Chapter 2

Business Problems and Data Science Solutions

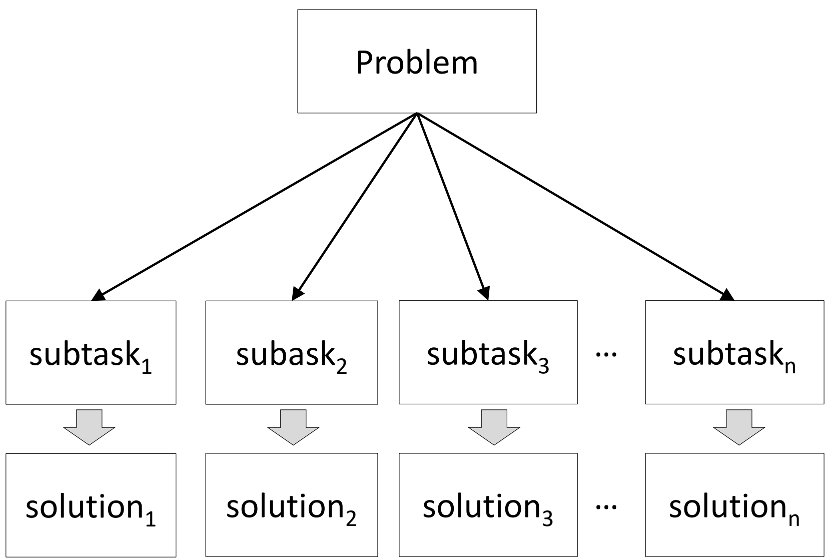
어떤

data mining technique
in high level

Dongchul Park
dpark@sookmyung.ac.kr

Data Science Process (1/2)

- A principle of data science (well-defined) data science의 기본원리
  - Data mining is a ***process*** with fairly well-understood stages
- Data science process 잘게 나누기
  - Data scientists decompose a real-world problem into ***subtasks***
  - The solutions to the subtasks are composed to solve the overall problem

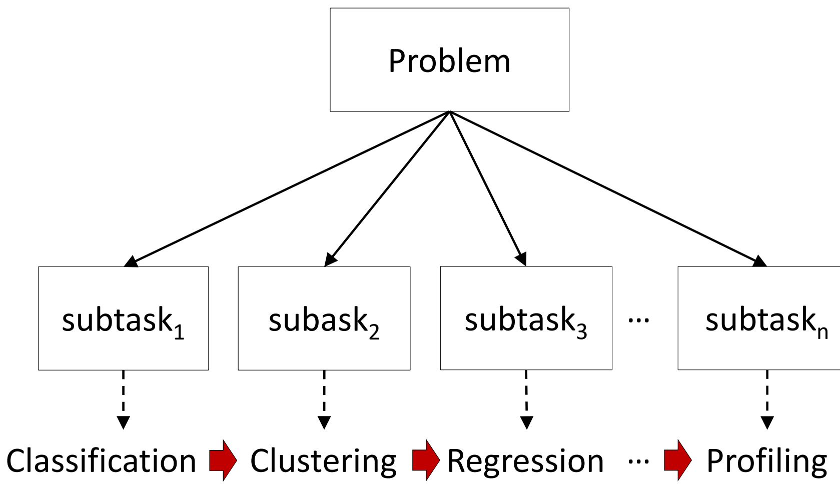


Data Science Process (2/2)

- There are ***common data mining tasks*** that underlie the problems
  - (ex) classification, regression, clustering, association rule discovery, …

회귀분석

- To be a good data scientist, you should
  - Know a lot about solving these common data mining tasks
  - Have the ability to decompose a problem into these common tasks



각각독립적으로

data science pipeline 병렬처리

Common Data Mining Tasks

- Despite the large number of specific data mining algorithms, there are only a few ***fundamentally different data mining tasks*** 많이 쓰이는 테크닉들.

분류 ① Classification

회귀분석 ② Regression (a.k.a. value estimation) : prediction

유사도 ③ Similarity matching

군집화 ④ Clustering

관계 A가 발생하면 B가 발생하는 경우

동시발생 ⑤ Co-occurrence grouping (a.k.a. association rule discovery)

⑥ Profiling (a.k.a. behavior description)

⑦ Link prediction 페북친구처럼.

사이즈↓ ⑧ Data reduction 형태..

인과관계 ⑨ Causal modeling

분류..?

1. Classification (1/2)

(≠ clustering)

- Predict (for each individual in a population) which of a set of ***classes*** this individual belongs to

↳ 전체 data 中 각각의 data 당

  – Usually the classes are mutually exclusive

상호 배타적 (서로 overlap이 안됨)

class의 set이 이미 존재
→ 정해져있다는 뜻 (ex. Yes/No)

세로줄(열)

Attributes

Classification Target

== each individual

Instances

가로줄(행)

Name	Salary	Sex	Age	Buy widget
Bloggs	15000	male	19	No
Jones	25000	male	33	Yes
Smit	23000	female	50	No
Smit	16000	male	40	No
Smit	200	male	10	No
Patel	30000	female	30	No
Steel	25000	male	23	Yes
Higgs	18000	female	55	No
Puggs	50000	male	57	Yes
Puggs	51000	female	57	No

...

model
(training data set)

New instance ---->

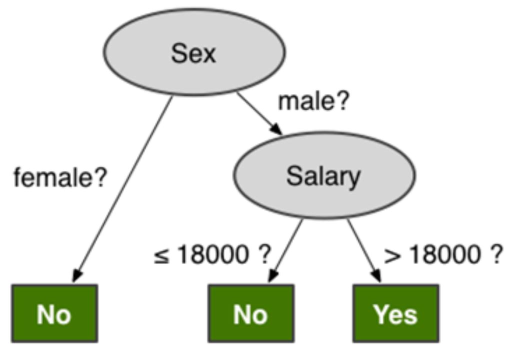
Lee	**42000**	**male**	**44**	**???**

1. Classification (2/2)

- General procedure
  - Given a training dataset, build a ***model*** that describes the classes of data
  - Given a new individual, apply the model to produce its ***estimated*** class

Name	Salary	Sex	Age	Buy widget
Bloggs	15000	male	19	No
Jones	25000	male	33	Yes
Smit	23000	female	50	No
Smit	16000	male	40	No
Smit	200	male	10	No
Patel	30000	female	30	No
Steel	25000	male	23	Yes
Higgs	18000	female	55	No
Puggs	50000	male	57	Yes
Puggs	51000	female	57	No

Training dataset



Model

yes/no같이 정해진 lable이 아니라

- A similar task: ***scoring*** or ***class probability estimation***
  - Produce the ***probability*** that a new individual belongs to each class
  - (ex) (Lee, 42000, male, 44) → (YES: **80%**, NO: **20%**)

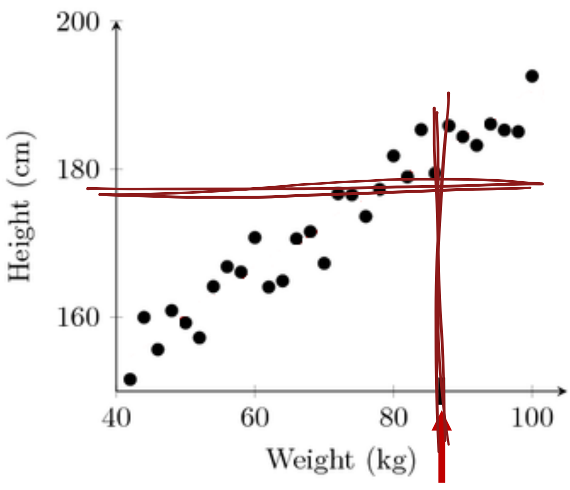
2. Regression (1/2)

회귀분석

- Estimate or predict, for each individual, the ***numerical value*** of some variable for that individual

  수치값을 추정함

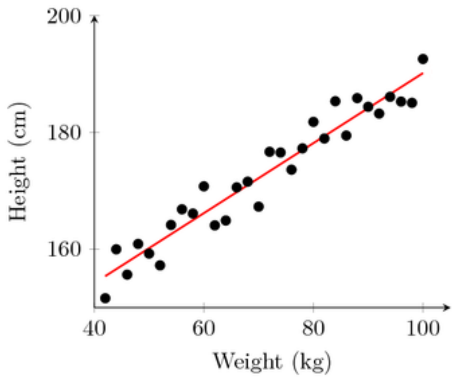
  – Also called "value estimation"



Given a weight of a person, what height will the person have?

2. Regression (2/2)

- General procedure
  - Given a training dataset, build a ***model*** that describes the value of the particular variable specific to each individual
  - Given a new individual, apply the model to produce its ***estimated*** value



Training dataset

$$\text{Height} = 0.9 \cdot \text{Weight} + 105.2$$

Model

- Regression vs. classification
  - Classification: predicts the ***class*** of an instance (e.g., YES/NO)
  - Regression: predicts the ***value*** associated with an instance (e.g., 178)

3. Similarity Matching (1/2)

- Identify ***similar*** individuals based on data known about them

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1 	8	1	?	2	7
User 2 	2	?	5	7	5
User 3 	5	4	7	4	7
User 4 	7	1	7	3	8
User 5 	1	7	4	6	?
User 6 	8	3	8	3	7

What users are similar to User 1?

What items are similar to Item 3?

3. Similarity Matching (2/2)

- General procedure '유사함'에 대한 정의(science적인 측면)
  - Define a ***distance measure*** between two individuals → '거리'라는 관점으로 재정의
  - Given an individual, find the individuals that minimize the distance

거리가 가깝다 = 유사하다

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1	8	1	?	2	7
User 2	2	?	5	7	5
User 3	5	4	7	4	7
User 4	7	1	7	3	8
User 5	1	7	4	6	?
User 6	8	3	8	3	7

User3 = (5, 4, 7, 4, 7)

User4 = (7, 1, 7, 3, 8)

유클리드

The distance between User 3 and User 4 is:

$$\sqrt{(5-7)^2+(4-1)^2+(7-7)^2+(4-3)^2+(7-8)^2} \approx 3.87$$

distance 계산 중 가장 쉬운 예시 : data type에 따라 다름

4. Clustering

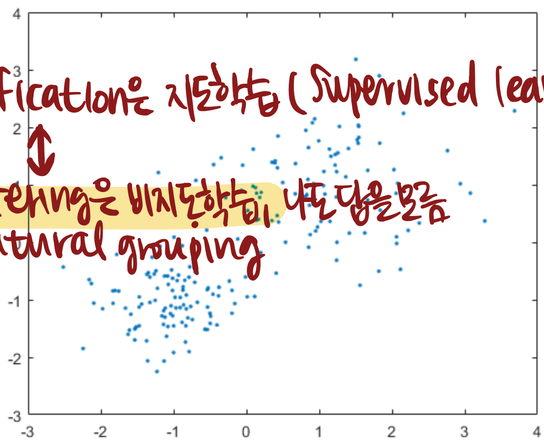
- Group ***similar*** individuals together
  - A ***distance measure*** is used to determine the similarity between two individuals

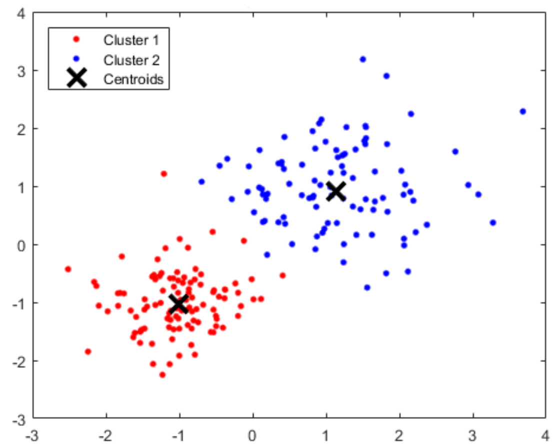
거리척도

* classification은 지도학습 (supervised learning)

clustering은 비지도학습, 나도 답을 모름

→ natural grouping





ex. K-means clustering

- Very useful to see which natural groups exist in the data
  - (ex) What types of customers do we have?

동시발생
연관성

5. Co-occurrence grouping

- Find ***associations*** or ***co-occurrence*** between entities
  - (ex) What items are commonly purchased together?

{Bread, Diapers, Beer, Coke}

{Milk, Diapers, Beer, Wine}

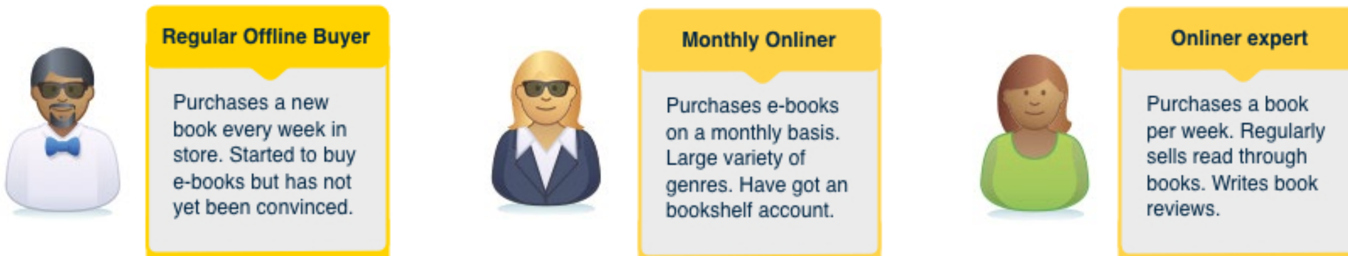
{Diapers, Beer, Coke}

An association rule: {Diapers} → {Beer}
(*"customers who buy diapers tend also to buy beer"*)

- Very useful for marketing
  - A special promotion, product display, or combination offer
  - Recommendation ("people who bought X also bought Y")

6. Profiling

- Characterize the ***typical*** behavior of an individual or group
  - Also called behavior description 행동특성을 기록



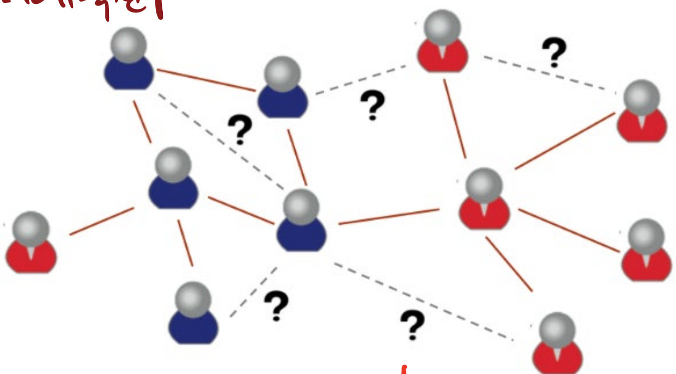
↔ normal

- Very useful for anomaly detection
  - A profile describes ***normal*** behaviors
  - If the current behavior is very different from the profile, issue an alarm
  - (ex) Fraud detection
    - Profile: the kind of purchases a person typically makes on a credit card
    - Alarm if a new charge on the card does not fit the profile

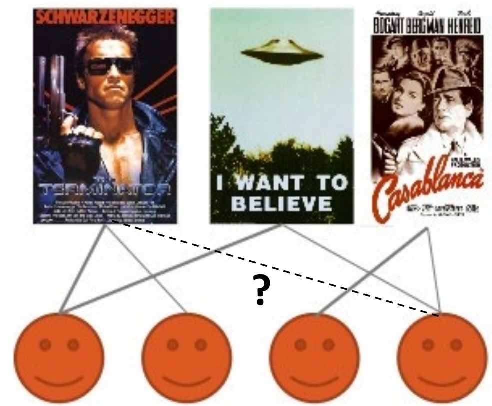
7. Link Prediction

- Predict ***connections*** between data items
  - Usually by suggesting that a link should exist and estimating its strength

ex. 페이스북 친구



영화



'A, B 사이에 link가 생겨야할 것 같다' 예측

- Very useful for recommendation
  - Recommending friends in social networking systems (SNS)
  - Recommending movies to customers

8. Data Reduction

- Reduce a dataset to a ***smaller*** dataset that contains much of the important information in the larger dataset

ㄴ가지고있도록

	Movie1	Movie2	Moive3	…	MovieN
User1	2	1	4	…	5
User2	3	3	2	…	4
User3	2	1	3	…	4
User4	2	2	1	…	4

	Action	Drama	Comedy
User1	4	2	2
User2	3	5	4
User3	2	4	4
User4	5	2	3

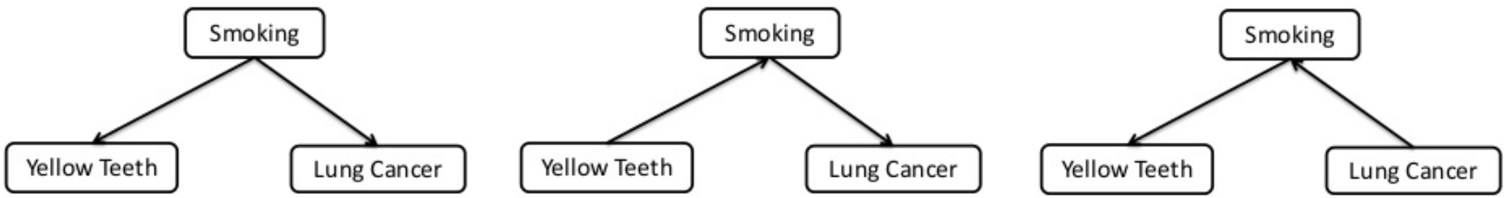
insight를 더 쉽게 찾기

- **Advantages**
  - The smaller dataset may be easier to deal with or to process
  - The smaller dataset may better reveal the information or improved insight
  - As a trade-off, data reduction usually involve loss of information

인과관계

9. Causal Modeling

- Understand what events or actions actually ***influence*** others



Does smoking cause lung cancer or vice versa?

ex) target marketing을 잘해서 많이 산건지
원래 많이 사는 고객인지?

- Example
  - Assume we observe that the targeted consumers purchase at a higher rate
  - Is this because of the targeting or are they just good customers?

머신러닝에서

Supervised vs. Unsupervised Methods (1/2)

지도학습 비지도학습

- Supervised data mining : data에 대한 정답을 알고있음
  - There is a specific ***target*** to specify
  - (ex) classification
    - "Buy widget" attribute is the target to define

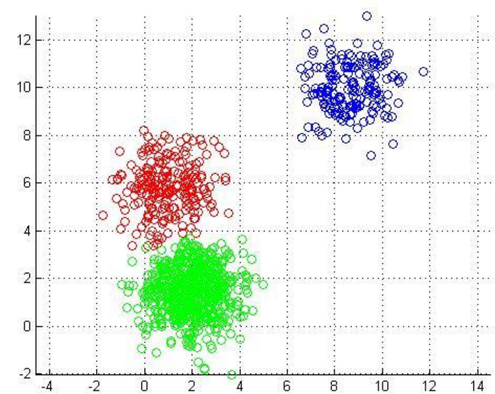
Classification Target

Name	Salary	Sex	Age	Buy widget
Bloggs	15000	male	19	No
Jones	25000	male	33	Yes
Smit	23000	female	50	No
Smit	16000	male	40	No
Smit	200	male	10	No
Patel	30000	female	30	No
Steel	25000	male	23	Yes
Higgs	18000	female	55	No
Puggs	50000	male	57	Yes
Puggs	51000	female	57	No

이미 알고있음

- Unsupervised data mining : 정답모름
  - There is ***no*** specific target to specify
    ↳ 딱 그 집에서 알고싶은 무언가 X
  - (ex) clustering
    - We have no specific target to define

자연스러운 grouping을 보자



Supervised vs. Unsupervised Methods (2/2)

- Supervised data mining
  - The purpose is to predict the ***target***
  - There must be a training dataset (i.e., the target value for each individual)
    - The target value is often called a ***label***

yes/no
high/middle/low
⋮

  - (e.g.) classification, regression, casual modeling
    - Similarity matching, link prediction, and data reduction (likely)

완전하는X

- Unsupervised data mining
  - The purpose is to find some patterns ***without*** any specific target
  - We don't need a training dataset
  - (e.g.) clustering, co-occurrence grouping, and profiling

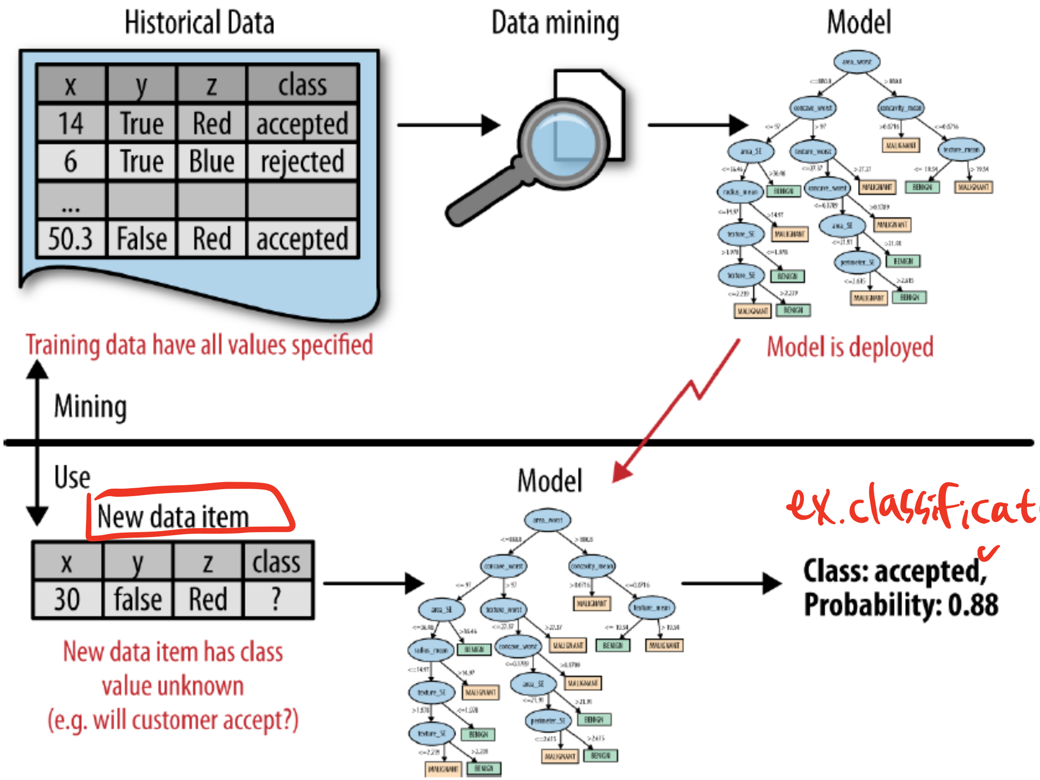
Classification vs. Regression

- Both are ***supervised*** data mining tasks
  - Distinguished by the type of target
- Classification
  - Target: a ***categorical*** (often binary) value (e.g., Yes/No, High/Mid/Low)
  - Example
    - "Will this customer purchase service S1 if given incentive I?" → Yes/No
    - "Which service package (S1,S2,S3) will a customer likely purchase if given incentive I?" → S1, S2, S3
- Regression
  - Target: a ***numerical*** value (e.g., 2.5)
  - Example
    - "How much will this customer use the service?" → $2,500

Data Mining and Its Results

단계

- Two phases of data mining



ex. classification

Mining phase

Find patterns or build models from existing data

training data set

Use phase

Apply the patterns or models to new data

하나정도의 framework
(rule)

Data Mining Process

범산업적으로

- Cross Industry Standard Process for Data Mining (CRISP-DM)
  - A well-understood process that places a structure on the problem

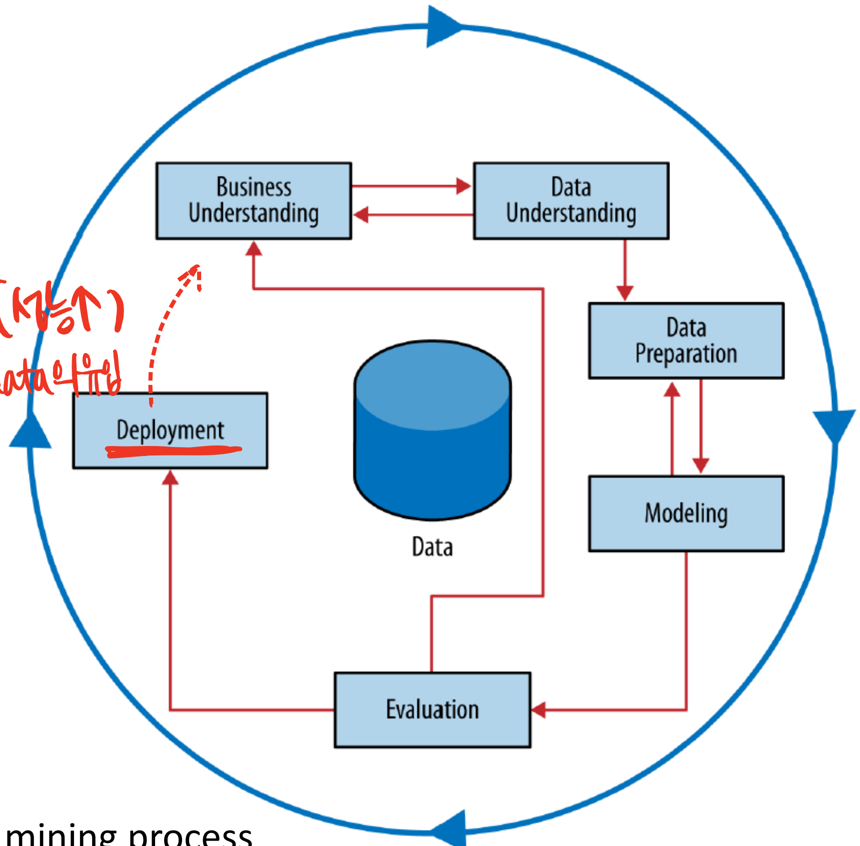


Fig. CRISP data mining process

1. Business Understanding

- Understand the business problem to be solved
  - Data science projects ***seldom*** come pre-packaged as clear problems
- Cast the business problem as one or more data science problems (sub-tasks)
  - The key to a success is a creative problem formulation by data scientists
- Design a solution for each data science problem
  - Classification, regression, clustering, …
  - A set of powerful tools can be used for each problem
- Recasting the problem and designing a solution is an ***iterative*** process of discovery

2. Data Understanding

- Data
  - The available raw material from which the solution will be built
  - (ex) a customer database, a transaction database, a marketing response database

이 data를 이용해 할 수 있는것과 없는것

- Understand the ***strengths*** and ***limitations*** of each data
  - Because rarely is there an exact match with the problem
  - (ex) For classification, we need labeled data (e.g., default = “yes” or “no”)

충분한지.

- Decide whether further investment in data is needed
  - Some data are virtually free, some data require effort to obtain, and some data may be purchased

3. Data Preparation

가공

- Clean and convert the data into more usable forms
  - Because some data analytic tools require data to be in a certain form

- Typical examples
  - Converting data to tabular format
  - Removing or inferring missing values
  - Converting data to different types (e.g., "Male", "Female" → 0, 1)
  - Normalizing or scaling numerical values (e.g., [-100, 100] → [0, 1])
  - Cleaning data (e.g., Age: 999 → ?)

data의 퀄리티가 중요

- The quality of data mining results heavily depends on this stage
  - (ex) missing values, abnormal values, non-normalized values, …

data mining의
실질적 technique 시작

4. Modeling

- The ***primary stage*** where data mining techniques are applied to the data

- Output
  - Some sort of model or pattern capturing regularities in the data

- It is very important to understand the ***fundamental ideas*** of data mining
  - i.e., data mining techniques and algorithms that exist

- We will discuss this subject throughout the course

분석

5. Evaluation (1/2)

엄격하게

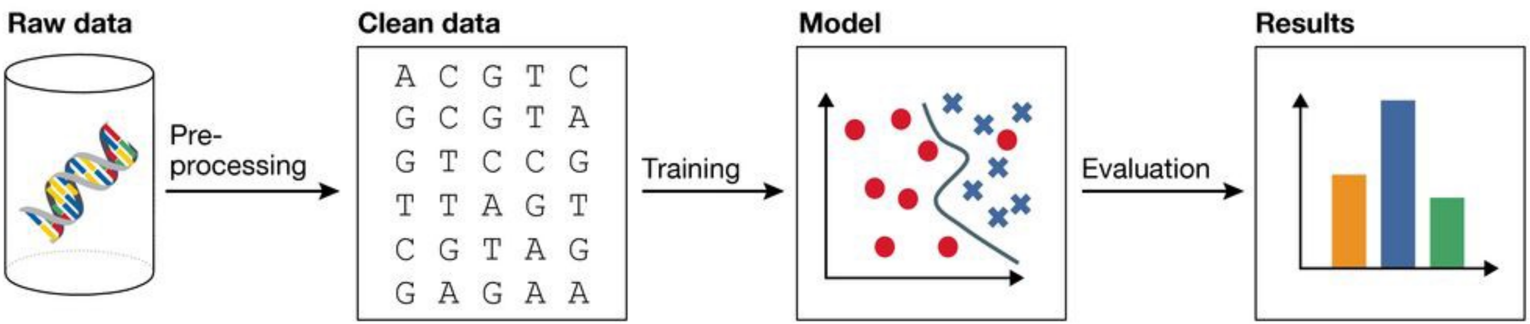
- Assess the data mining results ***rigorously***
  - To gain confidence that they are valid and reliable before moving on (deployment단계)
- Examples
  - Estimate the prediction accuracy of the model (e.g., 90%?)
  - Check the generality of the model beyond the training data (다른 data에도 적용가능한지
  - Estimate the rate of false alarms
- Instead of deploying the results immediately, it is usually advisable to test a model first in a controlled lab (i.e., testbed)
  - Because it is easier, cheaper, quicker, and safer

5. Evaluation (2/2)

- A data scientist should be able to explain the model and its evaluation results ***easily*** to stakeholders
  - Not just to data scientists

    이해관계자들

  - e.g., managers, executives, programmers, …



6. Deployment 적용, 배치

- Put the results of data mining (or systems) into ***real*** use
- Usual scenario
  - A new predictive model (or system) is ***implemented***
  - The model (or system) is integrated with existing information systems
- In many cases
  - ← 모델 만들고 평가
  - Data science teams: produce a working prototype and evaluate it
  - Data engineering teams: deploy the model into a production system
  - ← 모델 받아서 실제 system에 deploy
- After deployment, the process often returns to the first phase
  - The next iteration can yield an improved solution by using the insight and experience obtained in the previous iteration

계속 upgrade

Other Analytics Techniques & Technologies

- Besides data mining, there are various technologies for the analysis of data
  – Statistics, database systems, machine learning, …

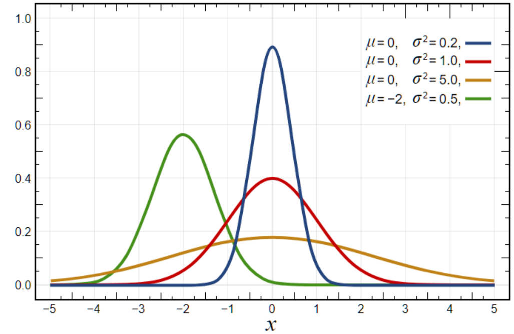
- It is important to be acquainted with these technologies
  – What are their goals?
  – What role do they play?
  – What differences do they have?

- An important skill for a data scientist is …
  – To be able to recognize what sort of analytic technology is appropriate for addressing a particular problem

1. Statistics

- Provides us with a huge amount of knowledge that ***underlies*** analytics 분석의 근간

- Examples
  - Data summary (e.g., means, median, variance, …)
  - Understanding different data distributions
  - Testing hypotheses
  - Quantifying uncertainty
  - Measuring correlation



- Many techniques for extracting models or patterns from data have their ***roots*** in Statistics

2. Database Querying (1/2)

질의, 질문

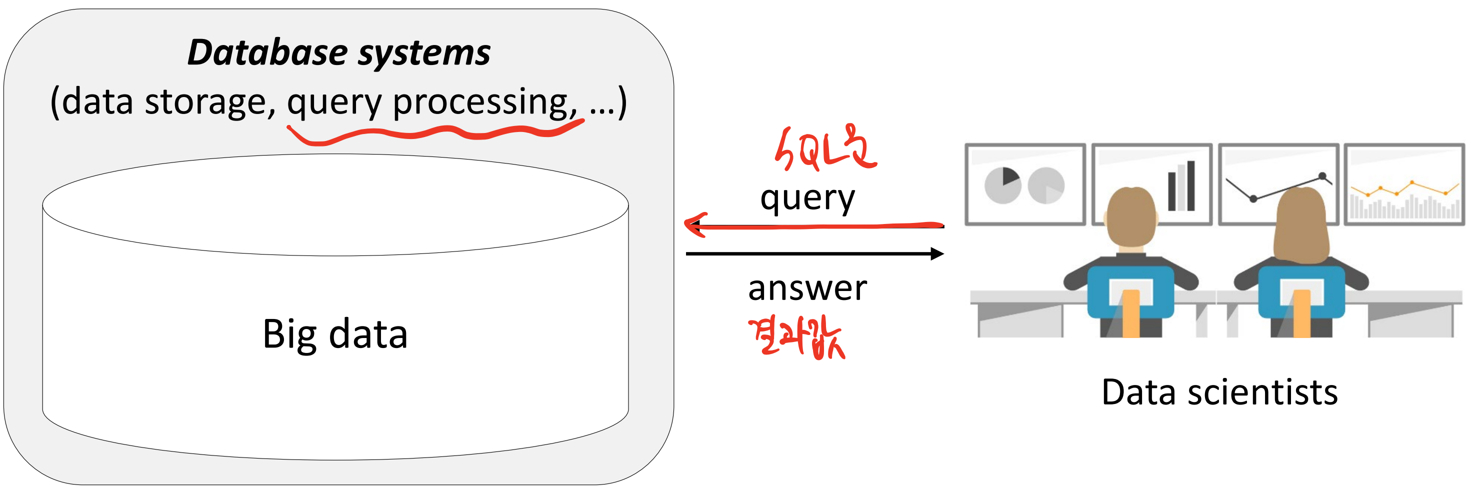
- Database system
  - A software application that allows the insertion, ***querying***, update, and management of data
- Database query
  - A specific request for data or statistics about data
    - Retrieving specified data, sorting, computing summary statistics, …
  - Formulated in a technical language and posed to a database system
    - (ex) SQL (Structured Query Language) 질의문

구조

```
SELECT name, address
FROM customers
WHERE age > 25 AND gender = 'Female' AND domicile = 'CA'
```

2. Database Querying (2/2)

- Data science vs. databases technologies
  - Data science can ***use*** database technologies to find or examine the data of interest stored in a database system



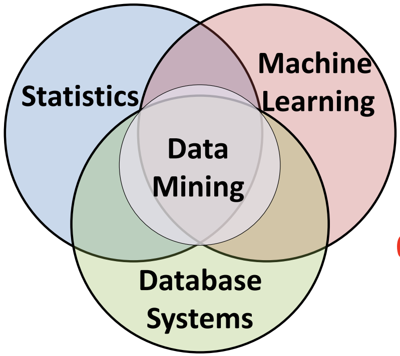
알아서학습

3. Machine Learning (1/2)

- Gives computer systems the ability to "***learn***" with data, without being explicitly programmed
  - A ***subfield*** of Artificial Intelligence (AI)

- Develops models and improves the models using data
  - Decision tree, artificial neural networks (deep learning), support vector machines, clustering, Bayesian networks, …

- However, the separation among these fields has blurred



overlapped

3. Machine Learning (2/2)

- ***Data mining*** and ***machine learning*** are closely linked
  - The filed of data mining started as an offshoot of machine learning
    - KDD (Knowledge Discovery and Data mining) 머신러닝의 한 가지
    - Techniques and algorithms are shared between the two
  - Find useful and informative pattern from data
    머신 : 성능개선(학습)
    데이터 : 패턴, 규칙
- Nevertheless, ***machine learning*** is more concerned with
  - Many types of performance improvement
    - (ex) robotics and computer vision
  - Issues of agency and cognition
    - (ex) how will an agent use learned knowledge to act in its environment
- ***Data mining*** is more concerned with
  - Finding patterns and regularities from data
  - Commercial applications and business issues

Examples of Applying These Techniques

- "Who are the most profitable customers?"
  - Database systems (if "profitable" can be calculated from existing data)

    query날려서.

- "Is there really a difference between the profitable customers and the average customer?"
  - Statistics (hypothesis testing)

    가설

- "But who really are these customers? Can I characterize them?"
  - Data mining (profiling)

- "Will some particular new customer be profitable? How much?"
  - Data mining (classification, regression)

    yes/no. 얼마만큼

Summary

- There is a well-defined ***data mining process*** (e.g., CRISP-DM)
  - Business understanding → data understanding → data preparation → modeling → evaluation → deployment
- A data scientist typically decomposes a problem into one or more ***common data mining tasks***
  - Classification, regression, similarity matching, clustering, association rule discovery, profiling, link prediction, data reduction, causal modeling
  - You should understand the fundamentals of these tasks
- Other related data analytics technologies
  - Statistics, database querying, machine learning
  - Though their boundaries are not always sharp, it is important to know about other techniques' capabilities to know when they should be used