differences)
- ความแตกต่างจากค่า (Value Differences)
- ความแตกต่างจากความหมาย (Semantic Differences)
- ความแตกต่างจากค่าข้อมูลขาดหาย (Missing Values)

Example

- บริษัท AutoX ขายรถยนต์ยี่ห้อหนึ่ง มีตัวแทนจำหน่าย 1000 แห่ง แต่ละแห่งมีฐานข้อมูลสต็อกรถยนต์ของตนเอง บริษัท AutoX ต้องการสร้างฐานข้อมูลกลาง โดยรวมข้อมูลจากฐานข้อมูลของตัวแทนจำหน่าย 1000 แห่ง มาไว้ที่เดียวกัน ประโยชน์ของการสร้างฐานข้อมูลกลาง ได้แก่
  - ช่วยตัวแทนจำหน่ายจัดหารถยนต์ตามความต้องการของลูกค้า ในกรณีที่ไม่มีรถในสต็อก
  - ช่วยนักวิเคราะห์ของบริษัทในการวิเคราะห์ทำนายตลาด
  - ช่วยฝ่ายผลิตในการปรับกำลังผลิตรถรุ่นที่เป็นที่ต้องการของตลาดได้ทันท่วงที

- ตัวแทนจำหน่ายทั้ง 1000 แห่ง ไม่ได้จัดเก็บข้อมูลด้วยโครงสร้างฐานข้อมูลเดียวกัน: ตัวแทนจำหน่าย A จัดเก็บ option อุปกรณ์รถยนต์ของรถแต่ละรุ่นด้วยตัวแปร boolean ในขณะที่ตัวแทนจำหน่าย B จัดเก็บ option อุปกรณ์รถยนต์แยกต่างหากในอีกตาราง
  - ตัวแทนจำหน่าย A : Cars (serialNo, model, color, autoTrans, cdPlayer, ...)
  - ตัวแทนจำหน่าย B : Autos (serial, model, color)

    Options (serial, option)
- การตั้งชื่อตารางที่ต่างกัน เช่น Cars กับ Autos หรือการตั้งชื่อคุณลักษณะที่ต่างกัน ได้แก่ serialNo กับ serial เป็นต้น
- ประเภทข้อมูลที่ต่างกัน เช่น ประเภทข้อมูลของ serial number ที่อาจเป็นเลขจำนวนเต็ม หรือสายอักษร
- ค่าข้อมูลที่ต่างกัน เช่น สีดำ ซึ่งอาจแทนด้วยค่า “black” หรือ “BL” ซึ่งอาจหมายถึง สีฟ้าในฐานข้อมูลตัวแทนจำหน่ายอีกแห่งหนึ่ง
- ศัพท์หรือค่าที่ใช้มีความหมายต่างกัน เช่น ฐานข้อมูลของตัวแทนจำหน่ายแห่งหนึ่ง Autos หมายถึง รถยนต์นั่งธรรมดาเท่านั้น ในขณะที่ฐานข้อมูลของอีกตัวแทนจำหน่าย หมายถึง รถยนต์นั่งธรรมดา และรถขับเคลื่อนสี่ล้อ
- ค่าบางค่าอาจไม่ถูกจัดเก็บในฐานข้อมูลหนึ่งๆ ตัวอย่างเช่น รุ่น (model) ของรถ ในฐานข้อมูลหนึ่งจัดเก็บรายละเอียดว่าเป็น Civic DX หรือ LX หรือ EX .ในขณะที่อีก



DM as a step in a KDD process

Knowledge

มีสาระ (nontrivial)

มีความถูกต้องใช้ได้จริง (valid)

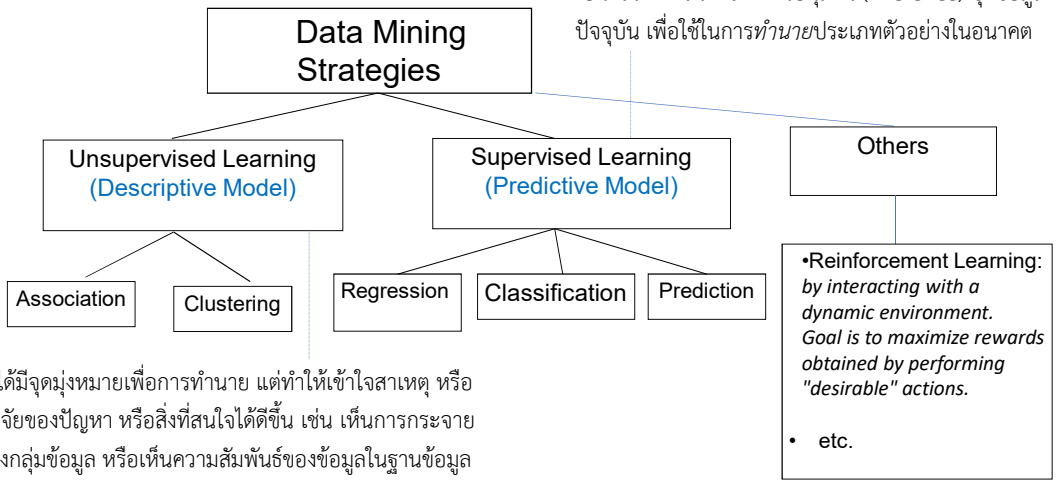
ไม่เคยทราบมาก่อน (novel / previously unknown)

นำไปใช้ให้เป็นประโยชน์ได้ (potentially useful)

น่าสนใจ (interesting)

สามารถทำความเข้าใจได้ (understandable)

Learning Category



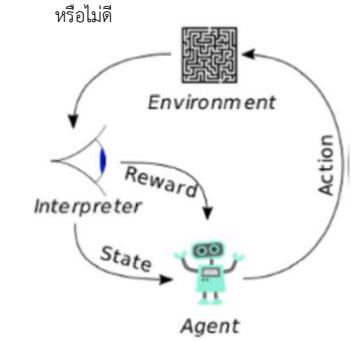
เป็นผลลัพธ์ที่สร้างจากการอนุมาน (inference) ชุดข้อมูลปัจจุบัน เพื่อใช้ในการ*ทำนาย*ประเภทตัวอย่างในอนาคต

ไม่ได้มีจุดมุ่งหมายเพื่อการทำนาย แต่ทำให้เข้าใจสาเหตุ หรือปัจจัยของปัญหา หรือสิ่งที่สนใจได้ดีขึ้น เช่น เห็นการกระจายของกลุ่มข้อมูล หรือเห็นความสัมพันธ์ของข้อมูลในฐานข้อมูล

Reinforcement Learning

การเรียนรู้แบบเสริมกำลัง

ML Algorithm ที่เรียนรู้บางสิ่งบางอย่างด้วยการลองผิดลองถูก โดยมีการเรียนรู้เกิดขึ้นระหว่างทางว่าการกระทำไหนดีหรือไม่ดี



en.wikipedia.org/wiki/Reinforcement_learning)

องค์ประกอบหลัก:

Agent – ผู้กระทำ Action

Action (a) – การกระทำของ Agent ที่ส่งผลบางอย่างต่อ Environment

Environment (e) – ระบบที่ Agent ต้องมีปฏิสัมพันธ์ด้วย

State (s) – สถานการณ์ของ Environment ที่ทาง Agent สามารถรับรู้ได้

Policy (π) – strategyที่ Agent ใช้ตัดสินใจเลือก Action: random or run a heuristic

Reward (R) – ตัวประเมินผลลัพธ์ที่เกิดจากการกระทำของ Agent เช่น คะแนน กำไรที่ได้รับ หรือ ผลแพ้ชนะ เป็นต้น

หลักการของ Reinforcement Learning คือ การเรียนรู้ของ Agent ที่เกิดจากปฏิสัมพันธ์แบบลองผิดลองถูกระหว่าง Agent กับ Environment โดย Agent จะสามารถรับรู้สถานการณ์ของ Environment ผ่าน State และใช้ Policy เลือกการกระทำ Action ที่ส่งผลต่อ Environment โดยหวังว่าจะได้ผลลัพธ์ Reward ที่ดีที่สุด รวมทั้งเรียนรู้ผ่านข้อผิดพลาดในอดีตที่เกิดขึ้น ตัวอย่างเช่น

การซื้อขายหุ้นให้ได้ผลตอบแทนมากที่สุด (Stock Trading Optimization) การค้นพบสูตรยารักษาโรคแบบใหม่ (Drug Discovery) หรือแม้กระทั่งระบบขับรถยนต์อัตโนมัติ (Self-driving car) เล่นเกมส์

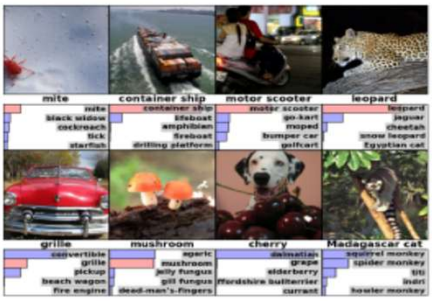
https://bigdata.go.th/big-data-101/introduction-to-reinforcement-learning/

Supervised learning: Classification

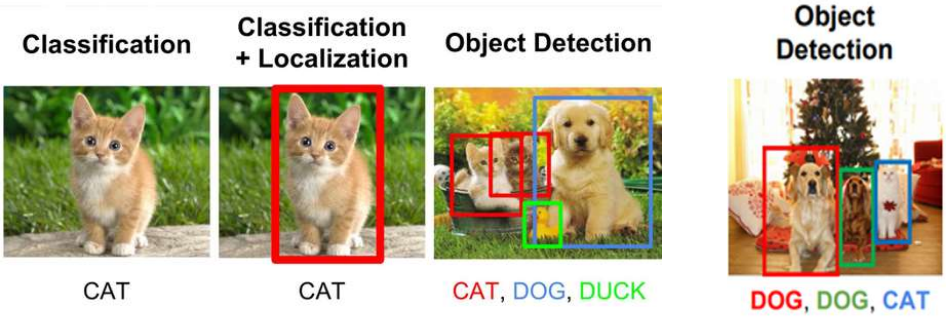
If the targets y_i represent categories, the problem is called classification.

Examples

- Handwritten digit recognition
- Transaction classification (fraud, valid)
- Object classification (cat, dog, hotdog, ...)
- Cancer detection



Find a function that generalizes relation
$f(x_i) \approx y_i$



Object Detection: a combination of image classification and object localization. It takes an image as input and produces one or more bounding boxes with the class label attached to each bounding box.

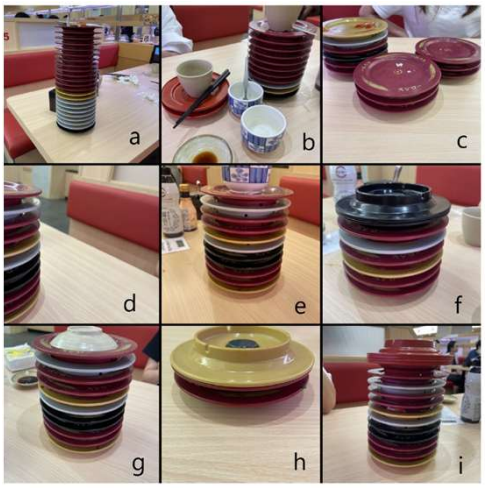
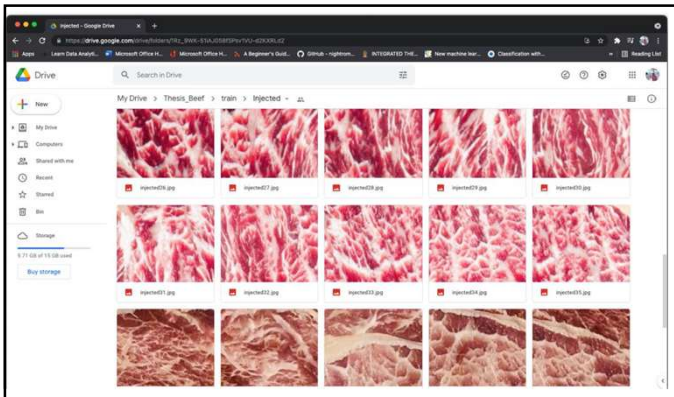
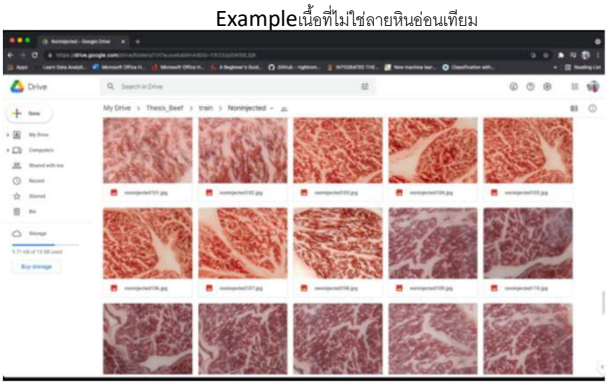


Figure 5: Example training images.



Exampleเนื้อลายหินอ่อนเทียม

Exampleเนื้อที่ไม่ใช่ลายหินอ่อนเทียม



Supervised learning: Regression

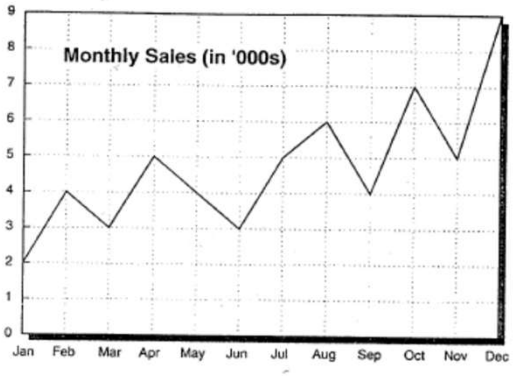
If the targets y_i represent continuous numbers, the problem is called regression.

Examples

- Stock market prediction
- Demand forecasting
- User involvement measurement
- Revenue analysis

Time series – prediction of time stamped variable

- Predictive data mining – time stamped variable; typical example – stocks, financial data, production planning.



Supervised Learning Distinction

ประเภท	เป้าหมาย	ผลลัพธ์
Classification	กำหนดค่าของผลลัพธ์ที่ยังไม่รู้ โดยเน้นที่พฤติกรรมปัจจุบันของตัวอย่างค้นถาม	Categorical
Regression		Numeric
Prediction	ทำนายผลลัพธ์ในอนาคตมากกว่าพฤติกรรมปัจจุบัน	Categorical or Numeric

ตัวอย่างงาน Classification

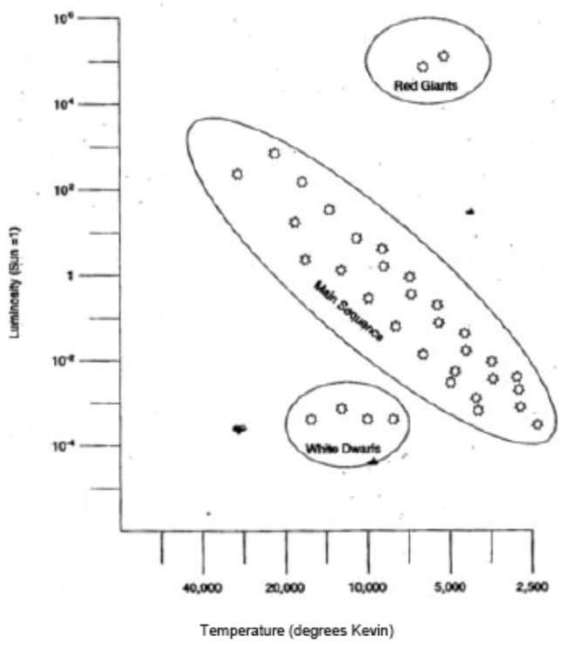
- ระบุคุณลักษณะของบุคคลที่อาบแดดแล้วผิวไหม้ง่ายกับผู้ที่ผิวไม่ไหม้
- จำแนกลักษณะผู้ยื่นขอสินเชื่อที่คาดว่าเป็นลูกค้าที่ดีหรือมีความเสี่ยง

ตัวอย่างงาน Estimation/Regression

- ประมาณความน่าจะเป็นที่ผู้ป่วยจะมีโอกาสหายจากโรค
- ประมาณความน่าจะเป็นที่รายการค่าใช้จ่ายหนึ่งจะเกิดจากบัตรเครดิตที่ถูกขโมย

ตัวอย่างงาน Prediction

- ทำนายราคาปิดเฉลี่ยตลาดดาวน์โจนส์สัปดาห์หน้า
- ทำนายว่าลูกค้ารายใดจะเปลี่ยนผู้ให้บริการโทรศัพท์มือถือในอีก 3 เดือนข้างหน้า



Clustering

- "Natural grouping" of examples similar in the sense of some distance measure
- Un-labeled data: the class of an example is not known
- Finding structure in unlabeled data

Association Mining

- Transaction data
- Market basket analysis
- {Cheese, Milk} → Bread [sup=50%, conf=80%]
- Association rule:

" 50% of customers buy all these products together

and 80% of customers who buy *cheese* and *milk* also buy *bread*"

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

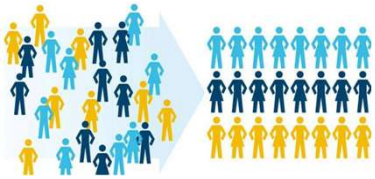
Major Data Mining Tasks

- **Classification:** predicting an instance class on the basis of its description.
- **Associations:** e.g. A & B & C occur frequently.
- **Clustering:** finding similarity groups in data.
- **Outlier Mining**: fraud detection
- **Trend and Evolution Analysis**: time-series mining or mining on time-related data
- **Mining Path Traversal Patterns** คือ การค้นหาแพตเทิร์นการท่องไปตาม web pages ต่างๆของผู้ใช้ที่บันทึกอยู่ใน Web Access Log

Why Data Mining? -- Potential Applications

- Database analysis and decision support
  - Market and customer analysis; analytical CRM
    - target marketing and advertising, customer relation management, market basket analysis, cross selling/marketing
  - Risk analysis and management
    - Forecasting, customer changes/ churn analysis, quality control, loan approval
  - Diagnostics (e.g. technical conditions of objects)
  - Fraud detection
- Other Applications:
  - Text mining (news group, email, documents) and Web analysis; Search engines; Biology; Sports, and e-commerce systems

Customer Segmentation



- Process of tagging and grouping customers based on shared characteristics
- Best way to reach the right customers at the correct time with the information they need
- This way, we can better understand the customers and meet their unique needs

Segmentation Model	How to Segment Customers
Demographic Segmentation	Age, gender, income, education, marital status
Geographic Segmentation	Country, state, city, town
Psychographic Segmentation	Personality, attitude, values, and interest
Technographic Segmentation	Mobile use, desktop use, apps and software
Behavioral Segmentation	Tendencies and frequent actions, feautre or product use, habits
Needs-based Segmentation	Product or service must-haves and needs of specific customer groups
Values-based Segmentation	Economic value of specific customer groups on the business

Data Mining Software/ Tools/ Datasets/ Tutorials

- WEKA
- RapidMiner (YALE)
- IBM: Intelligent Miner
- SPSS / Integral Solutions Ltd.: Clementine
- Datasets: Kdnuggets, Kaggle
- Tutorials: Scikit-Learn, coursera, edX
- Libraries: GitHub, TensorFlow, Keras