

INTRODUCTION TO DATA MINING

Week01

What is Data Mining?

- Data mining
 - ▣ Data mining is the process of discovering patterns, relationships, anomalies, or useful insights from large datasets
 - ▣ It involves statistical techniques, database management, and machine learning algorithms to extract meaningful information
- Key objectives of data mining
 - ▣ Identify hidden patterns and trends in data
 - ▣ Support decision-making through predictive and descriptive analytics
 - ▣ Improve business intelligence and automation

What is Data Mining?

Data Mining is
Discovering patterns or
extracting information

from data

to utilize results
for further use

The diagram illustrates the components of Data Mining. It features three horizontal yellow bars. The first bar contains the text 'Discovering patterns or extracting information'. The second bar contains 'from data'. The third bar contains 'to utilize results for further use'. To the right of these bars, three curly braces group the text into categories: a brace for 'Discovering patterns or extracting information' is labeled 'Action'; a brace for 'from data' is labeled 'Resource'; and a brace for 'to utilize results for further use' is labeled 'Purpose'.

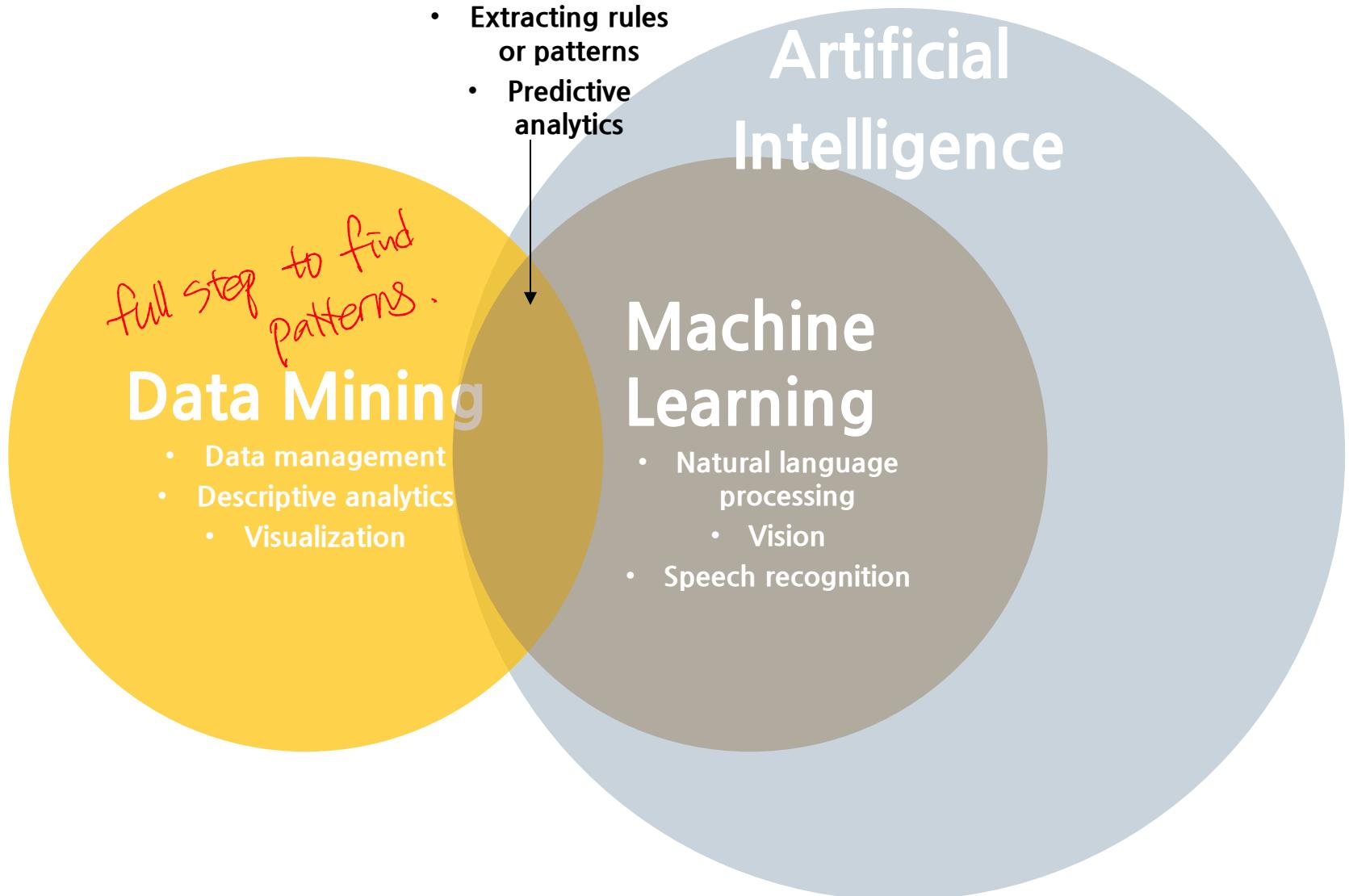
Are Machine Learning and Data Mining Different?

- Machine learning
 - ▣ Machine learning (ML) is a subset of AI that enables computers to learn patterns from data and make predictions without explicit programming.
 - ▣ Unlike traditional rule-based systems, ML algorithms improve performance over time as they process more data.
*find rule automatically
from data.*
- Key objectives of machine learning
 - ▣ Train models to recognize patterns in data
 - ▣ Make accurate predictions for unseen data
 - ▣ Automate decision-making based on learned patterns

What is Artificial Intelligence?

- Artificial intelligence
 - ▣ Artificial intelligence (AI) is a broad field of computer science that aims to create intelligent systems capable of mimicking human cognition, including learning, reasoning, and problem-solving.
 - ▣ Machine learning and data mining are part of AI, but AI encompasses more than just these techniques.
- Key objectives of artificial intelligence
 - ▣ Develop intelligent systems that can make autonomous decisions
 - ▣ Simulate human cognitive abilities such as reasoning, perception, and language processing
 - ▣ Enable automation across various industries

What is Data Mining?



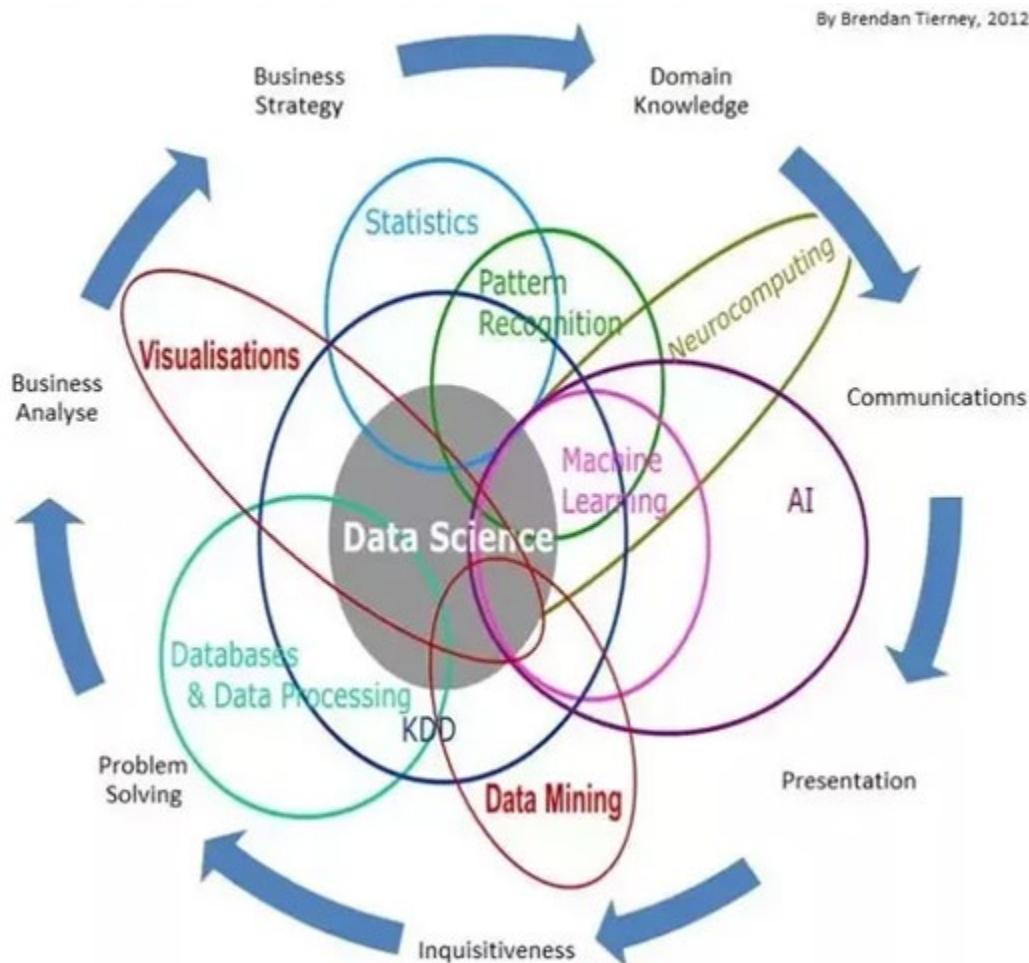
Data Mining vs. Machine Learning vs. Artificial Intelligence

□ Comparison

Feature	Data Mining	Machine Learning	Artificial Intelligence
Definition	Extracting patterns and knowledge from data	Training models to learn from data and make predictions	Creating intelligent systems that mimic human cognition
Main Goal	Discover hidden relationships in data	Enable computers to learn and adapt without explicit programming	Automate reasoning, learning, and problem-solving
Methods Used	Statistical analysis, rule-based systems, machine learning	Supervised, unsupervised, and reinforcement learning	Machine learning, Neural networks (deep learning), expert systems
Relationship	Uses ML techniques to analyze data	A subset of AI focused on learning patterns	The broadest concept, encompassing ML and data mining

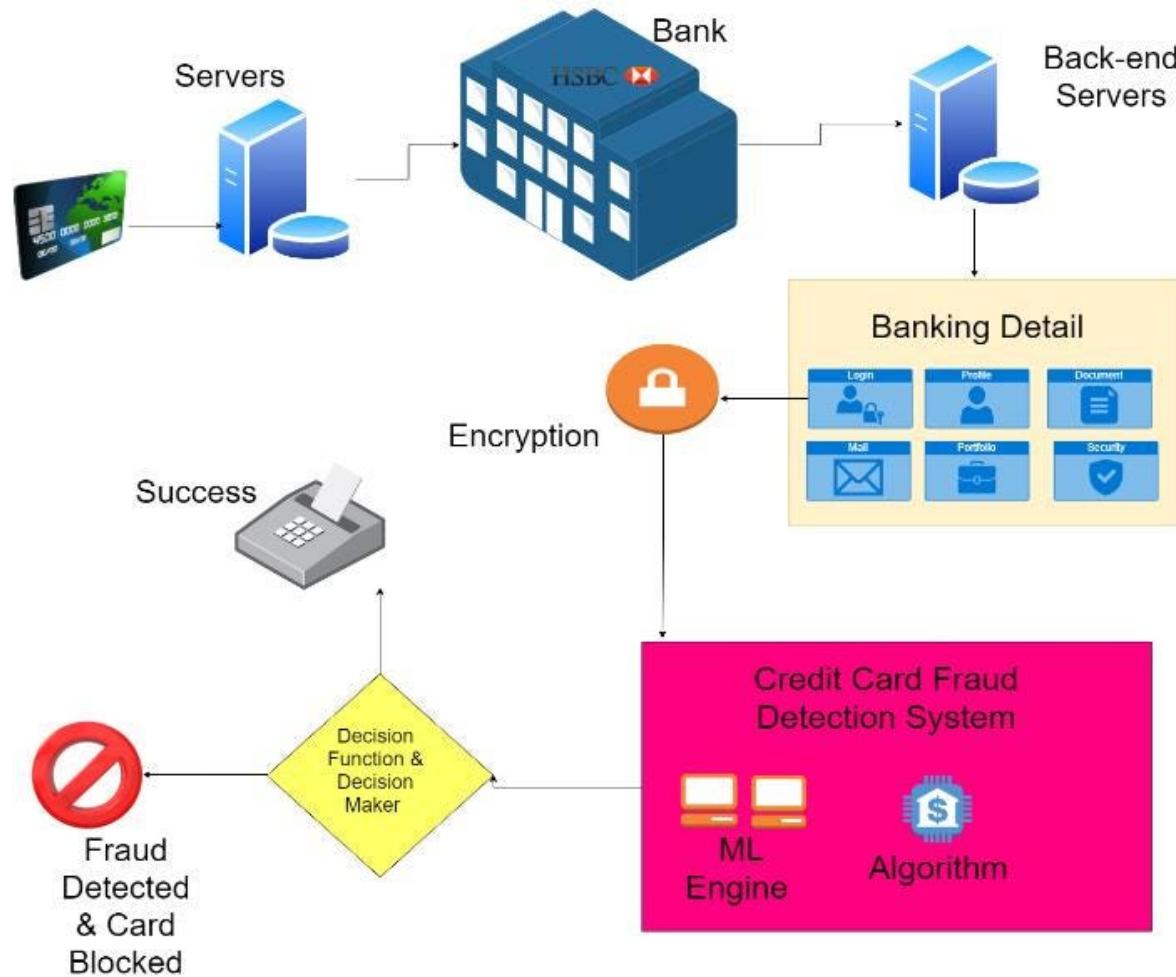
Data Mining, Part of Data Science

- Data Science if Multidisciplinary



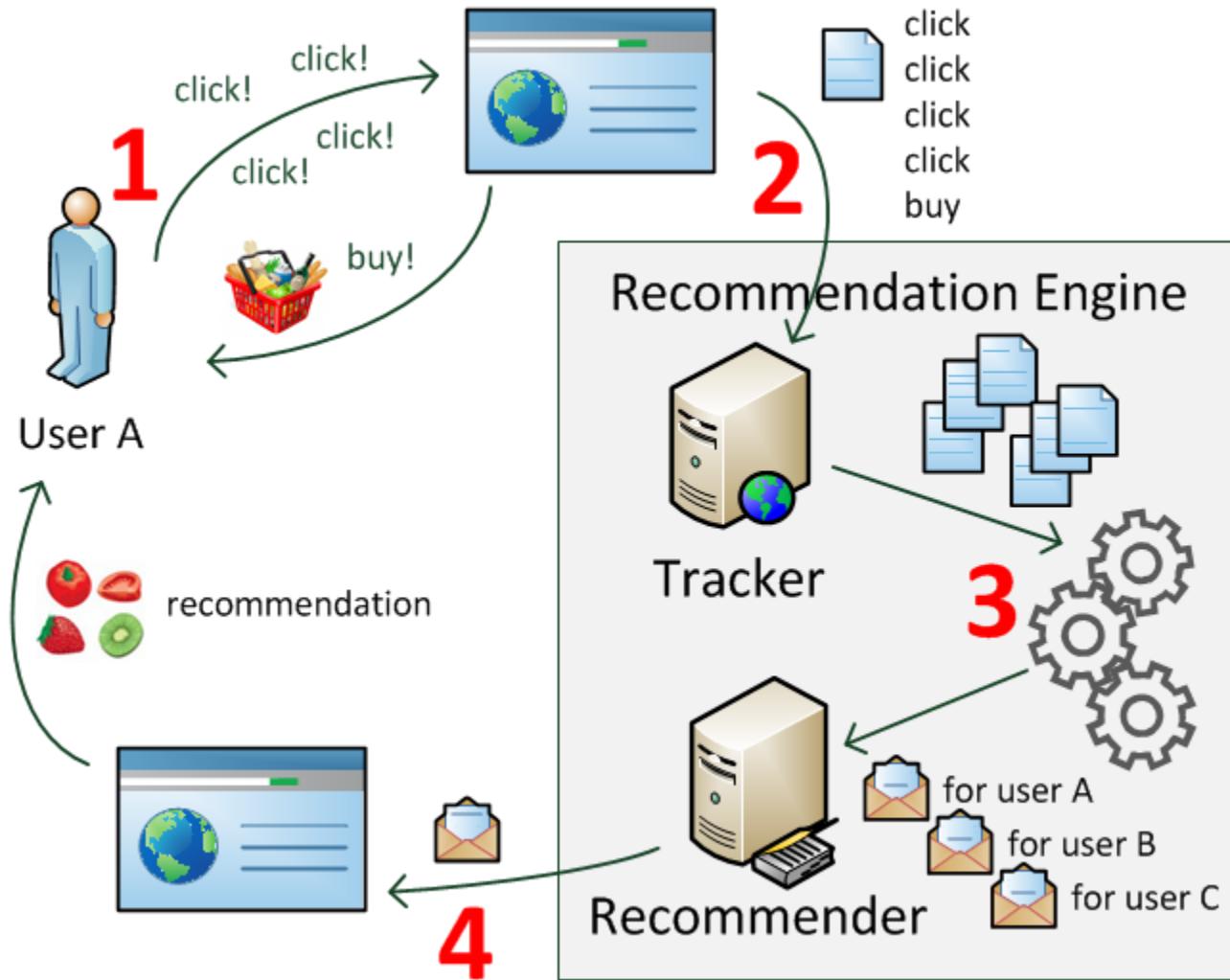
Application Areas of Data Mining

□ Fraud detection systems



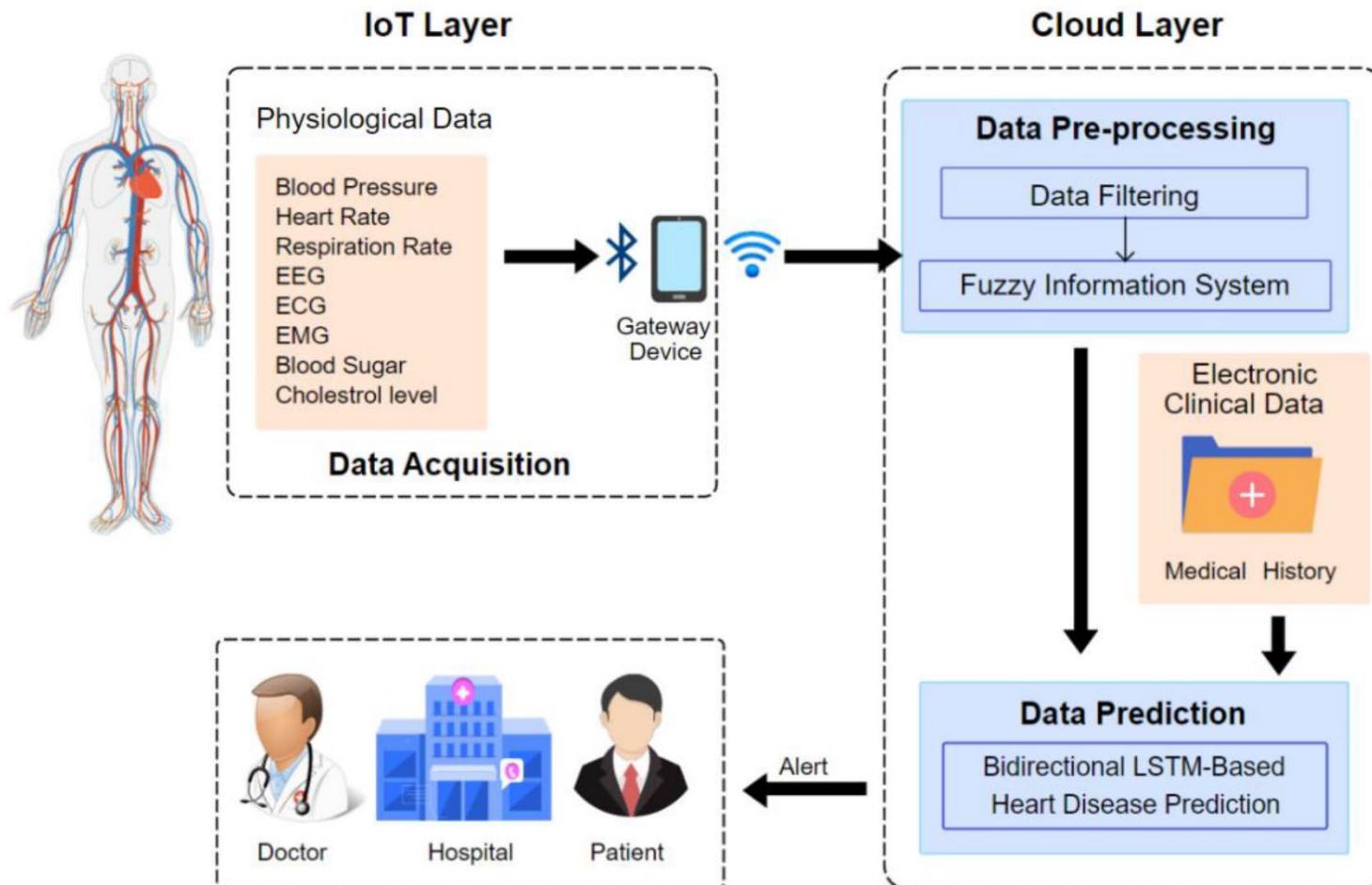
Application Areas of Data Mining

- Recommendation systems



Application Areas of Data Mining

- Healthcare monitoring systems

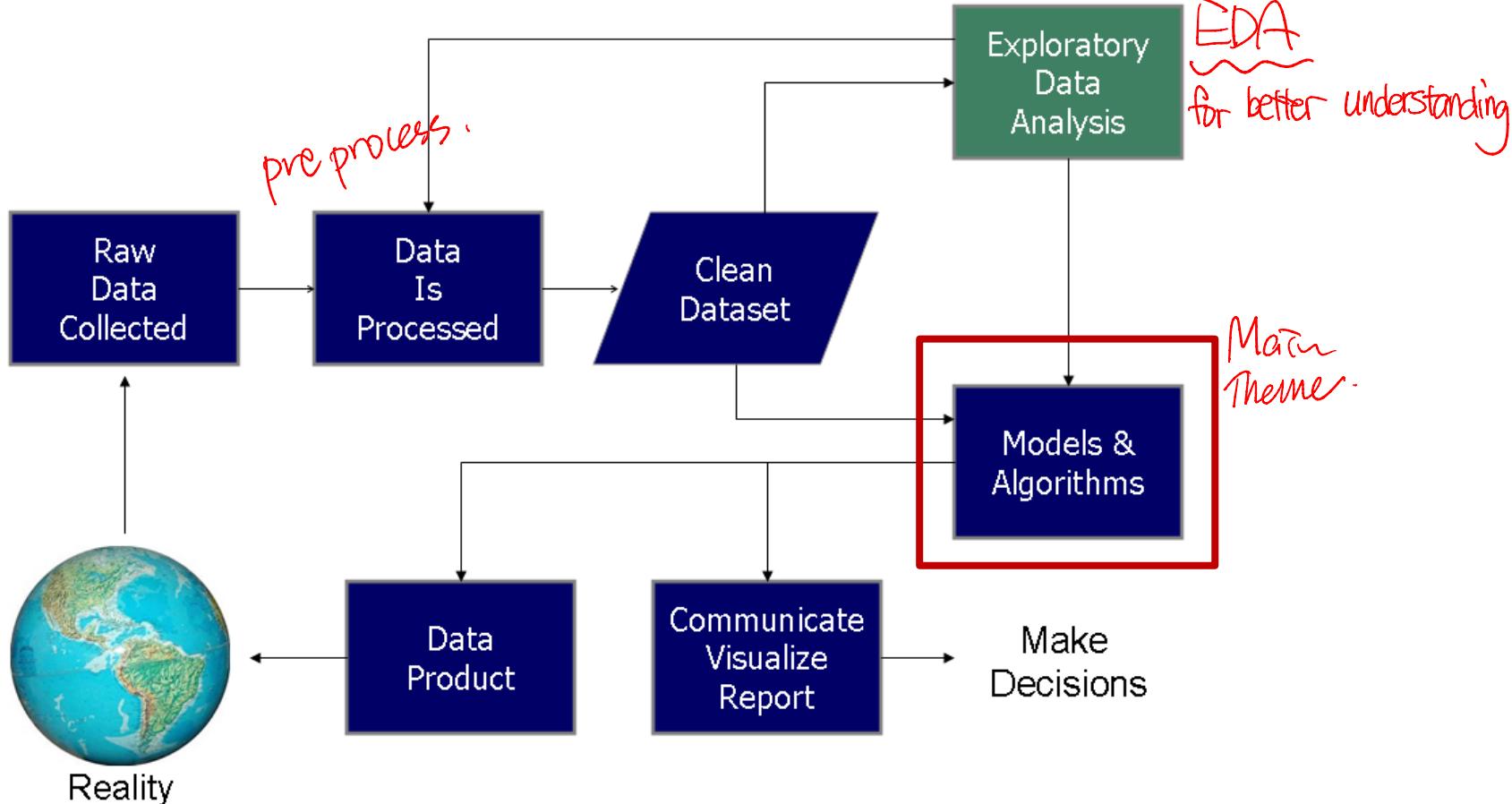


Application Areas of Data Mining

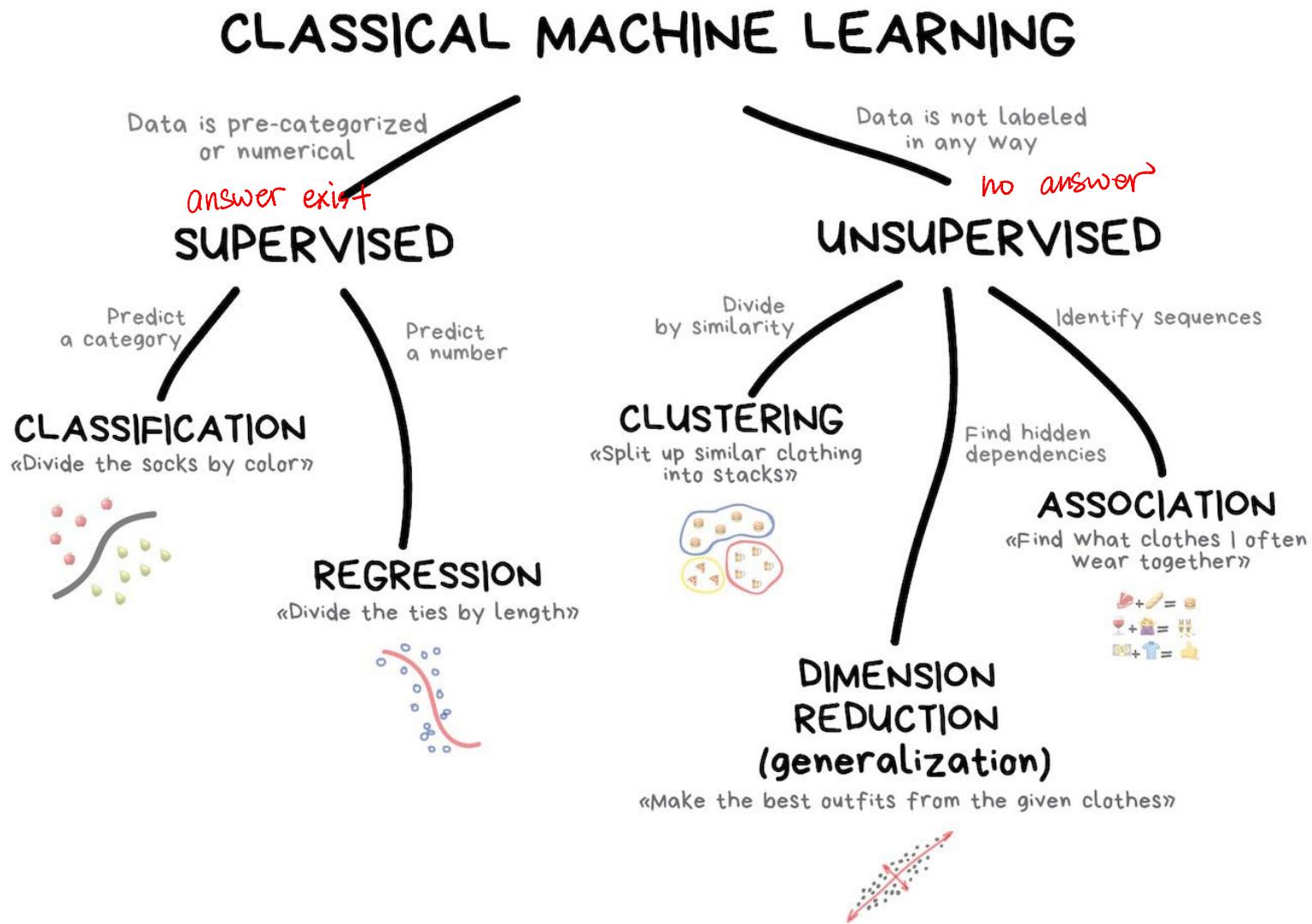
□ Smart factory



What We Will Learn in This Class



Topics Covered in This Class



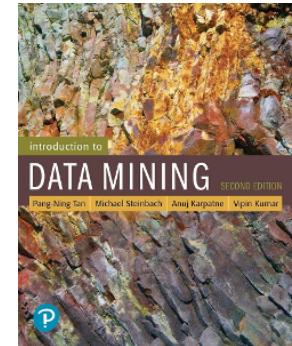
Topics Covered in This Class

- Supervised learning
 - ▣ Regression
 - Linear regression
 - Nearest neighbor methods
 - Decision tree
 - ▣ Classification
 - Logistic regression
 - Naïve Bayes
 - Nearest neighbor methods
 - Decision tree
- Unsupervised learning
 - ▣ Clustering
 - k -means
 - Hierarchical clustering
 - ▣ Dimension reduction
 - Principal component analysis (PCA)
 - ▣ Association rule mining

Recommended Data Mining Textbooks

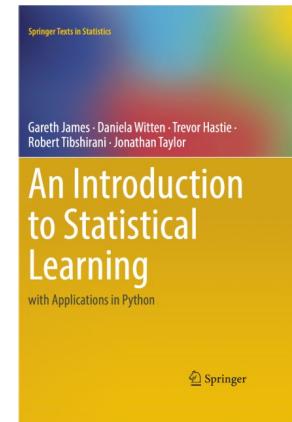
1. Introduction to Data Mining, 2nd edition

- <https://www-users.cse.umn.edu/~kumar001/dmbook/index.php>



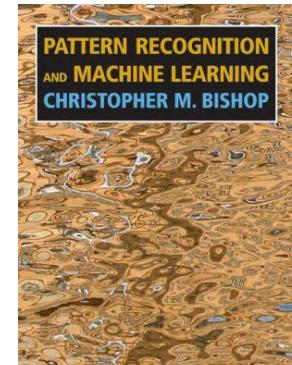
2. An Introduction to Statistical Learning: with Applications in Python

- <https://www.statlearning.com/>



3. Pattern Recognition and Machine Learning

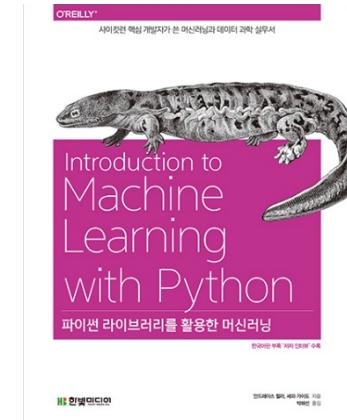
- <https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/>



Recommended Data Mining Textbooks

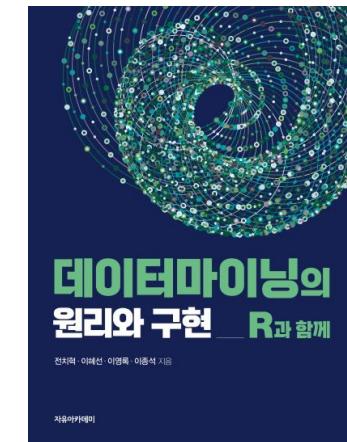
4. 파이썬 라이브러리를 활용한 머신러닝

- https://www.hanbit.co.kr/store/books/look.php?p_code=B6119391002



5. 데이터마이닝의 원리와 구현 : R과 함께

- <http://www.freeaca.com/new/book/MainBookView.aspx?bookid=31311&page=1&ca1=3&ca2=54&sword=&stype=>

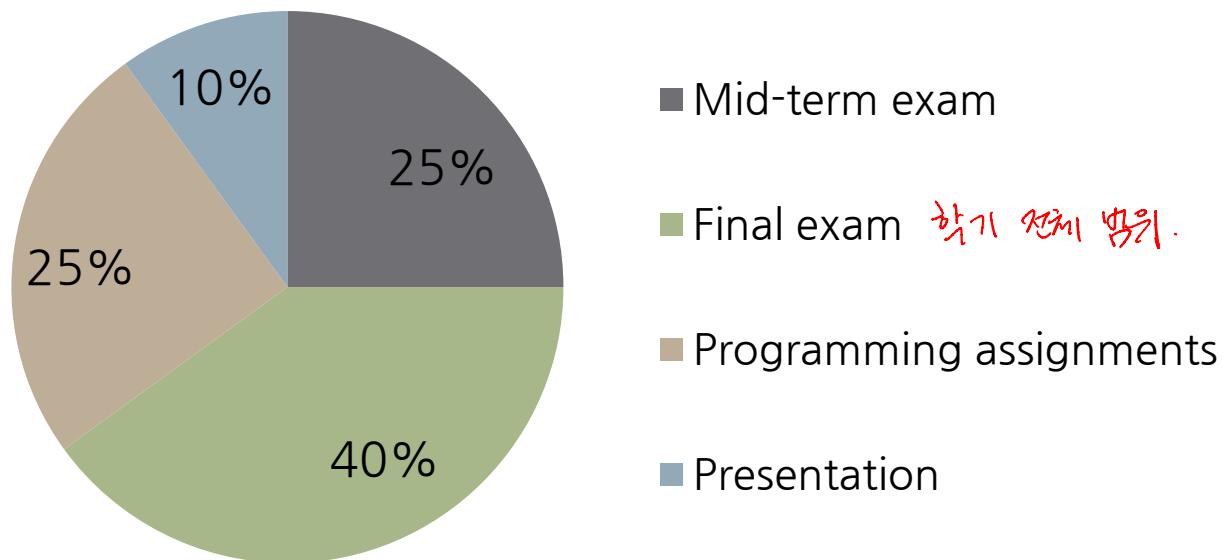


Principals of Lecture

- Understand main goal and basic principles of each data mining techniques
 - ▣ Why is an algorithm proposed?
 - ▣ What is a key point?
- Deliver principles as easy as possible without mathematics
- Understand detailed process of each data mining techniques
 - ▣ How are an algorithm working?
- Explain process step by step
- ※ Some equations will be introduced for explanation
- Exercise what you learned during lectures
 - ▣ Main programming language: Python
 - ▣ Confirm algorithms studied during lectures through programming exercises

Principals of Lecture: Assessment

- Course assessment
 - ▣ Exams will be held two times: mid-term and final exams
 - Final exam will cover the whole lectures
 - ▣ Programming assignments related with lectures
 - ▣ Team presentation: Case study
 - Topic proposal will be presented on the **9th** week
 - The final result will be presented on the **15th** week
 - Each team consist of 2~3 students (random)



Principals of Lecture: Assessment

- Exams
 - ▣ Assess the theoretical knowledge learned in class
 - Must understand principles and process of the data mining algorithms covered in class
 - ▣ No multiple choice questions
 - ▣ Can use a scientific calculator
 - ▣ **Schedule**
 - Mid-term exam: 8th Week, 4/24 (in the evening)
 - Final exam: 14th Week, 6/5 (in the evening)

Principals of Lecture: Assessment

- Team presentation
 - ▣ Case study using data mining
 - The purpose of data analysis
 - What is the problem?
 - Method
 - How did they solve the problem through data mining?
 - Result
 - What kinds of implication could be derived from the results of data analysis?

Schedule

Week	Date	Contents	Remarks
1	3/6	Introduction	
2	3/13	Background of data mining	
3	3/20	Explanatory data analysis (EDA)	
4	3/28	Linear regression: Theory Part 1 & Exercise	
5	4/3	Linear regression: Theory Part 2 & Exercise	
6	4/10	Logistic regression: Theory & Exercise	
7	4/17	Naïve Bayes classifier: Theory & Exercise	
8	4/24	Mid-term exam (in the evening)	
9	5/1	Nearest neighbor algorithm: Theory & Exercise Presentation: Case study topic proposal	<i>supervised</i>
10	5/8	Decision tree: Theory & Exercise	
11	5/15	Clustering: Theory & Exercise	
12	5/22	Dimensionality reduction: Theory & Exercise	
13	5/29	Association rule mining: Theory & Exercise	
14	6/5	Final exam (in the evening)	
15	6/12	Presentation: Case study	

Q & A

- If you want to ask a question related with lectures for data mining algorithms outside of class, please use the Q&A board of the e-class
 - ▣ Your question may be helpful to other students
 - Share your questions with other students
 - ▣ Do not ask individual questions by e-mail

Python: Installation

Installation

□ Python

- Visit <https://www.python.org/downloads/> and download Python installation file depending on your OS(Windows, Linux/UNIX, Mac OS X) and which version you want to install
 - This slide assumes that OS is Windows
- There are two stable versions of Python: 3.X, 2.X
 - Two versions are a little bit different, but different features do not matter in this course



The screenshot shows the Python website's homepage. At the top, there's a navigation bar with links for About, Downloads, Documentation, Community, Success Stories, News, and Events. Below the navigation bar, there's a search bar and a "Socialize" button. A large yellow "Donate" button is also visible. The main content area features a code snippet in a terminal window:

```
# Python 3: Fibonacci series up to n
>>> def fib(n):
    >>>     a, b = 0, 1
    >>>     while a < n:
    >>>         print(a, end=' ')
    >>>         a, b = b, a+b
    >>>         print()
    >>> fib(1000)
0 1 1 2 3 5 8 13 21 34 55 89 144 233 377 610 987
```

To the right of the code, there's a section titled "Functions Defined" with the following text:

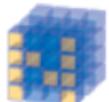
The core of extensible programming is defining functions. Python allows mandatory and optional arguments, keyword arguments, and even arbitrary argument lists. [More about defining functions in Python 3](#)

At the bottom of the page, there are five numbered buttons (1, 2, 3, 4, 5) and a footer message: "Python is a programming language that lets you work quickly and integrate systems more effectively. [»» Learn More](#)".

Installation Useful Packages



- SciPy
 - Python-based ecosystem of open-source software for mathematics, science, and engineering
 - <http://www.scipy.org/>



NumPy
Base N-dimensional array package



SciPy library
Fundamental library for scientific computing



Matplotlib
Comprehensive 2D Plotting



IPython
Enhanced Interactive Console



Sympy
Symbolic mathematics



pandas
Data structures & analysis

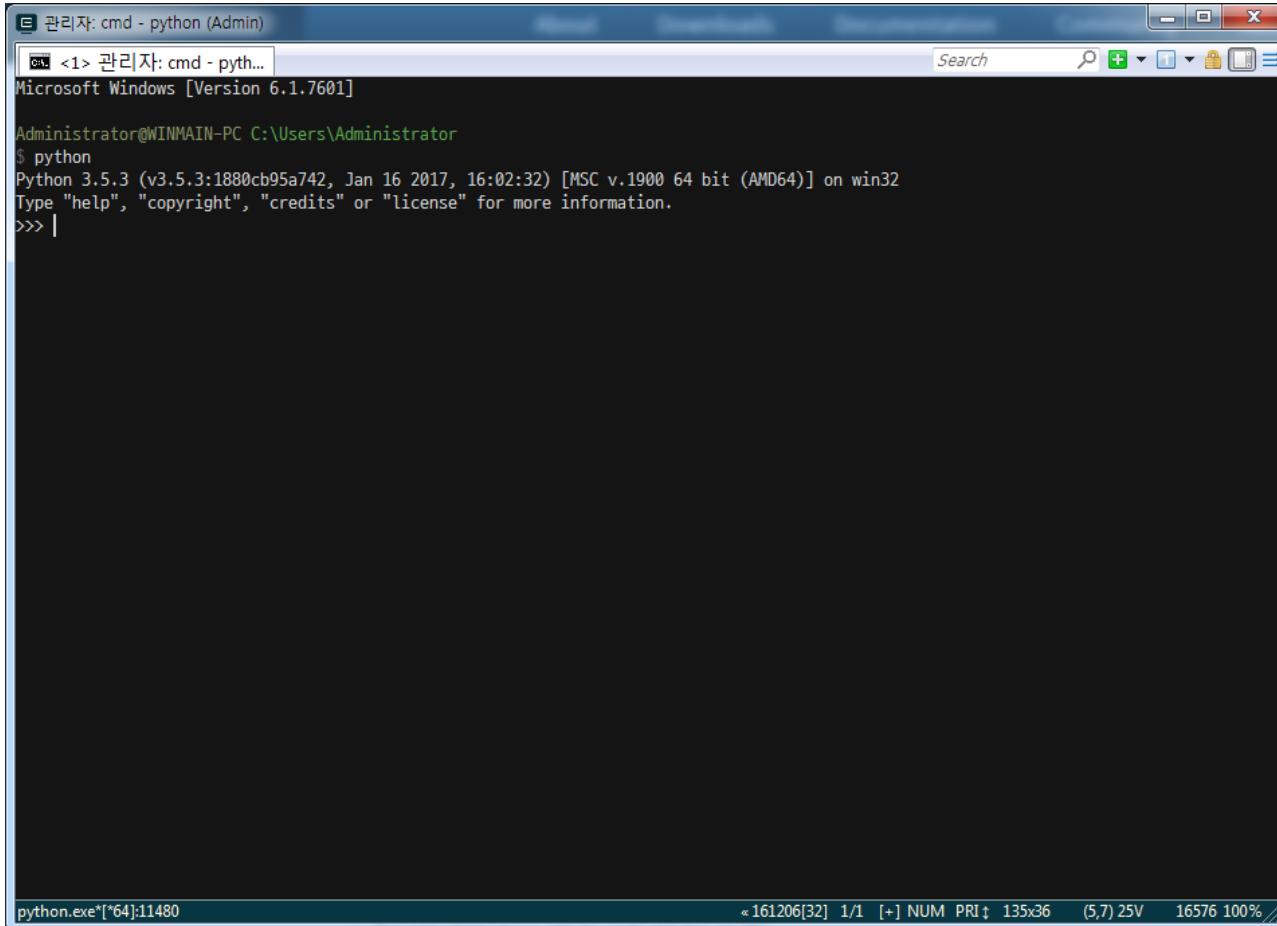
[Core packages]

Installation Useful Packages

- sci-kit learn
 - ▣ Free software machine learning library for the Python programming language
 - Simple and efficient tools for predictive data analysis
 - Built on Numpy, Scipy, and matplotlib
 - ▣ <https://scikit-learn.org/stable/index.html>

Start Python

- To start python, just type python at cmd prompt
 - ▣ Python is script language



The screenshot shows a Windows Command Prompt window titled "관리자: cmd - python (Admin)". The title bar includes the text "관리자: cmd - pyth..." and "Microsoft Windows [Version 6.1.7601]". The window content displays the Python 3.5.3 startup message:

```
Administrator@WINMAIN-PC C:\Users\Administrator
$ python
Python 3.5.3 (v3.5.3:1880cb95a742, Jan 16 2017, 16:02:32) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> |
```

The taskbar at the bottom shows the process name "python.exe[*64]:11480".

Start Python

- There are many Python IDEs(Integrated Development Environment)
 - ▣ However, notepad is also used for writing Python scripts
 - ▣ If you want to use better IDE than notepad
 - <http://pedrokroger.net/choosing-best-python-ide/>
 - ▣ There is also default IDE installed with Python

The image shows two windows side-by-side. The left window is titled "Python 3.4.3 Shell" and contains the Python interpreter prompt, showing the version and build information. The right window is titled "getAddress.py" and shows a Python script file with code for reading URLs using the urllib module.

```
Python 3.4.3 (v3.4.3:9b73f1c3e601, Feb 24 2015, 22:43:06) [MSC v.1600 32 bit (In tel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
```

```
#!/usr/bin/python27
# -*- coding: utf-8 -*-

import urllib, urllib2
from bs4 import BeautifulSoup
from unidecode import unidecode
import math
import csv
import datetime
import sys
import os.path
import shutil
import json
import xlrd

reload(sys)
sys.setdefaultencoding('utf-8')

def read_url(url):
    try:
        opener = urllib.FancyURLopener({})
        f = opener.open(url)
        document = f.read()
        opener.close()
        urllib.urlcleanup()
    except:
        pass
```

Python: Easy Installation

Scientific Python distributions

- The easiest way to install the packages of the SciPy stack is to download one of these Python distributions, which includes all the key packages
 - Anaconda: A free distribution for the SciPy stack. Supports Linux, Windows and Mac.
 - Enthought Canopy: The free and commercial versions include the core SciPy stack packages. Supports Linux, Windows and Mac.
 - Python(x,y): A free distribution including the SciPy stack, based around the Spyder IDE. Windows only.
 - WinPython: A free distribution including the SciPy stack. Windows only.
 - Pyzo: A free distribution based on Anaconda and the IEP interactive development environment. Supports Linux, Windows and Mac.

Scientific Python distributions

- Anaconda
 - ▣ URL: <https://www.anaconda.com/distribution/>
- WinPython
 - ▣ URL : <http://winpython.github.io/>
 - <https://sourceforge.net/projects/winpython/>

Scientific Python distributions

- Spyder
 - ▣ The Scientific PYthon Development EnviRonment

The screenshot shows the Spyder IDE interface with several windows open:

- Project explorer:** Shows the file structure of the current project, including files like `temp.py`, `interpolation.py`, and `__init__.py`.
- Editor:** Displays the code for `temp.py`. The code generates data for an ascending spiral in 3-space, performs spline calculations, and plots the results.
- Variable explorer:** A table showing variables and their values, such as `array_int8` (int8), `array_uint32` (uint32), `bars` (container.BarContainer), `df` (DataFrame), `filename` (str), `list_test` (list), `nrows` (int), `r` (float64), `radii` (float64), `region` (tuple), `rgb` (float64), `series` (Series), and `test_none` (NoneType).
- IPython console:** Shows the execution of code to generate a 3D surface plot and a 2D polar plot. The 3D plot shows a complex surface with axes labeled from -84.41 to 90°. The 2D plot is a polar chart with radial axes ranging from 36.70 to 700 and angular axes from 0° to 315°.

Python: Short Tutorial

Variable Types

□ List

- A list contains items separated by commas and enclosed within square brackets ([])

```
list = [ 'abcd', 786 , 2.23, 'john', 70.2 ]
```

```
tinylist = [123, 'john']
```

```
print list      # Prints complete list
```

```
print list[0]    # Prints first element of the list
```

```
print list[1:3]  # Prints elements starting from 2nd till 3rd
```

```
print list[2:]   # Prints elements starting from 3rd element
```

```
print tinylist * 2 # Prints list two times
```

```
print list + tinylist # Prints concatenated lists
```

Variable Types

- Tuples
 - ▣ A tuple consists of a number of values separated by commas. Unlike lists, however, tuples are enclosed within parentheses
 - ▣ The main differences between lists and tuples are
 - Lists are enclosed in brackets ([]) and their elements and size can be changed
 - Tuples are enclosed in parentheses (()) and cannot be updated (read-only)

```
tuple = ( 'abcd', 786 , 2.23, 'john', 70.2 )
tinytuple = (123, 'john')
```

```
print tuple      # Prints complete list
print tuple[0]    # Prints first element of the list
print tuple[1:3]   # Prints elements starting from 2nd till 3rd
print tuple[2:]    # Prints elements starting from 3rd element
print tinytuple * 2 # Prints list two times
print tuple + tinytuple # Prints concatenated lists
```

Variable Types

- Dictionary
 - ▣ They work like associative arrays or hashes found in Perl and consist of key-value pairs
 - ▣ Dictionaries are enclosed by curly braces ({}) and values can be assigned and accessed using square braces ([])

```
dict = {}  
dict['one'] = "This is one"  
dict[2]    = "This is two"
```

```
tinydict = {'name': 'john','code':6734, 'dept': 'sales'}
```

```
print dict['one']      # Prints value for 'one' key  
print dict[2]          # Prints value for 2 key  
print tinydict         # Prints complete dictionary  
print tinydict.keys()  # Prints all the keys  
print tinydict.values() # Prints all the values
```

Data Conversion

Function	Description
int(x [,base])	Converts x to an integer. base specifies the base if x is a string.
long(x [,base])	Converts x to a long integer. base specifies the base if x is a string
float(x)	Converts x to a floating-point number.
complex(real [,imag])	Creates a complex number.
str(x)	Converts object x to a string representation.
repr(x)	Converts object x to an expression string.
eval(str)	Evaluates a string and returns an object.
tuple(s)	Converts s to a tuple.
list(s)	Converts s to a list.
set(s)	Converts s to a set.
dict(d)	Creates a dictionary. d must be a sequence of (key,value) tuples.
frozenset(s)	Converts s to a frozen set.
chr(x)	Converts an integer to a character.
unichr(x)	Converts an integer to a Unicode character.
ord(x)	Converts a single character to its integer value.
hex(x)	Converts an integer to a hexadecimal string.
oct(x)	Converts an integer to an octal string.

Basic Operation

Operator	Description	Example
+ Addition	Adds values on either side of the operator.	$a + b = 30$
- Subtraction	Subtracts right hand operand from left hand operand.	$a - b = -10$
* Multiplication	Multiplies values on either side of the operator	$a * b = 200$
/ Division	Divides left hand operand by right hand operand	$b / a = 2$
% Modulus	Divides left hand operand by right hand operand and returns remainder	$b \% a = 0$
** Exponent	Performs exponential (power) calculation on operators	$a^{**}b = 10 \text{ to the power } 20$
//	Floor Division - The division of operands where the result is the quotient in which the digits after the decimal point are removed.	$9//2 = 4 \text{ and } 9.0//2.0 = 4.0$

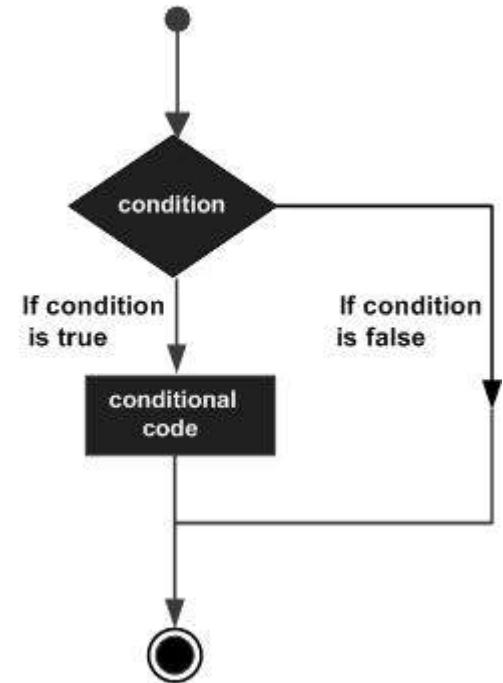
Comparison Operators

Operator	Description	Example
<code>==</code>	If the values of two operands are equal, then the condition becomes true.	$(a == b)$ is not true.
<code>!=</code>	If values of two operands are not equal, then condition becomes true.	
<code>◊</code>	If values of two operands are not equal, then condition becomes true.	$(a ◊ b)$ is true. This is similar to <code>!=</code> operator.
<code>></code>	If the value of left operand is greater than the value of right operand, then condition becomes true.	$(a > b)$ is not true.
<code><</code>	If the value of left operand is less than the value of right operand, then condition becomes true.	$(a < b)$ is true.
<code>>=</code>	If the value of left operand is greater than or equal to the value of right operand, then condition becomes true.	$(a >= b)$ is not true.
<code><=</code>	If the value of left operand is less than or equal to the value of right operand, then condition becomes true.	$(a <= b)$ is true.

Decision Making

- Decision making is anticipation of conditions occurring while execution of the program and specifying actions taken according to the conditions

```
var = 100
if ( var == 100 ):
    print("Value of expression is 100")
else:
    print("Value of expression is not 100")
```



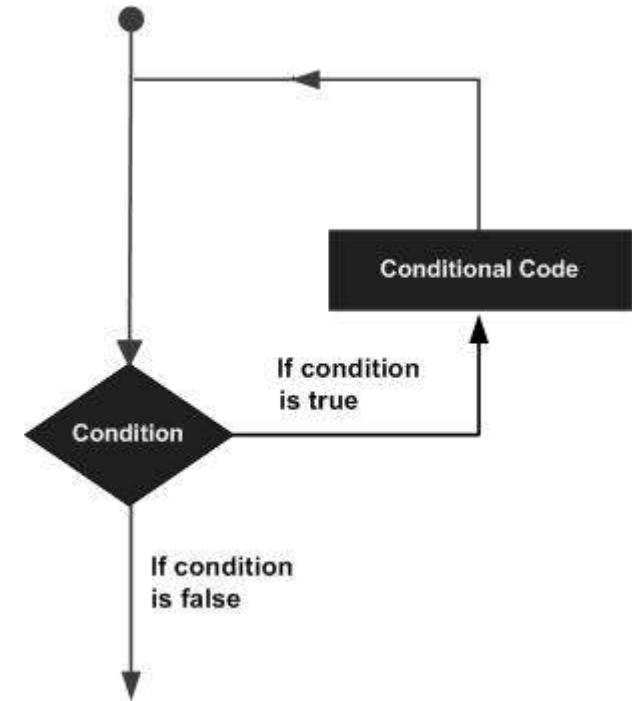
Loop

- A loop statement allows us to execute a statement or group of statements multiple times

```
primes = [2, 3, 5, 7]
for prime in primes:
    print(prime)
```

```
for x in range(5): # or range(5)
    print(x)
```

```
count = 0
while count < 5:
    print(count)
    count += 1 # This is the same as count = count + 1
```



Loop

- Loop control statements change execution from its normal sequence

Control Statement	Description
break statement	Terminates the loop statement and transfers execution to the statement immediately following the loop.
continue statement	Causes the loop to skip the remainder of its body and immediately retest its condition prior to reiterating.
pass statement	The pass statement in Python is used when a statement is required syntactically but you do not want any command or code to execute.

```
count = 0
while True:
    print(count)
    count += 1
    if count >= 5:
        break
```

List comprehensions

- Python supports a concept called “list comprehensions” used to construct lists in a very natural, easy way

```
>>> A=[x**2 for x in range(10)]
>>> print(A)
[0, 1, 4, 9, 16, 25, 36, 49, 64, 81]
>>> B = [x for x in S if x % 2 == 0]
>>> print(B)
[0, 4, 16, 36, 64]
>>> C = [x+3 for x in A]
>>> print(C)
[3, 4, 7, 12, 19, 28, 39, 52, 67, 84]
>>> D = [x+3 if x%2==0 else x for x in A]
>>> print(D)
[3, 1, 7, 9, 19, 25, 39, 49, 67, 81]
```

List comprehensions

- $A=[x**2 \text{ for } x \text{ in range}(10)]$
 - ▣ $\text{range}(10)$ creates list whose elements are from zero to nine
 $[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]$
 - ▣ $\text{for } x \text{ in range}(10)$: loop for elements in range(10)
 - x represents each element in range(10)
 - ▣ $x**2$ for x in range(10): for every element in range(10), calculate x^2
 - Results are stored in A as list

Index of Python

□ Python list

index	0	1	2	3	4	5	6
negative index	-7	-6	-5	-4	-3	-2	-1
	8	7	5	13	75	65	11

```
>>>A=[8,7,5,13,75,65,11]
```

```
>>>A[0]
```

```
8
```

```
>>>A[3]
```

```
13
```

```
>>>A[-1]
```

```
11
```

```
>>>A[1:4]
```

```
[7,5,13]
```

```
>>>A[:3]
```

```
[8,7,5]
```

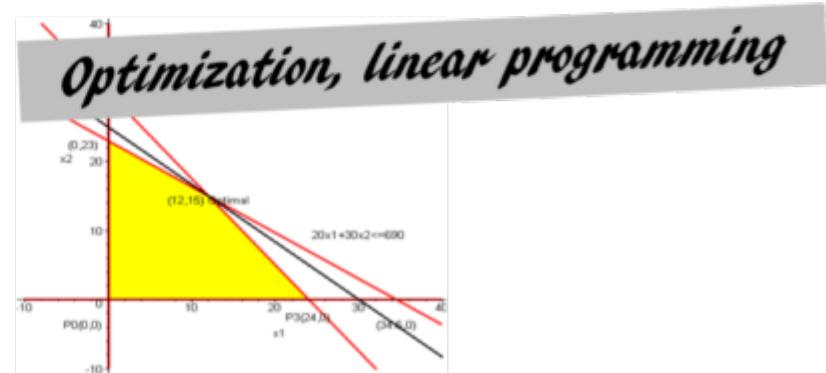
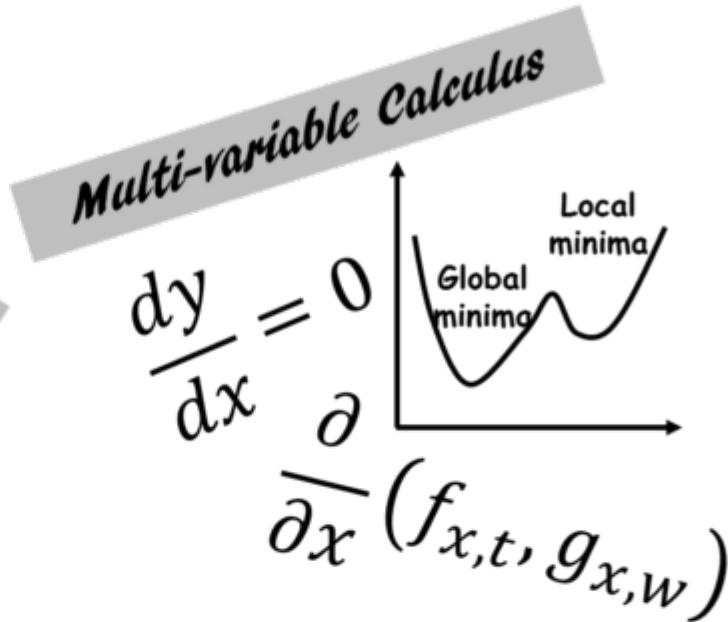
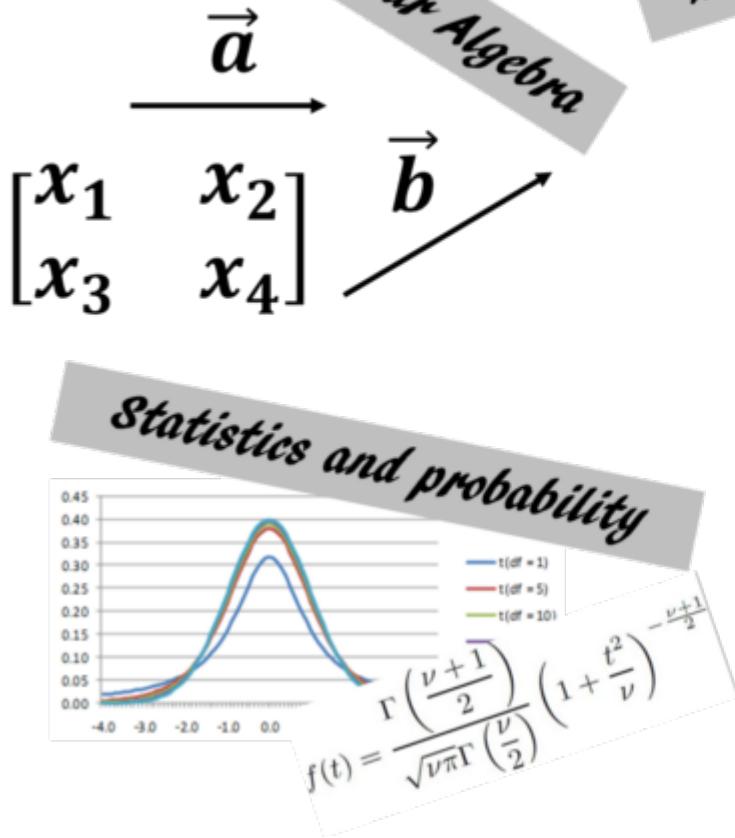
```
>>>A[4:]
```

```
[75,65,11]
```

BACKGROUND OF DATA MINING

Week02

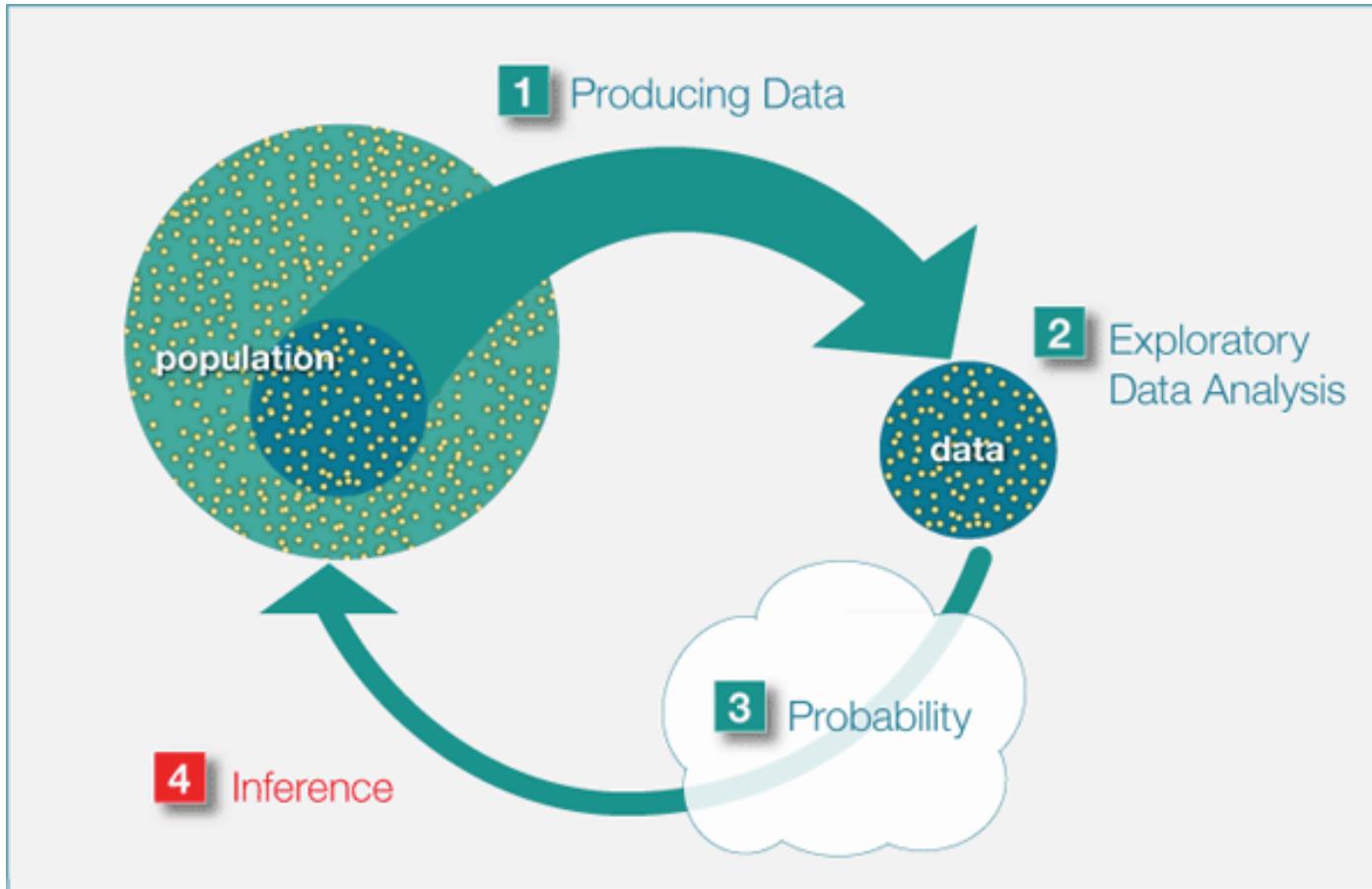
Essential Math for Data Mining



Essential Math for Data Mining: Statistics

- Two main branches of statistics
 - ▣ Descriptive statistics *characteristic, property . explore data*
 - Describe the basic features of data
 - Data summaries and descriptive statistics, central tendency, variance, covariance, correlation
 - ▣ Inferential statistics *estimate population based on sample-*
 - Deduce properties of an underlying distribution of probability
- Probability
 - ▣ Sampling, measurement, error, random number generation
 - ▣ Basic probability: basic idea, expectation, probability calculus, Bayes theorem, conditional probability
 - ▣ Probability distribution functions—uniform, normal, binomial, chi-square, student's t-distribution, Central limit theorem

Essential Math for Data Mining: Statistics



Essential Math for Data Mining: Linear Algebra

- Linear algebra
 - ▣ The study of linear sets of equations and their transformation properties
 - ▣ Concern linear equations, linear functions and their representations in vector spaces and through matrices
 - ▣ Used in most areas of science and engineering, because it allows modeling many natural phenomena, and efficiently computing with such models

$$\begin{aligned}3x + 5y &= 7 \\x - 2y &= 6\end{aligned}$$



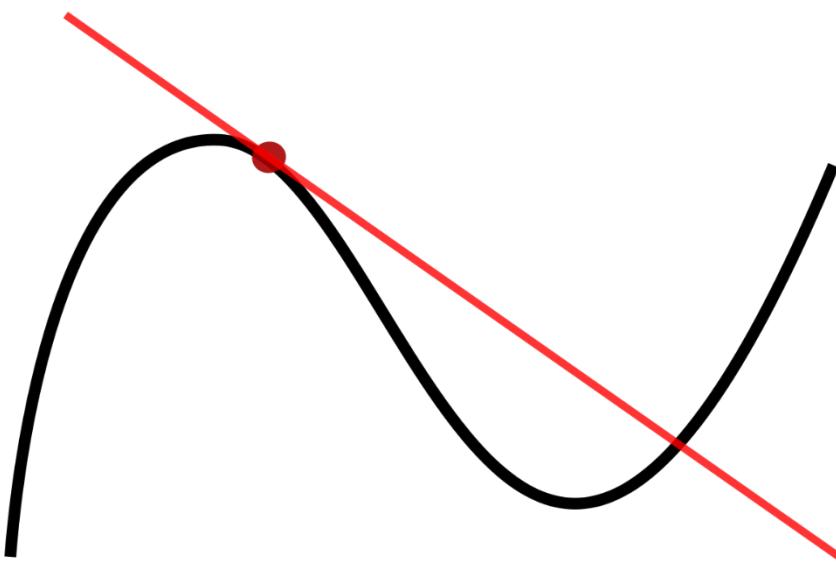
$$\begin{bmatrix} 3 & 5 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 7 \\ 6 \end{bmatrix}$$



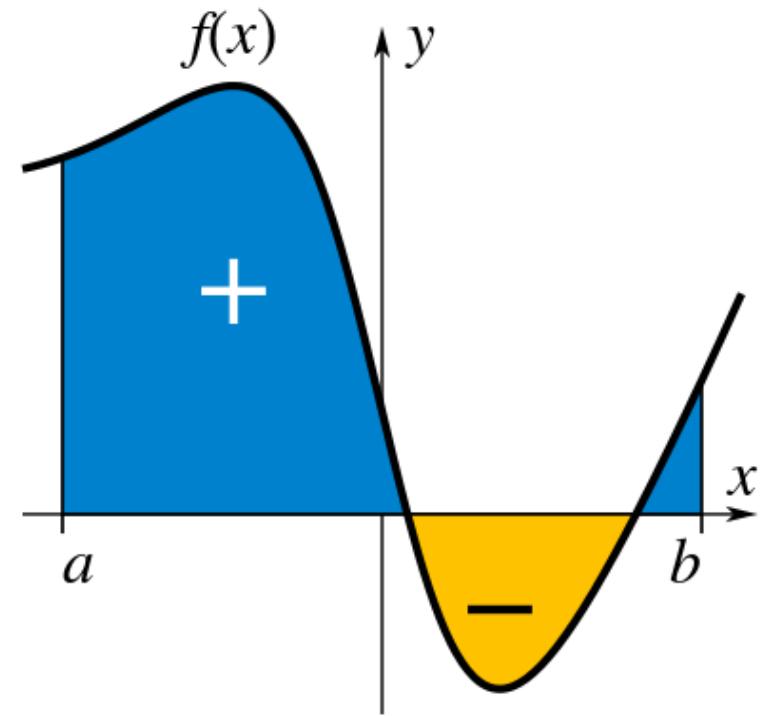
$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 & 5 \\ 1 & -2 \end{bmatrix}^{-1} \begin{bmatrix} 7 \\ 6 \end{bmatrix}$$

Essential Math for Data Mining: Calculus

- Calculus
 - ▣ Branch of mathematics concerned with the calculation of instantaneous rates of change (differential calculus) and the summation of infinitely many small factors to determine some whole (integral calculus)



❖ differential calculus



integral calculus

Essential Math for Data Mining: Optimization

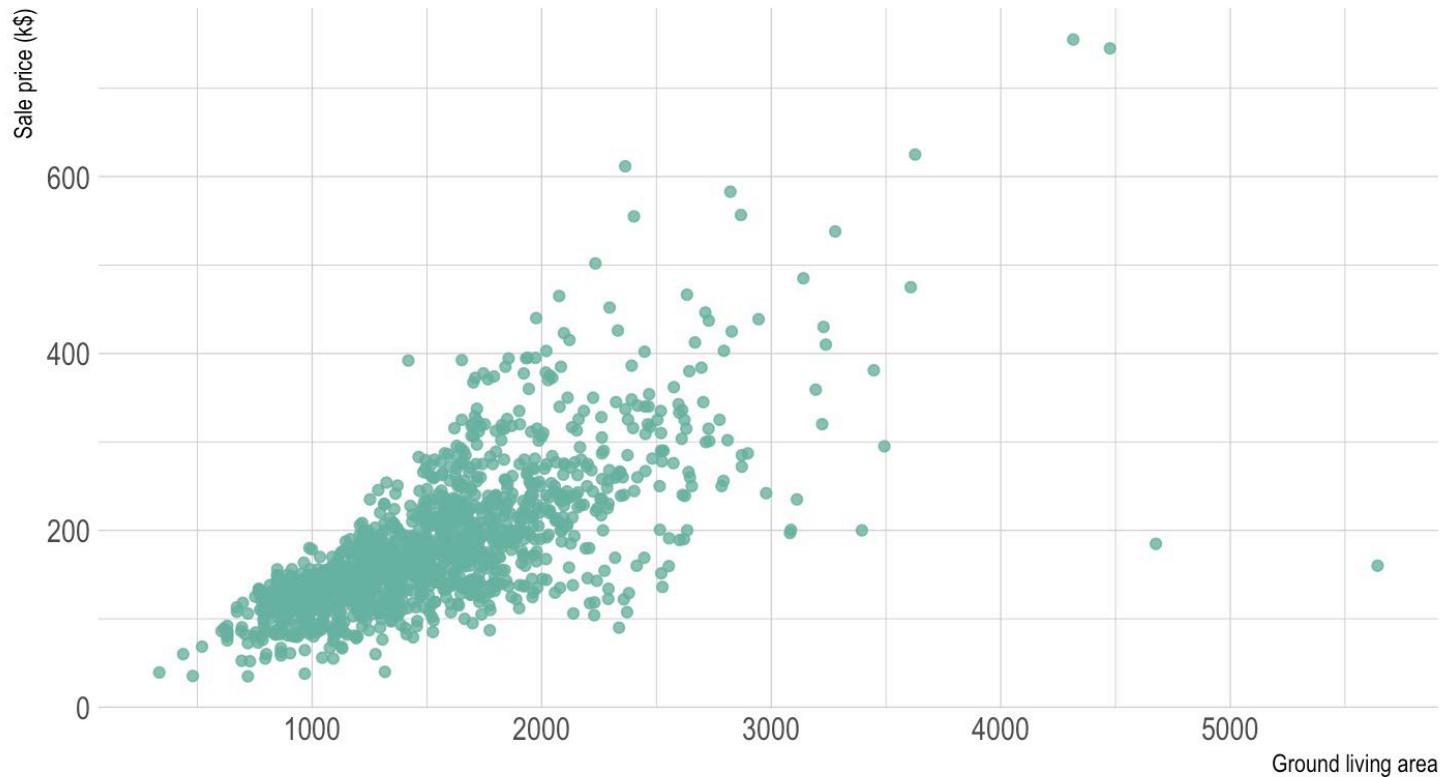
- Optimization
 - ▣ Collection of mathematical principles and methods used for solving optimization problems
 - ▣ Optimization problem is the problem of finding the best solution from all feasible solutions
 - In the simplest case, an optimization problem consists of maximizing or minimizing a real function

Statistics

- A vast set of tools for understanding data

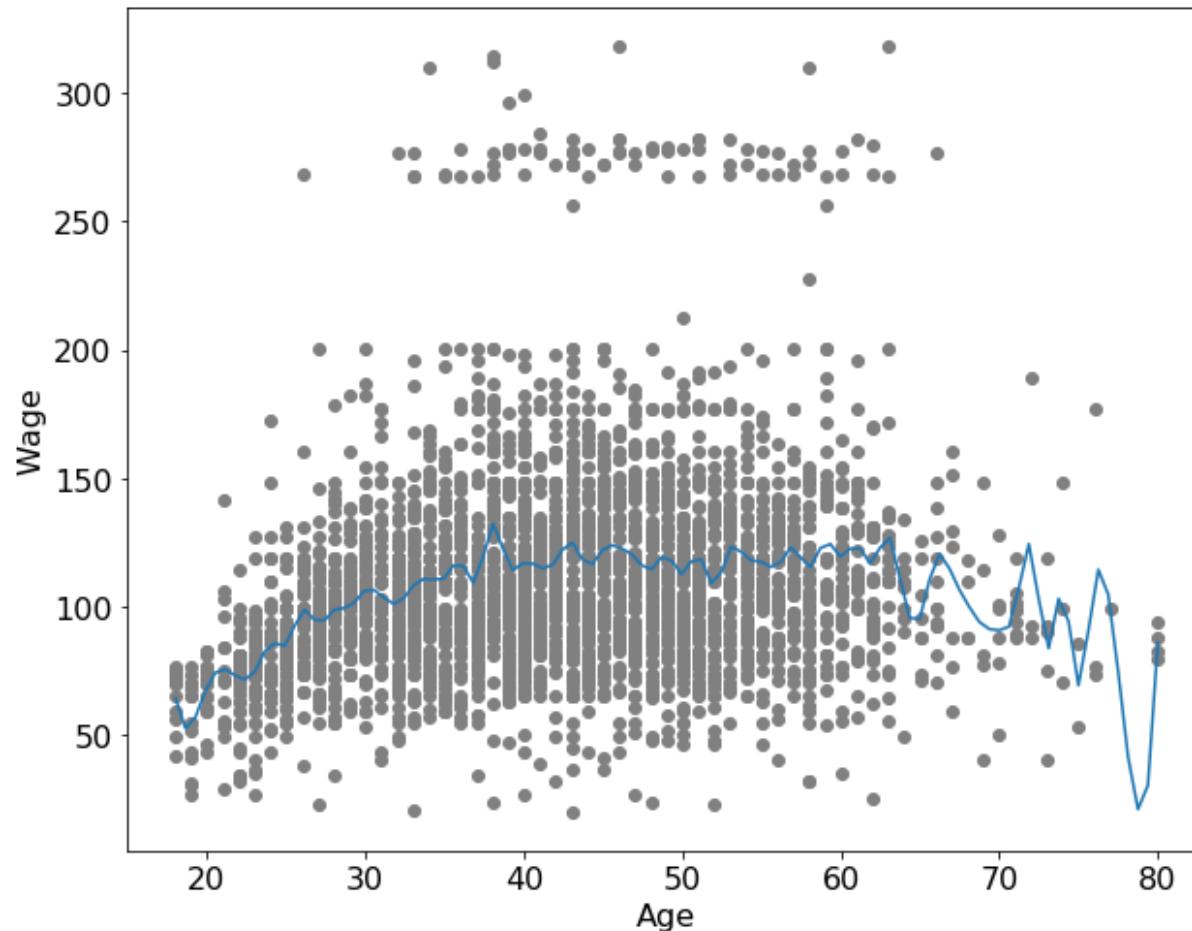
visualize
→ qualitative.

Ground living area partially explains sale price of apartments



Statistics

- A vast set of tools for understanding data

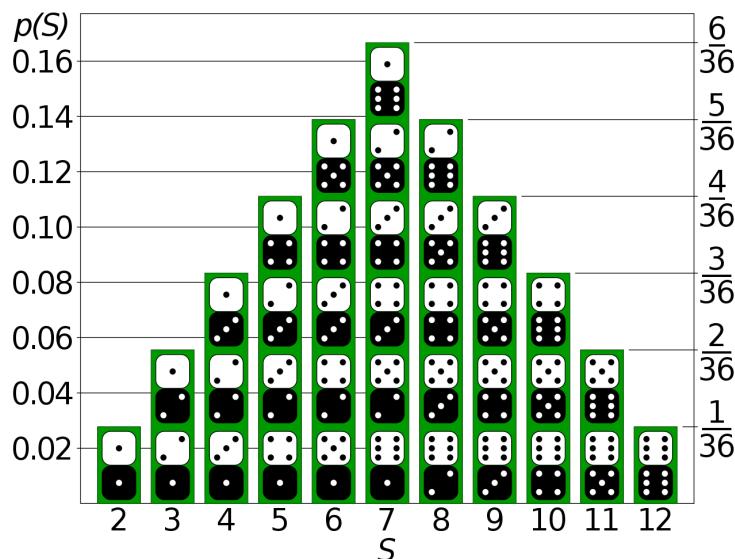


Statistics

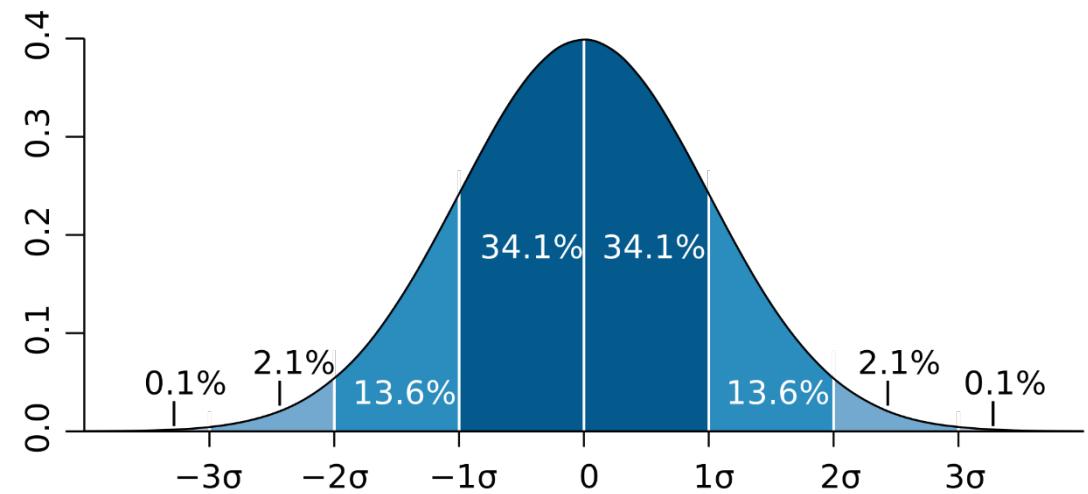
- Descriptive statistics
 - ▣ A **summary statistic** that quantitatively describes or summarizes features of a collection of information
 - ▣ Univariate
 - Mean, Median, Mode *平均数*
 - Variance, standard deviation, Percentile
 - Skewness, kurtosis *偏度* *峰度*
 - ▣ Bivariate or multivariate
 - Cross-tabulations and contingency tables
 - Graphical representation via scatterplots
 - Quantitative measures of dependence (covariance, correlation)

Statistics

- Probability distribution
 - ▣ A mathematical function that provides the probabilities of occurrence of different possible outcomes
 - The probabilities of occurrence of the specific observations



Discrete random variable →
probability mass function



Continuous random variable →
probability density function

Statistics: Discrete Probability Distributions

- Bernoulli distribution *... Yes/No, tossing coin*
 - ▣ The discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$
$$\Pr(X = 1) = p = 1 - \Pr(X = 0) = 1 - q$$
 - ▣ A special case of the binomial distribution where a single trial is conducted (so n would be 1 for such a binomial distribution)
 - ▣ Probability mass function
$$f(X = k; p) = p^k (1 - p)^{1-k}$$
- Binomial distribution *이항 분포*
 - ▣ The discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes-no question, and each with its own Boolean-valued outcome: success (with probability p) or failure (with probability $q = 1 - p$).
 - ▣ Probability mass function
$$f(X = k; n, p) = \Pr(X = k) = \frac{n!}{k!(n - k)!} p^k (1 - p)^{n-k}$$
$$n C_k p^k (1-p)^{n-k}$$

$$f(X = k; n, p) = \Pr(X = k) = \frac{n!}{k!(n - k)!} p^k (1 - p)^{n-k}$$
$$n C_k p^k (1-p)^{n-k}$$

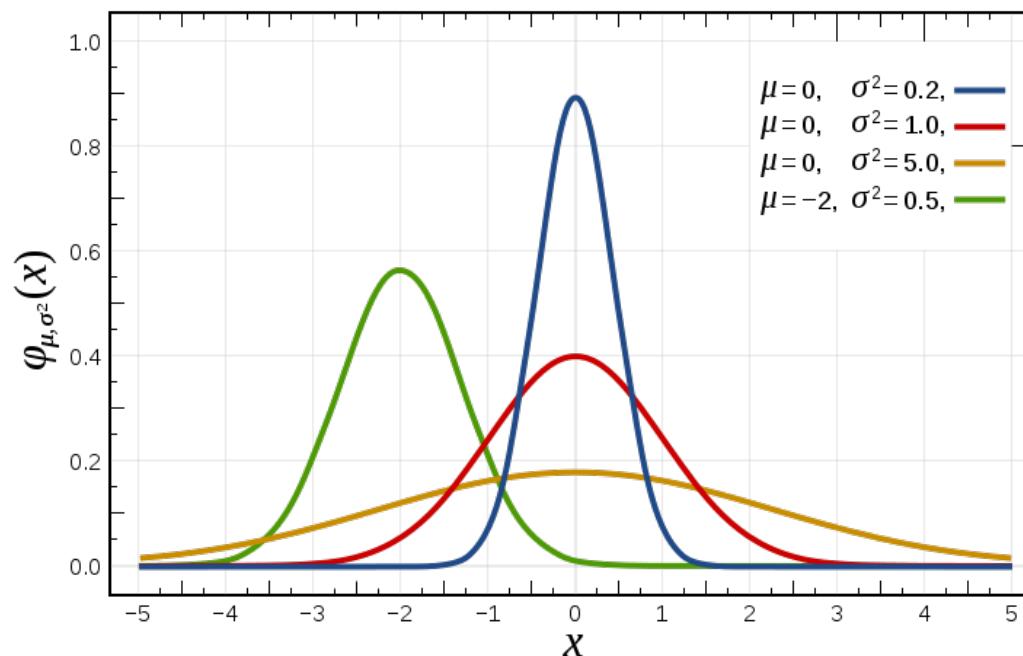
Statistics: Continuous Distributions

- Normal (Gaussian) distribution
 - Very common continuous probability distribution
 - Bell-shaped

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

■ μ : mean

■ σ : standard deviation



Statistics: Continuous Distributions

- Student's t -distribution (t -distribution)
 - ▣ Continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the **sample size is small** and **population standard deviation is unknown**
- Let X_1, \dots, X_n be independent and identically distributed (iid) as $N(\mu, \sigma^2)$
 - ▣ Sample mean

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- ▣ Sample variance

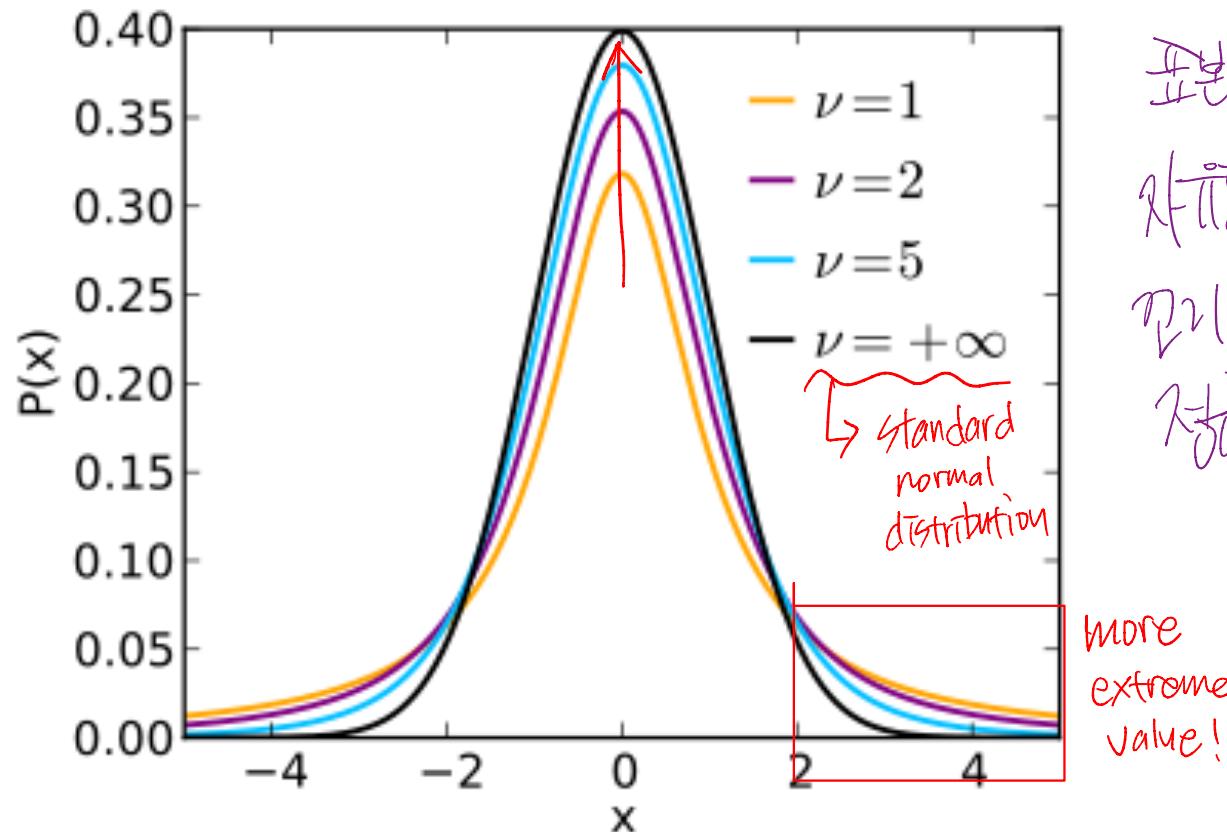
$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- ▣ The random variable $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution
- ▣ The random variable $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a Student's t -distribution with **$n-1$ degrees of freedom**

모집단의 분산 모를 때!

Statistics: Continuous Distributions

- The probability density function of t -distribution with varying degree of freedom



표본 수 증가
자연 분포
균일 분포
정규분포에 가까워지기

Statistics: Student's t -distribution

- Probability density function

$$f(x; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- ν : degree of freedom
- Γ : gamma function

$$\Gamma(n) = (n-1)! \quad \text{if } n \text{ is positive integer}$$

$$\Gamma(z) = \int_0^{-\infty} x^{z-1} e^{-x} dx$$

Statistics: Continuous Distributions

□ Chi-squared distribution (χ^2)

자수분포
제곱분산

- The distribution of a sum of the squares of k independent standard normal random variables

- Let X_1, \dots, X_k be independent, standard normal random variables

$$Y = \sum_{i=1}^k X_i^2$$

is distributed according to the chi-squared distribution with k degrees of freedom

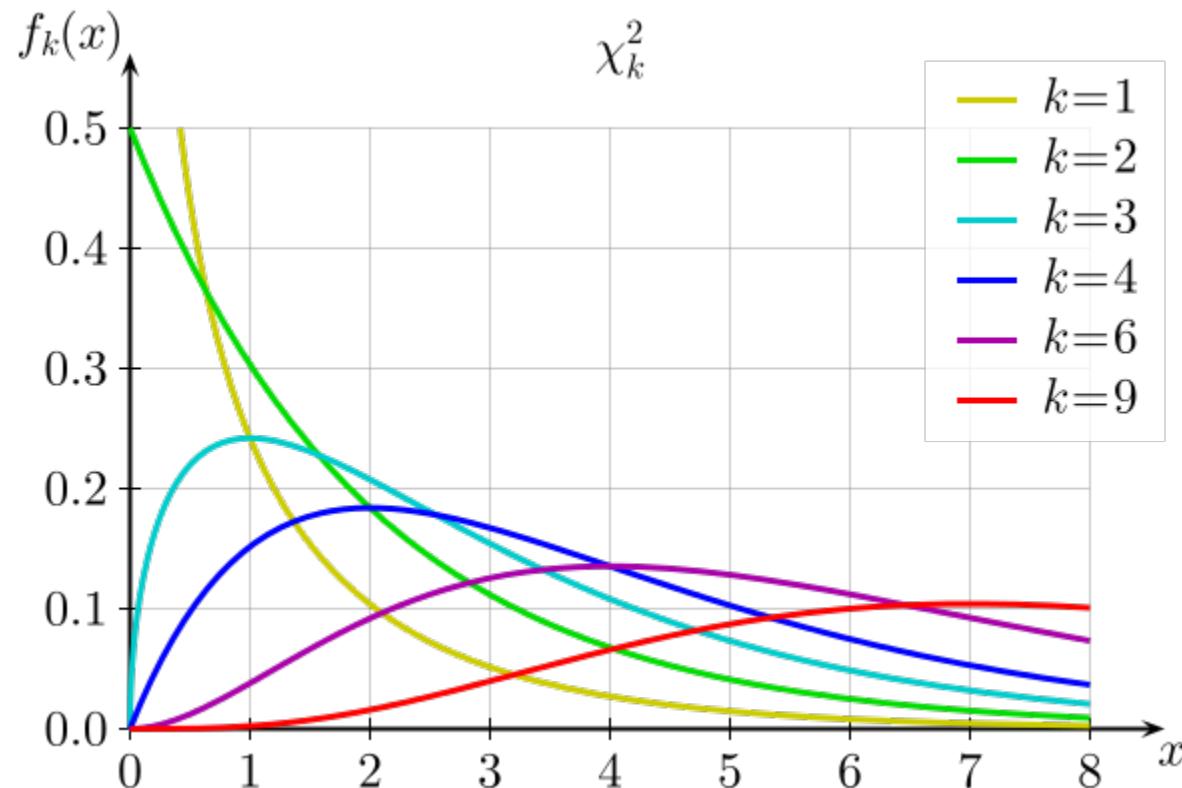
$$Y \sim \chi^2(k) \text{ or } Y \sim \chi_k^2$$

- Probability density function

$$f(x; k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

Statistics: Continuous Distributions

- The probability density function of chi-squared distribution with varying degree of freedom



Statistics: Continuous Distributions

□ **F-distribution**

- A random variate of the F -distribution with parameters d_1 and d_2 arises as the ratio of two appropriately scaled chi-squared variates
- Let X_1 and X_2 be two independent random variables and $X_1 \sim \chi^2(d_1)$ and $X_2 \sim \chi^2(d_2)$

$$Y = \frac{X_1/d_1}{X_2/d_2}$$

is distributed according to the F -distribution with d_1 and d_2 degrees of freedom

$$Y \sim F(d_1, d_2)$$

□ Probability density function

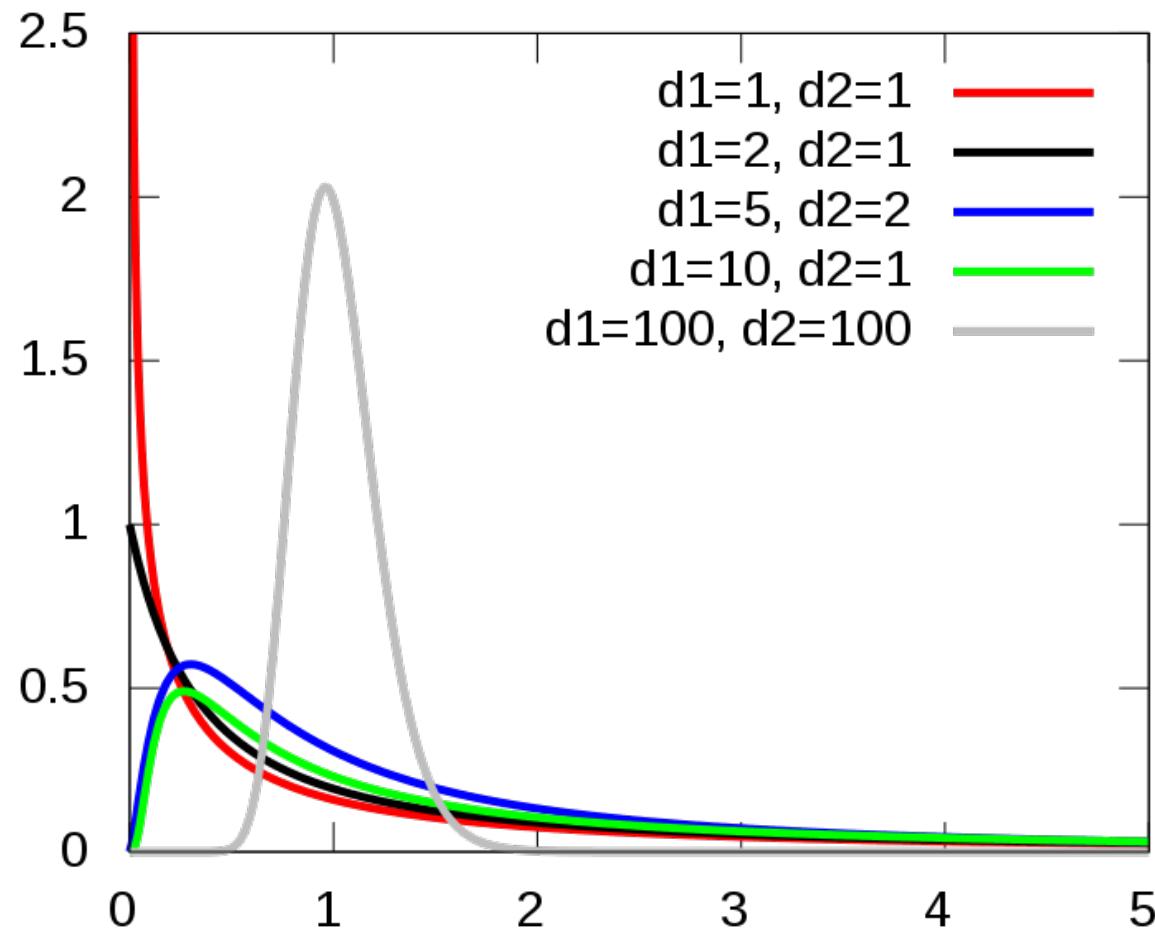
$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

- B: beta function

$$B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt$$

Statistics: Continuous Distributions

- The probability density function of F -distribution with varying degree of freedom



Linear Algebra

- Linear Algebra
 - ▣ Basic properties of matrix and vectors—scalar multiplication, linear transformation, transpose, conjugate, rank, determinant
 - ▣ Inner and outer products, matrix multiplication rule and various algorithms, matrix inverse
 - ▣ Special matrices—square matrix, identity matrix, triangular matrix, idea about sparse and dense matrix, unit vectors, symmetric matrix, Hermitian, skew-Hermitian and unitary matrices
 - ▣ Gaussian/Gauss-Jordan elimination, solving $Ax=b$ linear system of equation
 - ▣ Matrix factorization and decomposition
 - ▣ Vector space, basis, span, orthogonality, orthonormality, linear least square
 - ▣ Eigenvalues, eigenvectors, and diagonalization, singular value decomposition (SVD)

Linear Algebra

- Linear algebra is the study of vectors and linear functions
 - ▣ Scalar
 - A scalar is a number
 - ▣ Vector
 - A vector is a list of numbers
 - ▣ Matrix
 - A matrix is also a collection of numbers
 - The difference is that a matrix is a table of numbers rather than a list
- ▣ $\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$
- ▣ $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$
- ▣ Linear equation
 - $$a_1x_1 + \cdots + a_nx_n = b$$
- ▣ Linear function
 - $$(x_1, \dots, x_n) \mapsto a_1x_1 + \cdots + a_nx_n$$

Linear Algebra

- Vectors

- ▣ Addition

$$\mathbf{v} + \mathbf{w}$$

- Example

$$\mathbf{v} + \mathbf{w} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$

- ▣ Linear combination

$$a\mathbf{v} + b\mathbf{w}$$

- Example

$$3\mathbf{v} + 4\mathbf{w} = 3 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 4 \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 15 \\ 22 \end{bmatrix}$$

Linear Algebra

□ Vectors

▣ Transpose

- column vector \leftrightarrow row vector
- Example

$$\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \rightarrow \mathbf{v}^T = [1 \quad 2]$$

▣ Dot product, inner product

$$[a_1, a_2, \dots, a_n] \cdot [b_1, b_2, \dots, b_n] \\ = a_1b_1 + a_2b_2 + \dots + a_nb_n$$

- Example

$$\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \\ \mathbf{v} \cdot \mathbf{w} = (1)(3) + (2)(4) = 9$$

▣ Length

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}}$$

- Example

$$\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \|\mathbf{v}\| = \sqrt{(1)(1) + (2)(2)} = \sqrt{5}$$

Linear Algebra

- Matrix

- ▣ Addition

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}$$

- ▣ Multiplication

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

- ▣ Linear equation

$$A\mathbf{x} = \mathbf{b}$$

- Example

$$\begin{array}{rcl} x_1 & = & b_1 \\ -x_1 + x_2 & = & b_2 \\ -x_2 + x_3 & = & b_3 \end{array}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

Linear Algebra

□ Matrix

□ Inverse matrix

- An n -by- n square matrix, A is called invertible (or nonsingular) if there exists an n -by- n square matrix, B such that

$$AB = BA = I_n$$

where I_n denotes the n -by- n identity matrix which is a square matrix with ones on the main diagonal and zeros elsewhere

- B is the inverse of A (A^{-1})
- If A has no inverse, A is singular or non-invertible
- Example

$$A = \begin{bmatrix} -1 & \frac{3}{2} \\ 1 & -1 \end{bmatrix}, A^{-1} = \begin{bmatrix} 2 & 3 \\ 2 & 2 \end{bmatrix}$$

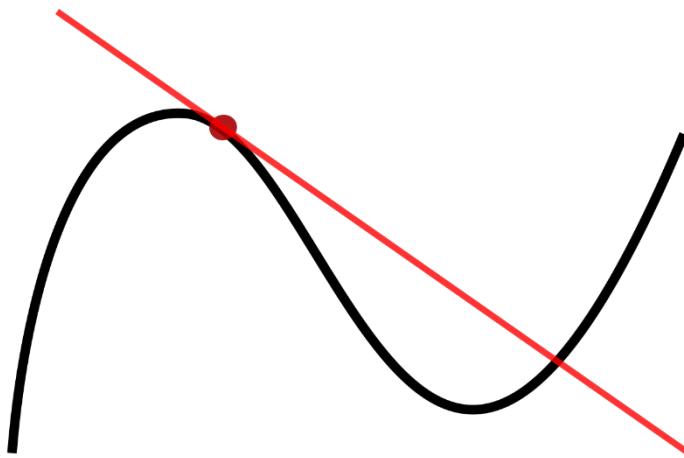
□ Solution of a linear equation

$$\mathbf{x} = A^{-1}\mathbf{b}$$

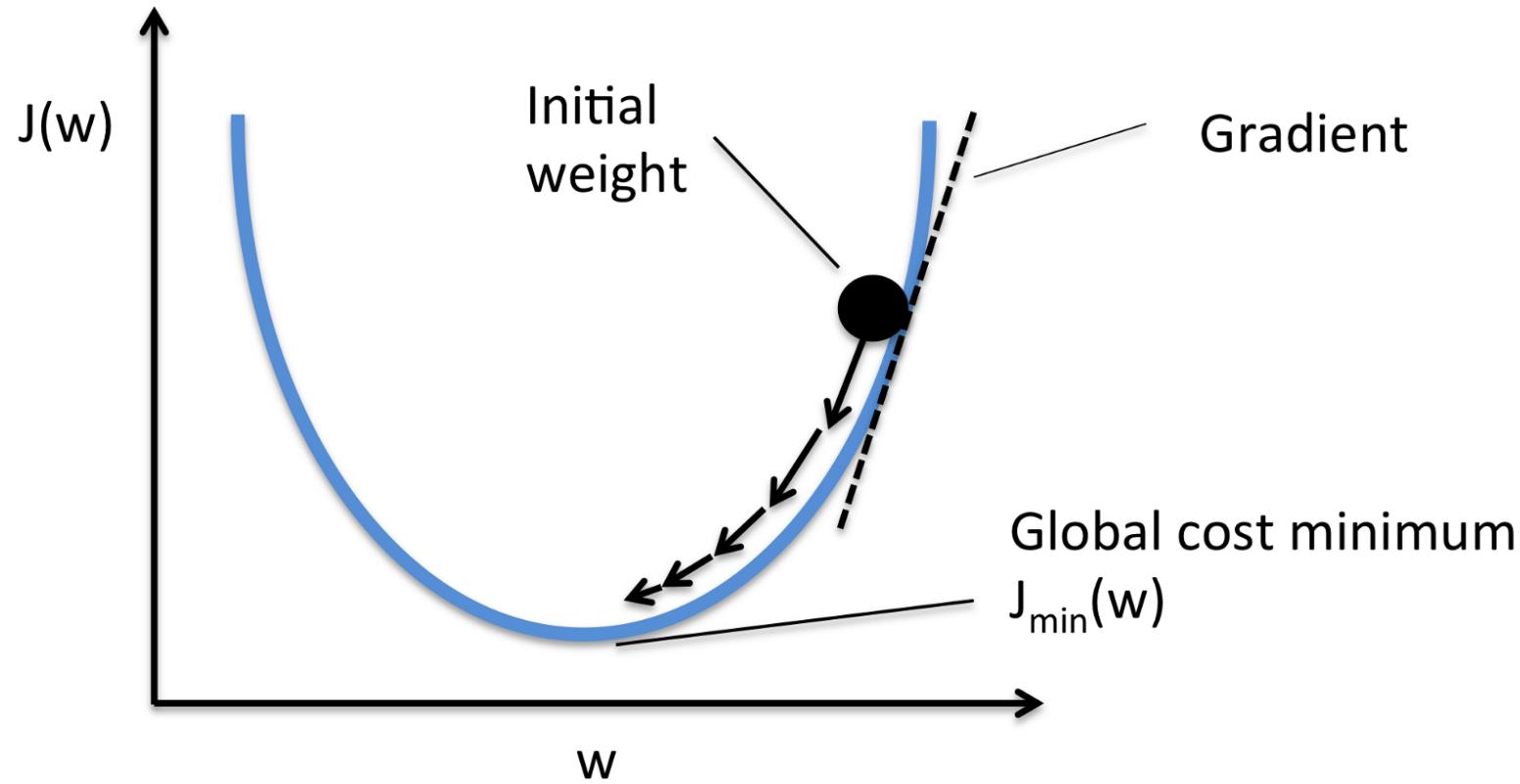
Calculus

- Derivative *기울기*
 - A function of a real variable measures the sensitivity to change of the function value (output value) with respect to a change in its argument (input value)

$$\frac{dy}{dx}$$



Calculus

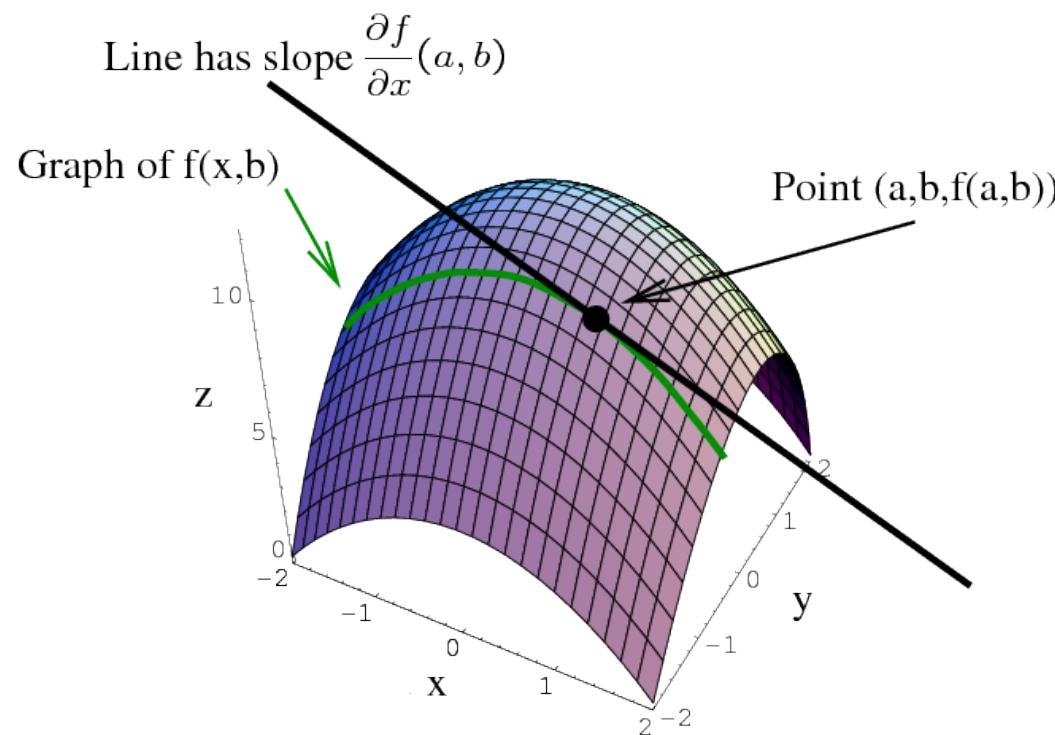


https://en.wikipedia.org/wiki/File:Gradient_Descent_in_2D.webm

Calculus

- Partial derivative *부분 미분*
 - A function of several variables is its derivative with respect to one of those variables, with the others held constant

$$\frac{\partial f}{\partial x}$$

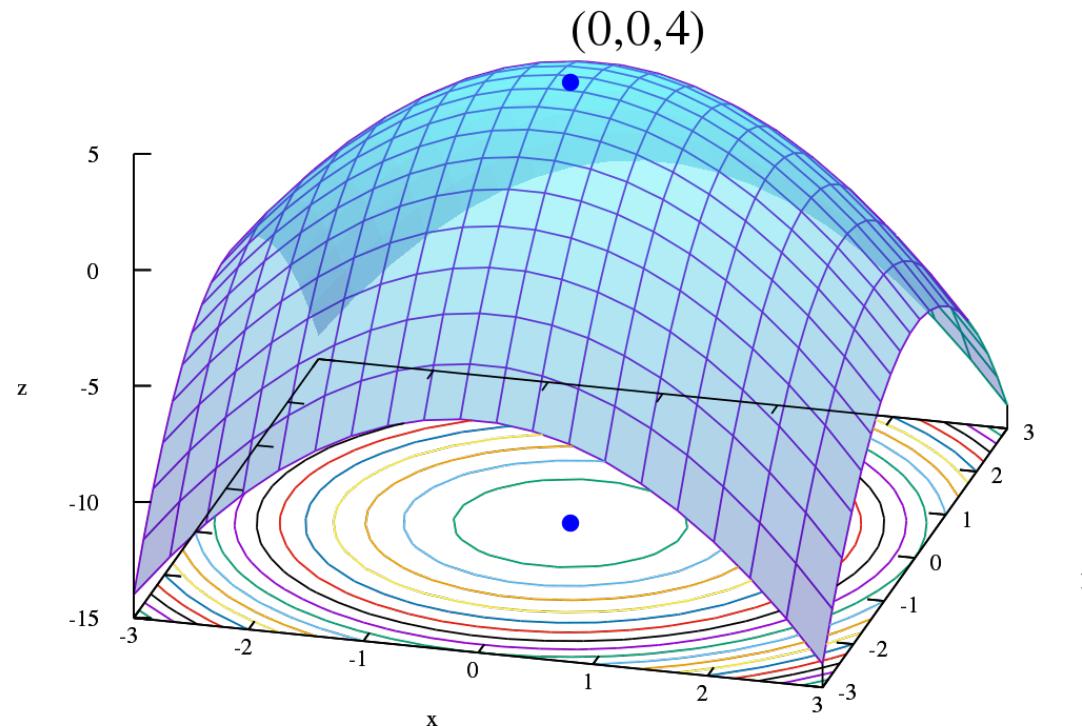


Optimization

- Optimization
 - ▣ Basics of optimization —how to formulate the problem
 - ▣ Linear programming, simplex algorithm
 - ▣ Integer programming
 - ▣ Constraint programming, knapsack problem
 - ▣ Randomized optimization techniques—hill climbing, simulated annealing, Genetic algorithms

Optimization

- Optimization problem
 - ▣ Maximizing or minimizing a real function by systematically choosing input values from within an allowed set and computing the value of the function



Optimization

Example

- For materials, the manufacturer has 750 m² of cotton textile and 1,000 m² of polyester. Every pair of pants (1 unit) needs 1 m² of cotton and 2 m² of polyester. Every jacket needs 1.5 m² of cotton and 1 m² of polyester.
- The price of the pants is fixed at \$50 and the jacket, \$40.
- What is the number of pants and jackets that the manufacturer must give to the stores so that these items obtain a maximum sale?**

- Variables to be determined

$$x = \text{number of pants}$$

$$y = \text{number of jackets}$$

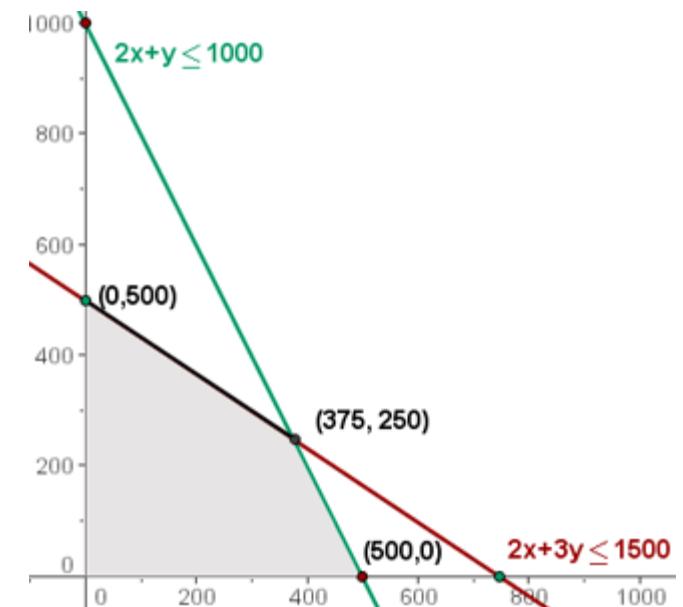
- Objective function

$$f(x, y) = 50x + 40y$$

- Constraints

$$x + 1.5y \leq 750$$

$$2x + y \leq 1000$$

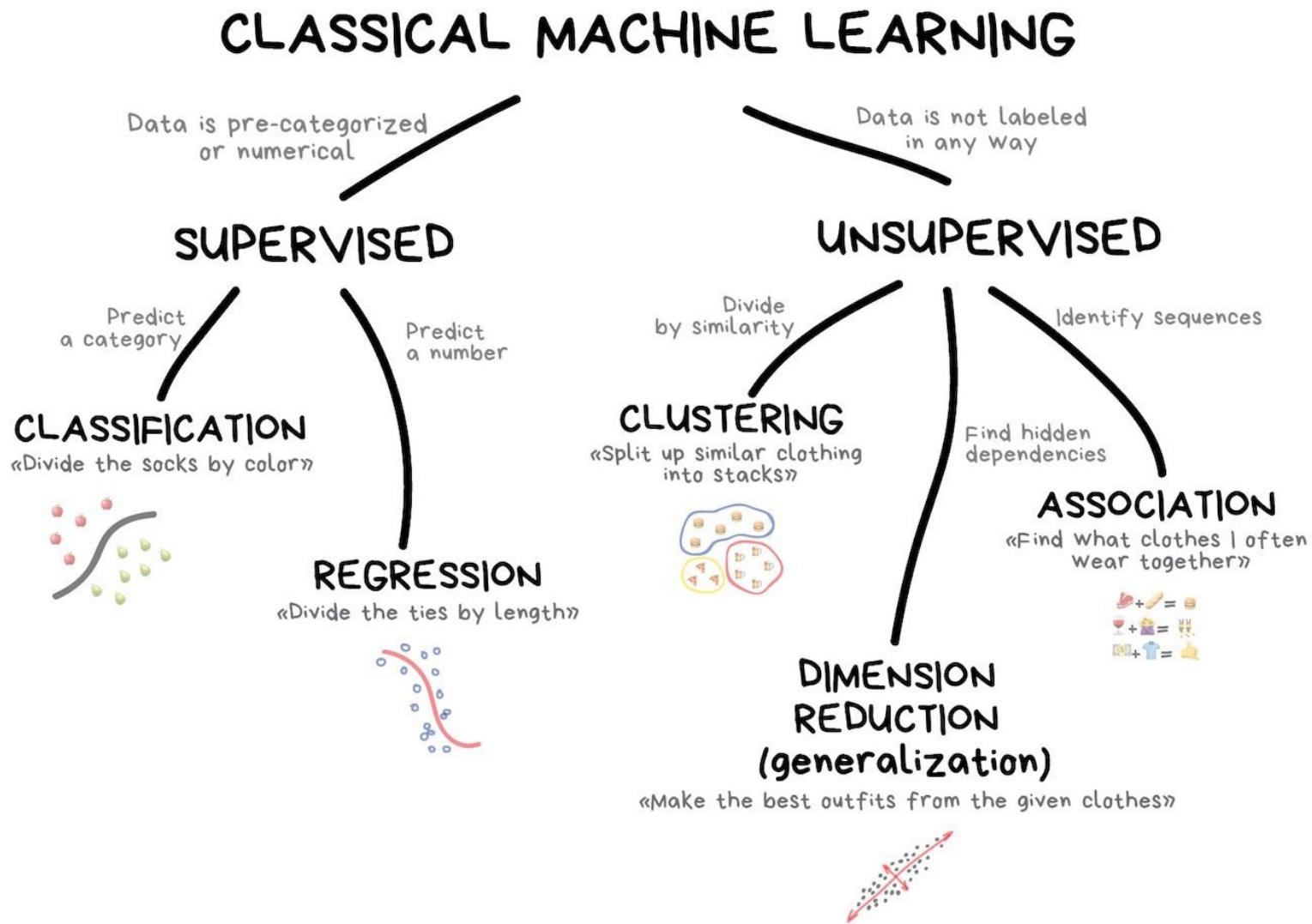


Optimization

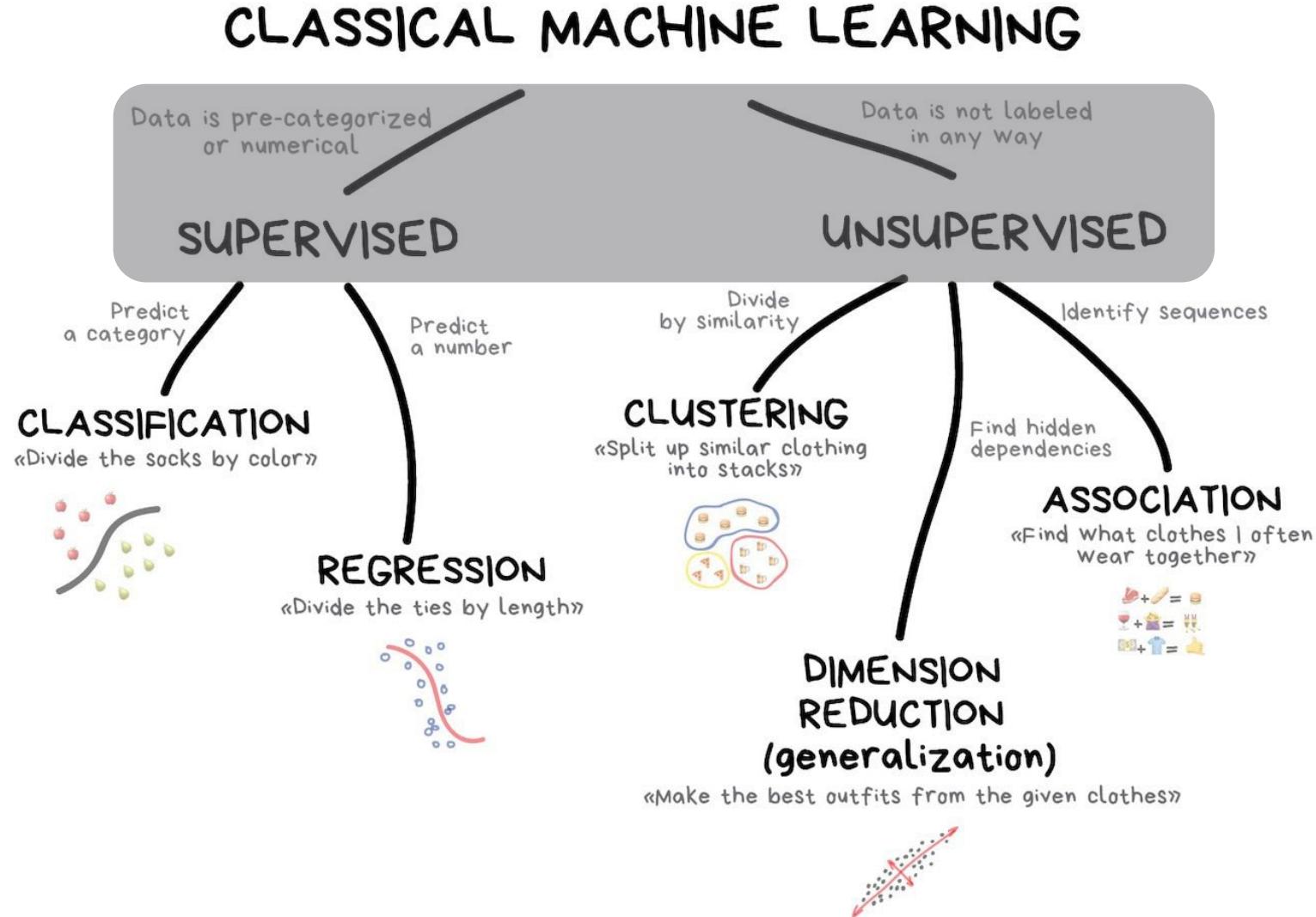
- Why are optimization algorithms important for data analysis?
 - ▣ One of fundamental data analysis tasks is to seek a function that approximately maps \mathbf{x}_i to y_i for each observation, i
$$y = f(\mathbf{x})$$
 - ▣ The process of finding f based on data is called learning or training
 - During learning, optimization algorithms provide a tool to find the most appropriate f

Basic Terminologies

Topics Covered in This Class



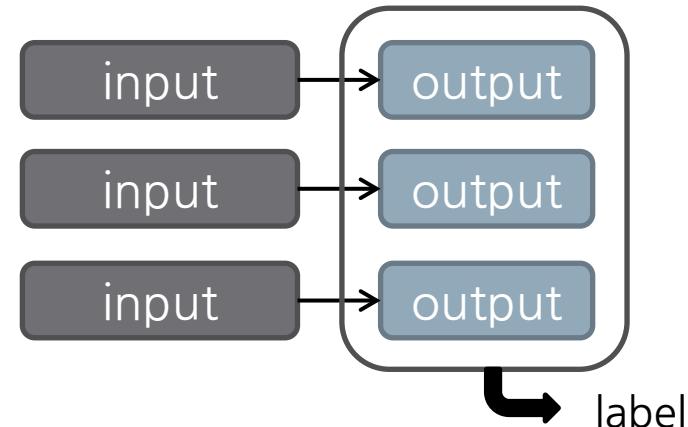
Topics Covered in This Class



Types of Learning

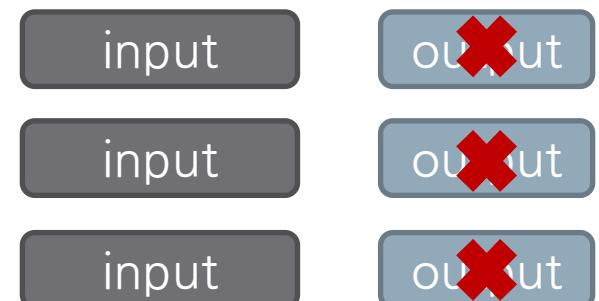
□ Supervised learning

- We have knowledge of output
 - We call such data **labeled**
 - **We know answer**
- Goal
 - Estimate output for unlabeled input

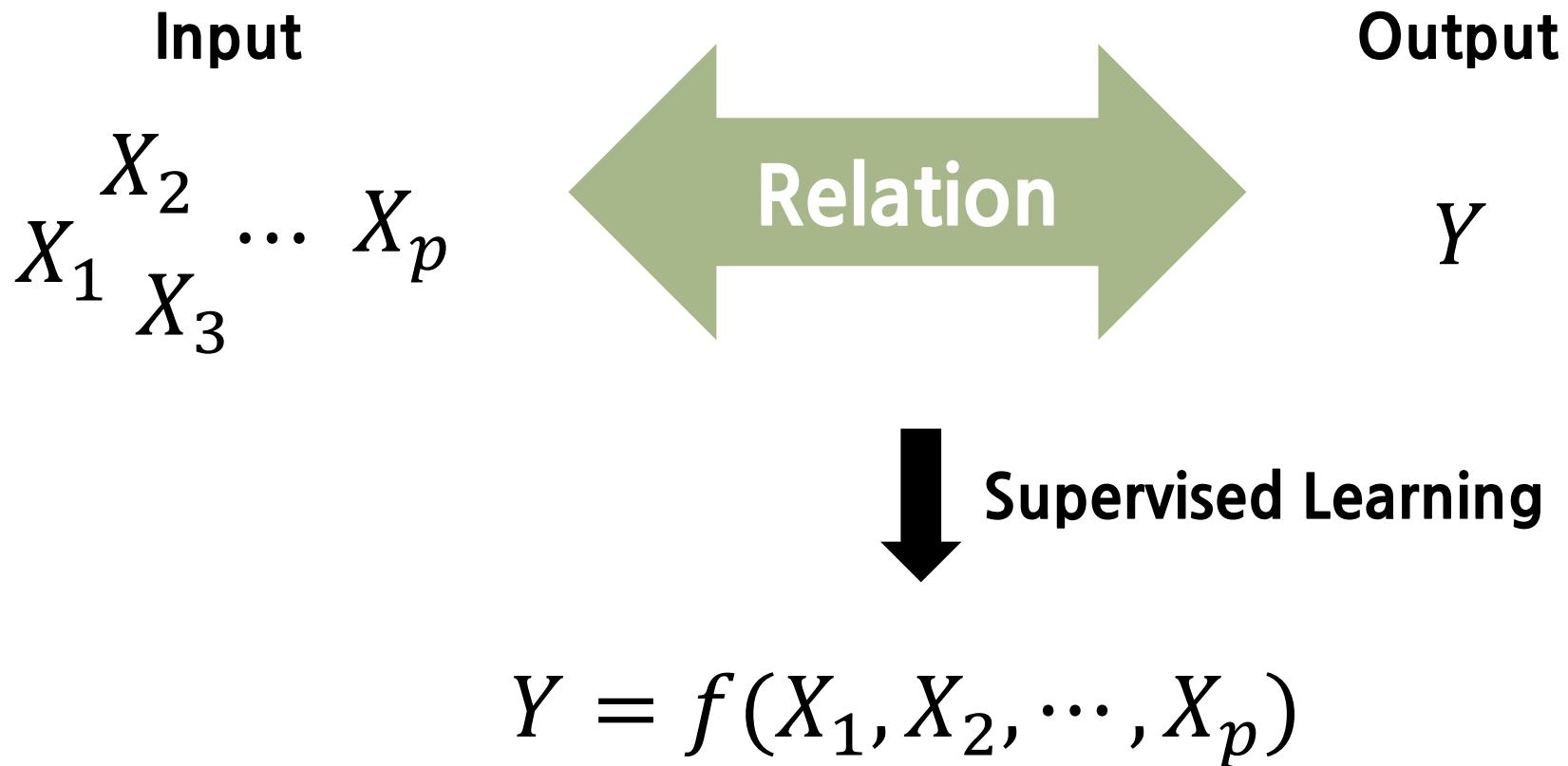


□ Unsupervised learning

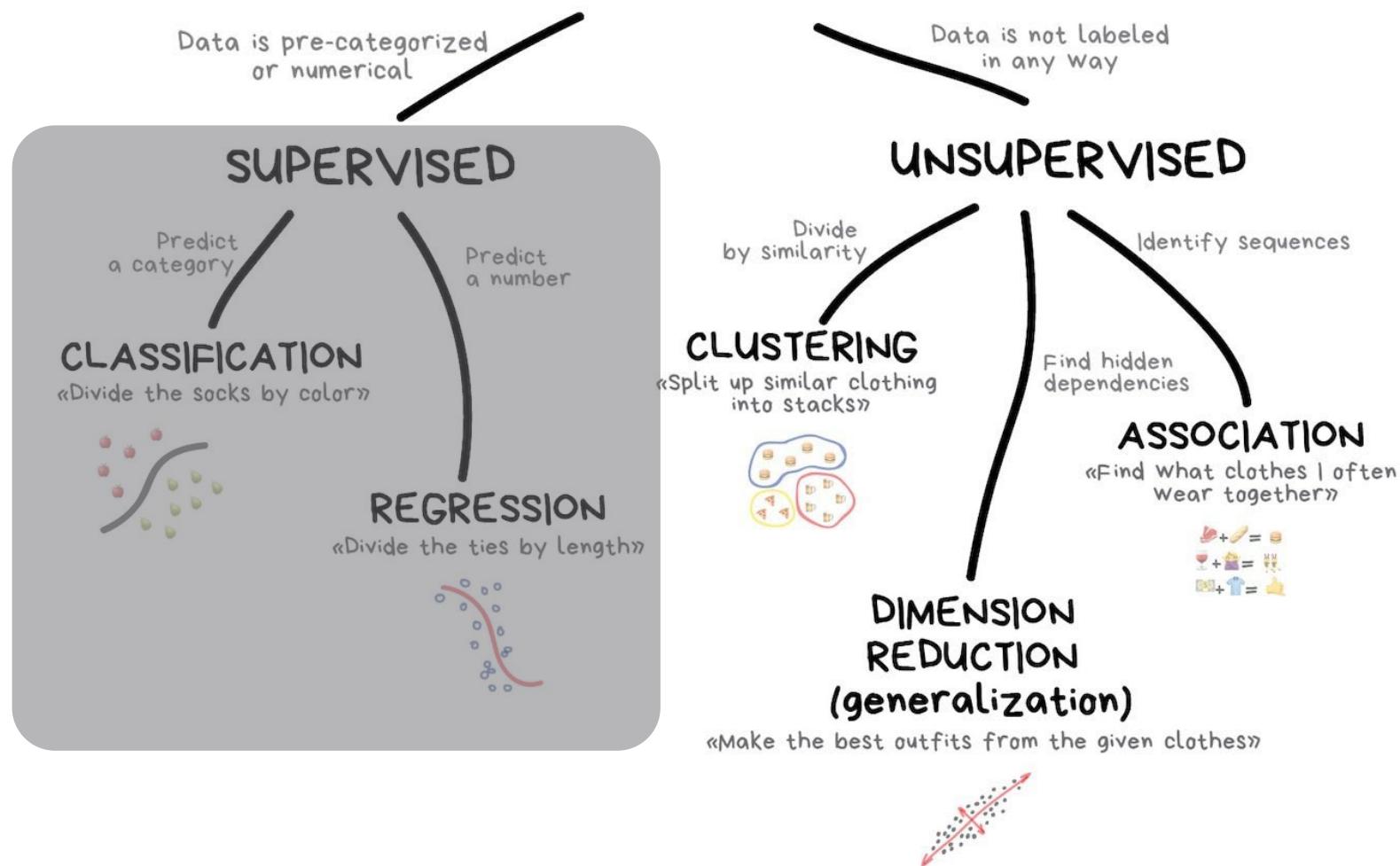
- **No output**
 - We call such data **unlabeled**
- Goal
 - Find patterns, groups, or relation



Supervised learning



Topics Covered in This Class



Data for Data Mining: Structured Data

- Example of data set
 - The input data set is usually expressed as a set of independent instances

Structured data.

instance,
sample,
example

Outlook	Temperature(°F)	Humidity(%)	Windy	Play Time(min)
Sunny	85	85	false	5
Sunny	80	90	true	0
Rainy	70	96	false	40
Rainy	68	80	false	65
Sunny	72	95	false	0
Sunny	69	70	false	70
Rainy	75	80	true	45

variable,
attribute,
feature

Types of Data

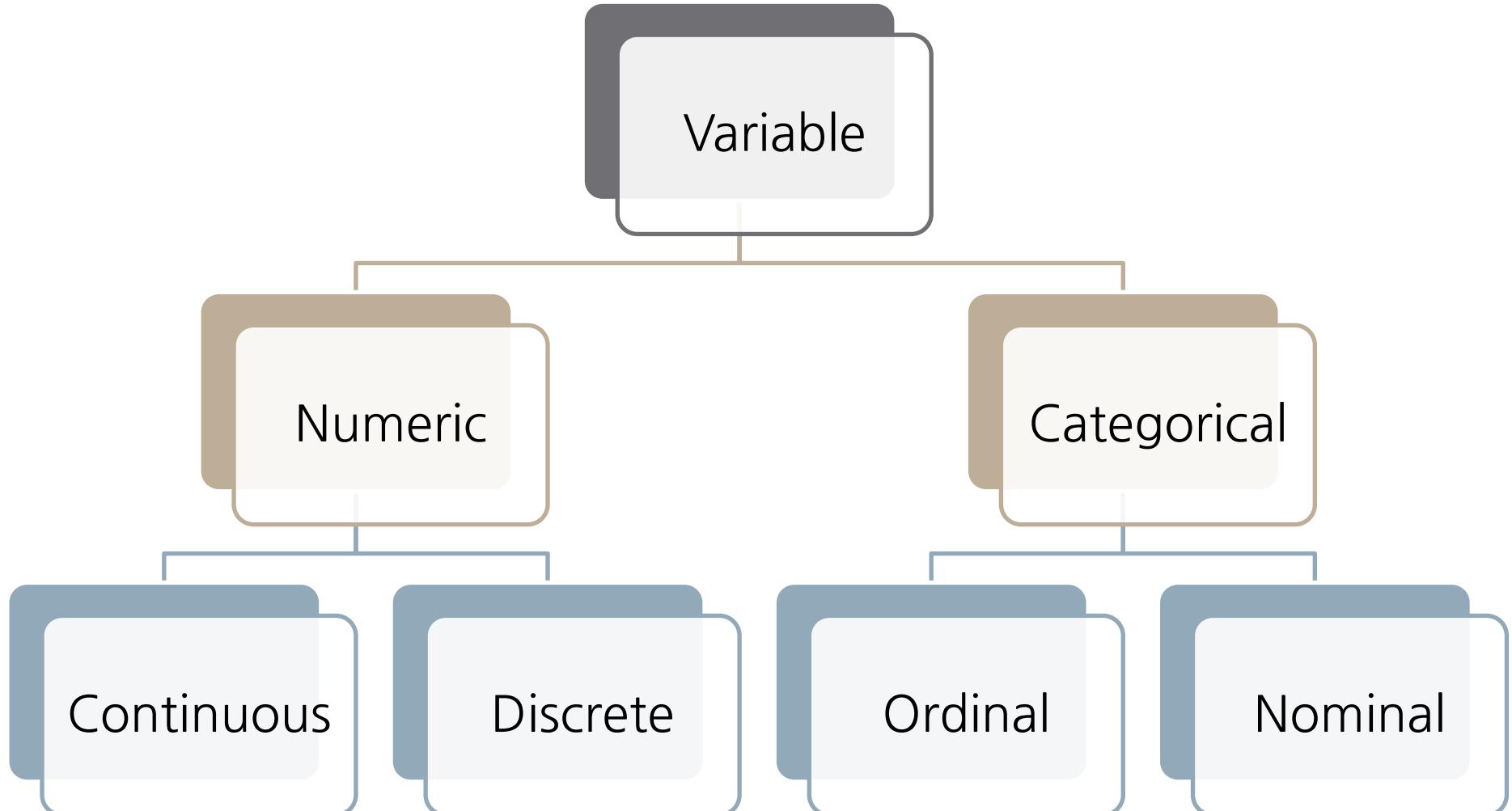
□ Structured

- Values of variable reside in a fixed field
- Examples
 - Numeric
 - Date
 - Restricted terms: (male, female), (Mr., Ms., Mrs.)
 - Address

□ Unstructured

- Values of variable do not reside in a fixed field
- Examples
 - Documents
 - Webpages
 - Images
 - Videos

Structured Data: Types of Variables



Structured Data: Types of Variables

□ Numeric (Quantitative)

- A broad category that includes any variable that can be counted, or has a numerical

Continuous

- A variable with infinite number of values
- Example
 - Many numeric variables: temperature, weight, height, pressure and etc.

Discrete

- A variable that can only take on a certain number of values or have a countable number of values between any two values
- Example
 - The number of cars in a parking lot
 - the number of flaws or defects

Structured Data: Types of Variables

□ Categorical

- A variable that contains a finite number of categories or distinct groups

□ Nominal

- A Variable that has two or more categories, but there is no intrinsic ordering to the categories.

■ Example

- (Male, Female), (Class 1, Class 2, Class 3), (Red, Yellow, Green)

□ Ordinal

- Similar to a nominal variable, but the difference between the two is that there is a clear ordering of the variables.

■ Example

- Score: A+, A, A-, B+, B, B-, C+, C, C-, D, F
 - Size: S, M, L, XL, XXL

Example: The Input to a Data Mining

- Example of data set

num-of-doors	body-style	wheel-base	length	make
2	convertible	88.6	168.8	Audi
2	convertible	88.6	168.8	BMW
2	hatchback	94.5	171.2	Chevrolet
4	sedan	99.8	176.6	BMW
4	sedan	99.4	176.6	Audi
2	sedan	99.8	177.3	Audi
4	wagon	105.8	192.7	Chevrolet

Types:

Discrete

Nominal

Continuous

Continuous

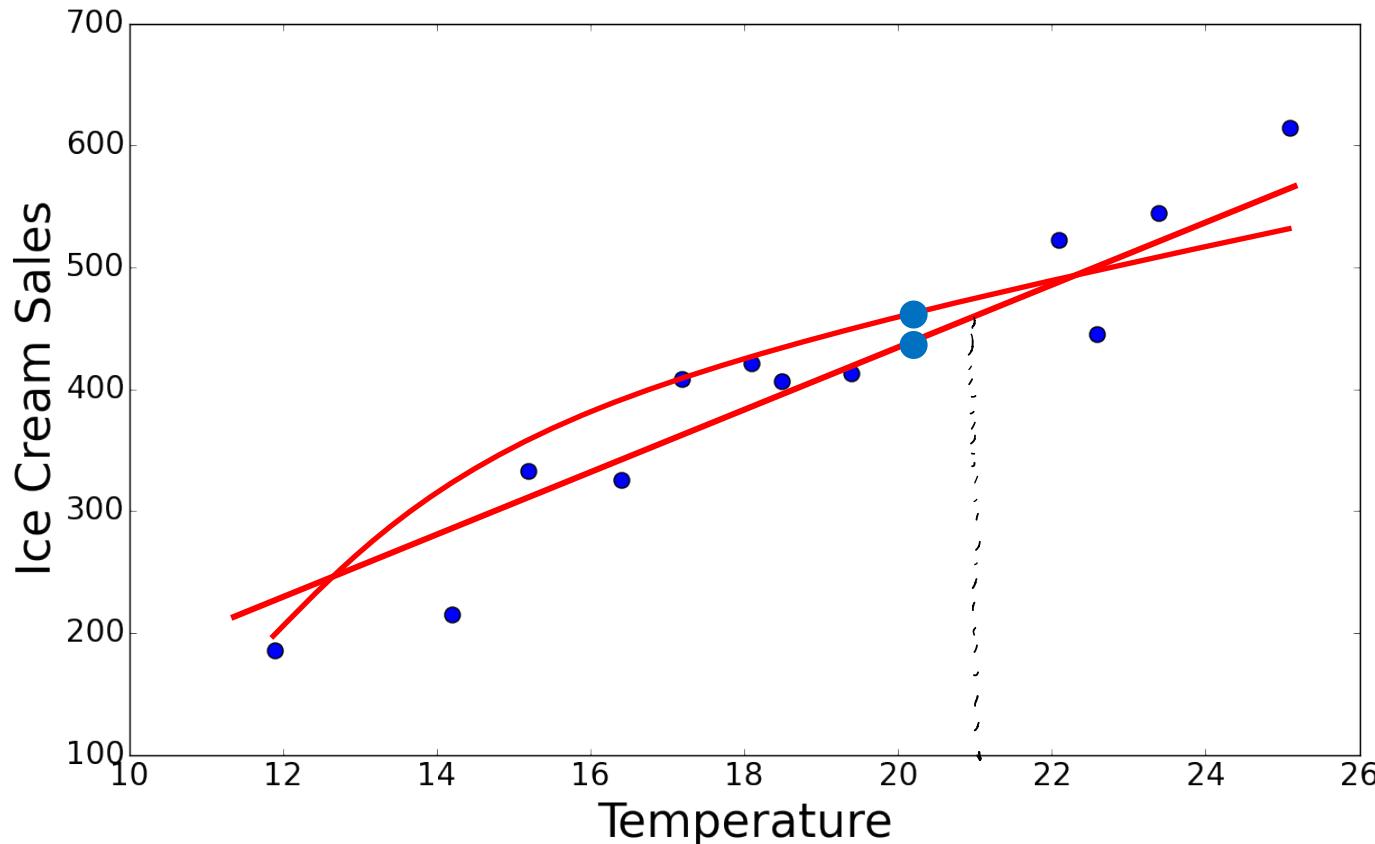
Nominal

Supervised Learning: Regression

eff

- Temperature vs. Ice Cream Sales
 - How about 21°C?

for continuous

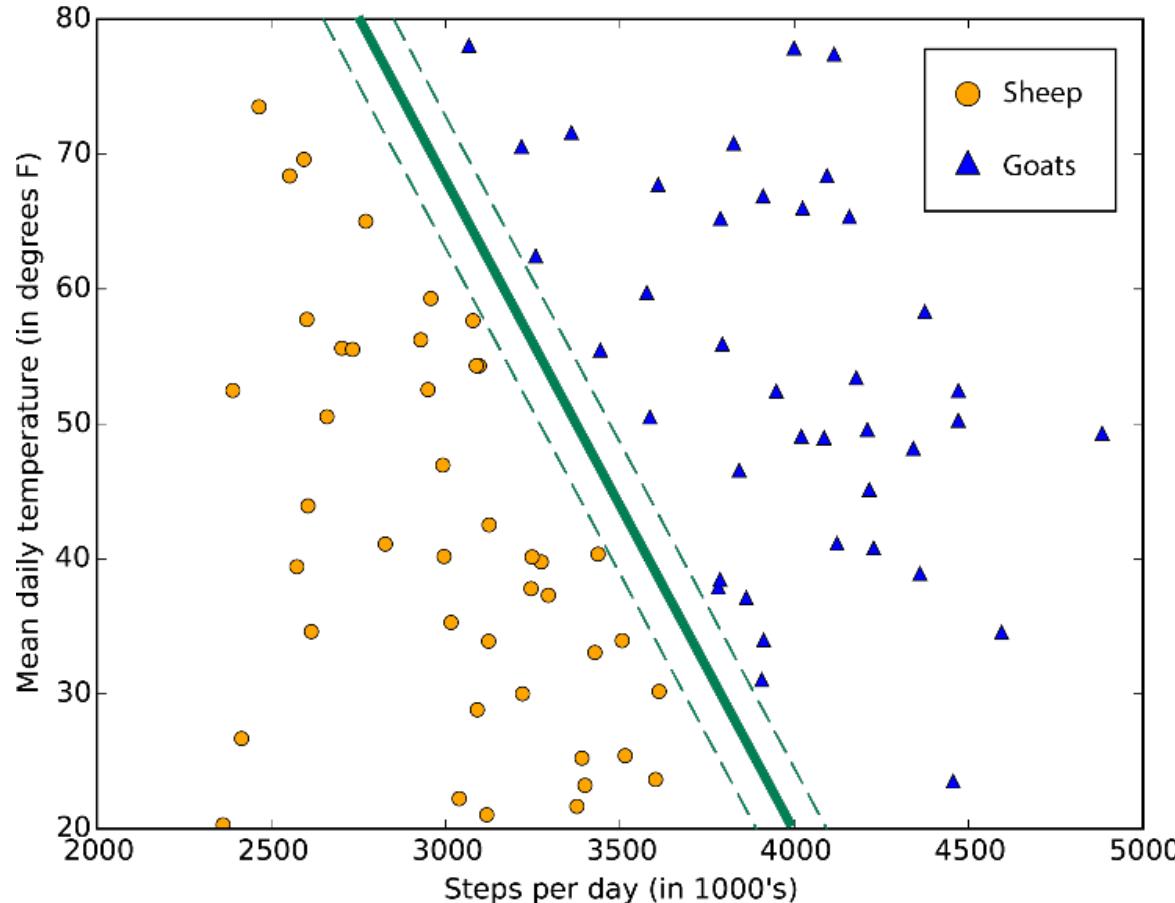


Supervised Learning: Classification

1d2
11.

- Which one is a sheep?

for discrete value · yes/no



Question

- Suppose you are working on weather prediction, and you would like to predict whether or not it will be raining at 5pm tomorrow. Y/N
You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?
 - ① Regression
 - ② Classification
- Suppose you are working on stock market prediction, and you would like to predict the price of the specific stock tomorrow (measured in dollars).
You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?
 - ① Regression
 - ② Classification

Overfitting vs. Underfitting

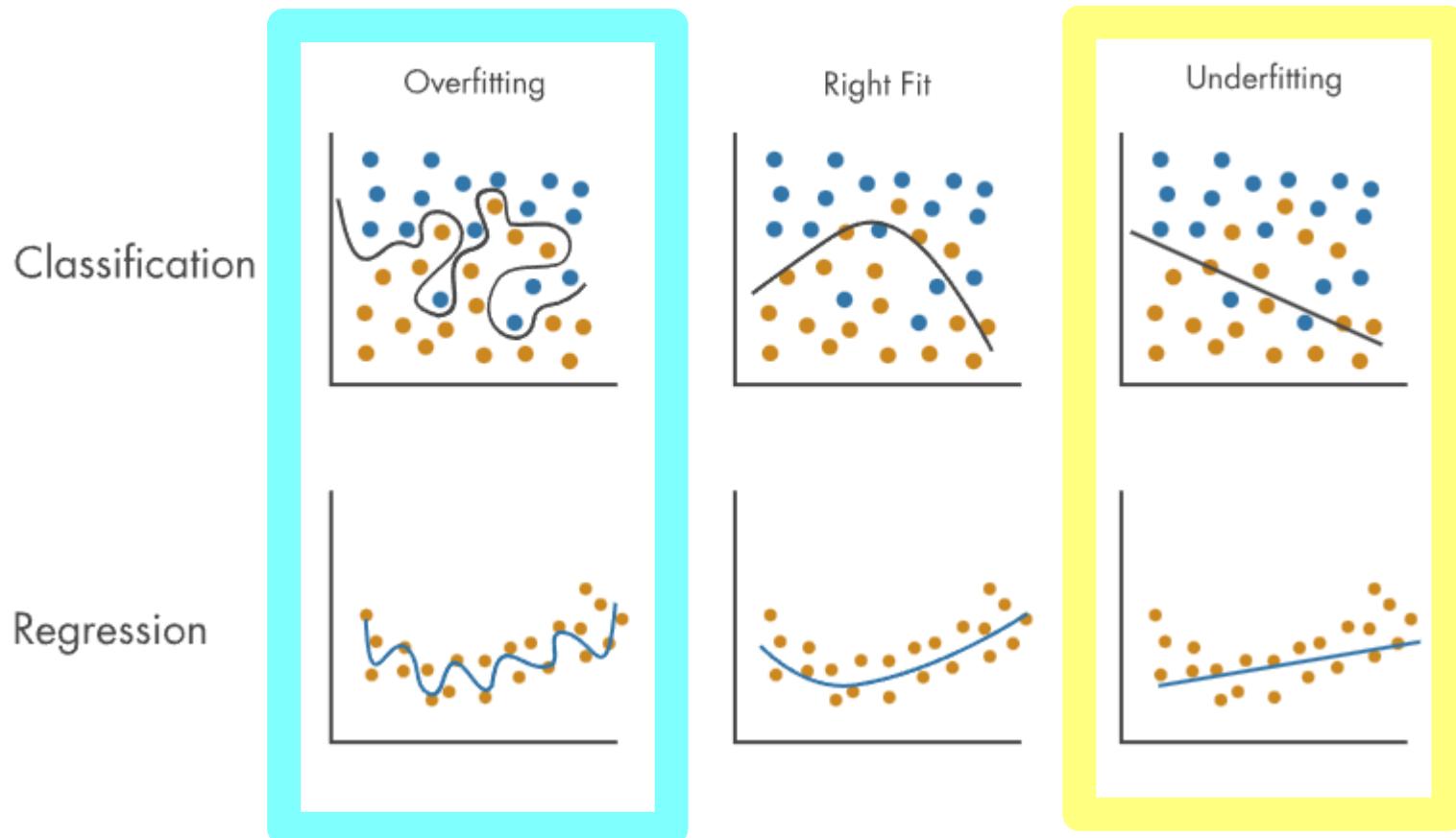
□ Overfitting

- Overfitting is a machine learning problem that occurs when a model is too closely aligned to training data, causing it to perform poorly on new data.
- How it happens
 - The model is too complex
 - The training data is too small or contains irrelevant information
 - The model memorizes subtle patterns in the training data
- Why it's a problem
 - An overfit model can't generalize well to new data
 - It can give inaccurate predictions
 - It can't perform well for all types of new data

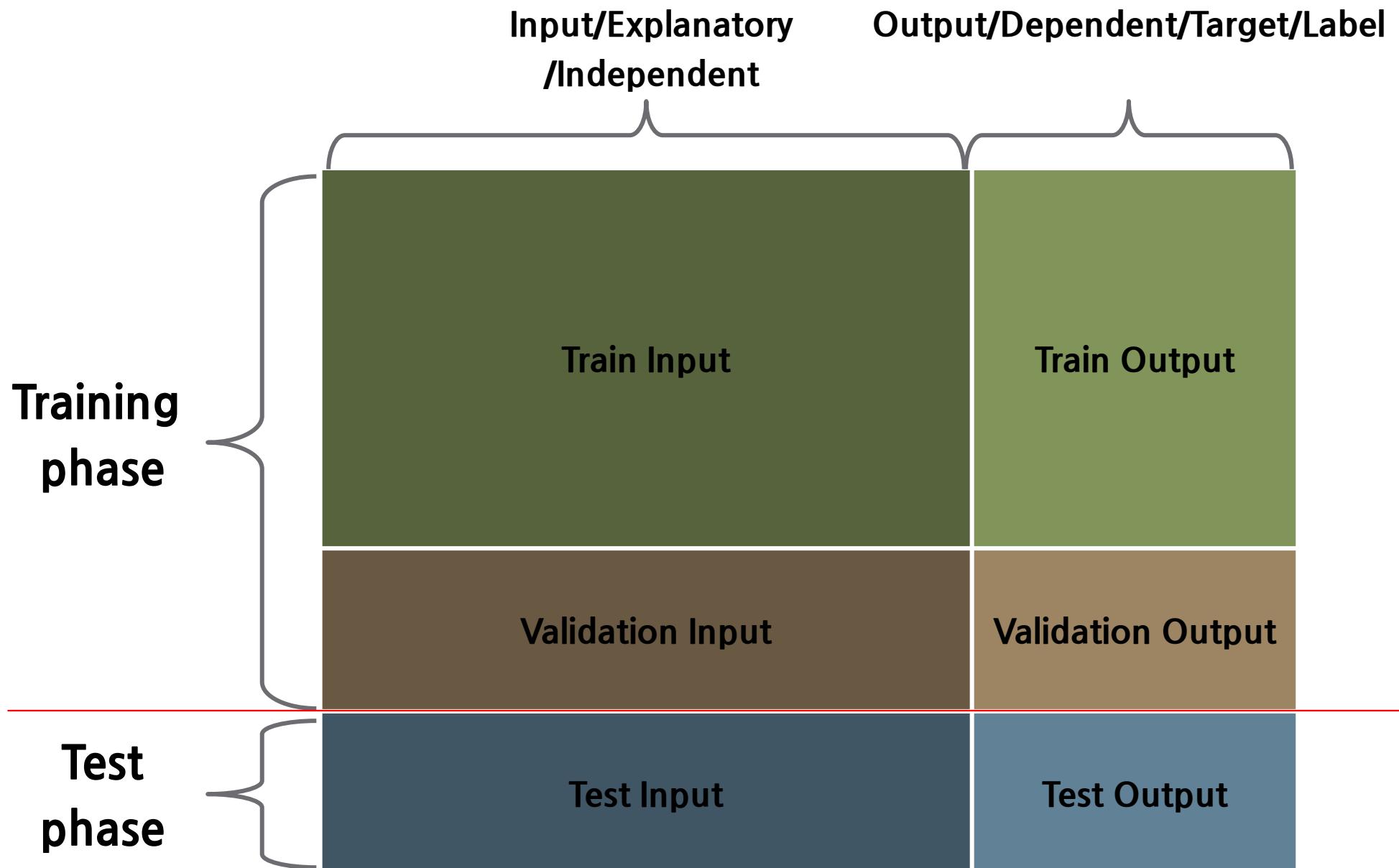
□ Underfitting

- Underfitting occurs when a machine learning model is too simple to capture the underlying patterns in data.

Overfitting vs. Underfitting



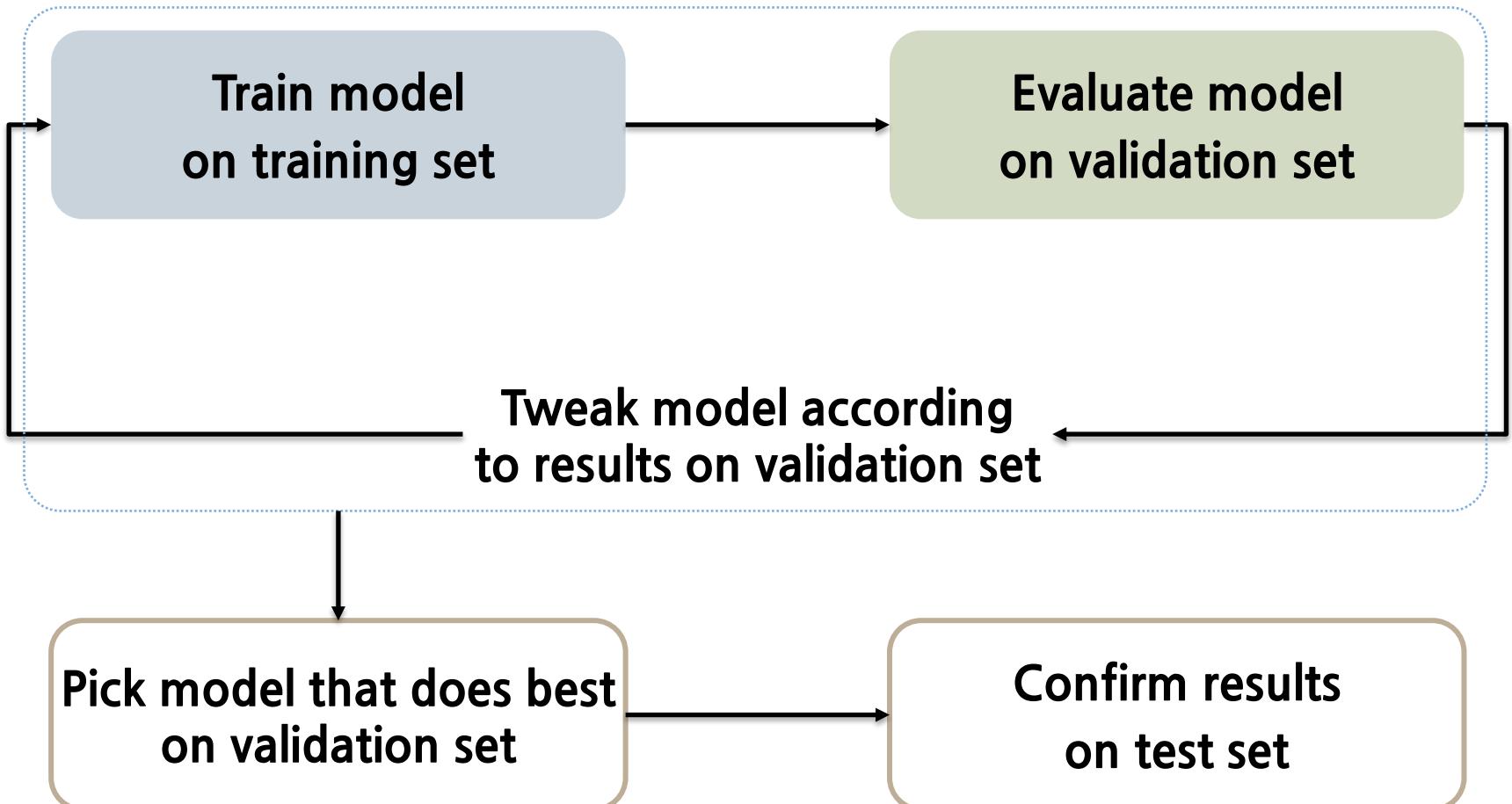
Data Partition



Data Partition

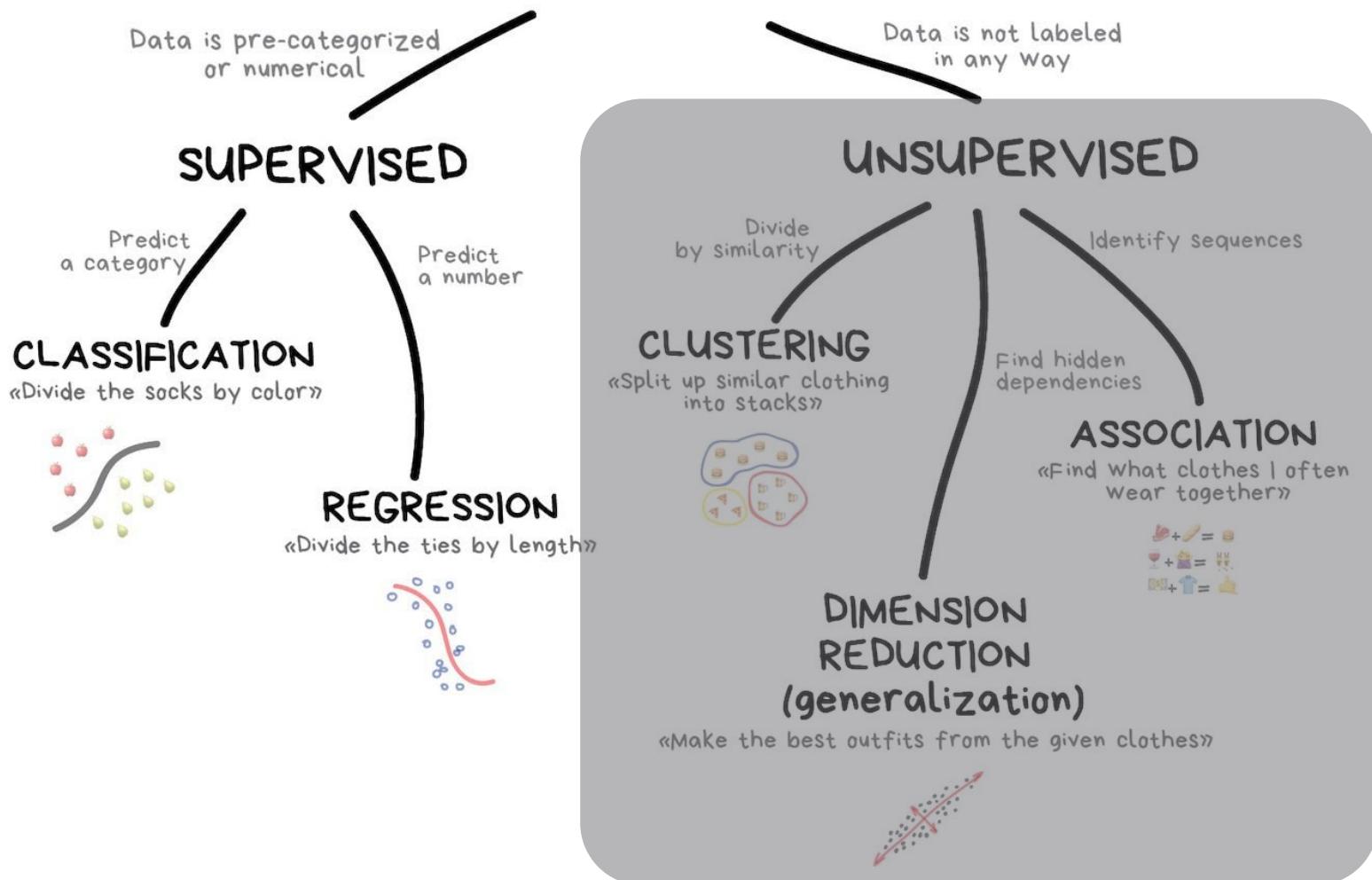
- Training set
 - ▣ Purpose: The training set is used to train the model. It contains the labeled examples the model will learn from. During the training phase, the model's parameters are adjusted based on the data in the training set.
- Validation set
 - ▣ Purpose: The validation set is used to tune hyperparameters and evaluate the model during training. It helps in selecting the best version of the model.
- Test set
 - ▣ Purpose: The test set is used to evaluate the model's final performance after training and validation. This set simulates new, unseen data, giving an unbiased estimate of how the model will perform in a real-world scenario.

Process of Supervised Learning with Partitioned Data



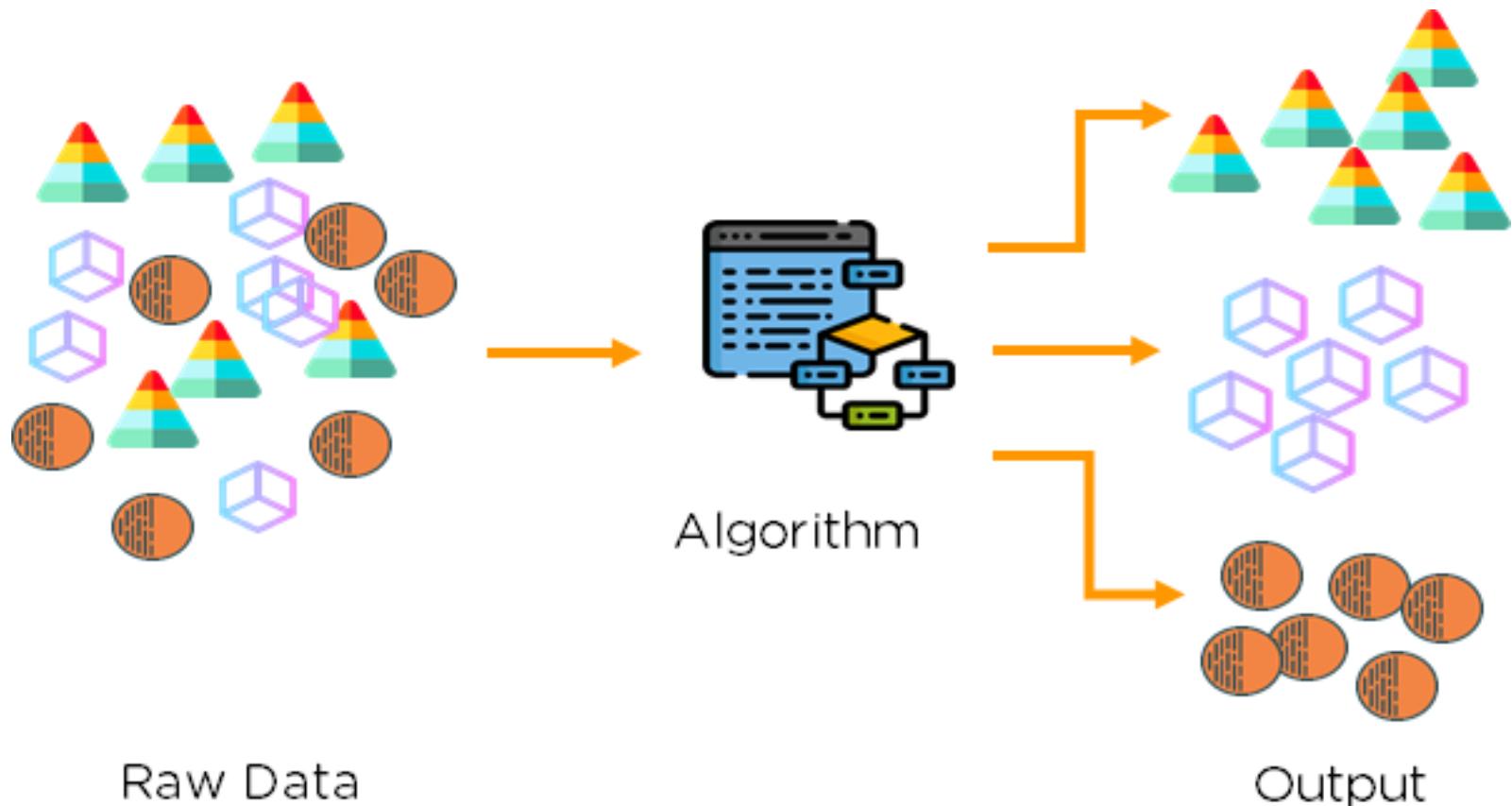
Topics Covered in This Class

CLASSICAL MACHINE LEARNING



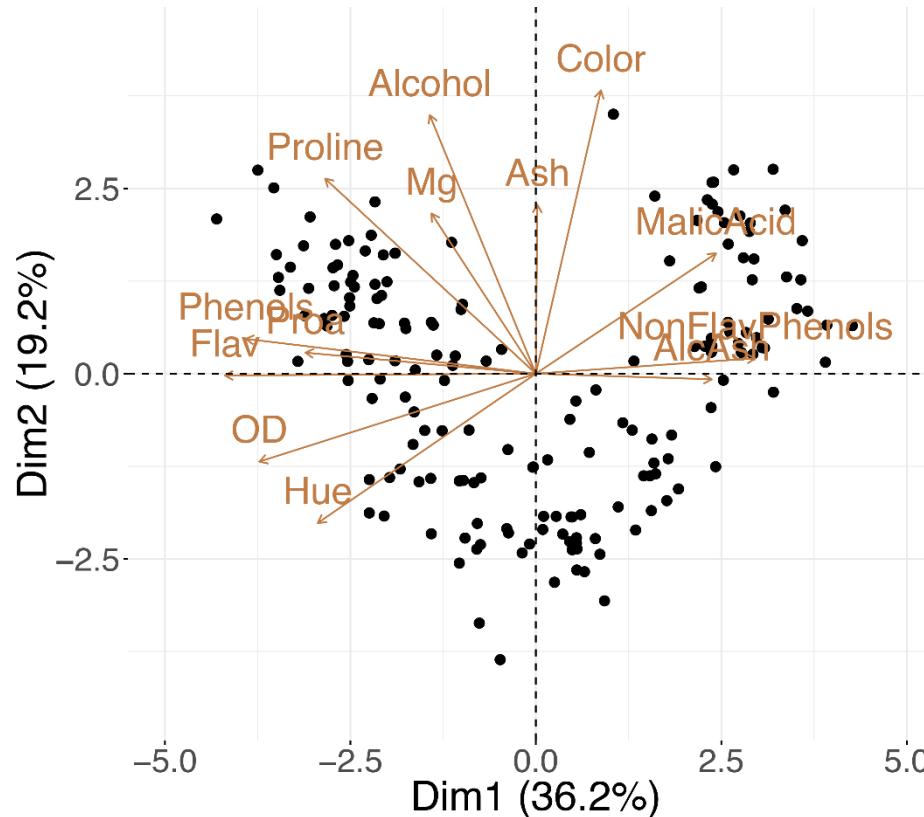
Unsupervised Learning: Clustering

- Grouping data points
 - How to determine which **group** does each data belongs to?



Unsupervised Learning: Dimensionality Reduction

- Dimensionality reduction
 - The process of reducing the number of random variables under consideration by obtaining a set of principal variables
 - High dimension → Low dimension



Unsupervised Learning: Association Rule Mining

similar with Search algorithm

- Find useful information from transactions

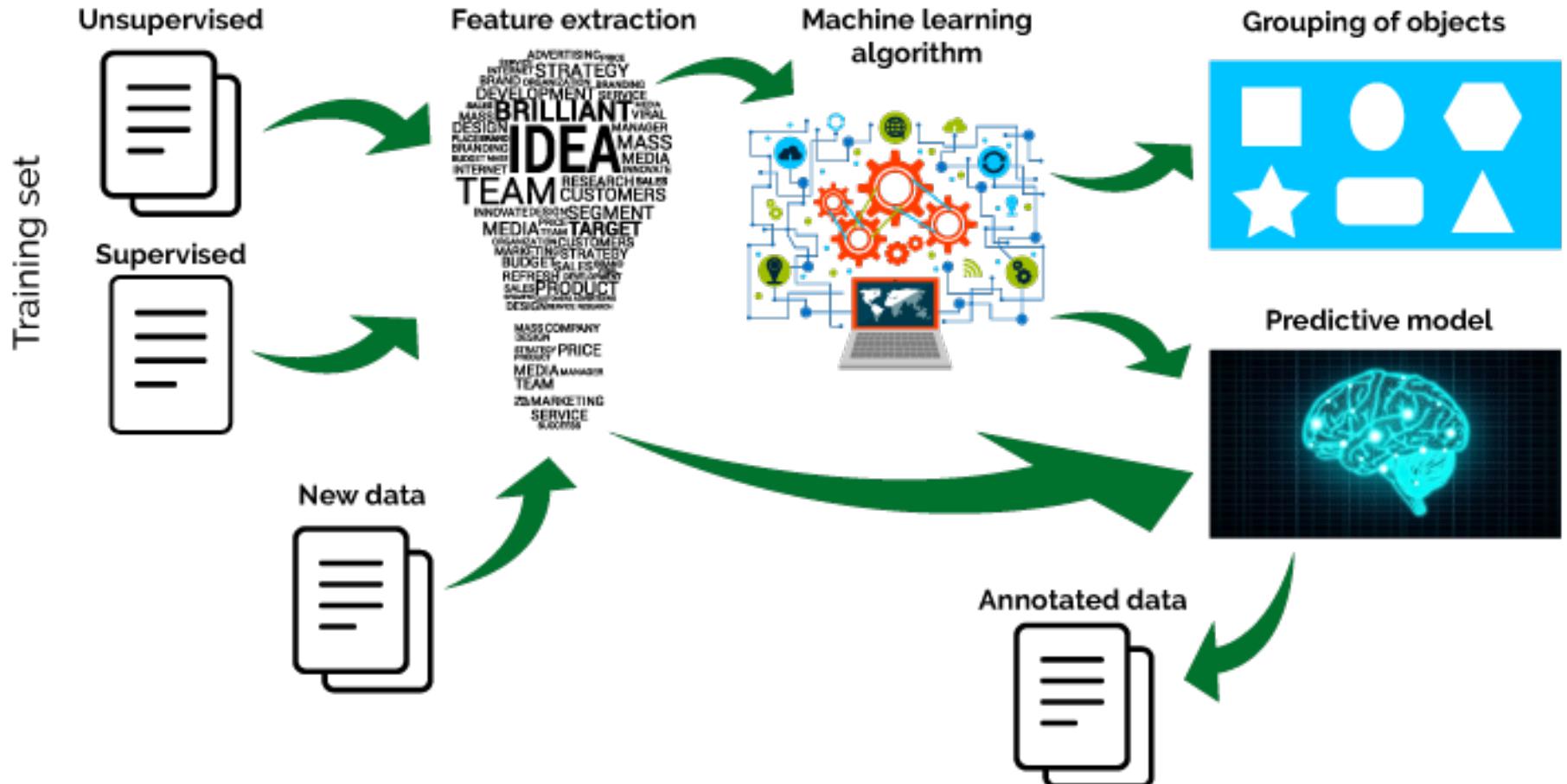
Datetime	Customer	Items
2015-07-15 14:03	1	orange juice, banana
2015-07-15 16:20	2	orange juice, milk
2015-07-16 10:14	3	detergent, banana, orange juice
2015-07-25 19:34	2	milk, bread, soda
2015-07-29 09:41	4	detergent, window cleaner
2015-08-01 20:55	1	bread, milk

- One of useful information is information like “If item A then item B”
 - This information is called association rule
- Find pair of items that are more likely to be purchased together based on transactions

Question

- Of the following examples, which would you address using an unsupervised learning algorithm? (Find all that apply.)
 - ① Given email labeled as spam/not spam, learn a spam filter. *supervised classify*
 - ② Given a set of news articles found on the web, group them into set of articles about the same story. *unsupervised clustering*
 - ③ Given a database of customer data, automatically discover market segments and group customers into different market segments. *unsupervised clustering*
 - ④ Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not. *supervised classification*

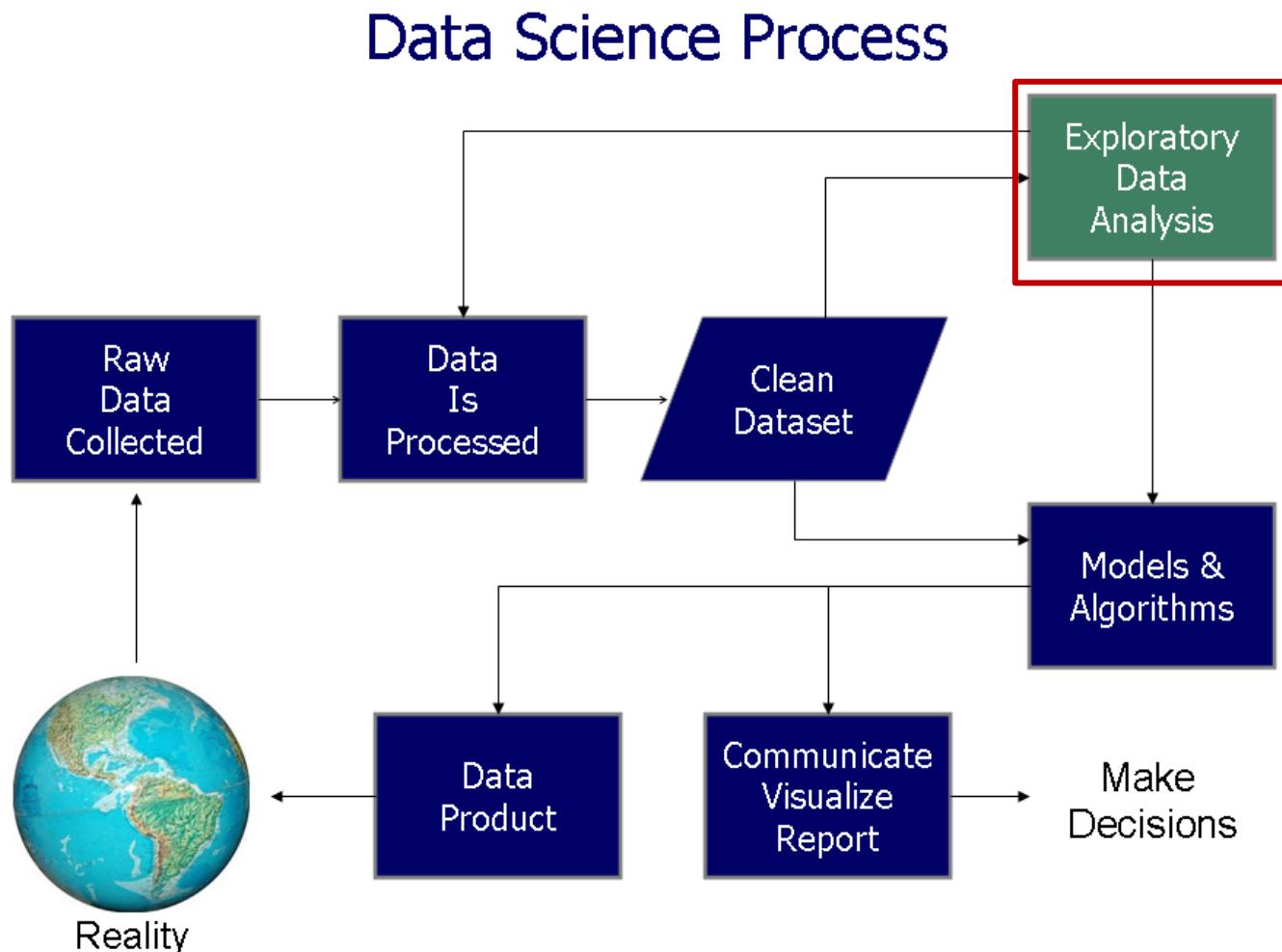
Overall Description



EXPLANATORY DATA ANALYSIS

Week03

Data Science Process



Explanatory Data Analysis

What is Exploratory Data Analysis (EDA)?

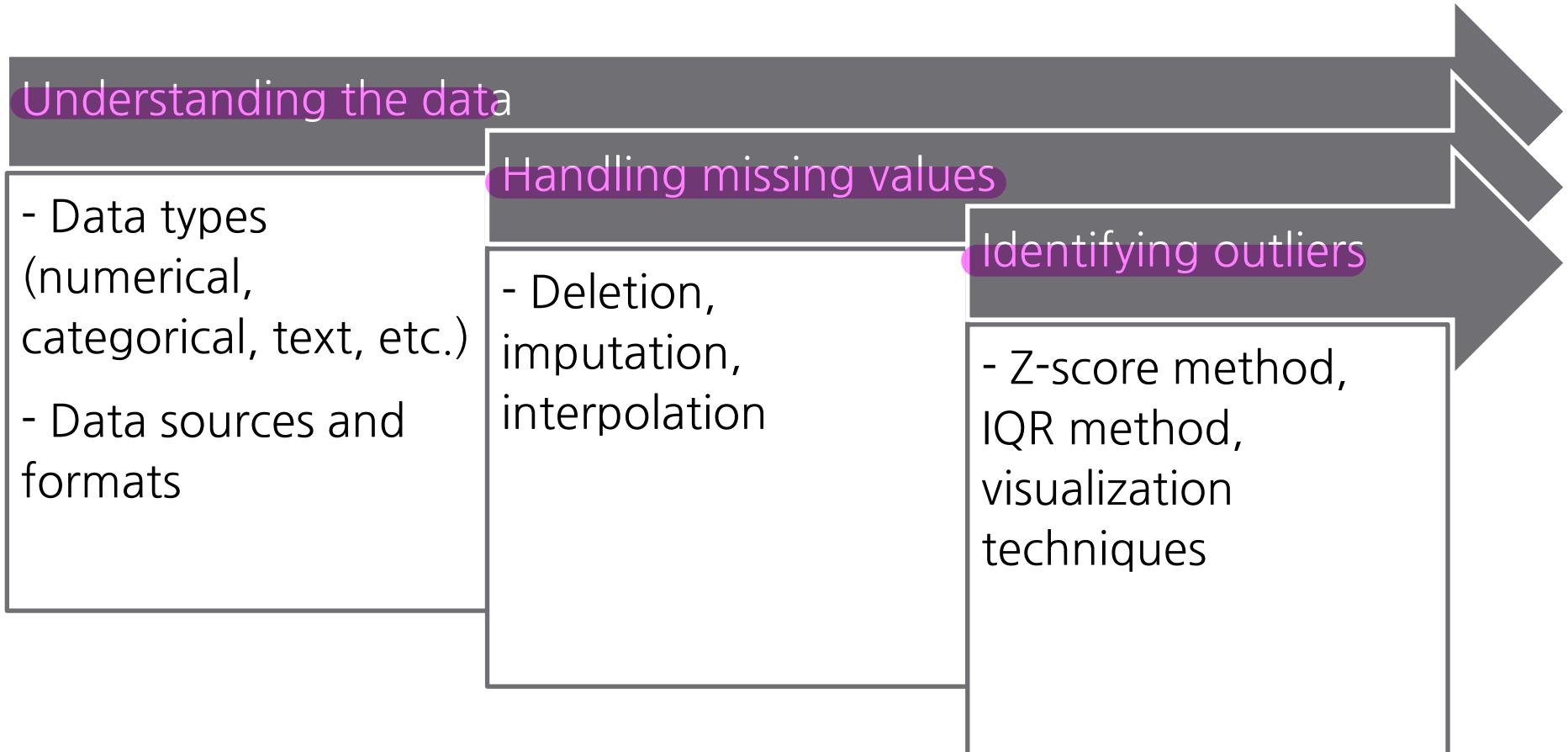
- Definition
 - ▣ EDA is an approach to analyzing data sets to summarize their main characteristics, often with visual methods
- Purpose
 - ▣ Identify patterns, detect anomalies, test hypotheses, and check assumptions with summary statistics and graphical representations
- Key Techniques
 - ▣ Data visualization, summary statistics, handling missing values, detecting outliers

Importance of EDA

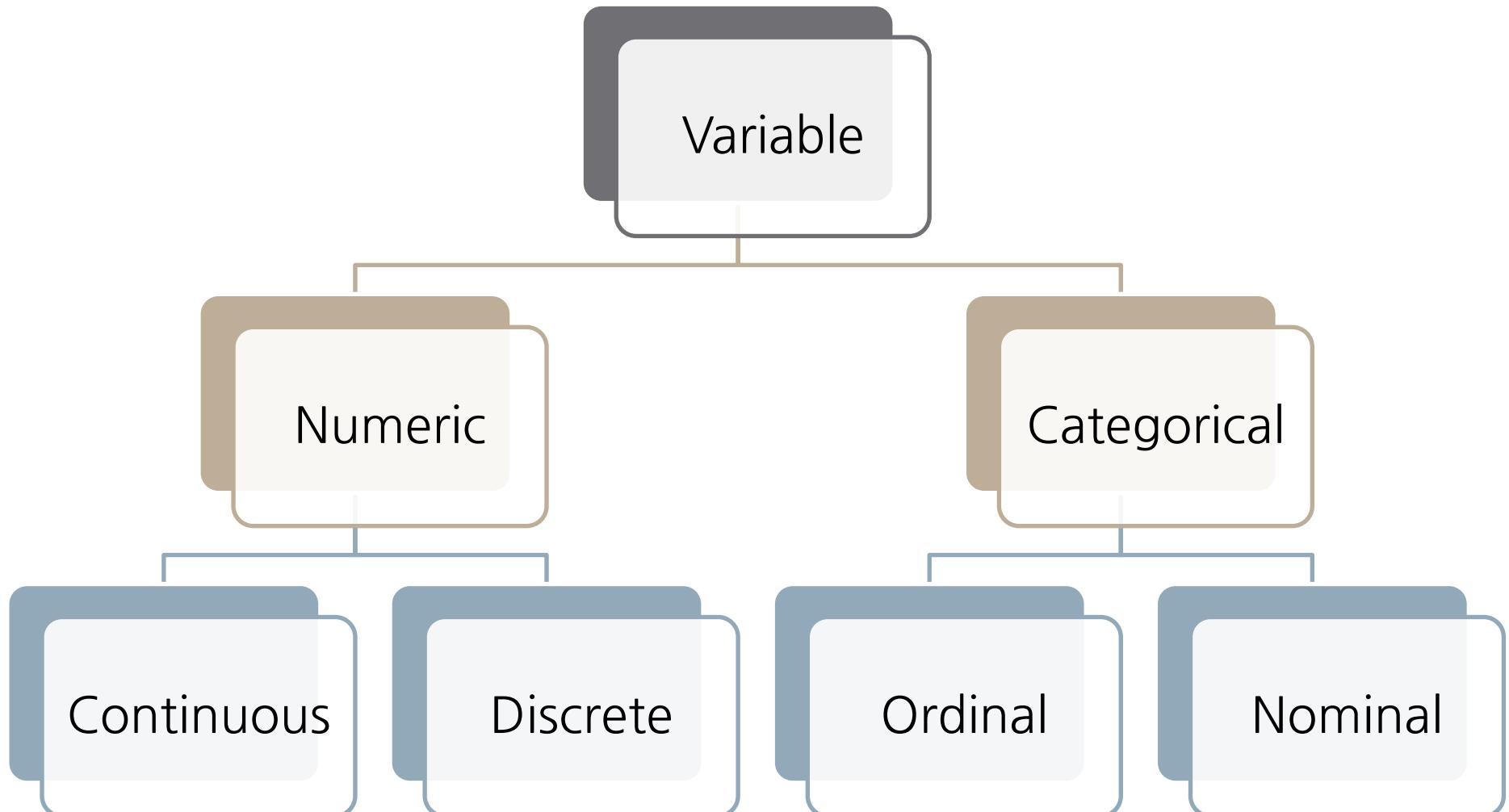
- Understanding the Data
 - Identify data distributions and relationships
 - Detect missing values and anomalies
- Data Preprocessing
 - Helps in feature selection and engineering
 - Assists in choosing appropriate machine learning models
- Insights & Decision Making
 - Provides actionable insights before model building
 - Supports better decision-making in data-driven projects.

Data Transformation?

Steps in EDA



Understanding Data: Types of Variables



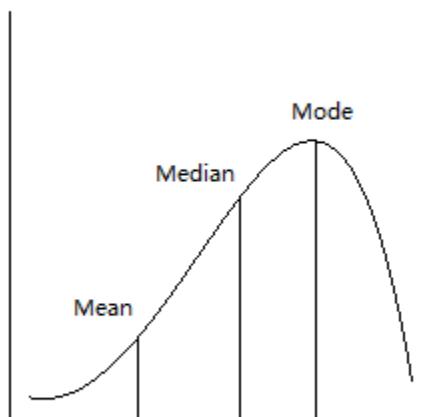
Understanding Data: Summary Statistics

- Measures of central tendency

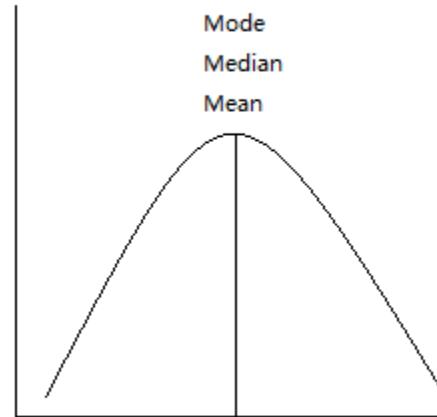
- Mean: The average value of a dataset

$$\mu = \frac{\sum_i x_i}{n}$$

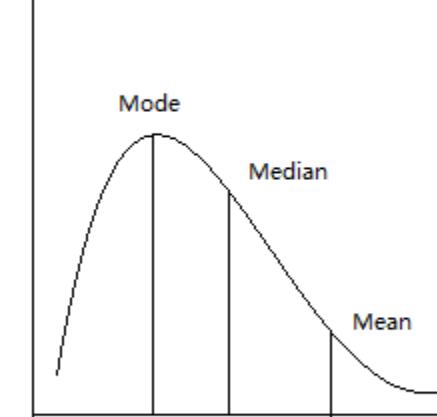
- Median: The middle value in a sorted dataset. If the number of observations (n) is odd, the median is the middle value. If n is even, the median is the average of the two middle values
 - Mode: The most frequently occurring value in the dataset



Left skew



Normal Distribution



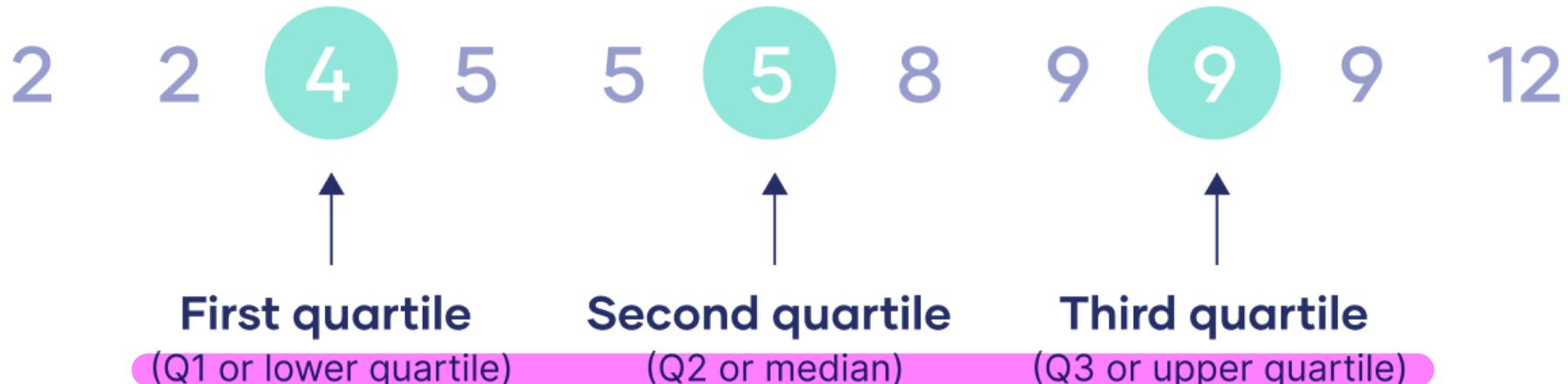
Right skew

Understanding Data: Summary Statistics

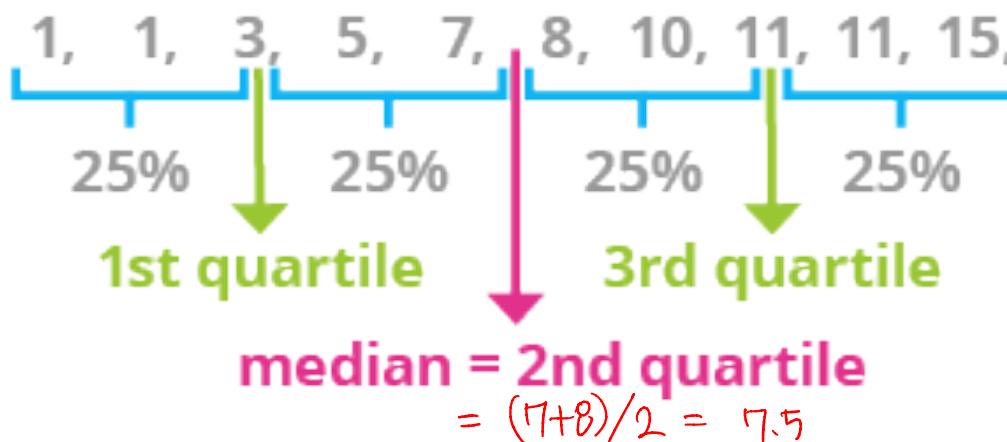
- Measures of spread
 - ▣ Variance: Measures the average squared deviation from the mean
 - Population
 - Sample
$$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n}$$
 - ▣ Standard Deviation: The square root of the variance, indicating dispersion
 - Population
 - Sample
$$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{n}}$$
 - ▣ Range: The difference between the maximum and minimum values in the dataset
$$Range = \max(x) - \min(x)$$
 - ▣ Inter-quartile range (IQR): Measures the spread of the middle 50% of data
$$IQR = Q_3 - Q_1$$

Understanding Data: Summary Statistics

odd case



even case



Understanding Data: Summary Statistics

- Skewness *entz*

- Skewness measures the symmetry of a distribution

- Population

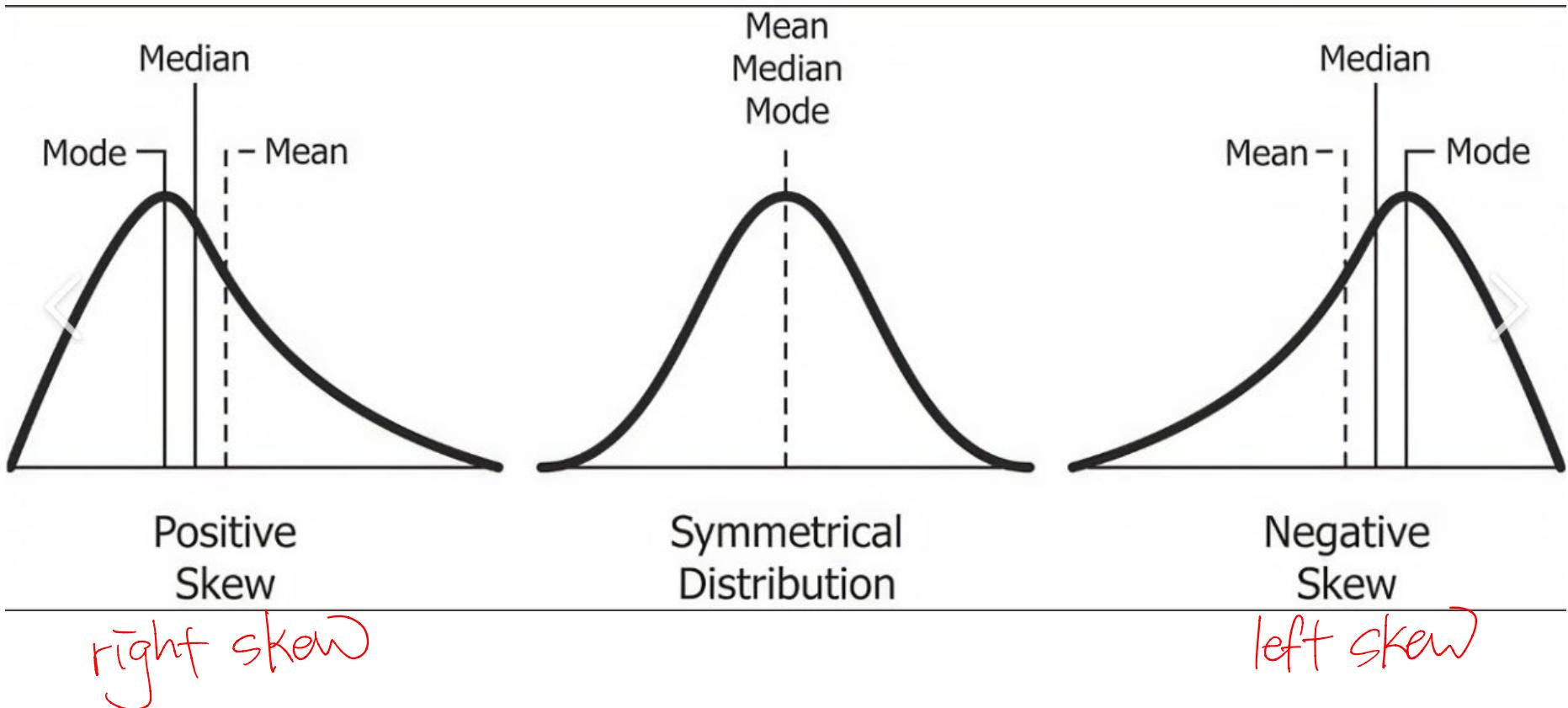
$$\gamma_1 = \mathbb{E} \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right]$$

- Sample

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}{\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$

- Positive skewness: The tail extends to the right (more values are on the left side of the mean).
 - Negative skewness: The tail extends to the left (more values are on the right side of the mean)
 - A skewness value of zero indicates a symmetric distribution

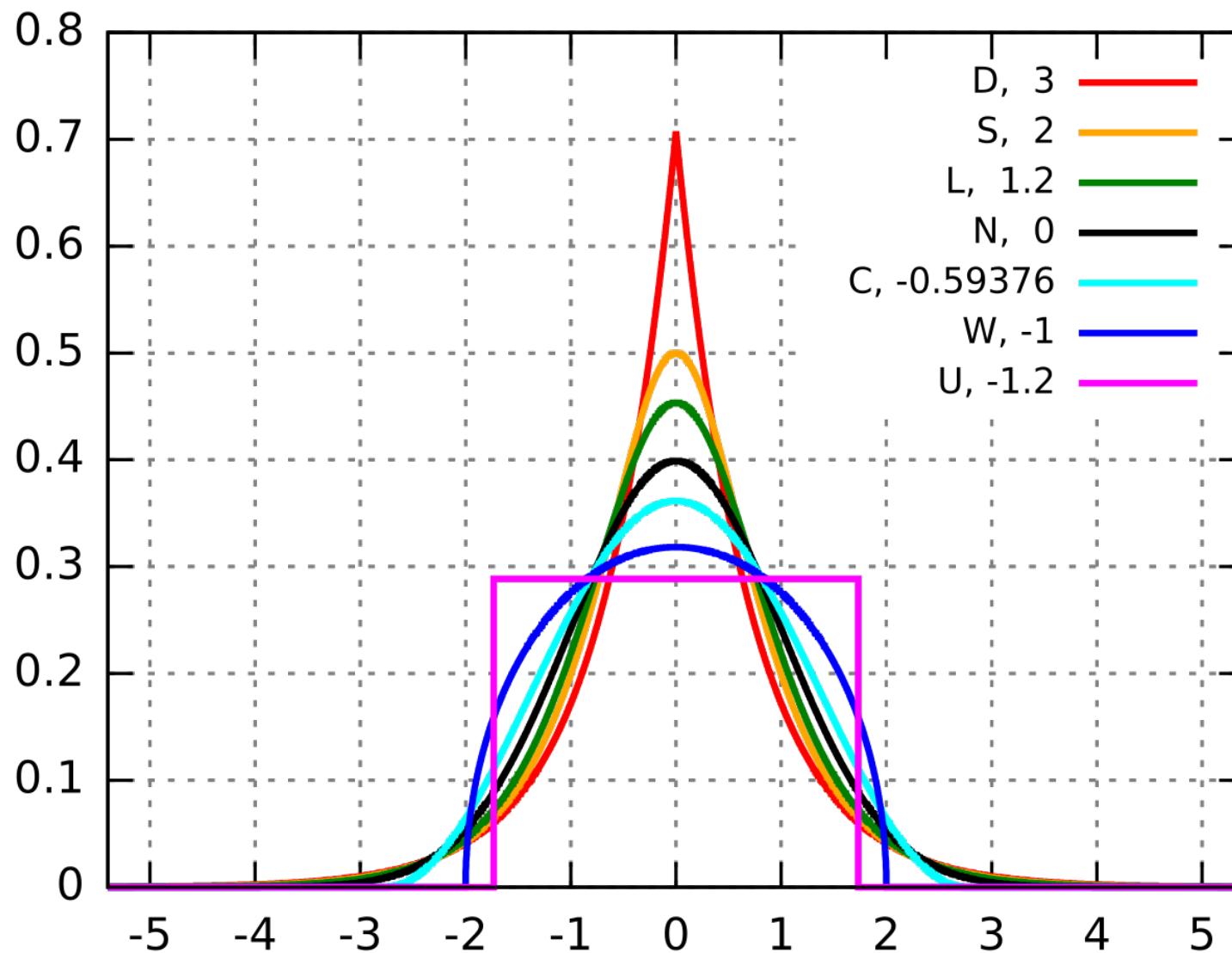
Understanding Data: Summary Statistics



Understanding Data: Summary Statistics

- Kurtosis *K_U*
 - ▣ Kurtosis measures the “tailedness” or “peakedness” of a distribution compared to a normal distribution
 - Population
 - Sample
 - Excess kurtosis (Fisher kurtosis)
$$Kurt[X] = \mathbb{E} \left[\left(\frac{X - \mu}{\sigma} \right)^4 \right]$$
$$\kappa = \frac{m_4}{s^4}$$
 - Positive kurtosis (leptokurtic): The distribution has heavier tails and a sharper peak than a normal distribution
 - Negative kurtosis (platykurtic): The distribution has lighter tails and a flatter peak than a normal distribution
 - A kurtosis value of 3 (or 0 for excess kurtosis) indicates a distribution similar to a normal distribution (mesokurtic)

Understanding Data: Summary Statistics



Understanding Data: Summary Statistics

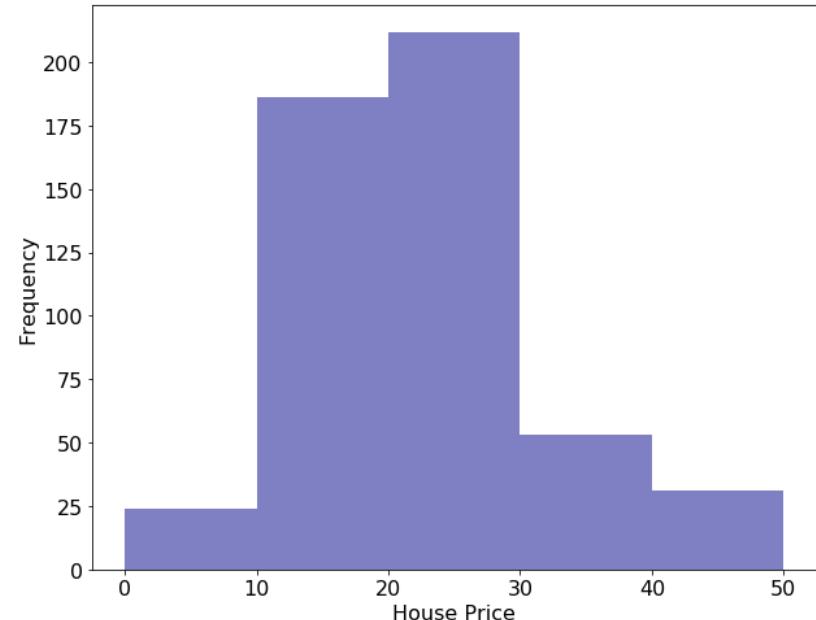
- Data distribution
 - ▣ Histograms: A graphical representation of the distribution of numerical data. It divides the dataset into bins and counts the number of observations within each bin, showing how data values are distributed across different ranges

24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9,
27.1, 16.5, 18.9, 15.0, 18.9, 21.7, 20.4,
18.2, 19.9, 23.1, 17.5, 20.2, 18.2, 13.6,
19.6, 15.2, 14.5, 15.6, 13.9, 16.6, 14.8
... ...



Bin	Count
0 to 10	24
10 to 20	186
20 to 30	212
30 to 40	53
40 to 50	31

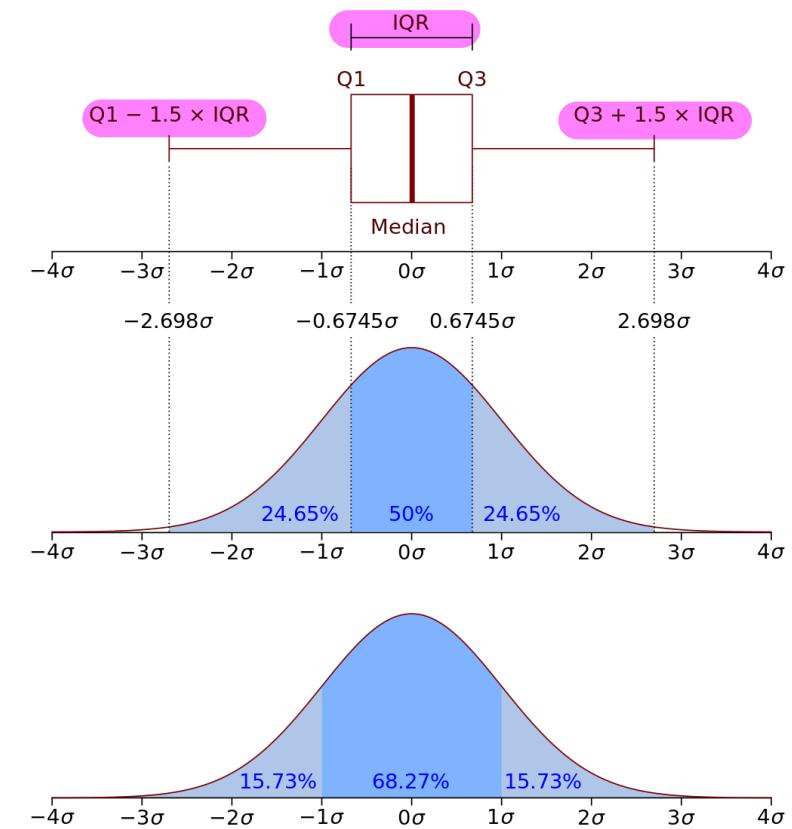
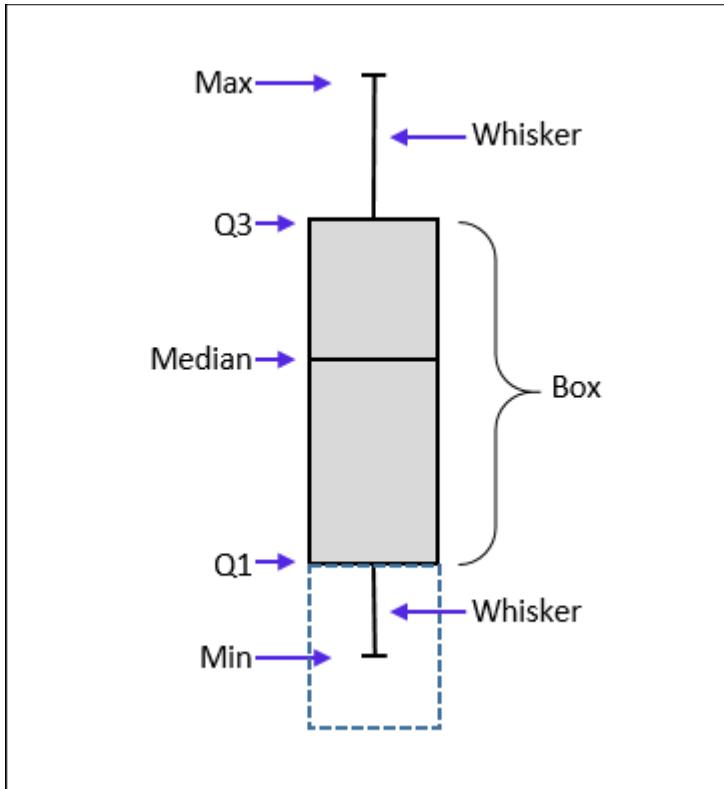
Histogram



Understanding Data: Summary Statistics

□ Data distribution

- Box plot: A visualization that summarizes the distribution of a dataset using five key summary statistics: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. It helps identify outliers and the spread of the data



Understanding Data: Summary Statistics

□ Correlations between variables

- ▣ Pearson correlation coefficient: Measures the linear relationship between two continuous variables

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

- ▣ Spearman rank correlation: Measures the strength and direction of the monotonic relationship between two variables \Rightarrow categorical variables (ordinal)

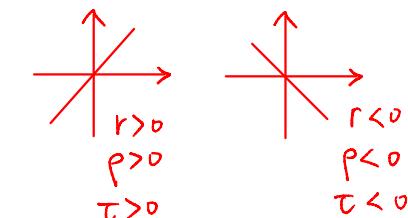
$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

- d_i : the difference between the ranks of corresponding values

- ▣ Kendall's tau: Measures the ordinal association between two variables

$$\tau = \frac{C - D}{\frac{1}{2} n(n - 1)}$$

- C : the number of concordant pairs
- D : the number of discordant pairs



Understanding Data: Summary Statistics

□ Spearman rank correlation

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

$n=5$

X	Y
10	40
20	30
30	50
40	20
50	60

Calculate ranks
and
differences in
ranks



X	<u>$Rank(X)$</u>	Y	<u>$Rank(Y)$</u>	d_i	d_i^2
10	1	40	3	-2	4
20	2	30	2	0	0
30	3	50	4	-1	1
40	4	20	1	3	9
50	5	60	5	0	0

$$1 - \frac{6 \times (4 + 0 + 1 + 9 + 0)}{5(5^2 - 1)} = 1 - \frac{6 \times 14}{5 \times 24} = 0.3$$

Understanding Data: Summary Statistics

- Kendall's tau

$$\tau = \frac{C - D}{\frac{1}{2} n(n-1)} = 0.2$$

$$\# \text{ pairs} = 5C_2 = \frac{5 \cdot 4}{2} = 10$$

X	Y
10	40
20	30
30	50
40	20
50	60



Pairs	Relationship	Pairs	Relationship
(10,40) & (20,30)	Discordant	(20,30) & (40,20)	Discordant
(10,40) & (30,50)	Concordant	(20,30) & (50,60)	Concordant
(10,40) & (40,20)	Discordant	(30,50) & (40,20)	Discordant
(10,40) & (50,60)	Concordant	(30,50) & (50,60)	Concordant
(20,30) & (30,50)	Concordant	(40,20) & (50,60)	Concordant

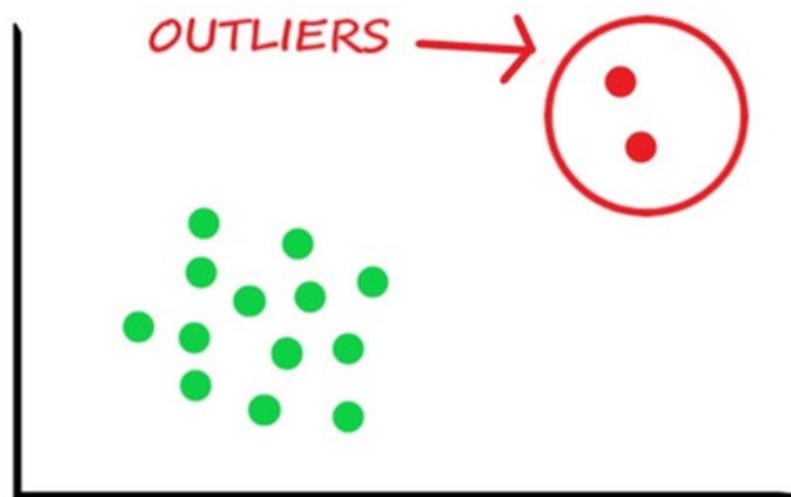
Handling Missing Values

- Methods to handle missing values
 - ▣ **Deletion**: Remove missing data points if they are few and do not significantly impact the dataset
 - ▣ Mean/Median/Mode **Imputation**: Replace missing values with statistical measures such as mean, median or mode
 - ▣ k -Nearest Neighbors (k NN) imputation: Use similar data points to estimate missing values
 - ▣ Regression imputation: Predict missing values based on relationships between variables
 - ▣ Multiple Imputation: Generate multiple datasets with estimated values and combine results

Identifying outliers

Outliers

- Outliers are data points that are significantly different from the rest of the data
- They can be unusually high or low compared to other observations in a dataset



Identifying outliers

- Characteristics of Outliers
 - ▣ Extreme values: Outliers fall far outside the normal range of data
 - ▣ Unexpected behavior: They may arise from errors, but they can also be valid data points
 - ▣ Impact on analysis: They can distort statistical measures such as mean, standard deviation, and regression models
- Types of outliers
 - ▣ Point outliers: A single data point that deviates significantly
 - Example: A person with an extremely high income compared to a group
 - ▣ Contextual outliers: Data points that are considered outliers only in certain contexts or subgroups
 - Example: A summer day temperature of 30°C in a region with a typical summer range of 20-25°C.
 - ▣ Collective outliers: A collection of data points that, together, show unusual behavior.
 - Example: A set of consecutive days with temperatures unusually high for a region.

Identifying outliers

- Why are outliers important?
 - ▣ Affect mean and standard deviation
 - Outliers can skew the mean and inflate the standard deviation, leading to misleading conclusions
 - ▣ Modeling issues
 - Outliers can affect the performance of machine learning models (e.g., linear regression) by making predictions less accurate
 - ▣ Data Integrity
 - Outliers may indicate data entry errors, measurement mistakes, or special cases that need further investigation

Identifying outliers

- Importance of removing outliers
 - ▣ Improved accuracy
 - Removing outliers can lead to more accurate models, ensuring that predictions are not disproportionately influenced by extreme values
 - ▣ Better data representation
 - The central tendency of the data (mean, median) becomes more representative of the general population
 - ▣ Enhanced model performance
 - Machine learning algorithms can perform better when outliers are removed, reducing noise and improving the fit of the model

Identifying outliers

assume normality

- Statistical method using Z-scores to identify outliers
 - ▣ Z-score measures how many standard deviations a data point is away from the mean. It is a standardized way of identifying outliers in normally distributed data
$$Z = \frac{X - \mu}{\sigma}$$
 - ▣ Typically, if a Z-score is greater than 3 or less than -3, the data point is considered an outlier (assuming a normal distribution)
 - This corresponds to values that are more than 3 standard deviations away from the mean

Calculate the
mean and
standard deviation

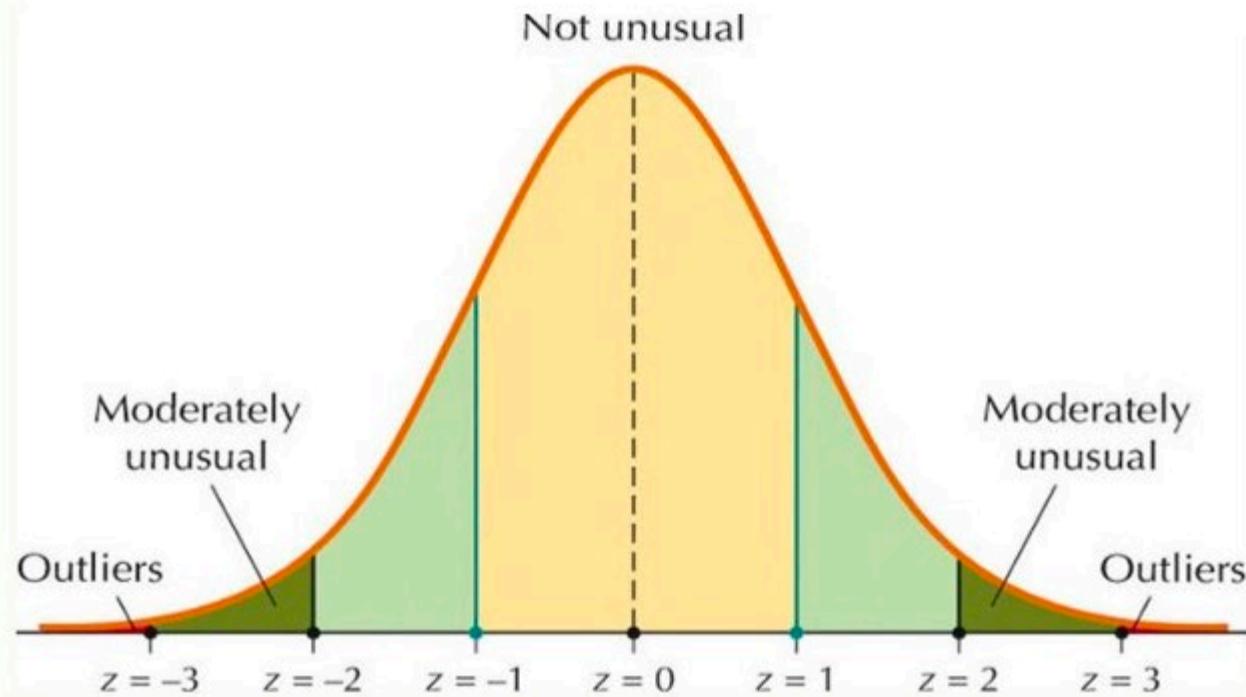
Calculate Z-scores
for each data
point

Set a threshold for
identifying outliers

Remove outliers

Identifying outliers

Detecting Outliers with z-Scores

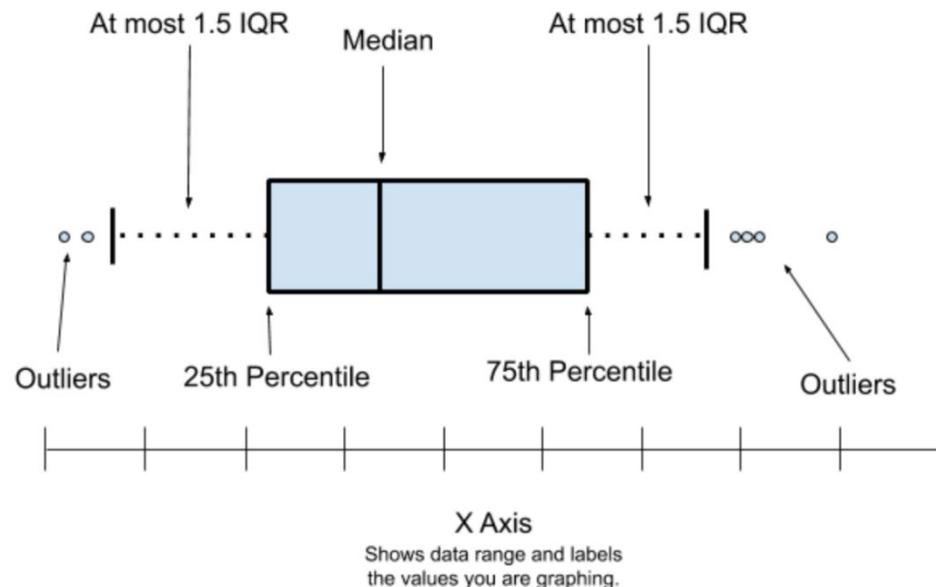


Identifying outliers

No normality assumption

- Statistical method using IQR to identify outliers
 - ▣ Calculate the lower and upper fences for outliers using IQR
 - Lower inner fence
$$\text{Lower inner fence} = Q_1 - 1.5IQR$$
 - Upper inner fence
$$\text{Upper inner fence} = Q_3 + 1.5IQR$$
 - Lower outer fence
$$\text{Lower outer fence} = Q_1 - 3IQR$$
 - Upper outer fence
$$\text{Upper outer fence} = Q_3 + 3IQR$$
 - ▣ Points beyond the inner fences in either direction are mild outliers; points beyond the outer fences in either direction are extreme outliers

Identifying outliers



Visualization: Why is Visualization Important in EDA?

1. Understanding Data Distributions

- ❑ Helps in identifying patterns, trends, and underlying structures in the data.
- ❑ Enables detection of skewness, multimodal distributions, and irregularities.

2. Identifying Outliers and Missing Data

- ❑ Visual tools like boxplots and scatter plots can highlight outliers.
- ❑ Heatmaps can reveal missing values in a dataset.

3. Discovering Relationships Between Variables

- ❑ Scatter plots and correlation heatmaps help identify strong or weak relationships.
- ❑ Pair plots allow simultaneous visualization of multiple variable interactions.

4. Enhancing Interpretability

- ❑ Complex numerical summaries can be difficult to interpret.
- ❑ Visualizations simplify insights and make data-driven decisions easier.

5. Guiding Feature Engineering

- ❑ Helps determine transformations like log-scaling, binning, or normalization.
- ❑ Aids in selecting important features for modeling.

Visualization: Types of Plots

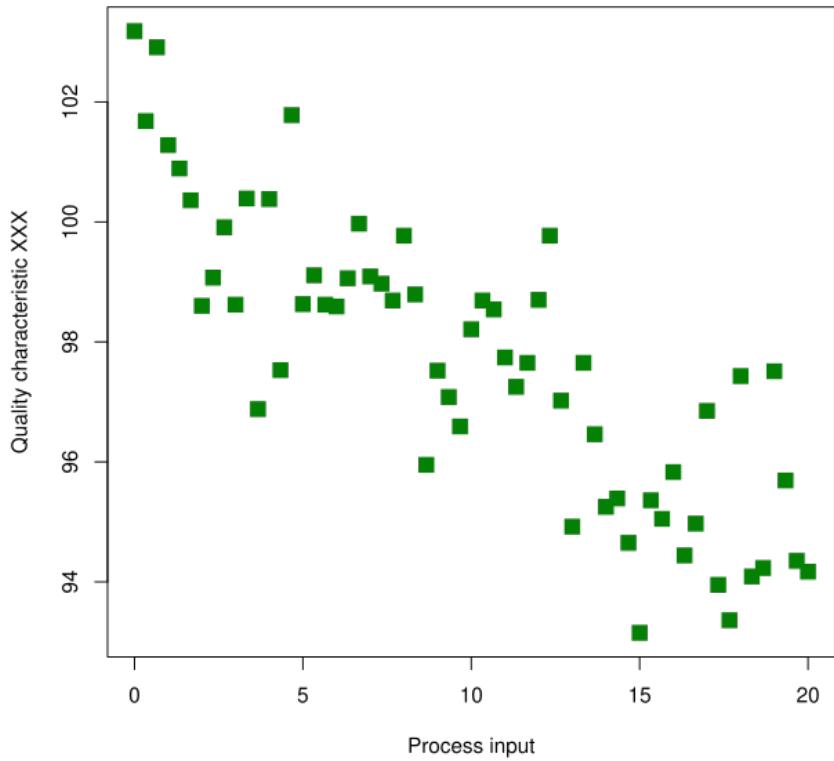
- Histograms
 - ▣ Use when: You want to analyze the distribution of a single continuous variable
 - ▣ Example: Examining the frequency of customer purchase amounts in an e-commerce dataset
 - ▣ Key insights: Shape of the distribution (normal, skewed, multimodal), outliers, and spread
- Boxplots
 - ▣ Use when: You need to compare distributions across categories or detect outliers
 - ▣ Example: Comparing salaries across different job roles in a company
 - ▣ Key insights: Median, quartiles, spread, and presence of outliers

Visualization: Types of Plots

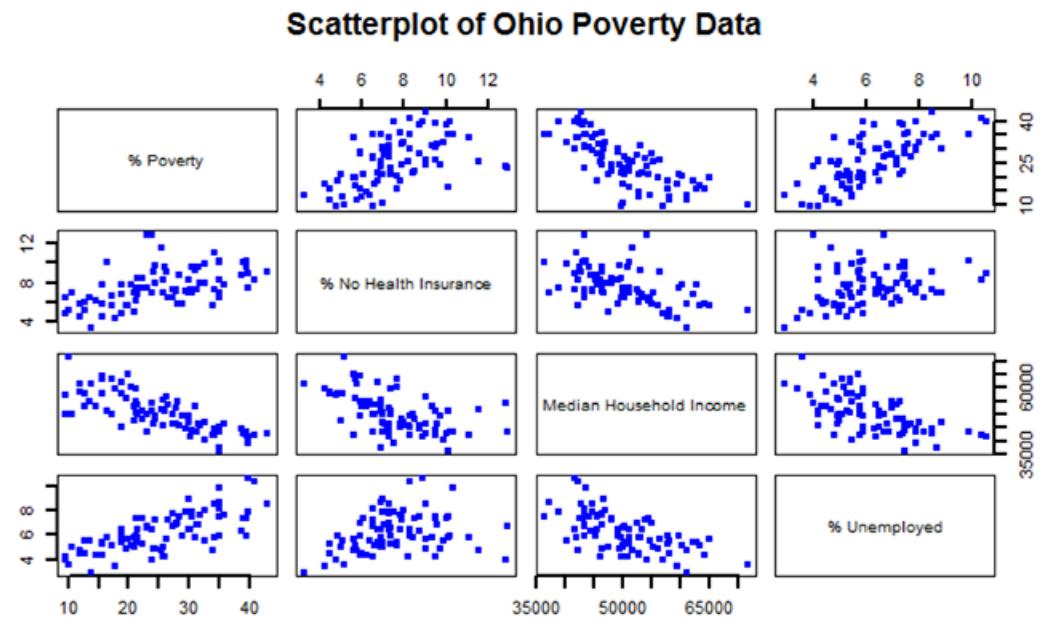
- Scatter plots
 - ▣ Use when: You want to examine the relationship between two numerical variables
 - ▣ Example: Checking if there is a correlation between advertising spend and sales revenue
 - ▣ Key insights: Strength and direction of relationships (linear, non-linear, or no correlation)
- Pair plots (Pairwise scatter plots)
 - ▣ Use when: You need to visualize relationships between multiple numerical variables in a dataset
 - ▣ Example: Analyzing relationships between weight, height, age, and cholesterol levels in a health study
 - ▣ Key insights: Trends, clusters, and correlations across multiple features

Visualization: Types of Plots

Scatterplot for quality characteristic XXX



Scatter plot



Pair plot

Feature Engineering and Transformation

□ Feature Engineering

- ❑ The process of creating new features from raw data to improve model performance
- ❑ Involves domain knowledge to extract meaningful information from raw data
- ❑ Examples: Extracting "day of the week" from a timestamp, creating interaction terms, or encoding categorical variables.

□ Feature Transformation

- ❑ Techniques used to modify features to make them more suitable for analysis and modeling
- ❑ Helps normalize, standardize, or reduce skewness in the data
- ❑ Examples: Log transformation to reduce skewness, min-max scaling, and principal component analysis (PCA) for dimensionality reduction.

Feature Engineering and Transformation

- Encoding categorical variables
 - ▣ One-Hot Encoding (OHE) categorical variable

- Converts categorical variables into a series of binary columns
 - Suitable for nominal (unordered) categories
 - Example: Category=[Red, Blue, Green]

Color	X_{Red}	X_{Blue}	X_{Green}
Red	1	0	0
Blue	0	1	0
Green	0	0	1

Feature Engineering and Transformation

- Encoding categorical variables
 - ▣ Label encoding
 - Assigns a unique integer to each category without considering any ranking.
 - Suitable for nominal categorical data where categories have no natural order.
 - Can sometimes be problematic if a machine learning model incorrectly interprets the numerical values as having an order.
 - Example: Category=[Red, Blue, Green]
Encoding: Red→0, Blue→1, Green→2
 - ▣ Ordinal encoding
 - Used for ordinal categorical data where categories have a clear order or ranking
 - Assigns numbers to categories based on their position in the order (e.g., "low" = 1, "medium" = 2, "high" = 3)
 - May introduce misleading information if the order between categories is not well-defined

Feature Engineering and Transformation

- Scaling and normalization
 - ▣ Min-max scaling: Rescales data to a fixed range [0,1]
- ▣ Standardization (Z-score normalization): Transforms data to have zero mean and unit variance

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- ▣ Robust scaling: Uses median and IQR to reduce the effect of outliers

$$X' = \frac{X - \text{median}(X)}{\text{IQR}(X)}$$

Feature Engineering and Transformation

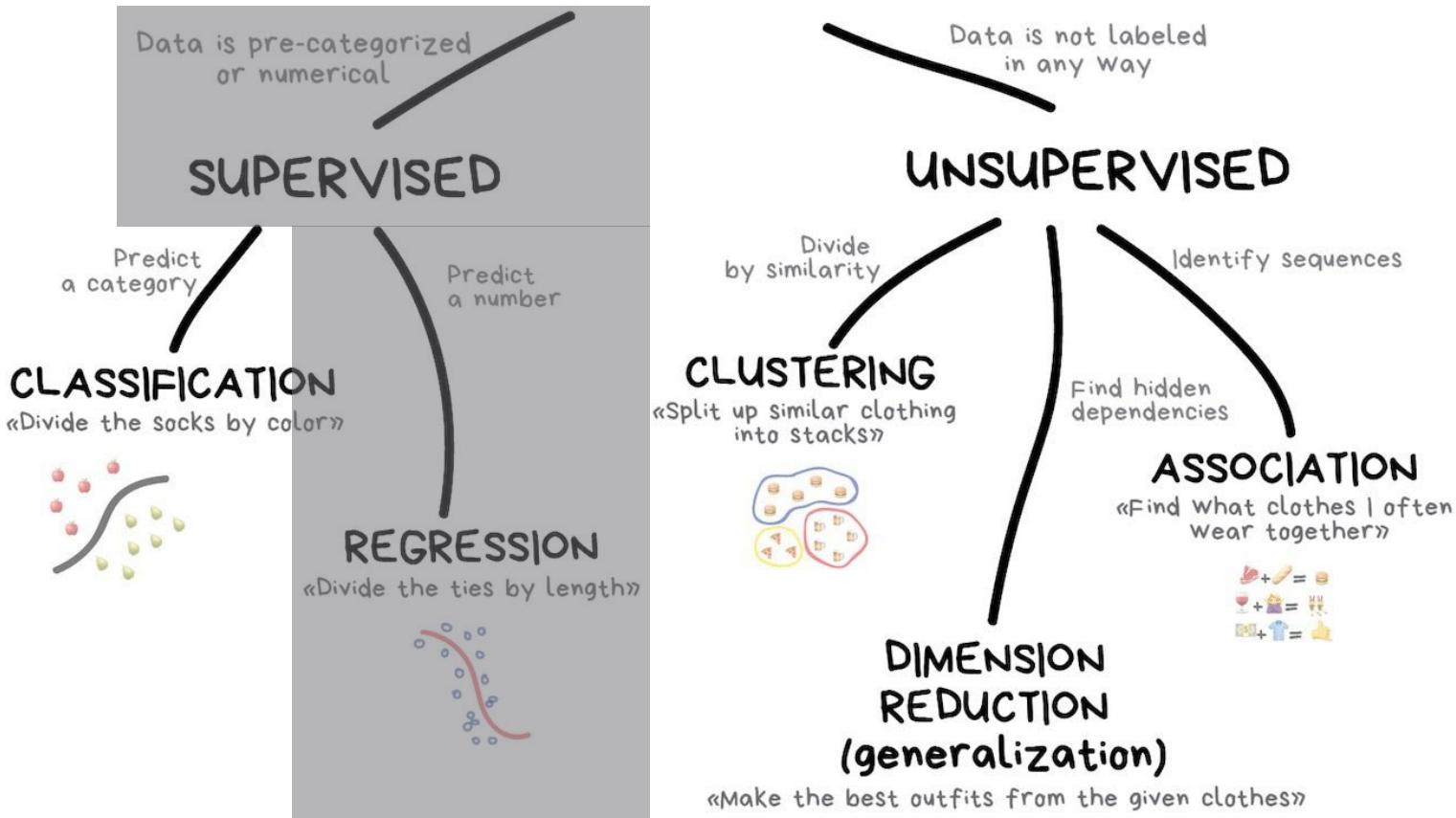
- Purposes of scaling and normalization
 - ▣ Ensuring **fair comparisons** between features
 - Different features in a dataset may have different units and scales (e.g., age in years vs. income in dollars). Scaling ensures that all features contribute equally to analysis and modeling
 - ▣ **Improving convergence** in machine learning models
 - Many machine learning algorithms (e.g., gradient descent-based models like logistic regression and neural networks) perform better when input data is normalized, leading to faster convergence and better optimization
 - ▣ Enhancing the **interpretability of data**
 - Some models (e.g., distance-based models like K-Nearest Neighbors and clustering algorithms) rely on numerical distances. Features with larger scales can dominate, leading to biased results. Scaling ensures that no single feature disproportionately influences outcomes
 - ▣ Reducing the **impact of outliers**
 - Robust scaling techniques (e.g., Robust Scaler using median and interquartile range) mitigate the influence of extreme values, improving model stability.
 - ▣ Facilitating **data visualization**
 - When visualizing multiple features together (e.g., in scatter plots or heatmaps), unscaled features can distort patterns. Scaling helps provide clearer insights.

LINEAR REGRESSION

Week04

Topics Covered in This Class

CLASSICAL MACHINE LEARNING

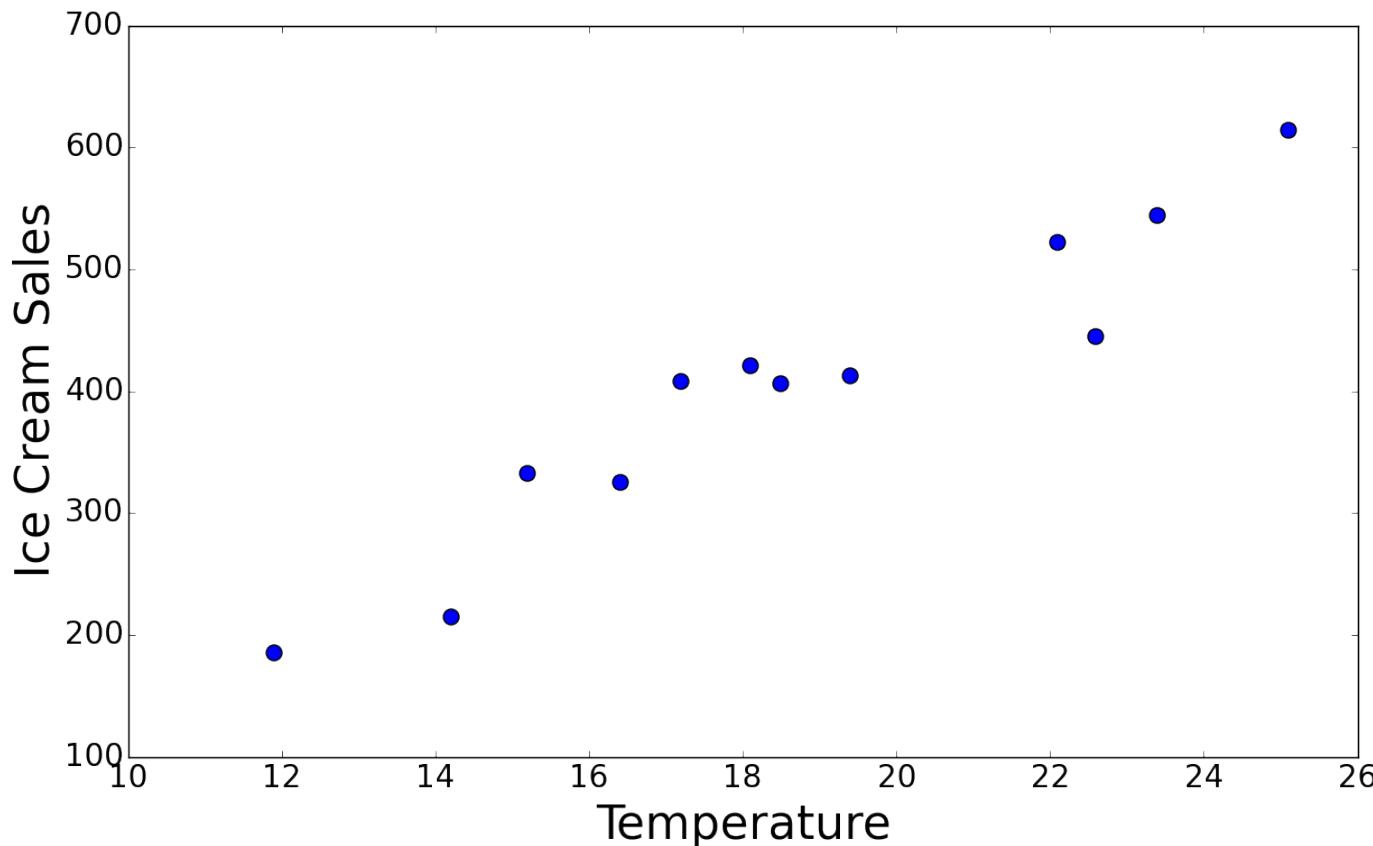


Linear Regression

Supervised Learning: Regression

- Prediction ice cream sales over given temperature

$$\text{ice cream sales} = f(\text{temperature})$$



Temperature (°C)	Ice Cream Sales (\$)
14.2	215
16.4	325
11.9	185
15.2	332
18.5	406
22.1	522
19.4	412
25.1	614
23.4	544
18.1	421
22.6	445
17.2	408
12.0	185
14.5	215
15.0	332
16.5	325
17.5	406
18.0	412
18.5	421
19.0	445
21.5	544
22.0	550
22.5	555
23.0	560
23.5	614
24.0	620
24.5	625
25.0	630

Linear Regression

- Linear regression
 - ▣ Based on the assumption that
the relationship between a scalar dependent variable y and explanatory(independent) variables X is linear
 - ▣ $X = [x_1, x_2, x_3, \dots, x_n]$
Explanatory variables: print run(x_1), page number(x_2)

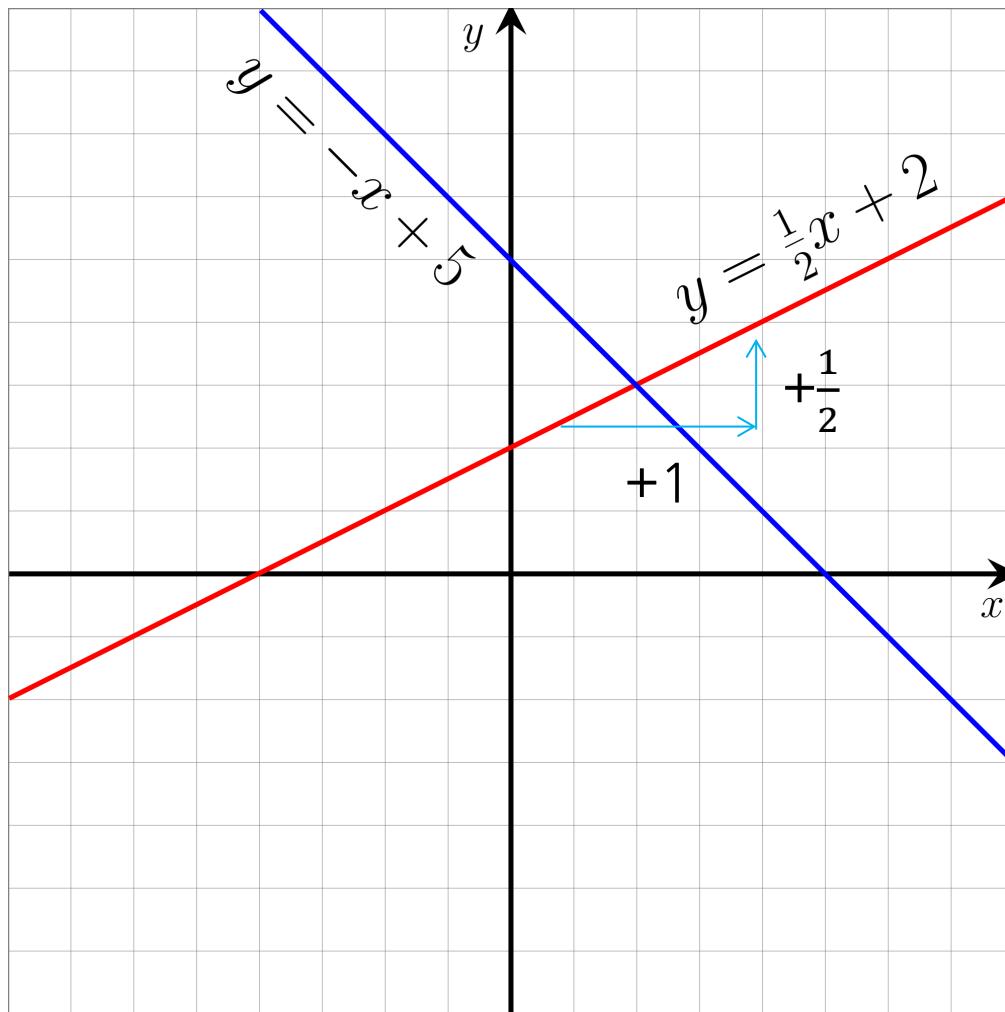
x_1	x_2
2800	22
2670	14
2800	37
2784	15
2800	38

→ X

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$
$$\epsilon \sim N(0, \sigma^2)$$

* Linear function

- You studied linear function when you are high school student!



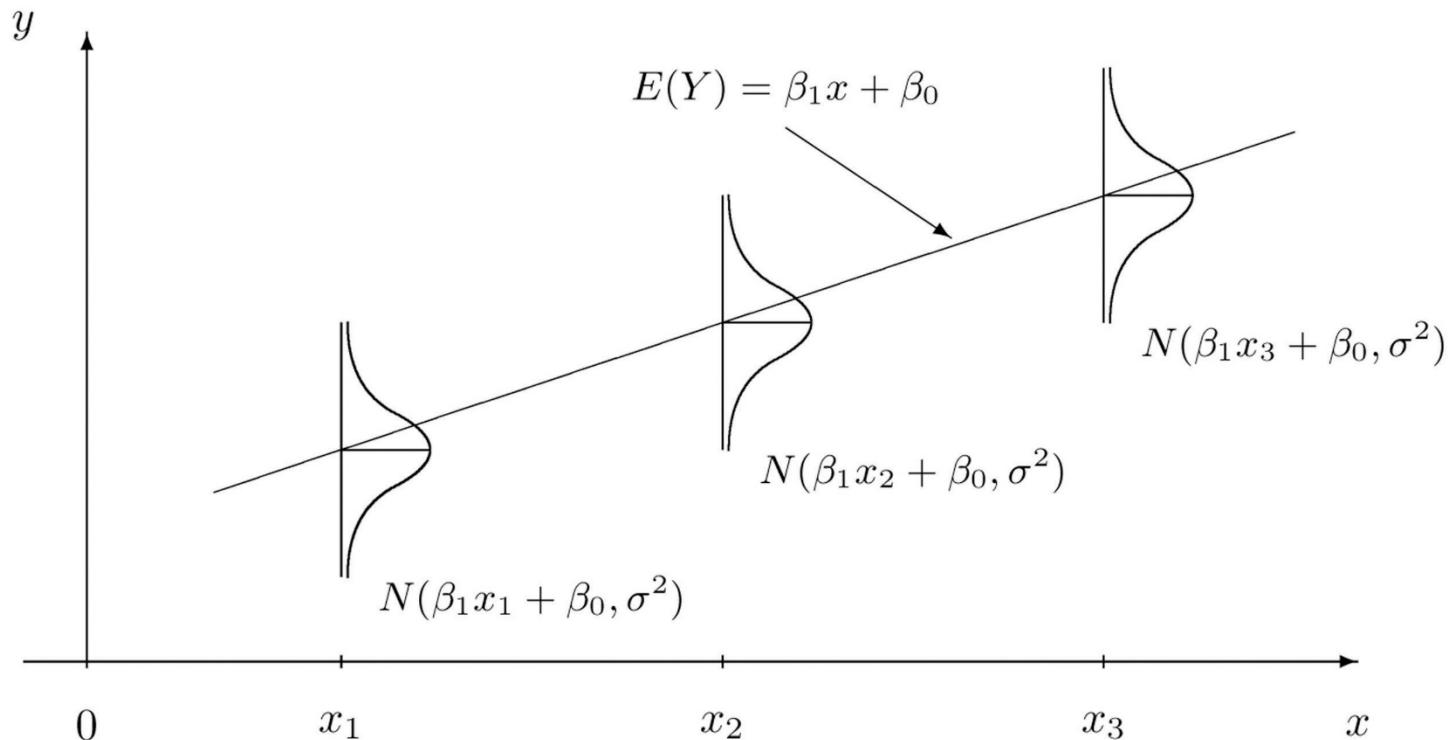
$$y = \frac{1}{2}x + 5$$

slope
→ coefficient

intercept

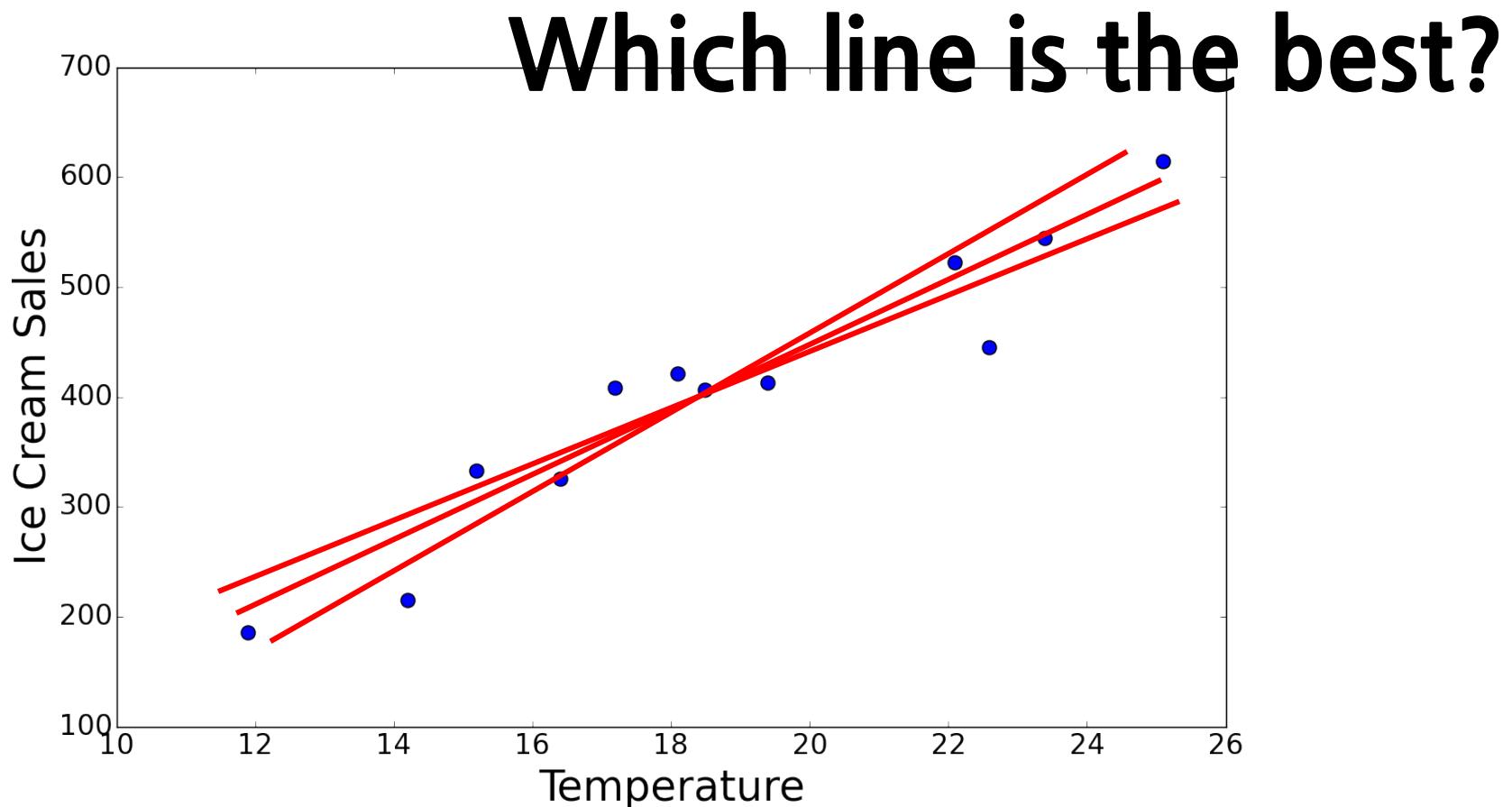
Main Assumptions of Linear Regression

- Linear regression analysis makes several key assumptions
 - Linear relationship
 - Homoscedasticity
 - Normality
 - No or little multicollinearity



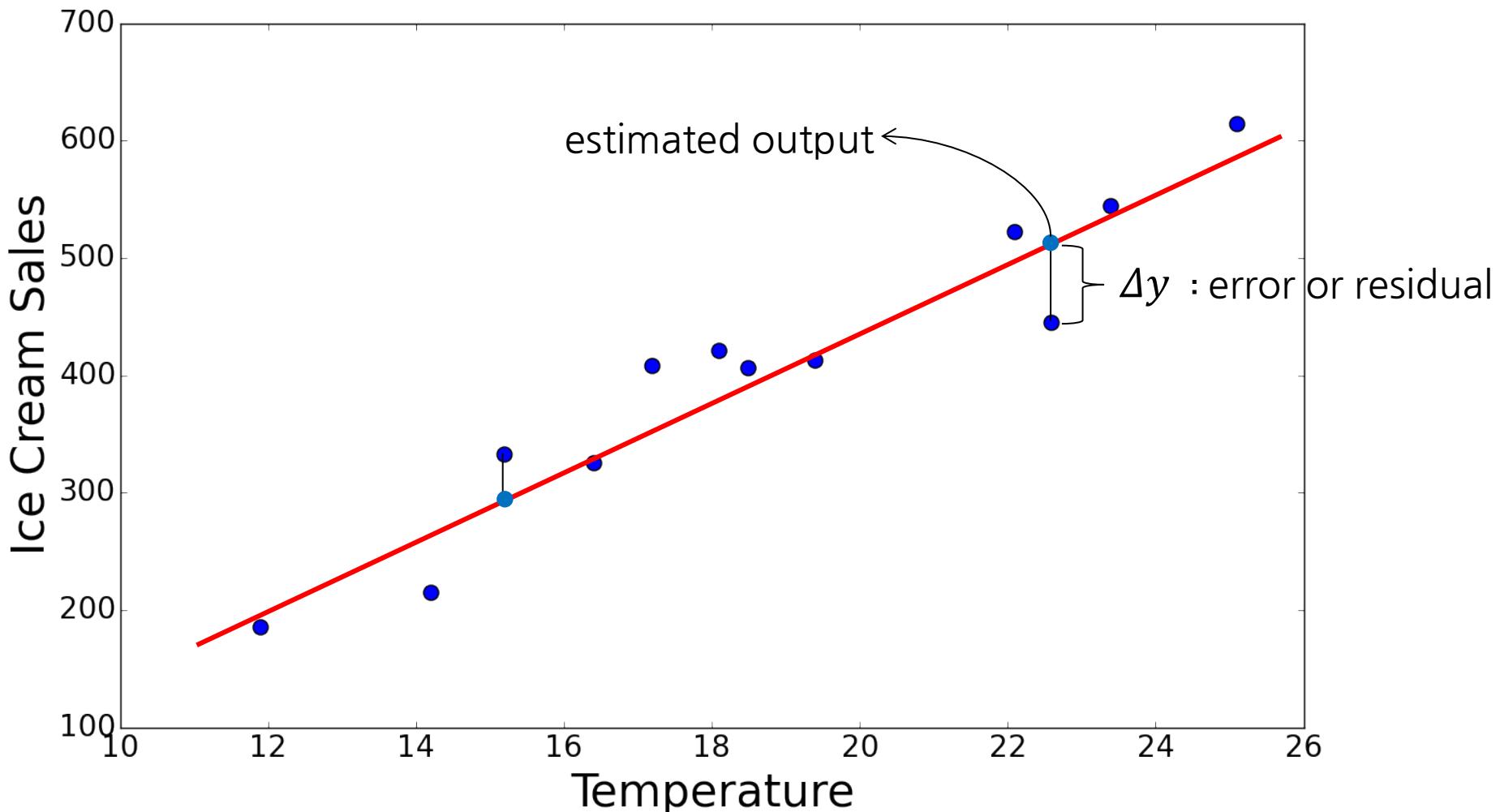
How to Determine Relationship between x and y

- Need a criterion



Least Square Method

- Minimize summation of squared error
 - ▣ Squared error=(estimated output-real output)²=error²



Least Square Method

- Minimize summation of squared error

$$\sum_i (y_i - \hat{y}_i)^2 \quad \text{GSE}$$

Sum for all data points in train set

estimated output

real output

- In the simple case: Only one independent variable
 - Estimated output $\hat{y}_i = \beta_0 + \beta_1 x_i$

$$\min \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

Unknown values

Known values

※ Summary for Notation

- Hat, ($\hat{}$)
 - ▣ Represents estimation
 - ▣ β_1 is unknown true value, $\hat{\beta}_1$ is estimation for β_1 through model learning
 - ▣ y_i is known output value of i -th sample, \hat{y}_i is estimated output by learned model
- Bar, ($\bar{}$)
 - ▣ Represents sample mean
 - ▣ Arithmetic average of the observed values of variable

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- ▣ If the number of input variables is more than one, elements of sample mean vector consist of average of each variable

$$\bar{\mathbf{x}} = \left(\frac{\sum_{i=1}^n x_{1i}}{n}, \frac{\sum_{i=1}^n x_{2i}}{n}, \dots, \frac{\sum_{i=1}^n x_{pi}}{n} \right) = \frac{\sum_{i=1}^n \mathbf{x}_i}{n}$$

- Bold character usually represents vector

Least Square Method

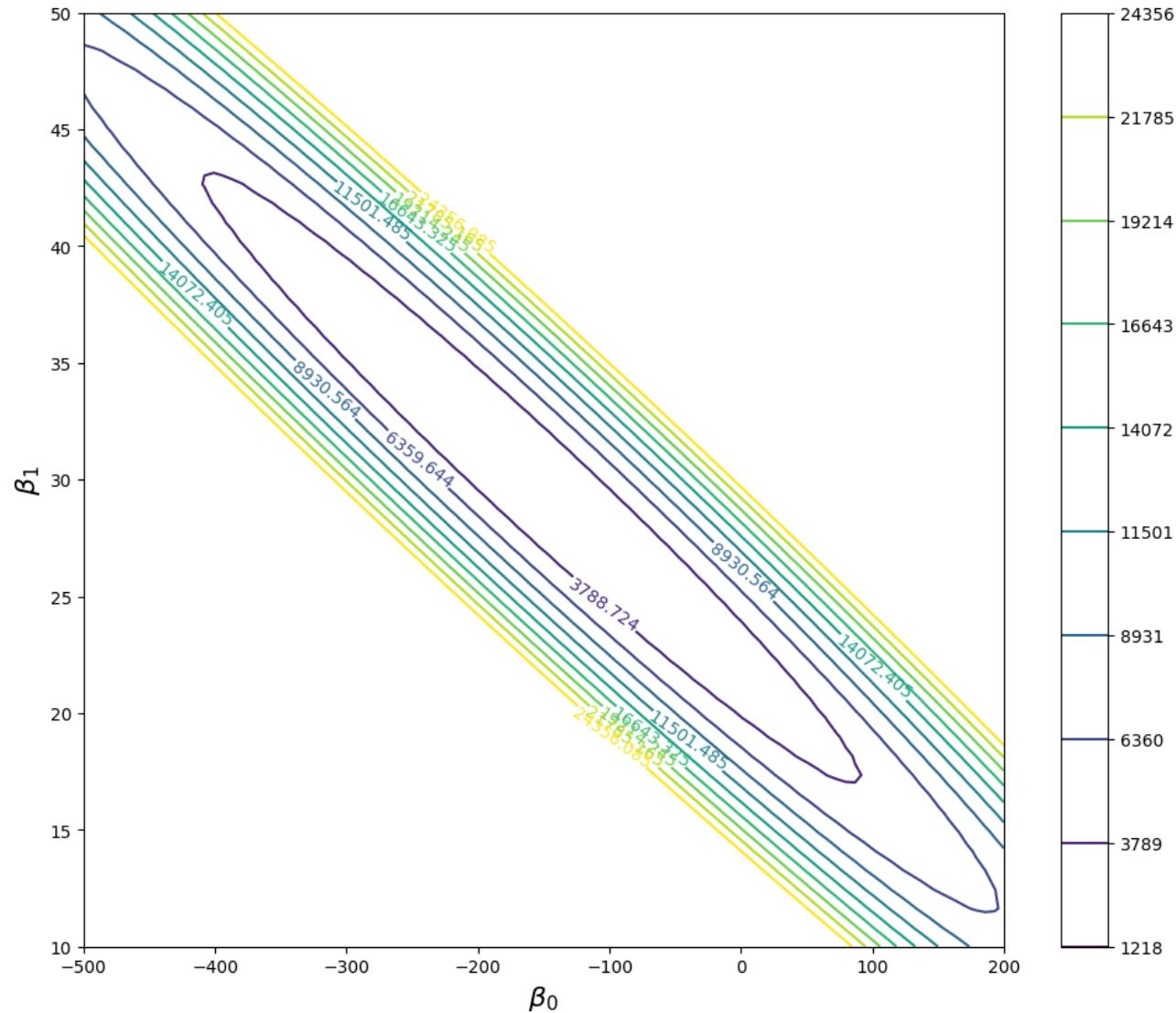
- Which one is better?

Temperature (°C)	Ice Cream Sales (\$)	$\beta_1 = 20, \beta_0 = -80$		$\beta_1 = 30, \beta_0 = -160$	
		Estimated Sales	Squared error	Estimated Sales	Squared error
14.2	y_i	215	\hat{y}_i	204	$(y_i - \hat{y}_i)^2$
16.4		325		248	5929
11.9		185		158	729
15.2		332		224	11664
18.5		406		290	13456
22.1		522		362	25600
19.4		412		308	10816
25.1		614		422	36864
23.4		544		388	24336
18.1		421		282	19321
22.6		445		372	5329
17.2		408		264	20736
sum				174901	14594

better

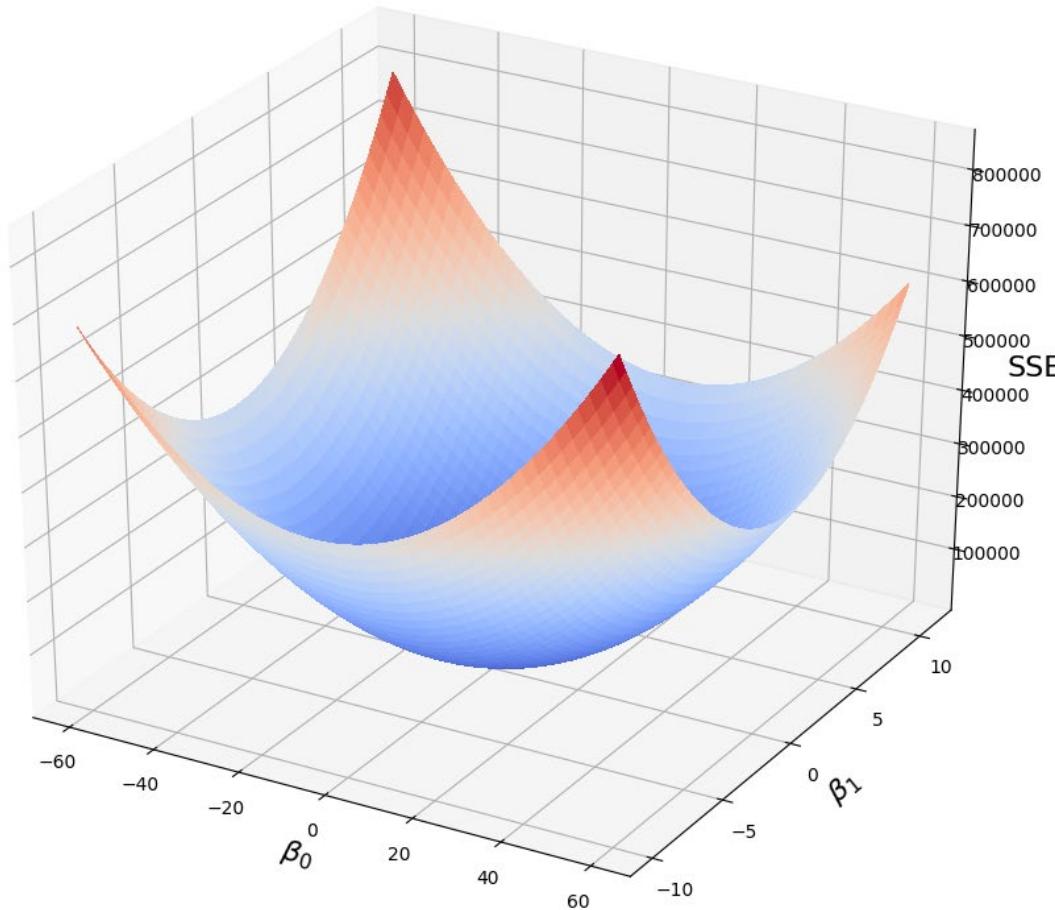
Least Square Method

- Summation of squared error with different β_0 and β_1



Least Square Method

- Summation of squared error with different β_0 and β_1 for simulated data from $y = x + 1$



Optimization for Linear Regression

- Variables to be determined

$$\beta_0, \beta_1 \quad \hat{y}_i = \beta_0 + \beta_1 x_i$$

- Objective function

$$\min f(\beta_0, \beta_1) = \min \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

- Constraints

- No constraint

Optimization for Linear Regression

- Solution

- ▣ Calculate partial derivatives with respect to β_0, β_1

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^n -2(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^n -2x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

- ▣ Solve linear equations

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Multiple Input Variables

- More than one input variable
 - Want to predict consumption of petrol

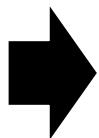
Petrol Tax(\$)	Average Income (\$)	Paved Highways (miles)	Proportion of population with driver's license	Consumption of petrol (M of gallons)
9	3571	1976	0.525	541
9	4092	1250	0.572	524
9	3865	1586	0.58	561
7.5	4870	2351	0.529	414
8	4399	431	0.544	410
10	5342	1333	0.571	457
8	5319	11868	0.451	344
8	5126	2138	0.553	467
8	4447	8577	0.529	464
7	4512	8507	0.552	498
...

Multiple Input Variables

- Estimation based on petrol tax, average income, length of paved highways, proportion of population with driver's license

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

- y = consumption of petrol
- x_1 = petrol tax
- x_2 = average income
- x_3 = length of paved highways
- x_4 = proportion of population with driver's license
- ϵ is random error which follows Gaussian distribution with 0 mean, σ^2 variance



$$\min \sum_i (y_i - \hat{y}_i)^2$$

Same as the simple case!

Optimization for Linear Regression: Multivariate

- multivariate linear regression

$$\min f(\beta_0, \dots, \beta_p) = \min \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_p x_{pi})^2$$

- Estimated parameters are obtained by setting partial derivatives zero

$$\left. \begin{aligned} \frac{\partial f(\beta_0, \dots, \beta_p)}{\partial \beta_0} &= \sum_{i=1}^n -2(y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_p x_{pi}) = 0 \\ \frac{\partial f(\beta_0, \dots, \beta_p)}{\partial \beta_1} &= \sum_{i=1}^n -2x_{1i}(y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_p x_{pi}) = 0 \\ &\vdots \\ \frac{\partial f(\beta_0, \dots, \beta_p)}{\partial \beta_p} &= \sum_{i=1}^n -2x_{pi}(y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_p x_{pi}) = 0 \end{aligned} \right\}$$

Multiple Input Variables

- Matrix approach to multiple regression model

$$y_1 = \beta_0 \cdot 1 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_p x_{k1} \quad n \text{ samples, } p \text{ input variables}$$

$$\begin{aligned} y_2 &= \beta_0 \cdot 1 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_p x_{k2} \\ &\vdots \end{aligned}$$

$$y_n = \beta_0 \cdot 1 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_p x_{kn}$$



$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$



$$= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\epsilon} \sim MVN(0, \sigma^2 \mathbf{I})$$



$$E = \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2$$

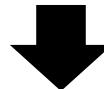
Optimization for Linear Regression: Multivariate

$$\min E = \min \|y - X\beta\|^2$$

- Solution is obtained by setting $\frac{\partial E}{\partial \beta} = 0$

$$\frac{\partial (\mathbf{x} - A\mathbf{s})^T W (\mathbf{x} - A\mathbf{s})}{\partial \mathbf{s}} = -2A^T W (\mathbf{x} - A\mathbf{s})$$

$X=Y$ $S=\beta$
 $A=X$ $W=I$



$$\frac{\partial (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta)}{\partial \beta} = -2X^T (\mathbf{y} - X\beta) = \mathbf{0}$$



$$\underline{X^T X \beta - X^T y = 0} \quad X = n \times (p+1)$$

$$X^T = (p+1) \times n$$

$$X^T X = (p+1) \times (p+1)$$

- Reference

- Matrix Cookbook

- <https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>

Estimation of Regression Coefficients: Multivariate

- Use least square methods as same as simple linear regression

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- \mathbf{X}^T : transpose matrix of \mathbf{X}
- \mathbf{X}^{-1} : inverse matrix of \mathbf{X}

$$\hat{\mathbf{y}} = \mathbf{X} \hat{\beta} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{H} \mathbf{y}$$

- Residual(error) terms

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H}) \mathbf{y} = [e_1 \ e_2 \ \dots \ e_n]^T$$

- Covariance

$$\begin{aligned} \text{Cov}[\mathbf{e}] &= \text{Cov}[\mathbf{y} - \hat{\mathbf{y}}] = \text{Cov}[(\mathbf{I} - \mathbf{H}) \mathbf{y}] = \text{Cov}[(\mathbf{I} - \mathbf{H})(\mathbf{X} \beta + \epsilon)] \\ &= \text{Cov}[(\mathbf{I} - \mathbf{H}) \epsilon] = (\mathbf{I} - \mathbf{H}) \text{Cov}[\epsilon] (\mathbf{I} - \mathbf{H})^T = \sigma^2 (\mathbf{I} - \mathbf{H}) (\mathbf{I} - \mathbf{H})^T = \sigma^2 (\mathbf{I} - \mathbf{H}) \end{aligned}$$

- SSE

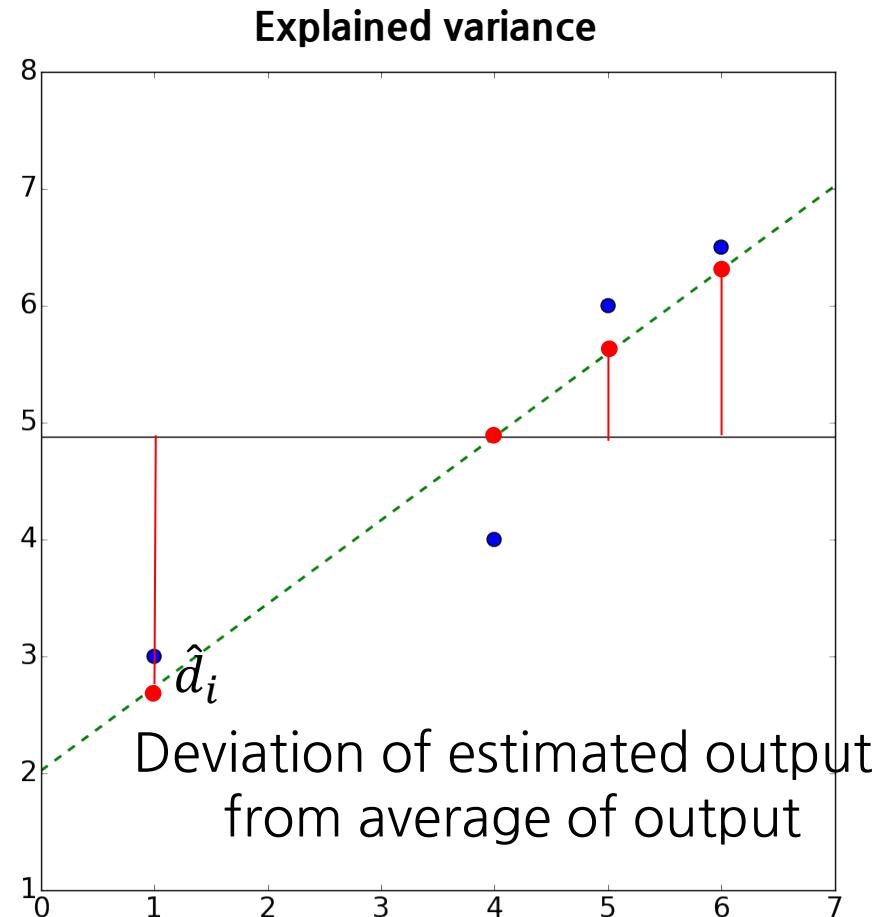
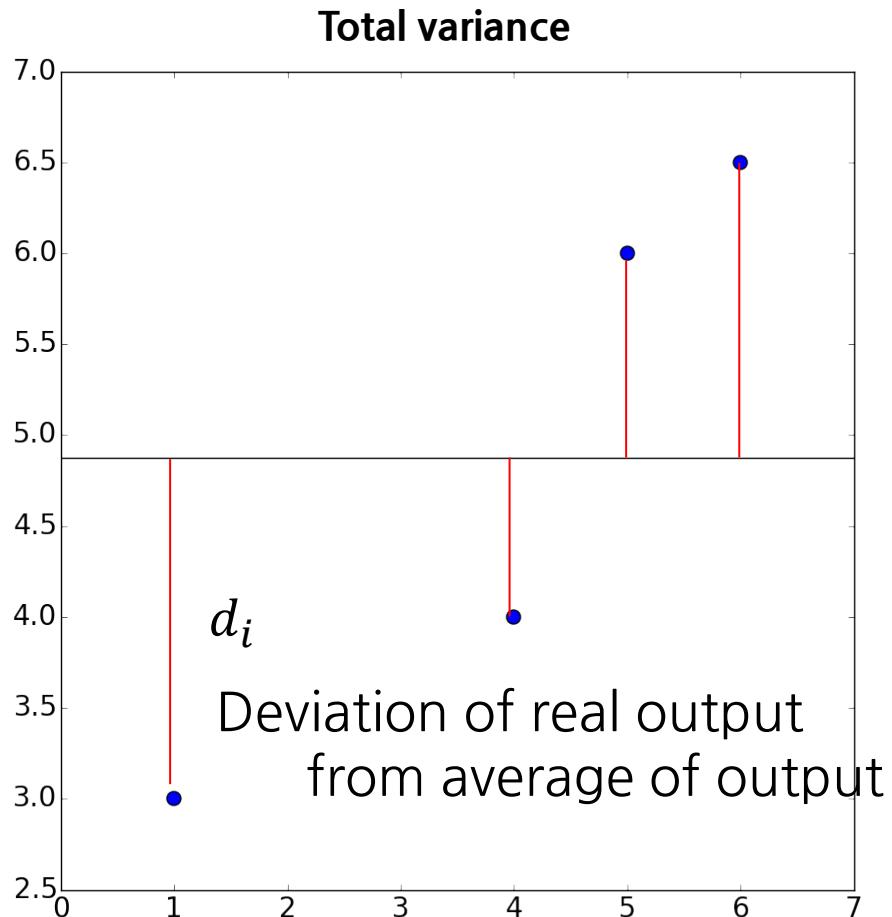
$$SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2 = \mathbf{e}^T \mathbf{e}$$

Is the Regression Model Significant?

- Modeling learning is not the end of the analysis
 - ▣ Check overall significance in regression models
 - Whether the regression model is overall significant for predicting a target
 - ▣ Check significance of regression coefficients
 - Whether the specific variable is significant for predicting a target
- In the case of simple linear regression, testing overall significance of the model is the same as testing significance of regression coefficients
 - ▣ Because only one explanatory variable is used

Sum of Square

- Explained variance(SSR) and Total variance(SST)



Sum of Square

Chi-squared distribution

- Total variance: the total sum of squares

$$SST = \sum_i (y_i - \bar{y})^2$$

- Explained variance: the regression sum of squares, also called the explained sum of squares

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

- Residual variance: the sum of squares of residuals, also called the residual sum of squares

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

- Relationship among three values

$$SST = SSR + SSE$$

Test of Model Significance

전체 회귀 모형 유의성 검증 or 분산 차이

- F -test for general regression models

- ▣ Hypothesis

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_1: \text{not all } \beta_i (i = 1, 2, \dots, p) \text{ equal zero}$$

- ▣ Test statistic

$$\frac{(MSR_p)}{(MSE_{n-p-1})}$$

$$F^* = MSR/MSE$$

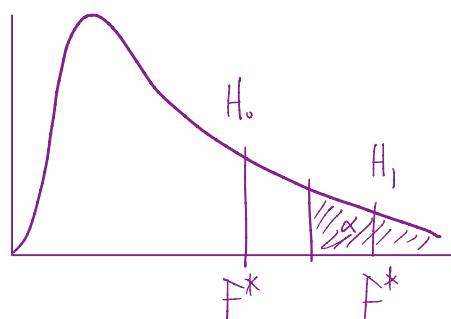
- F follows F -distribution with $(p, n - p - 1)$ degree of freedom

- ▣ Decision rule

If $F^* \leq F(1 - \alpha; p, n - p - 1)$, conclude H_0

If $F^* > F(1 - \alpha; p, n - p - 1)$, conclude H_1

- α : significance level



⌘ Statistical Test

- A statistical test provides a mechanism for making quantitative decisions about a process or processes
 - ▣ The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process
 - ▣ The procedure is based on how likely it would be for a set of observations to occur if the null hypothesis were true
- Null hypothesis
 - ▣ A general statement or default position that there is no relationship between two measured phenomena, or no association among groups
- Alternative hypothesis
 - ▣ It is the hypothesis used in hypothesis testing that is contrary to the null hypothesis

⌘ Statistical Test

- Steps in testing for statistical significance

State the research hypothesis



State the null hypothesis & alternative hypothesis



Select a probability of error level (alpha level)



Select and compute the test for statistical significance



Interpret the results

※ Statistical Test

두 집단의 평균이 유의하게 다른가?,

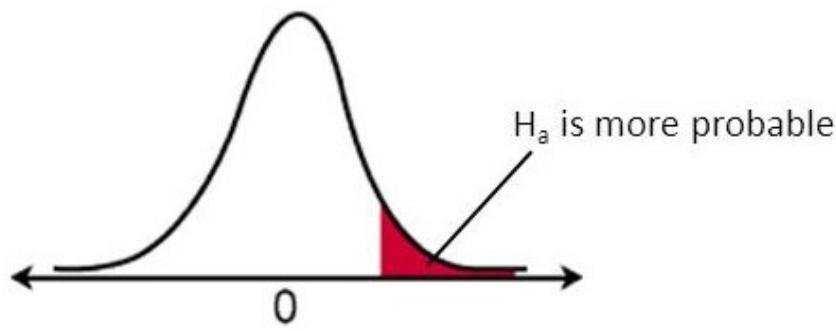
- Consider 20 first year resident female doctors drawn at random from one area
 - ▣ resting systolic blood pressures measured using an electronic sphygmomanometer
 - Sample mean = 130.05
 - ▣ Research hypothesis is that a resting systolic blood pressure of 120 mm Hg is predicted as the population mean
 - ▣ Null Hypothesis

$$H_0: \mu = 120$$

$$H_1: \mu \neq 120$$

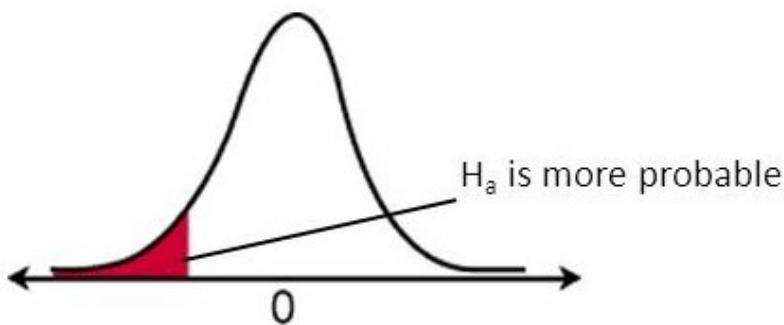
- ▣ Set significance level as 0.05
- ▣ Determine test statistics and underlying distribution
 - $t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}}$
 - t follows t -distribution with the degree of freedom as $n - 1$

* Statistical Test



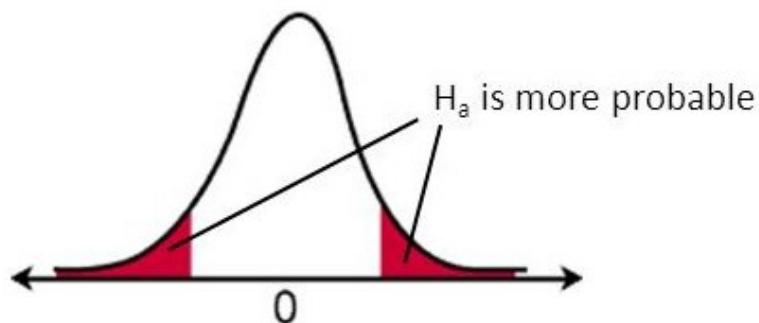
Right-tail test

$$H_a: \mu > \text{value}$$



Left-tail test

$$H_a: \mu < \text{value}$$



Two-tail test

$$H_a: \mu \neq \text{value}$$

Test of Model Significance

- ANOVA table for multiple regression model with p input variables

Factor	Sum of square	Degree of freedom	Mean square	F-value	p-value
Model	SSR	p	$MSR = SSR/p$	$F_0 = MSR/MSE$	$P\{F_{p,n-p-1} > F_0\}$
Residual	SSE	$n - p - 1$	$MSE = SSE/(n - p - 1)$		
Total	SST	$n - 1$			

- Analysis of Variance (ANOVA)

Degree of Freedom

- The number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary
 - ▣ The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it
- Sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

- ▣ The reason that denominator is $n - 1$ is that degree of freedom of sample mean, \bar{y} is $n - 1$
- ▣ Another reason is that in the case of that denominator is $n - 1$, S^2 is unbiased estimator of variance of population
- Mean squared error for simple linear regression

$$MSE = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

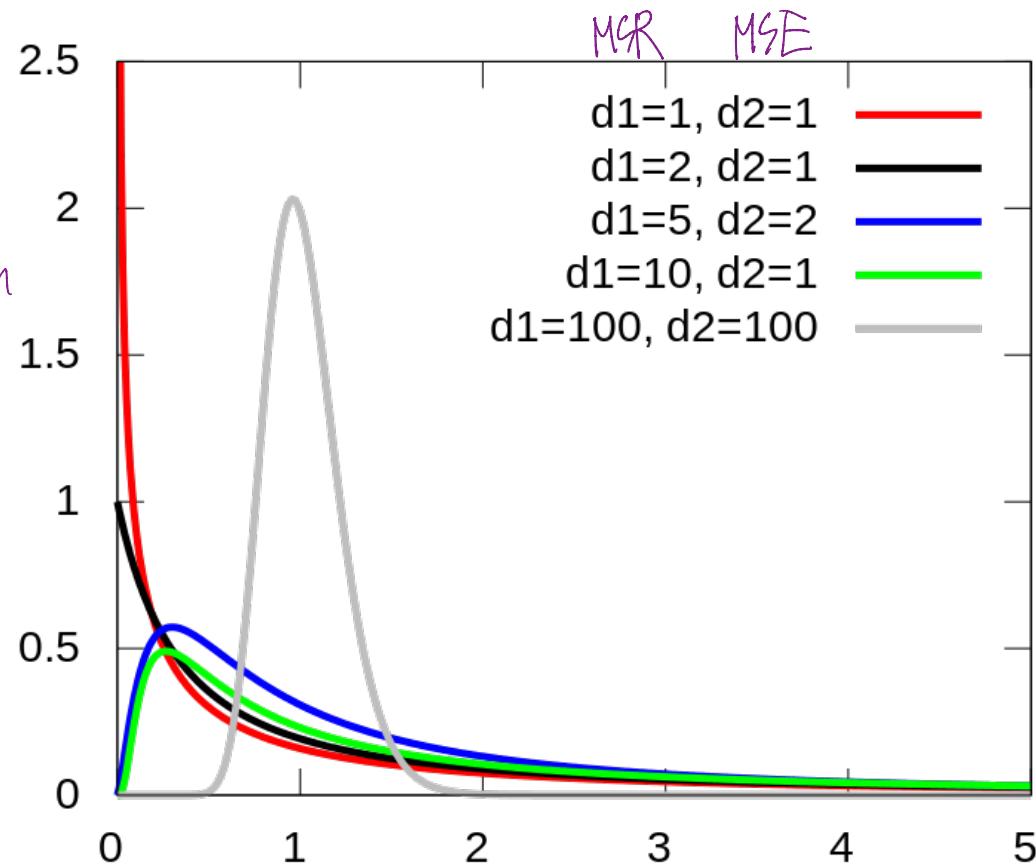
$P=1$

- ▣ The reason that denominator is $n - 2$ is that \hat{y}_i is calculate from $\hat{\beta}_0 + \hat{\beta}_1 x_i$ and it depends on two estimators $\hat{\beta}_0, \hat{\beta}_1 \rightarrow$ Decrease two degrees of freedom

※ F distribution

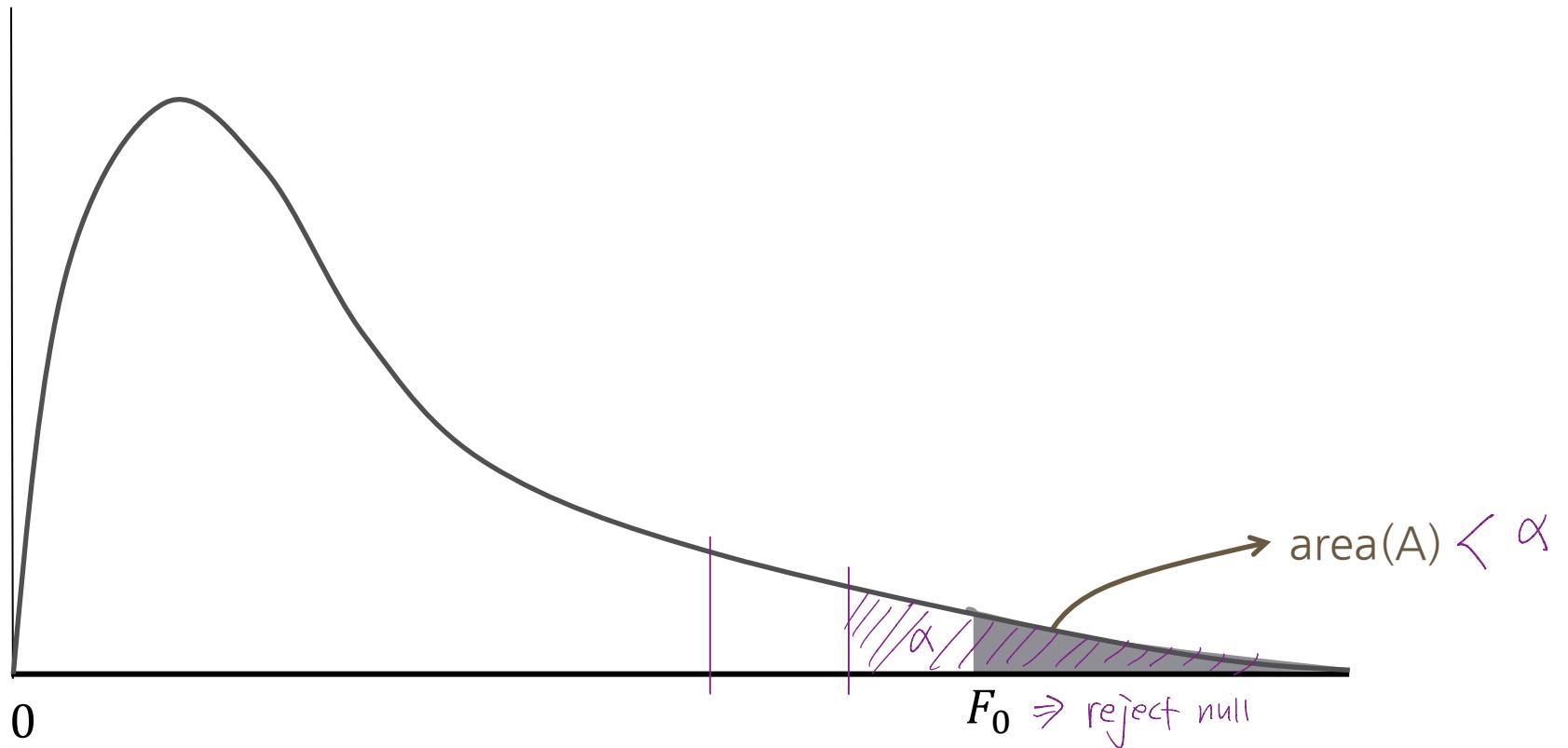
- F statistics follows F distribution with $(p, n - p - 1)$ degree of freedom
 - Probability density function of F distribution with different parameters
 - F distribution is determined by two parameters

df 4
Normal Distribution



Test of Model Significance $\Rightarrow F$

- If (area under density function from F_0 to ∞) $< \alpha$
 - Reject null hypothesis → not all $\beta_i (i = 1, 2, \dots, p)$ equal zero
 - ▣ α is significance level
 - ▣ significance level is usually set to 0.1, 0.05, or 0.01
 - The higher significance level, the higher probability to reject null hypothesis



※ Table for Distributions

F-Distribution, Continued
Upper 0.01 Critical Points

		$F_{0.01}(r_1, r_2)$								
		r_1								
r_2		10	15	20	25	30	40	60	120	∞
1	6055.9	6157.3	6208.7	6239.8	6260.7	6286.8	6313.0	6339.4	6365.9	
2	99.40	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50	
3	27.23	26.87	26.69	26.58	26.50	26.41	26.32	26.22	26.13	
4	14.55	14.20	14.02	13.91	13.84	13.75	13.65	13.56	13.46	
5	10.05	9.72	9.55	9.45	9.38	9.29	9.20	9.11	9.02	
6	7.87	7.56	7.40	7.30	7.23	7.14	7.06	6.97	6.88	
7	6.62	6.31	6.16	6.06	5.99	5.91	5.82	5.74	5.65	
8	5.81	5.52	5.36	5.26	5.20	5.12	5.03	4.95	4.86	
9	5.26	4.96	4.81	4.71	4.65	4.57	4.48	4.40	4.31	
10	4.85	4.56	4.41	4.31	4.25	4.17	4.08	4.00	3.91	
11	4.54	4.25	4.10	4.01	3.94	3.86	3.78	3.69	3.60	
12	4.30	4.01	3.86	3.76	3.70	3.62	3.54	3.45	3.36	
13	4.10	3.82	3.66	3.57	3.51	3.43	3.34	3.25	3.17	
14	3.94	3.66	3.51	3.41	3.35	3.27	3.18	3.09	3.00	
15	3.80	3.52	3.37	3.28	3.21	3.13	3.05	2.96	2.87	
16	3.69	3.41	3.26	3.16	3.10	3.02	2.93	2.84	2.75	
17	3.59	3.31	3.16	3.07	3.00	2.92	2.83	2.75	2.65	
18	3.51	3.23	3.08	2.98	2.92	2.84	2.75	2.66	2.57	
19	3.43	3.15	3.00	2.91	2.84	2.76	2.67	2.58	2.49	
20	3.37	3.09	2.94	2.84	2.78	2.69	2.61	2.52	2.42	
21	3.31	3.03	2.88	2.79	2.72	2.64	2.55	2.46	2.36	
22	3.26	2.98	2.83	2.73	2.67	2.58	2.50	2.40	2.31	
23	3.21	2.93	2.78	2.69	2.62	2.54	2.45	2.35	2.26	
24	3.17	2.89	2.74	2.64	2.58	2.49	2.40	2.31	2.21	
25	3.13	2.85	2.70	2.60	2.54	2.45	2.36	2.27	2.17	
26	3.09	2.81	2.66	2.57	2.50	2.42	2.33	2.23	2.13	
27	3.06	2.78	2.63	2.54	2.47	2.38	2.29	2.20	2.10	
28	3.03	2.75	2.60	2.51	2.44	2.35	2.26	2.17	2.06	
29	3.00	2.73	2.57	2.48	2.41	2.33	2.23	2.14	2.03	
30	2.98	2.70	2.55	2.45	2.39	2.30	2.21	2.11	2.01	
40	2.80	2.52	2.37	2.27	2.20	2.11	2.02	1.92	1.80	
60	2.63	2.35	2.20	2.10	2.03	1.94	1.84	1.73	1.60	
120	2.47	2.19	2.03	1.93	1.86	1.76	1.66	1.53	1.38	
∞	2.32	2.04	1.88	1.77	1.70	1.59	1.47	1.32	1.00	

LINEAR REGRESSION

Week05

Linear Regression

Is the Regression Model Significant?

- Modeling learning is not the end of the analysis
 - ▣ Check overall significance in regression models
 - Whether the regression model is overall significant for predicting a target
 - ▣ Check significance of regression coefficients
 - Whether the specific variable is significant for predicting a target
- In the case of simple linear regression, testing overall significance of the model is the same as testing significance of regression coefficients
 - ▣ Because only one explanatory variable is used

F-test		모든 자료가 예측에 유의한가
T-test		개별 변수가 예측에 유의한가

Test Concerning Regression Coefficients

- Test for $\beta_j (j = 0, 1, 2, \dots, p)$

- Hypothesis

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

- Test statistic

$$t_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

- $se^2(\hat{\beta}) = MSE(\mathbf{X}^T \mathbf{X})^{-1} \rightarrow se^2(\hat{\beta}_j) = [MSE(\mathbf{X}^T \mathbf{X})^{-1}]_{j,j}$

- Decision rule

If $|t_j| \leq t\left(1 - \frac{\alpha}{2}; n - p - 1\right)$, conclude H_0

If $|t_j| > t\left(1 - \frac{\alpha}{2}; n - p - 1\right)$, conclude H_1

Test Concerning Regression Coefficients

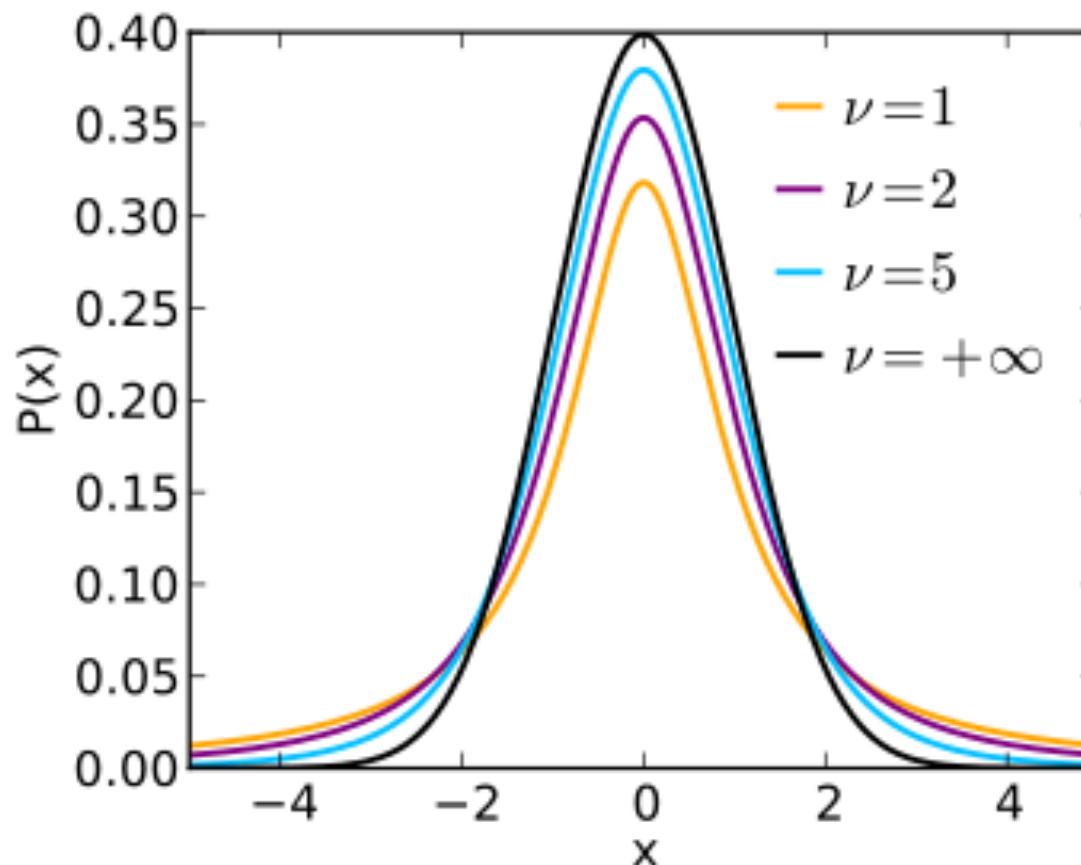
- $se^2(\hat{\beta}_i)$
 - ▣ Ex) two input variables

$$(X^T X)^{-1} = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 \\ x_{00} & & \\ & x_{11} & \\ & & x_{22} \end{bmatrix}$$

- $se^2(\hat{\beta}_0) = MSE \cdot x_{00} \rightarrow se(\hat{\beta}_0) = \sqrt{MSE \cdot x_{00}}$
- $se^2(\hat{\beta}_1) = MSE \cdot x_{11} \rightarrow se(\hat{\beta}_1) = \sqrt{MSE \cdot x_{11}}$
- $se^2(\hat{\beta}_2) = MSE \cdot x_{22} \rightarrow se(\hat{\beta}_2) = \sqrt{MSE \cdot x_{22}}$

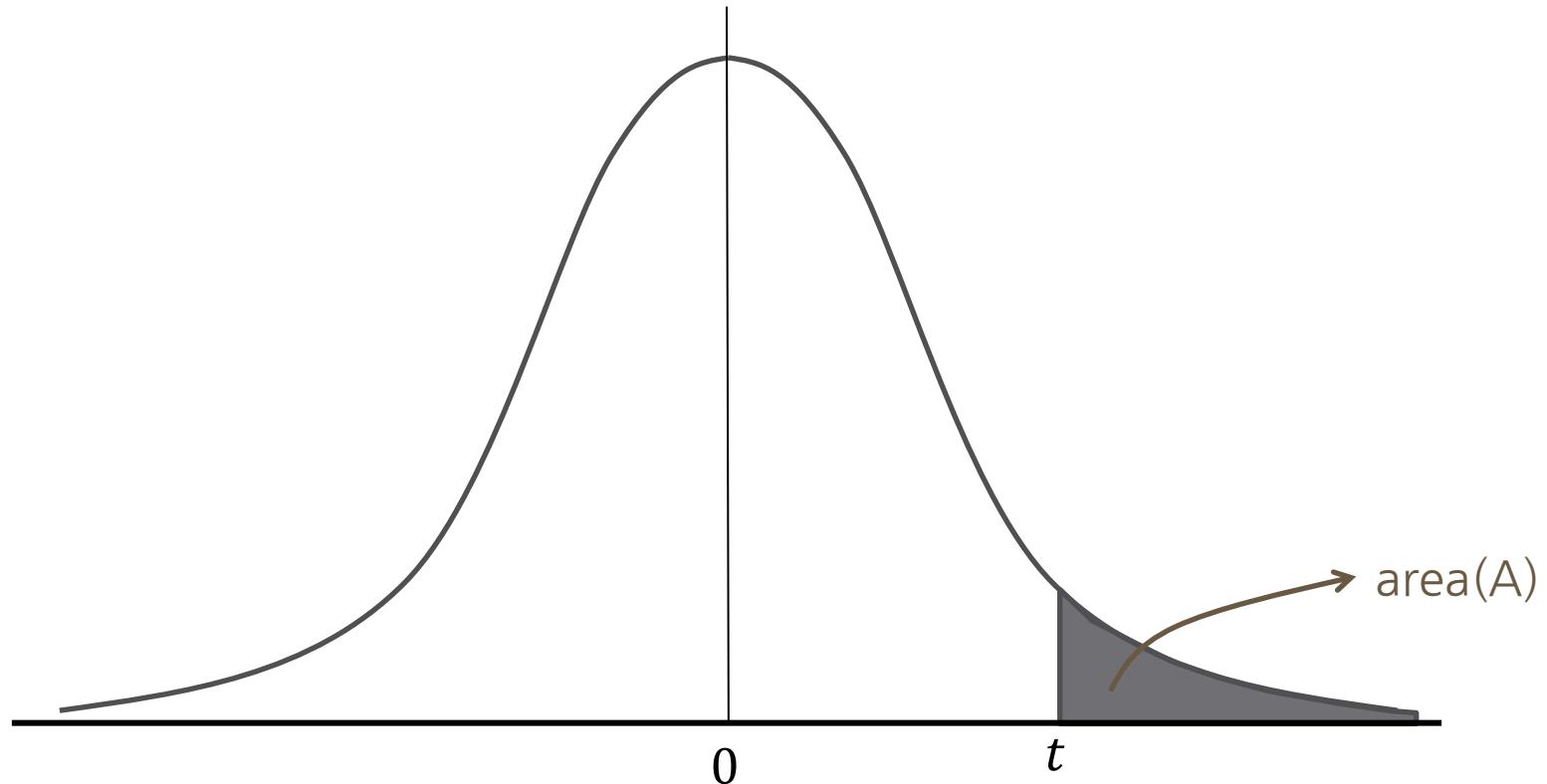
Test Concerning Regression Coefficients

- Test statistics of t -test follows student's t distribution with $n - p - 1$ degree of freedom
 - ▣ Probability density function of student's t distribution with different parameters(degree of freedom)

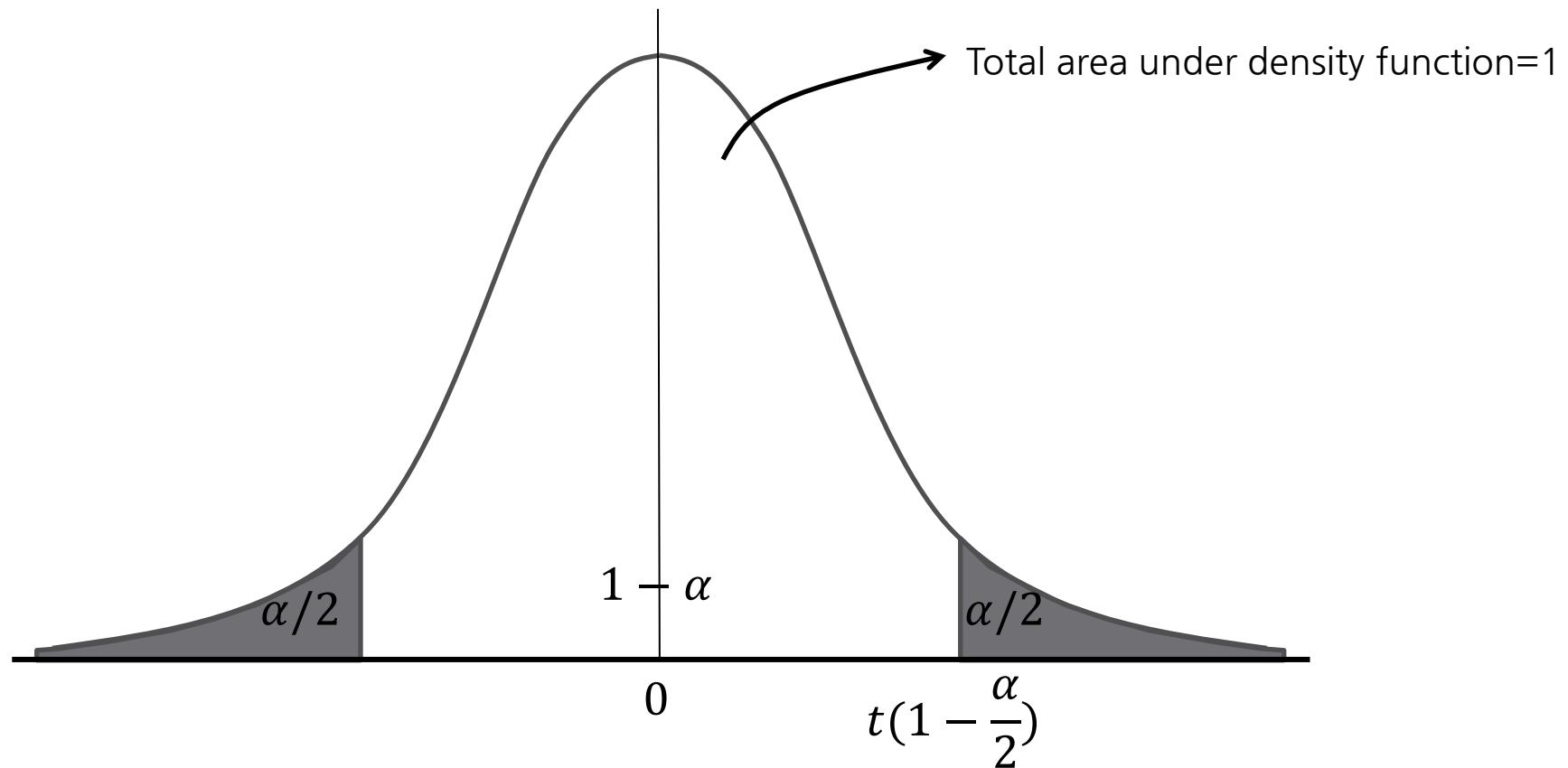


Test Concerning Regression Coefficients

- If $(\text{area under density function from } |t| \text{ to } \infty) < \frac{\alpha}{2}$
→ Reject null hypothesis → β_i is not zero
 - ▣ α is significance value
 - ▣ significance level is usually set to 0.1, 0.05
 - The higher significance level, the higher probability to reject null hypothesis



Test Concerning Regression Coefficients



Test Concerning Regression Coefficients

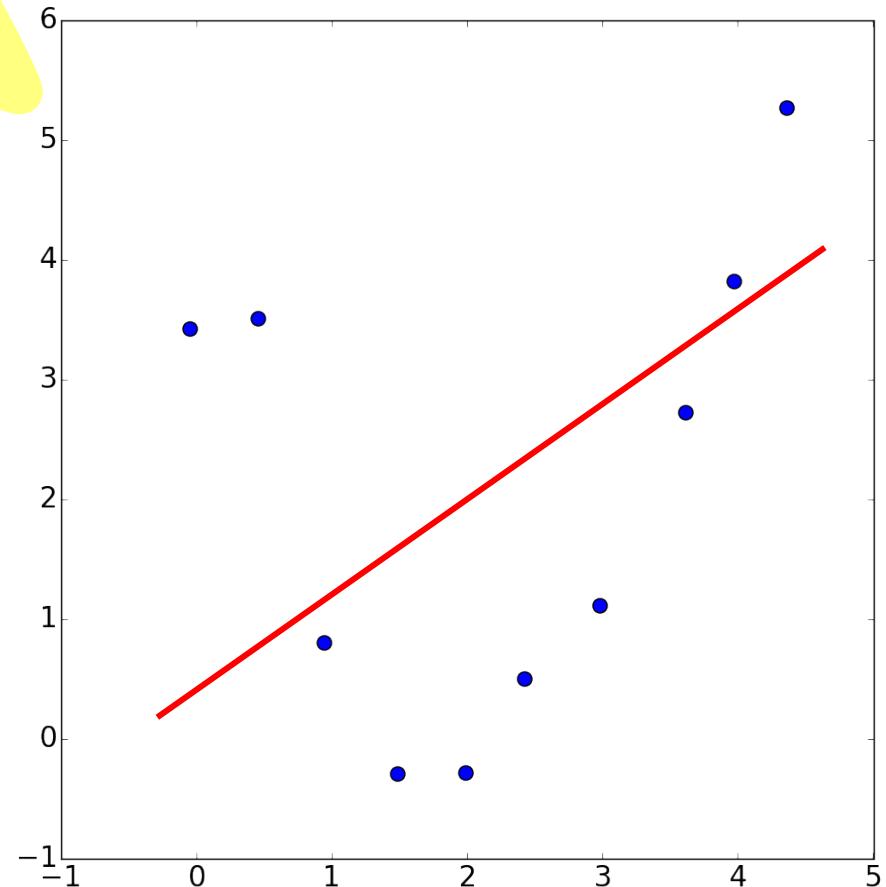
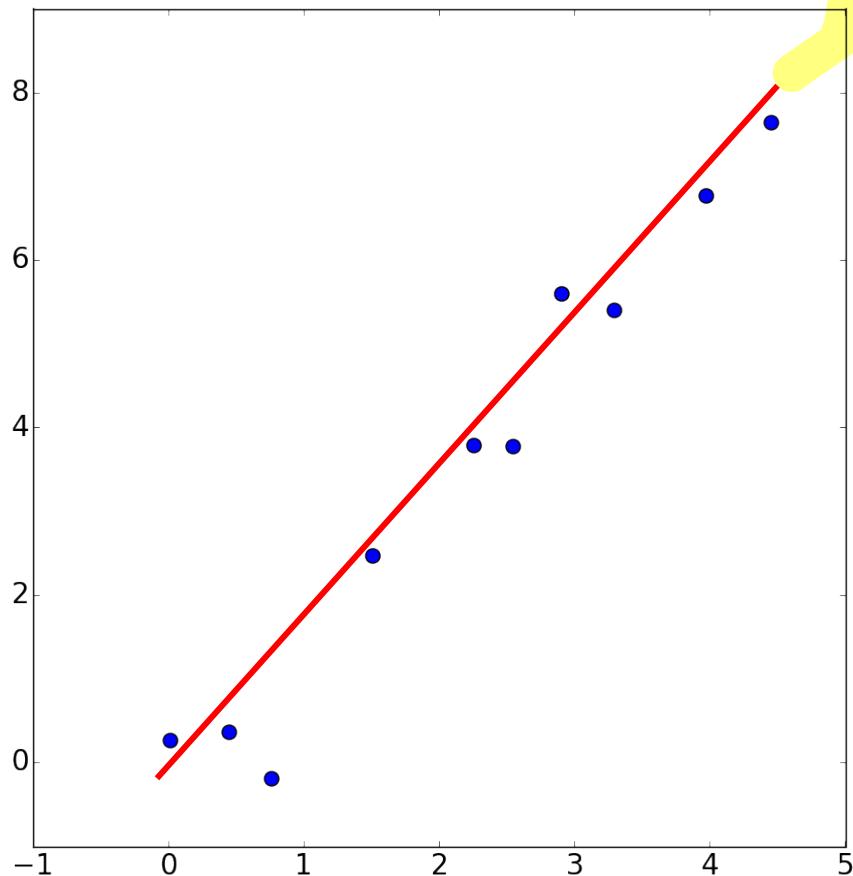
- How to calculate area?
 - ▣ Don't worry. There is pre-calculated table!

Student t-Table									
Alpha	0.250	0.200	0.150	0.100	0.050	0.025	0.010	0.005	0.0005
df									
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.656	636.578
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.600
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850

t value that area is 0.25 with 20 degree of freedom

Goodness-of-fit

- How to measure quantitatively performance of fitted models?
 - Calculate goodness-of-fit



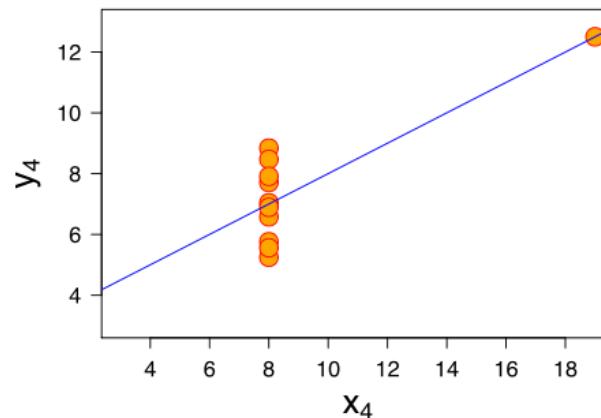
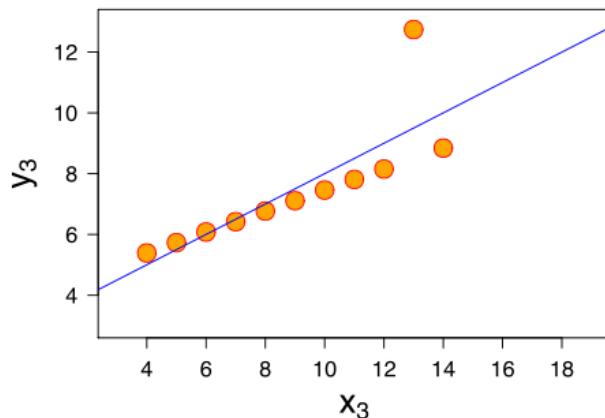
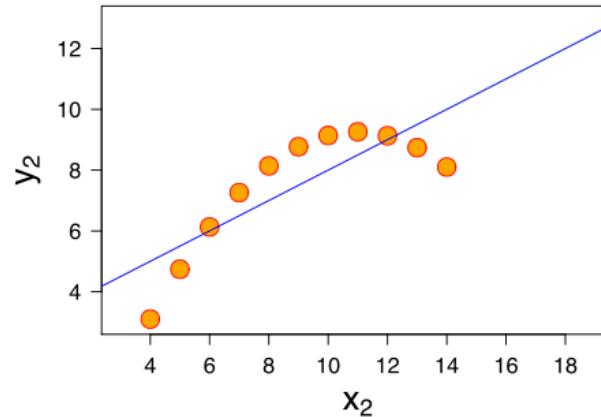
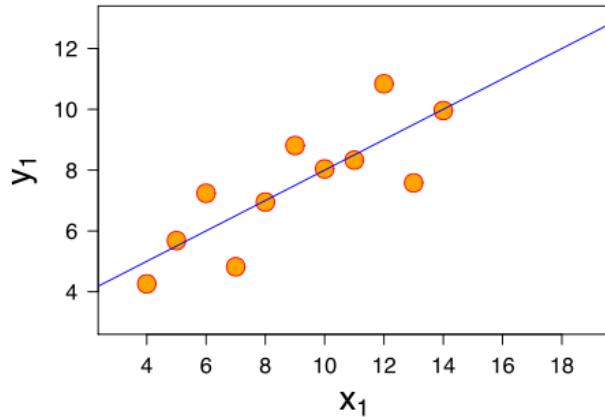
R^2

- Statistical measures for goodness-of-fit
 - R^2 ($0 \leq R^2 \leq 1$)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R^2 is NOT All-around Player

- Anscombe's quartet
 - ▣ The same linear regression line but are themselves very different.



Adjusted R^2

- Adding more input variables to the regression model increases R^2 and never reduce it
 - ▣ Tend to add more input variables to the model

Is always right to add more variables?

- Adjusted R^2

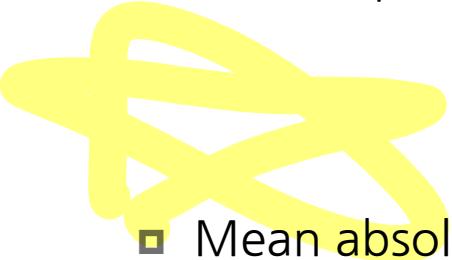
$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-1}{n-p-1} \right) (1 - R^2)$$

Depend on the number of input variables 

- ▣ Penalty on the number of input variable by $n - p - 1$
- ▣ Adjusted R^2 may actually become smaller when another input variable is introduced into the model

Performance metrics

- Functions to measure regression performance
 - ▣ Mean squared error


$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

- ▣ Mean absolute error

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

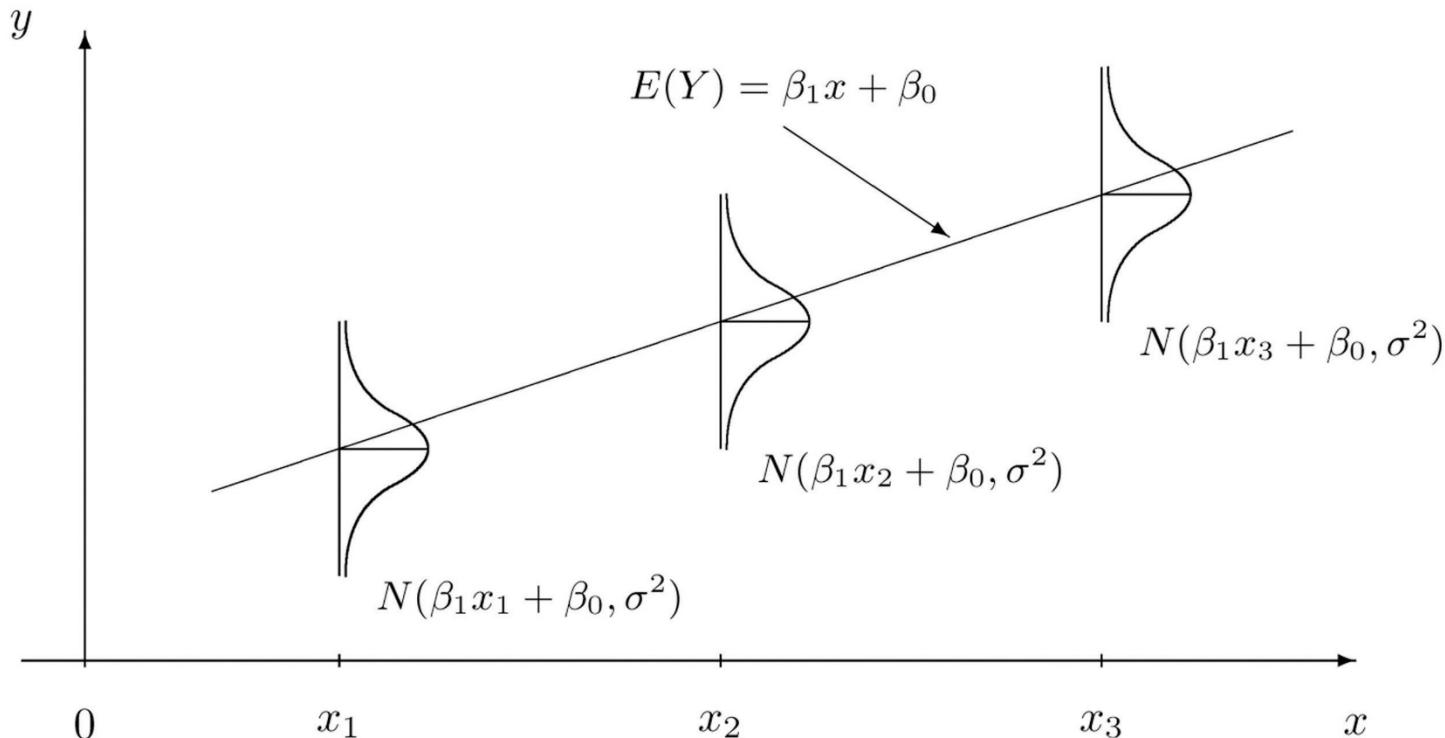
- ▣ Median absolute error

- robust to outliers

$$MedAE = \text{median}(|y_1 - \hat{y}_1|, \dots, |y_n - \hat{y}_n|)$$

Check Appropriateness of Linear Regression

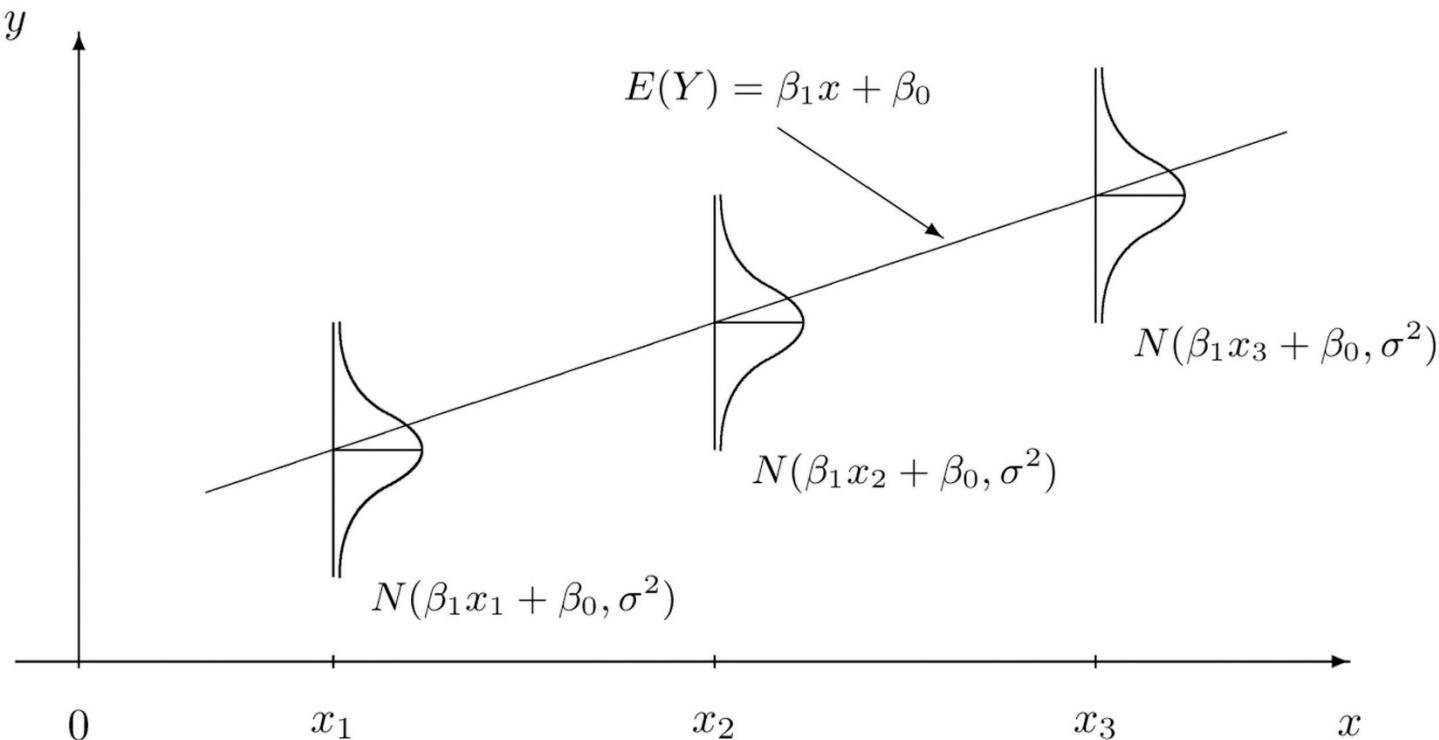
- Do you remember main assumptions of linear regression?



Main Assumption of Linear Regression

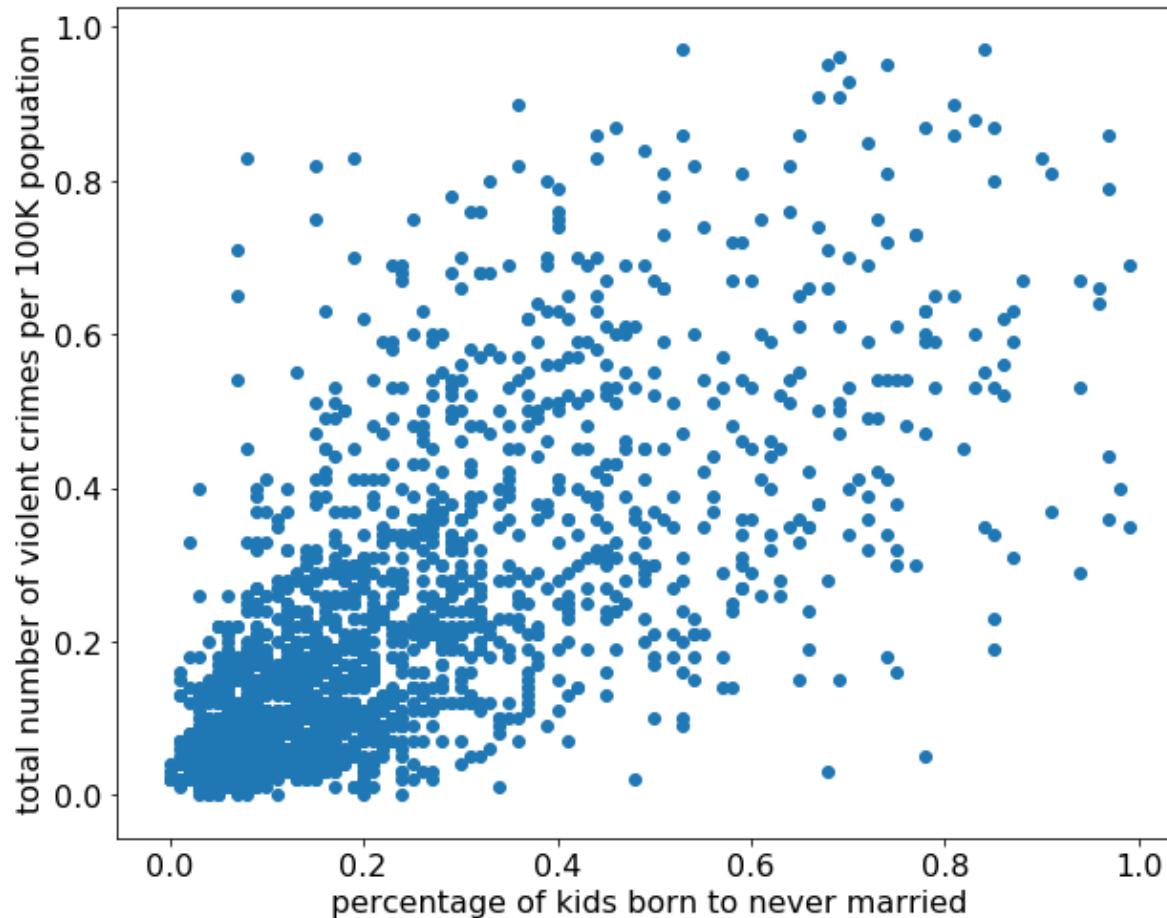
- Linear regression analysis makes several key assumptions

- Linear relationship
- Homoscedasticity
- Normality
- No or little multicollinearity



Check Appropriateness of Linear Regression

- **Linear relationship**
 - ▣ Check relationships between input variables and a responsive variable



Relationships between Input Variables

- If some of input variables are highly correlated, regression coefficients are unstable

	1	2	3	4	5	6	7	8	9	10
x_1	98	120	140	195	181	128	107	106	88	77
x_2	24	35	36	51	45	30	29	24	22	19
x_3	21	11	31	42	57	82	67	13	55	36

- Correlation matrix

$$corr = \begin{bmatrix} 1.00 & 0.98 & 0.17 \\ 0.98 & 1.00 & 0.11 \\ 0.17 & 0.11 & 1.00 \end{bmatrix}$$

- x_1 and x_2 are highly correlated

※ Covariance

- Variance of a random variable X is the expected value of the squared deviation from the mean ($\mu = \mathbb{E}[X]$)

$$Var(X) = \mathbb{E}[(X - \mu)^2]$$

- Sample variance is calculated by

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n - 1)}$$

- Covariance is a measure of how much two random variables change together

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

- Variance is the covariance of a random variable with itself

$$Var(X) = Cov(X, X)$$

- Sample covariance is calculated by

$$q_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

※ Correlation

- Any statistical relationship, whether causal or not, between two random variables or bivariate data
 - ▣ Pearson's correlation coefficient
 - The most popular correlation

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Sample correlation coefficient

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^n (x_i - \bar{x})^2)(\sum_{i=1}^n (y_i - \bar{y})^2)}}$$

Relationships between Input Variables

- Two difference cases

	1	2	3	4	5	6	7	8	9	10
y_1	295	310	404	567	574	532	442	283	366	285
y_2	282	311	402	581	573	523	446	277	374	274

- Output values of two cases are quite similar
- Regression coefficient for y_1 and y_2

$$\text{Case 1: } [\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3] = [2.16 \quad 0.14 \quad 2.88]$$
$$\text{Case 2: } [\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3] = [1.73 \quad 2.18 \quad 2.97]$$

- Because x_1 and x_2 are highly correlated, explained variance by x_2 is also explained by $x_1 \rightarrow$ Coefficient of x_2 is quite unstable

Why This Situation Happens

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- To estimate regression coefficients, inverse matrix of $\mathbf{X}^T \mathbf{X}$ should be calculated
 - Ill-conditioned matrices
 - If a small change in the coefficient matrix results in a large change in the solution, the coefficient matrix is called ill-conditioned
- $$\begin{cases} x + y = 2 \\ x + 1.001y = 2 \end{cases} \quad \text{and} \quad \begin{cases} x + y = 2 \\ x + 1.001y = 2.001 \end{cases}$$
- Left: $x = 2, y = 0$
 - Right: $x = 1, y = 1$

$\begin{bmatrix} 1 & 1 \\ 1 & 1.001 \end{bmatrix}$ is ill-conditioned

Variance Inflation Factor



- Variance inflation factor(VIF) quantifies the severity of multicollinearity in a least square method

[Multicollinearity]

A Phenomenon in which two or more input variables in a multiple regression model are highly correlated

→ In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data

- Variance of estimated coefficients for j – th input variable

$$var(\hat{\beta}_j) = se^2(\hat{\beta}_j) = [MSE(\mathbf{X}^T \mathbf{X})^{-1}]_{j,j} = \frac{MSE}{(n - 1)se^2(x_j)} \frac{1}{1 - R_j^2}$$

- R_j^2 is the R^2 for the regression of the x_j on the other input variables

- VIF

$$\frac{1}{1 - R_j^2}$$

Variance Inflation Factor

- Calculate VIF
 - ▣ Step 1) Apply least square method to regression problem that i -th input variable is regressed by the remained input variables

$$x_i = \alpha_1 x_1 + \cdots + \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + \cdots + \alpha_p x_p + \alpha_0 + \epsilon$$

- ▣ Step 2) Calculate R^2 for above regression problem and set the value as R_i^2
- ▣ Step 3) Calculate VIF from R_i^2

$$VIF = \frac{1}{1 - R_i^2}$$

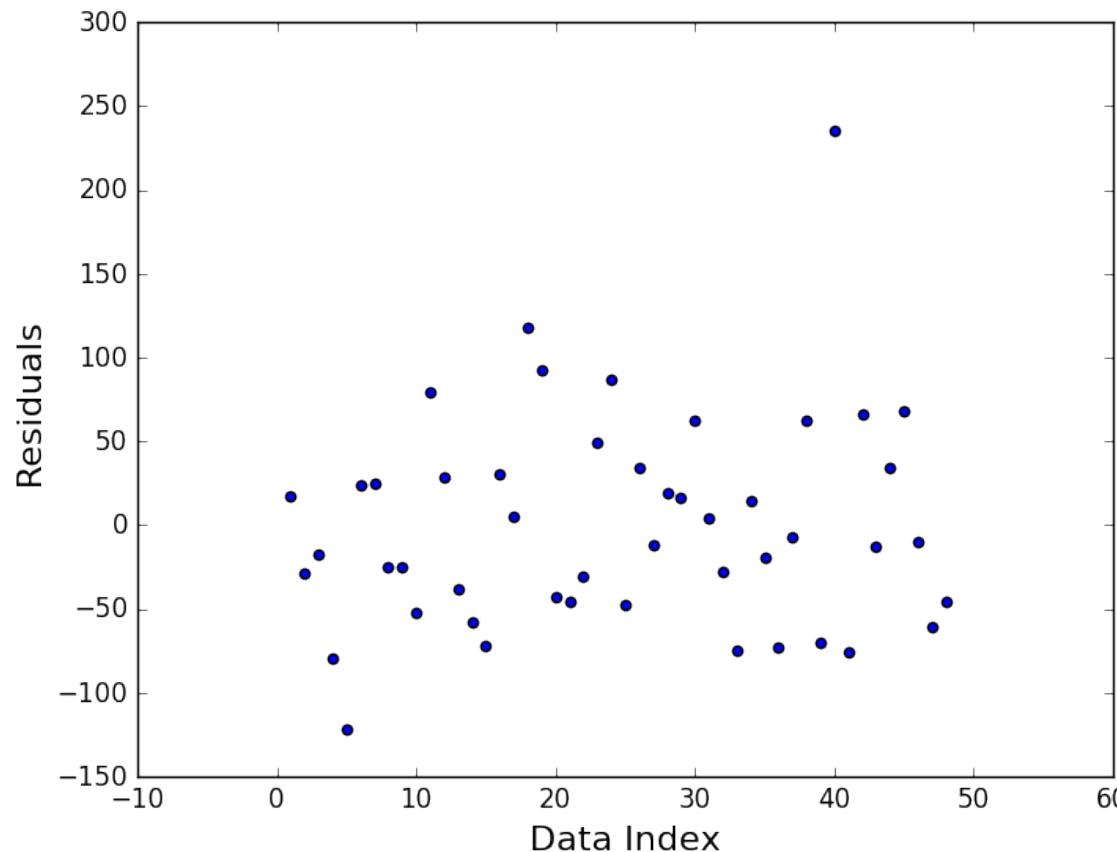
- A rule of thumb is that if $VIF(\hat{\beta}_i) > 10$ then multicollinearity is high
 - ▣ In this case, do not use x_i as explanatory variable to estimate output

Check Appropriateness of Linear Regression

□ Normality

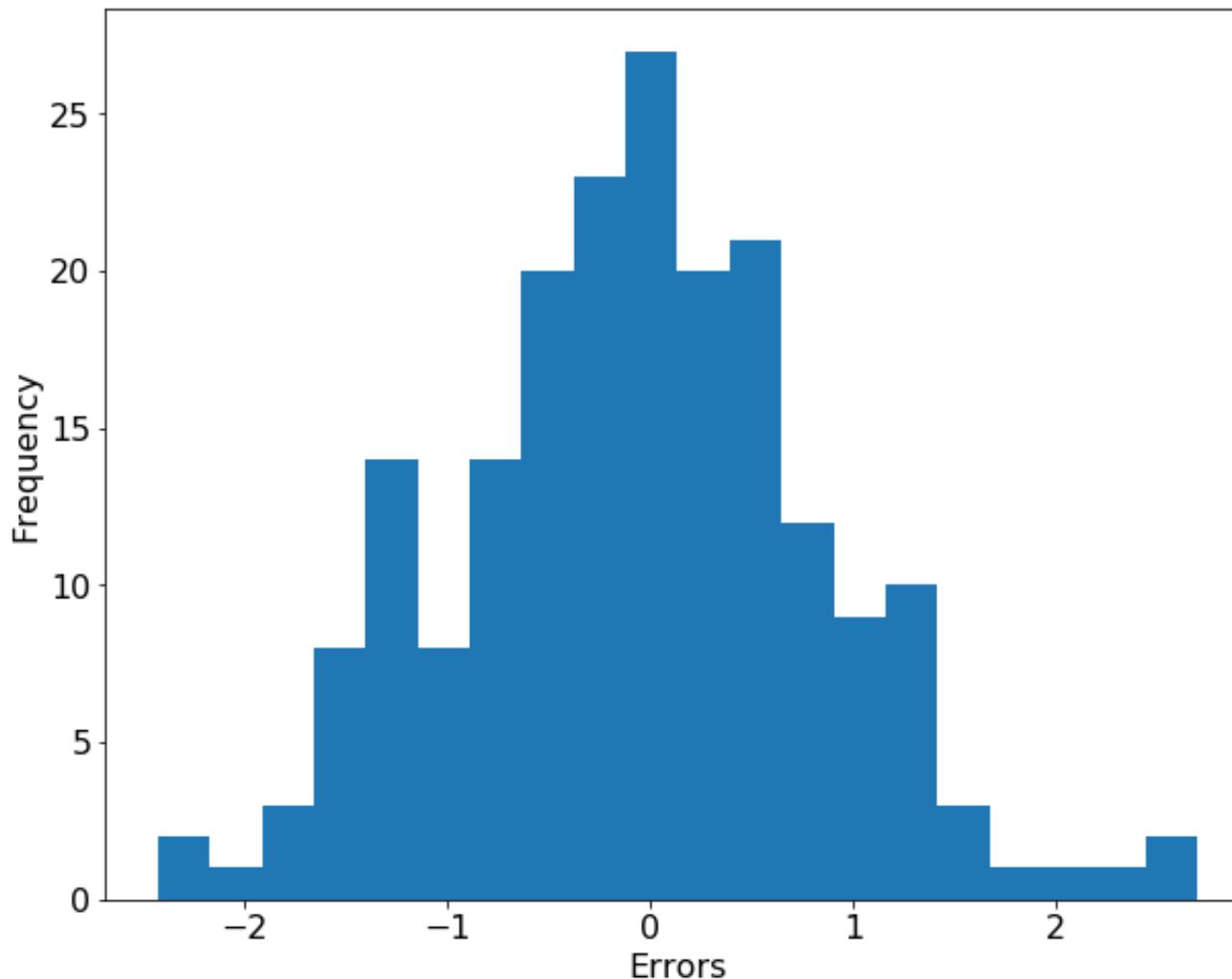
- Errors should follow normal distribution
- Calculate errors (residuals) and check normality

$$e_i = y_i - \hat{y}_i$$



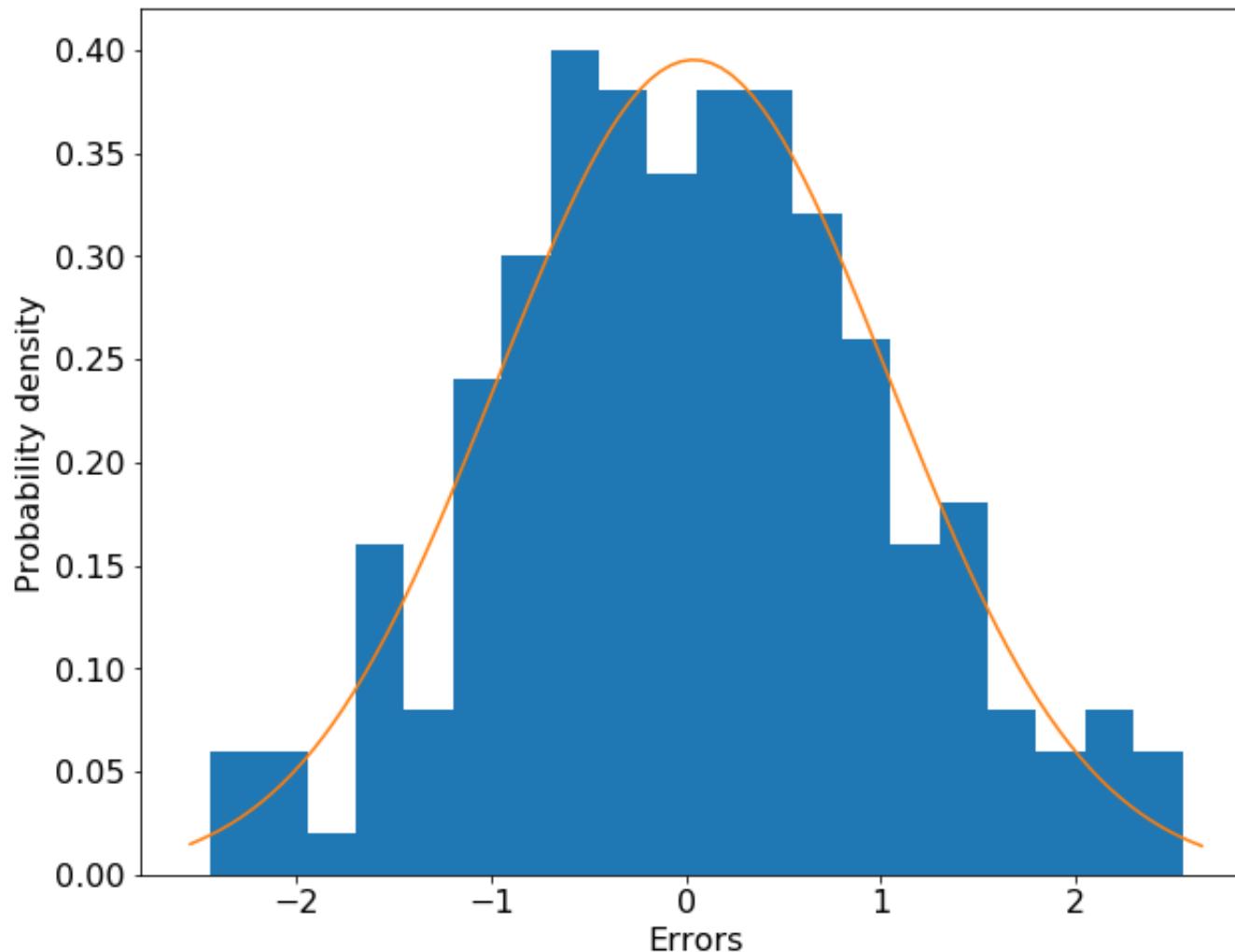
Check Appropriateness of Linear Regression

- Histogram



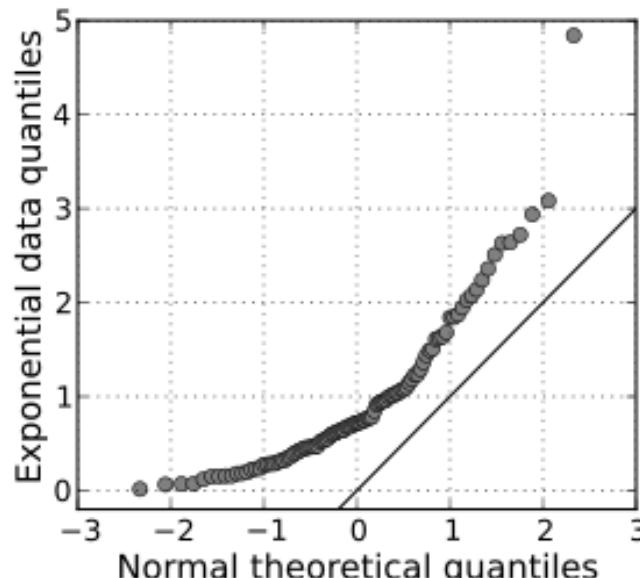
Check Appropriateness of Linear Regression

- Histogram

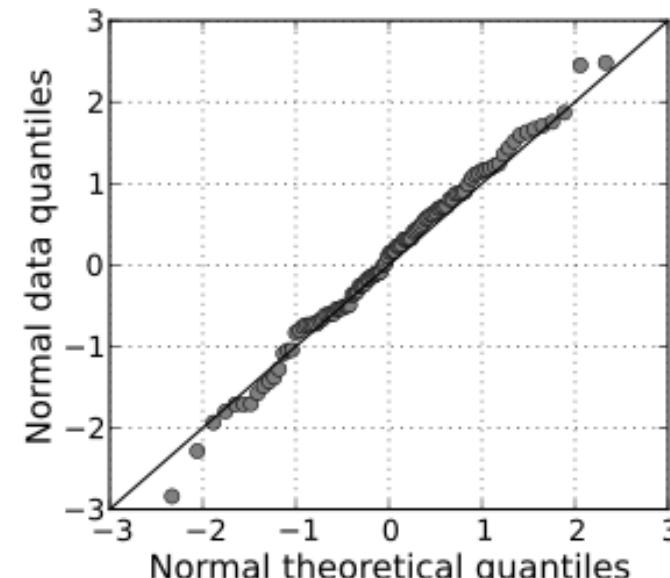


Check Appropriateness of Linear Regression

- Q-Q plot *Normality*
 - A probability plot, which is a graphical method for **comparing two probability distributions** by plotting their quantiles against each other
 - Quantiles are cutpoints dividing a set of observations into equal sized groups
 - q -Quantiles are values that partition a finite set of values into q subsets of (nearly) equal sizes
 - Median is 2-quartile, 0.5 quantile and 50 percentile

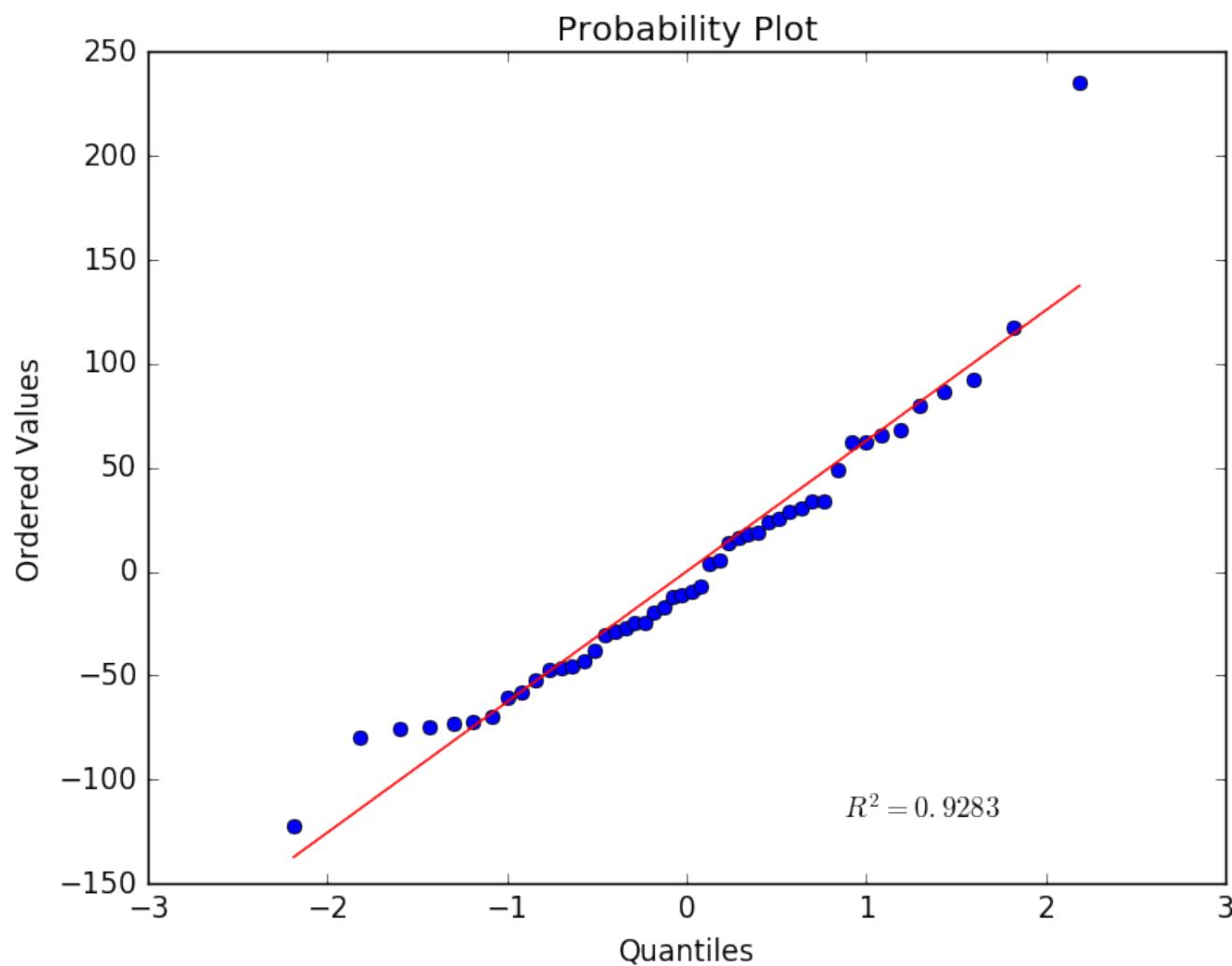


$X \sim Exp(1)$



$X \sim N(0,1)$

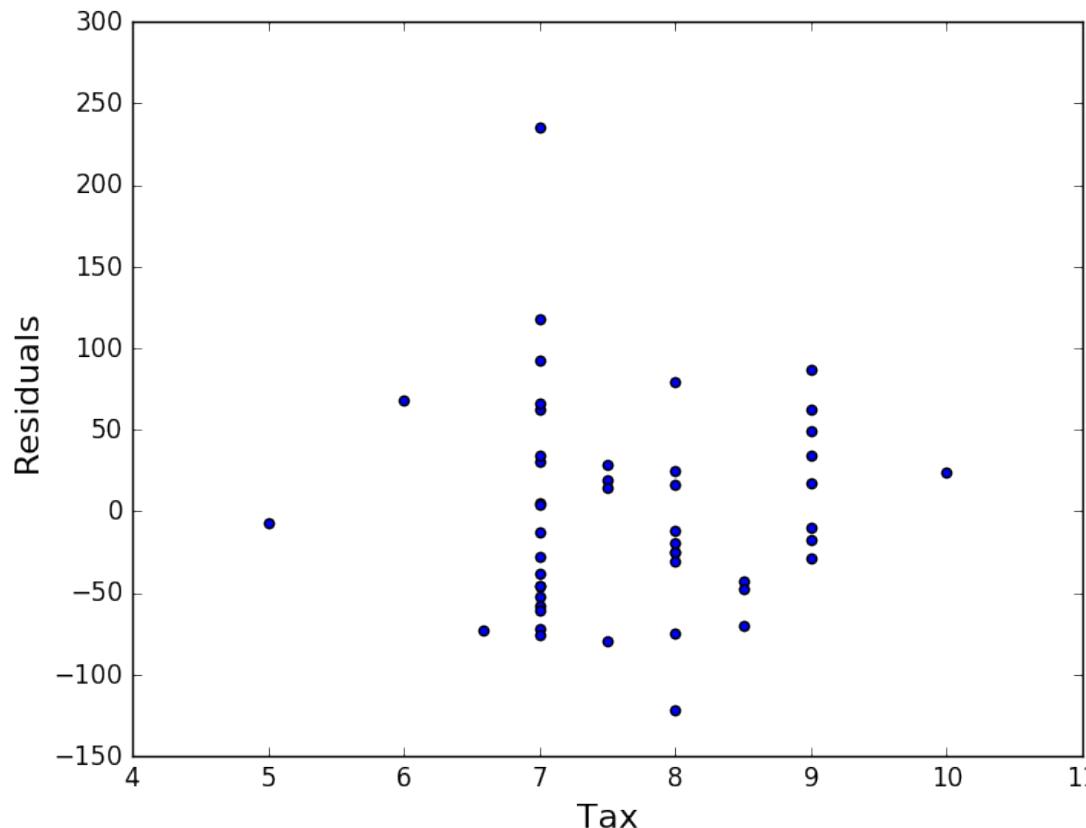
Q-Q Plot



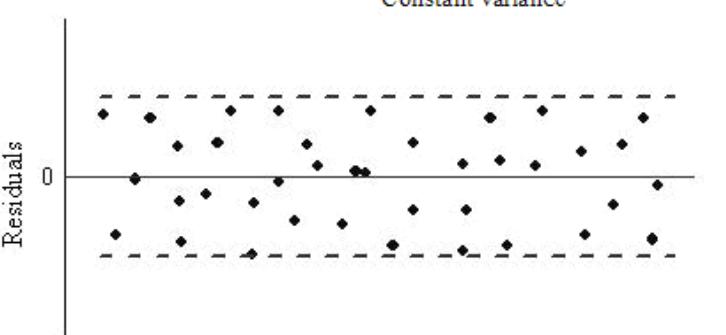
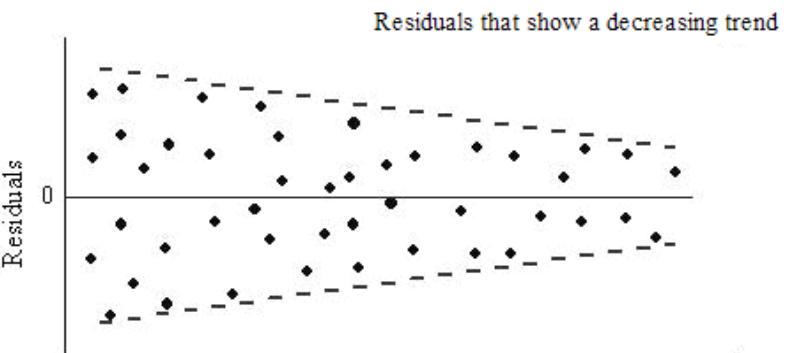
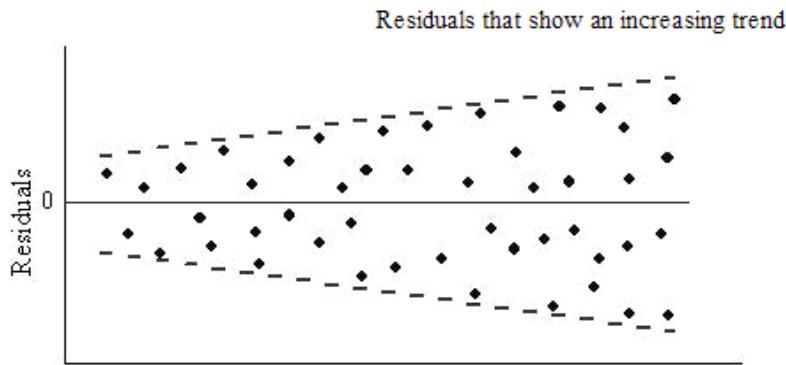
Check Appropriateness of Linear Regression

□ Homoscedasticity↔Heteroscedasticity

- Check whether all random variables in the sequence or vector have the same finite variance

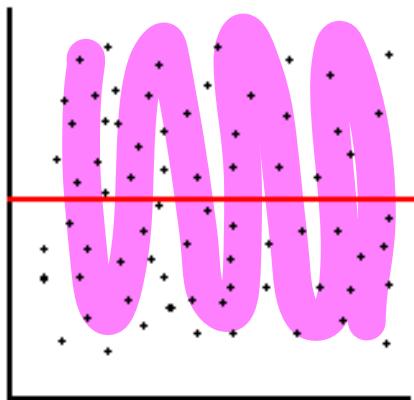


Check Appropriateness of Linear Regression

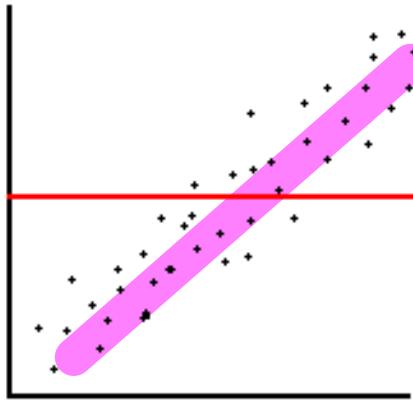


Check Appropriateness of Linear Regression

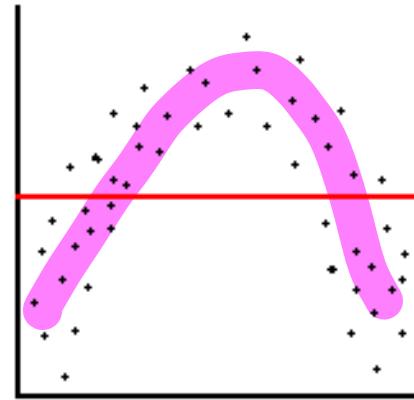
- Residual plot



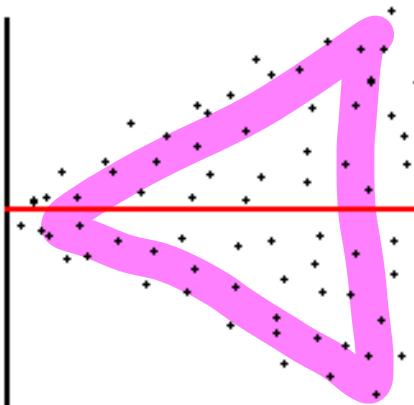
(a) Unbiased and Homoscedastic



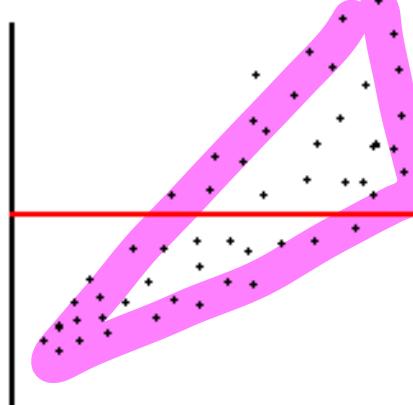
(b) Biased and Homoscedastic



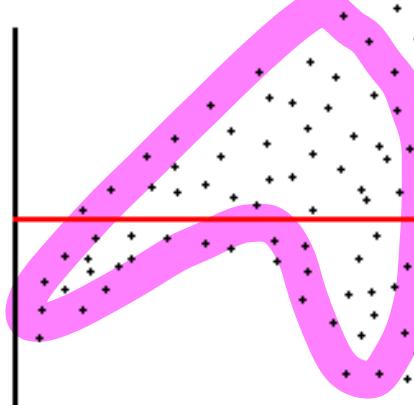
(c) Biased and Homoscedastic



(d) Unbiased and Heteroscedastic



(e) Biased and Heteroscedastic

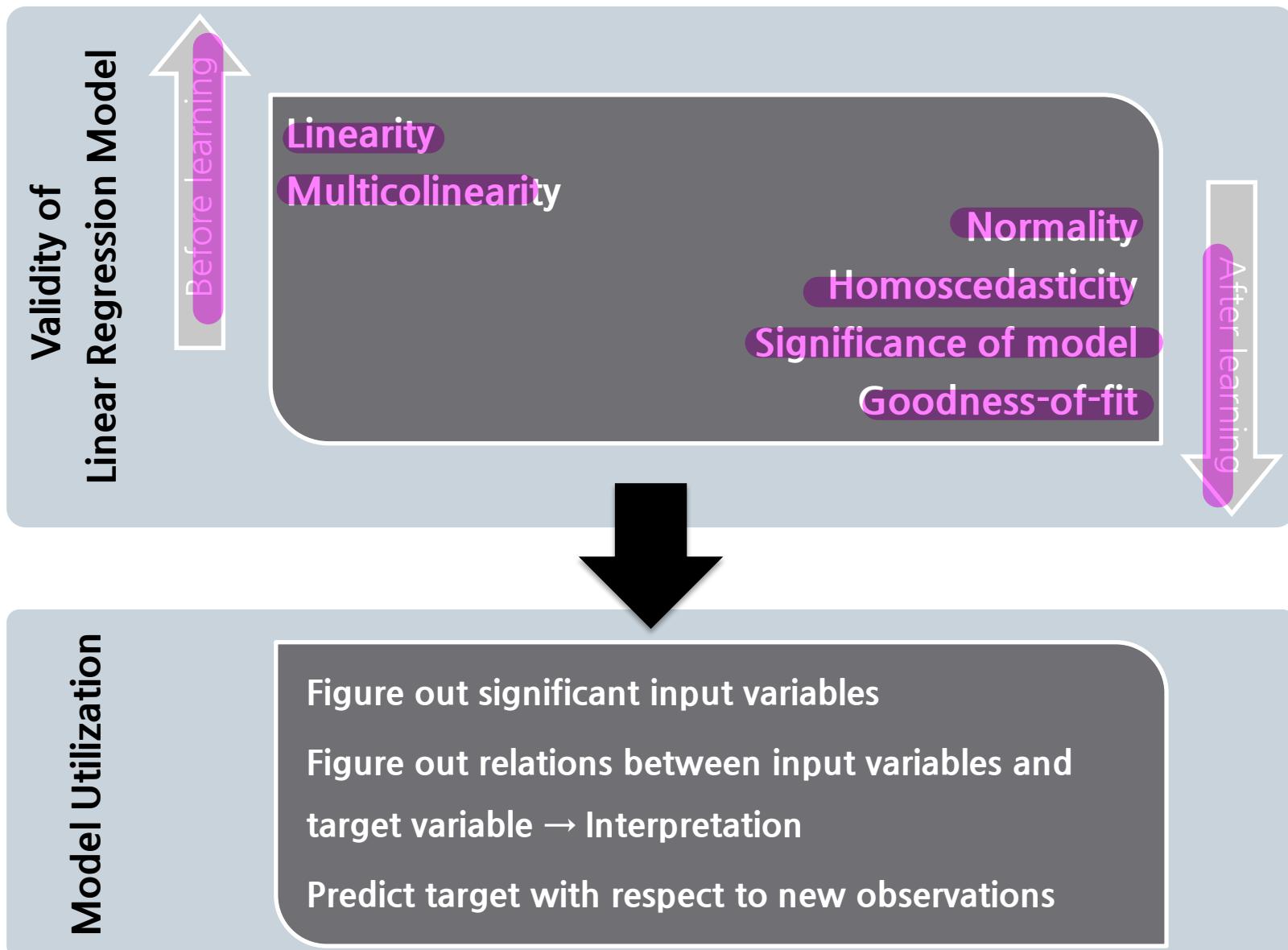


(f) Biased and Heteroscedastic

Interpretation & Prediction

- If the fitted regression model is appropriate and significant you can use the model for future use
 - ▣ Linear regression models have strength in interpretation
 - Each coefficient explain relationship between each explanatory variable and the target variable
 - ▣ Based on the fitted model, predict the target on test samples

Overall Process for Linear Regression



Feature Scaling

- Predict consumption of petrol
 - ▣ Linear model by least square method

$$y = -34.8x_1 - 0.0666x_2 - 0.002x_3 + 1336x_4 + 377.3$$

Petrol Tax(\$)	Average Income (\$)	Paved Highways (miles)	Proportion of population with driver's license	Consumption of petrol (M of gallons)
9	3571	1976	0.525	541
9	4092	1250	0.572	524
9	3865	1586	0.58	561
7.5	4870	2351	0.529	414
...

How about changing scale of variable?

Feature Scaling

- Change unit of paved highways from mile to cm

$$1 \text{ mile} = 160934.4 \text{ cm}$$

Petrol Tax(\$)	Average Income (\$)	Paved Highways (cm)	Proportion of population with driver's license	Consumption of petrol (M of gallons)
9	3571	31683974.4	0.525	541
9	4092	20043000	0.572	524
9	3865	25430558.4	0.58	561
7.5	4870	37696874.4	0.529	414
...

- Linear regression on new data

$$y = -34.8x_1 - 0.0666x_2 - 1.5 \times 10^{-7}x_3 + 1336x_4 + 377.3$$

Feature Scaling

- Scale change only affects on the changed variable
 - ▣ Coefficients of other variables are not changed
 - ▣ If variable x is replaced with ax , coefficient of x , β by linear regression is changed to β/a
 - ▣ If scale of certain variable is too large, coefficient of the variable might be too small
→ It is better to change scale

Variable Transformation

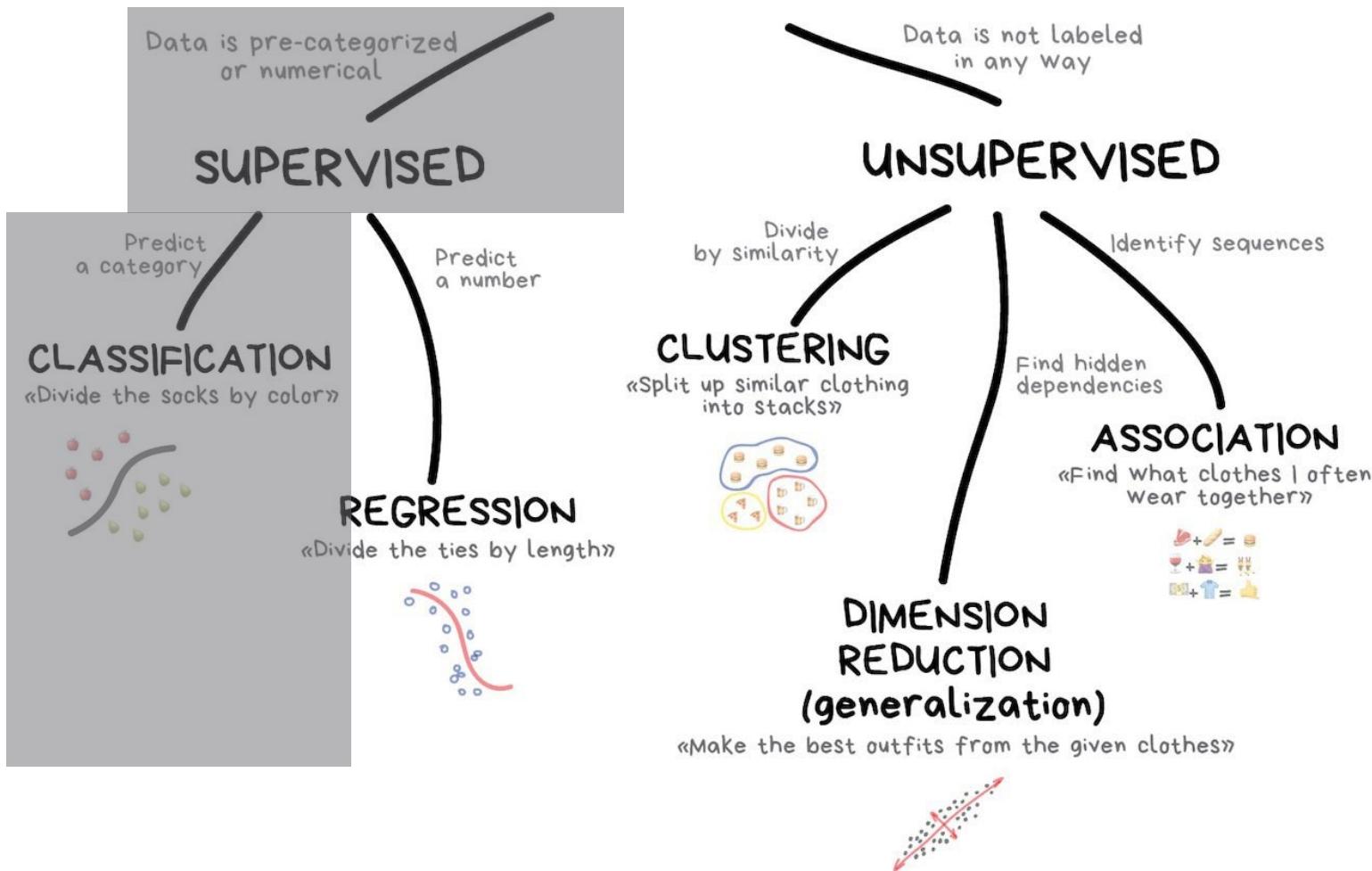
- Linear regression algorithm is quite simple, but it can be extended using transformation
 - $x \rightarrow x^2$
 - $x \rightarrow \log x$
 - $x \rightarrow \sqrt{x}$

LOGISTIC REGRESSION

Week06

Topics Covered in This Class

CLASSICAL MACHINE LEARNING



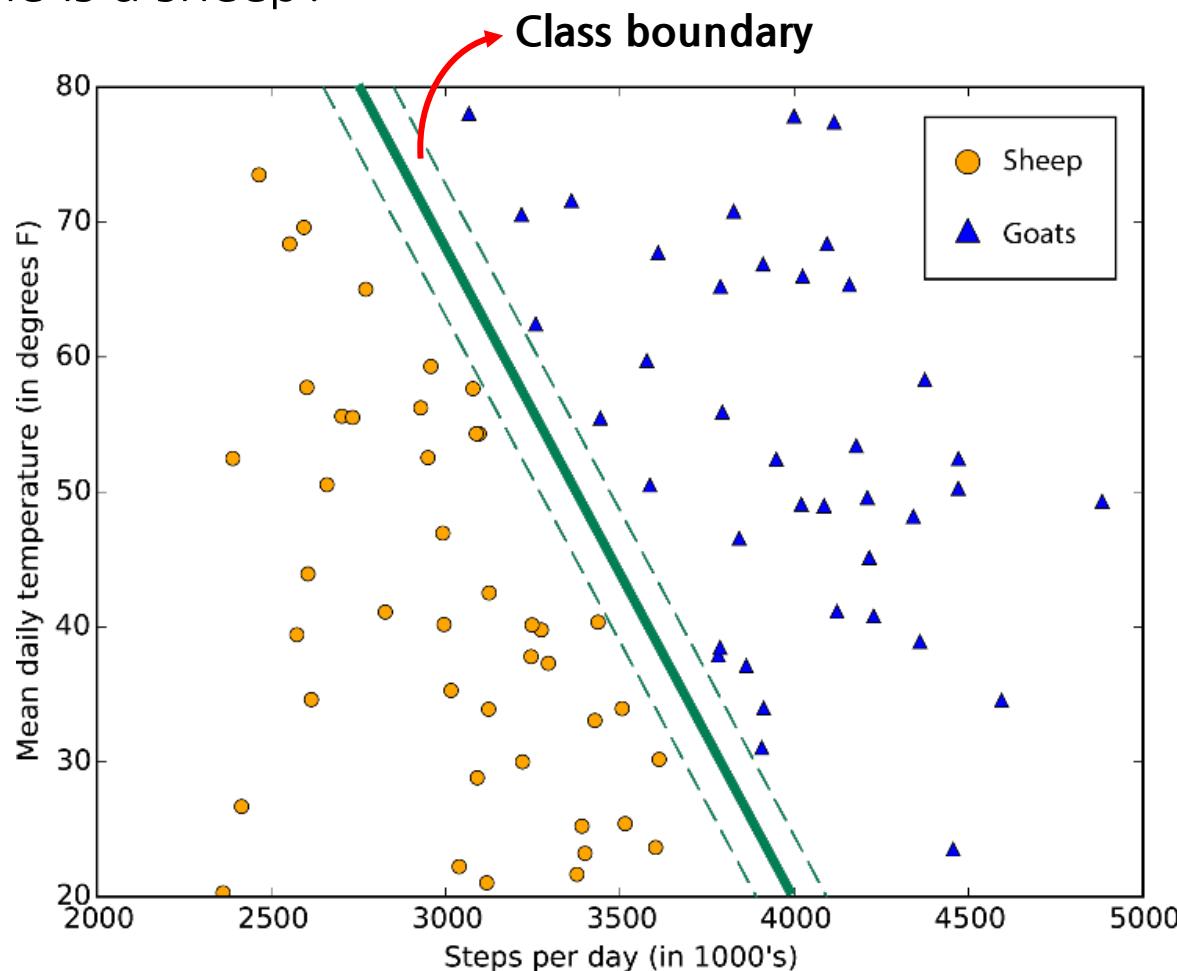
Logistic Regression

Supervised: Classification

- Classification problem
 - ▣ Output is categorical variable
 - Spam/Non Spam
 - Male/Female
 - Long/Medium/Short
 - O/X
 - Binary classification
 - ▣ The number of categories is 2
 - ▣ Generally, these two categories are denoted as 0 and 1
 - 0 and 1 are not integer in this case
 - Multi-class classification
 - ▣ More than two classes
- $$y \in \{0,1\}$$
- $$y \in \{1,2,\dots,C\}, \quad C > 2$$

Supervised: Classification

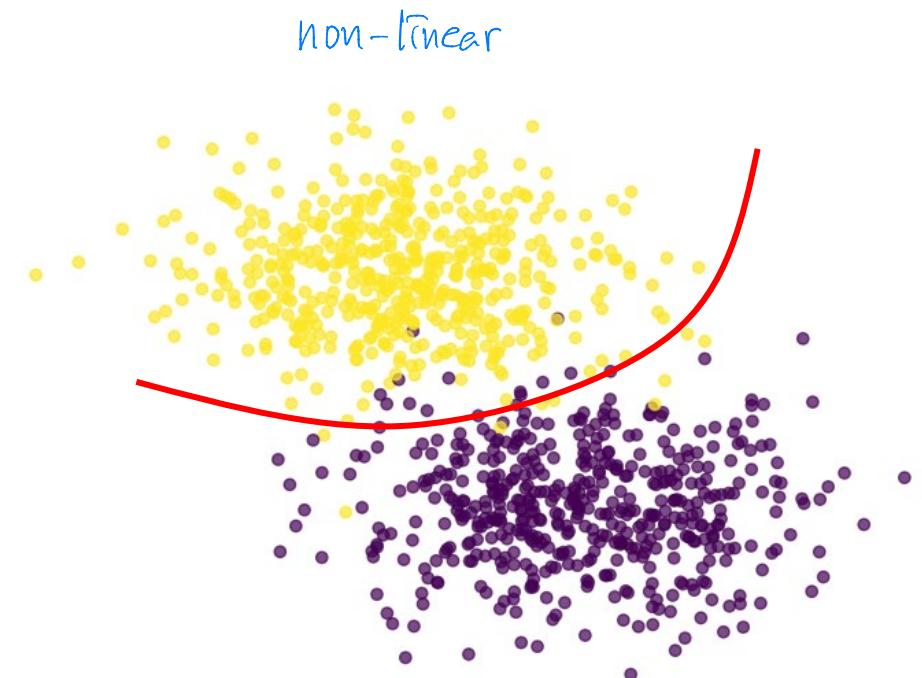
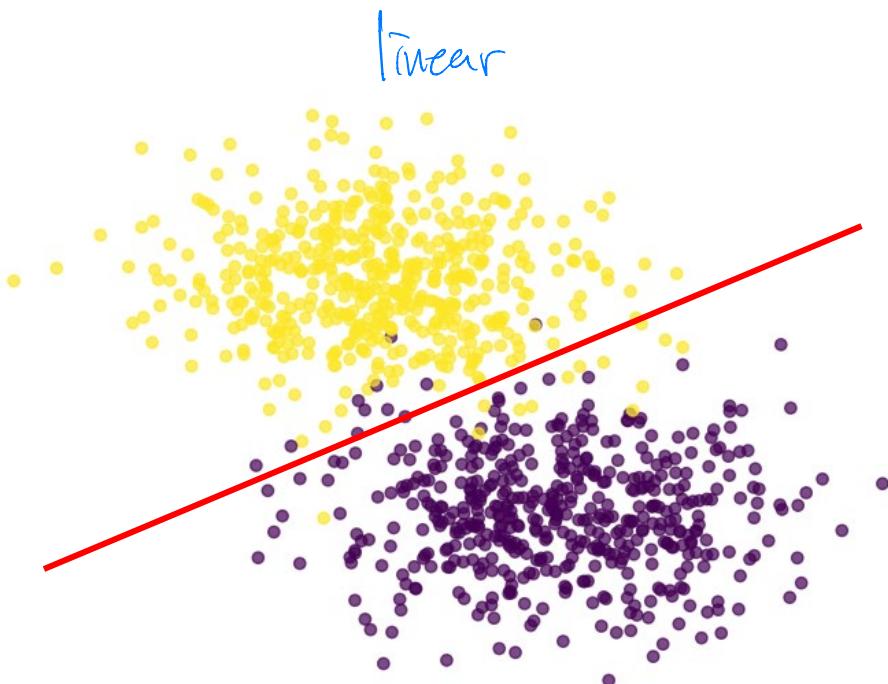
- Which one is a sheep?



The Decision Boundary of Classifiers

- Decision boundary

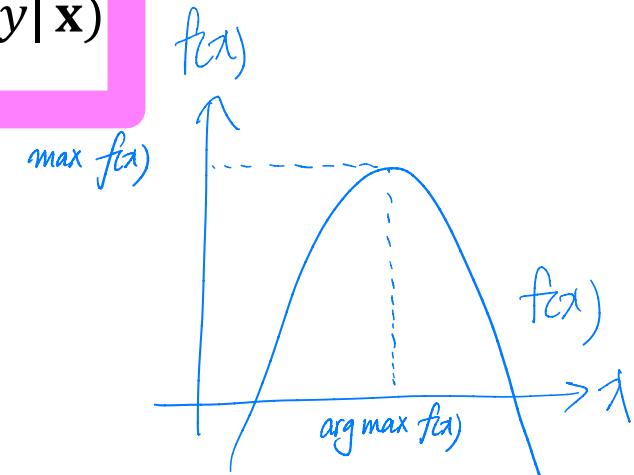
$$y = f(X), \quad y \in \{1, 2, \dots, C\}$$



Types of Classifiers

- A classifier is a function that assigns to a sample, \mathbf{x} a class label \hat{y}
$$\hat{y} = f(\mathbf{x})$$
- A **probabilistic classifier** obtains **conditional distributions** $\Pr(Y|\mathbf{x})$, meaning that for a given $\mathbf{x} \in \mathcal{X}$, they assign probabilities to all $y \in Y$
 - Hard classification

$$\hat{y} = \arg \max_y \Pr(Y = y | \mathbf{x})$$



Logistic Regression

- Logistic regression
 - ▣ Regression model where the dependent variable is categorical
 - ▣ The probabilities of possible outcomes are modeled using explanatory variables

$$f(x) = P(Y|X)$$

- $0 \leq f(x) \leq 1$

How can we ensure that $f(x)$ remains within [0,1]?

$$P(Y=0|X) + P(Y=1|X) = 1$$

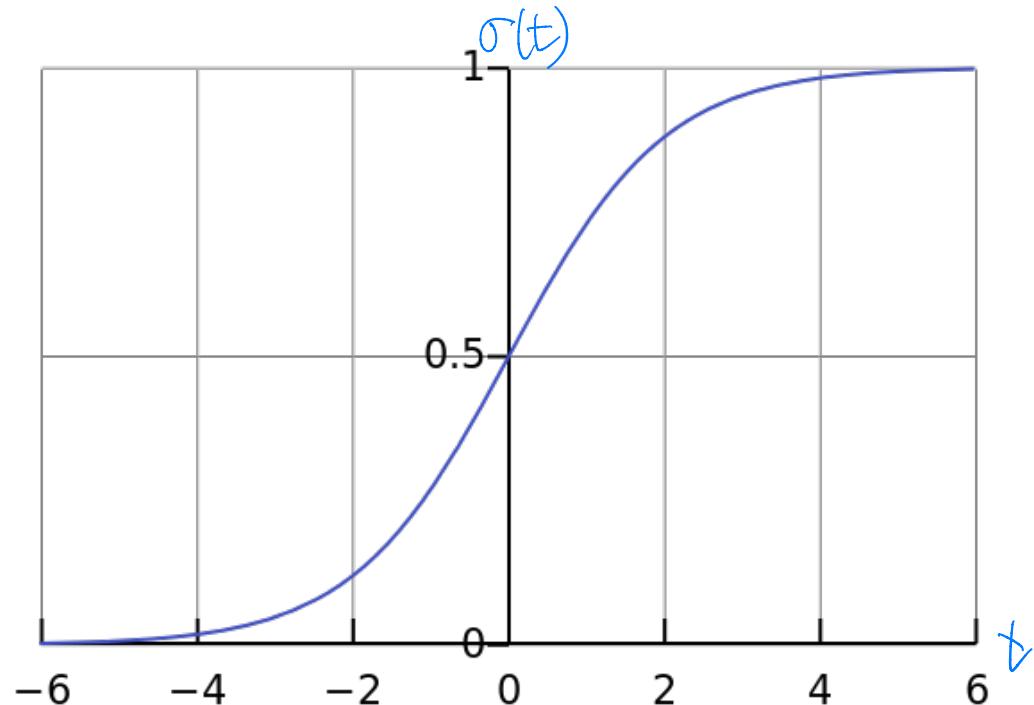
Logistic Regression: Logistic function

- Logistic function is the function that can take an input with any value from negative to positive infinity, whereas the output always takes values between zero and one

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

$$\lim_{t \rightarrow -\infty} \sigma(t) = 0$$

$$\lim_{t \rightarrow \infty} \sigma(t) = 1$$



- In logistic regression, t is determined by explanatory variables

Logistic Regression

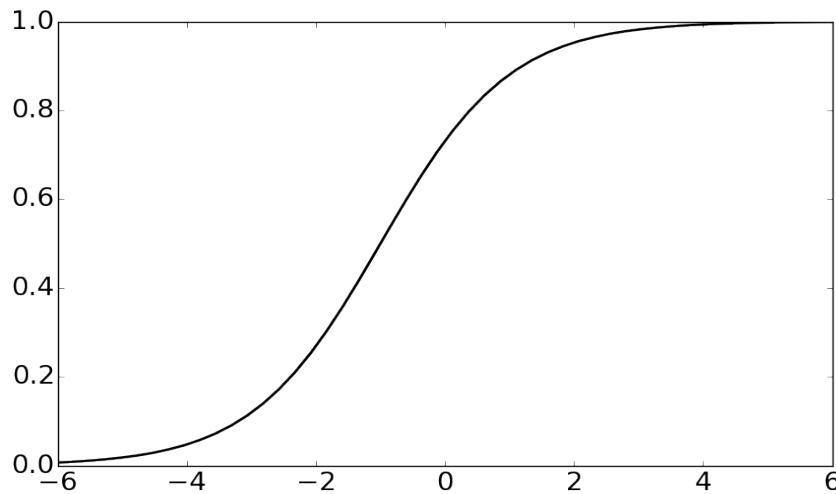
- t is determined by linear combination of explanatory variables

$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (-\infty, \infty)$$

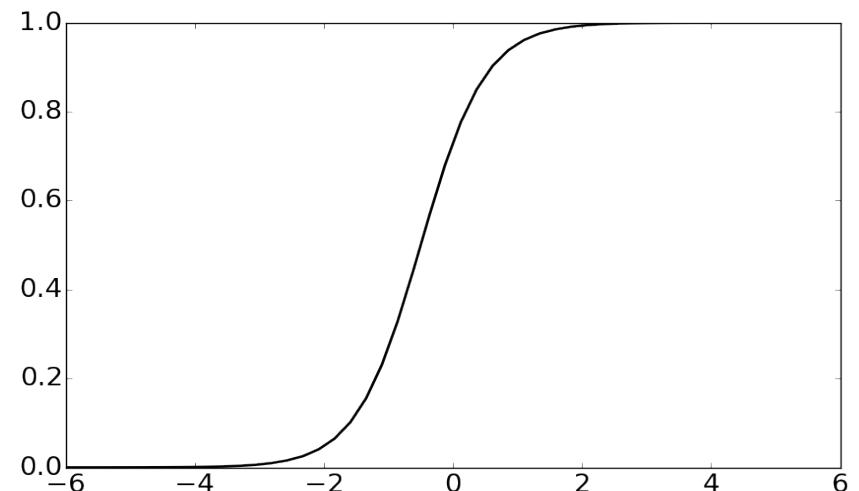


$$f(x) = P(Y = 1 | \mathbf{x}) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p}} \quad [0, 1]$$

$$t = 1 + x$$

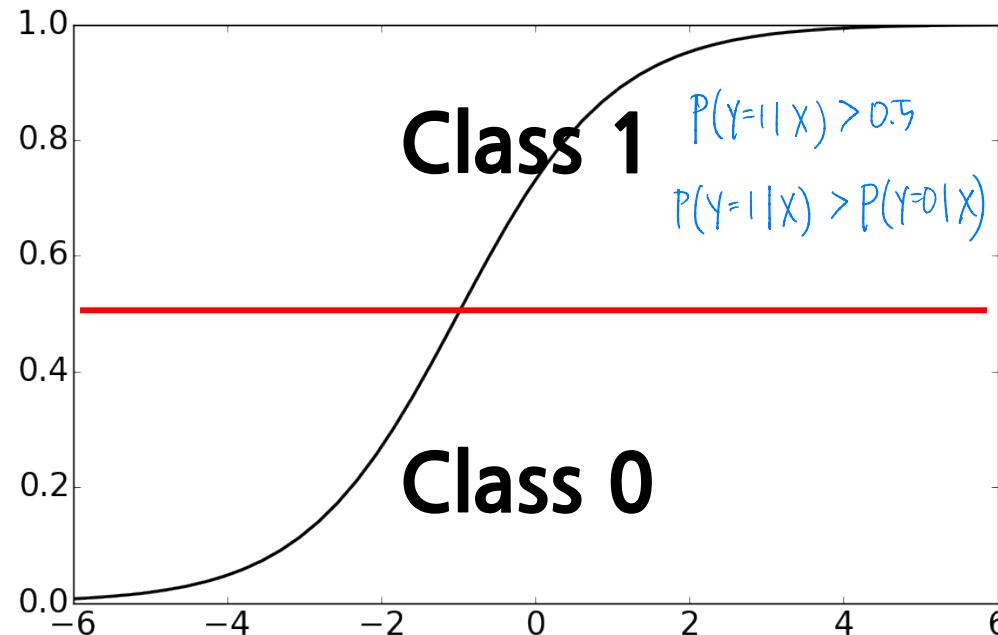


$$t = 1 + 2x$$



Logistic Regression

- Determine class
 - ▣ Set class boundary
 - Without any prior knowledge about class, set 0.5



- If you have some knowledge about class distribution, class boundary can be determined based on the knowledge

Logistic Regression: Parameter Estimation

$$f(x) = P(Y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p}}$$

- Unknown parameters
 - $\beta_0, \beta_1, \dots, \beta_p$
- Logistic regression should estimate $\beta_0, \beta_1, \dots, \beta_p$ based on the given observations

Maximum Likelihood Estimation



- Maximum likelihood estimation (MLE)
 - ▣ Method of estimating the parameters of statistical model
 - ▣ Given a statistical model, maximize likelihood
- Example of maximum likelihood estimation
 - ▣ Suppose that data set $D = \{x_1, x_2, \dots, x_n\}$ consists of n independent and identically distributed(iid) samples coming from a distribution with an unknown probability density function $f(x)$
 - ▣ Assume $f(x)$ belongs to a certain type of distributions with parameters θ
 - ▣ Joint density function for all observations

$$f(x_1, x_2, \dots, x_n | \theta) = f(x_1 | \theta) \times f(x_2 | \theta) \times \cdots \times f(x_n | \theta)$$

because x_i is iid sample

- ▣ Likelihood

$$\mathcal{L}(\theta; x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta)$$

Likelihood

- Imagine the situation that a ball is drawn from the bag consisting of three blue balls and five white balls with replacement
 - Drawing is repeated five times and output is color of ball

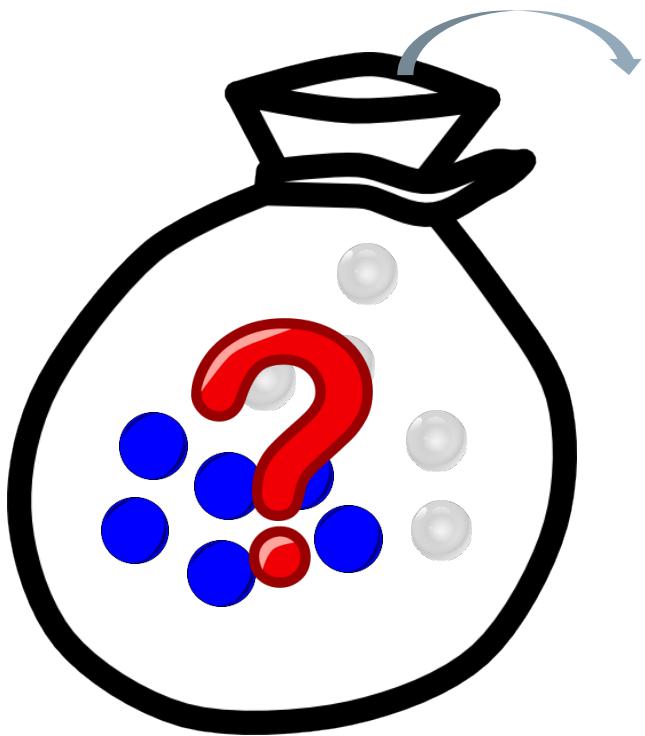
	1	2	3	4	5
Case 1	blue	white	blue	white	white
Case 2	blue	blue	blue	blue	blue

Which case is more probable?

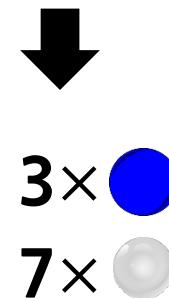


Likelihood represents how much probable is observed data samples given statistical model

Example of Likelihood Function



Sampling with replacement



- Want to estimate p_{blue} and p_{white} based on the sampling result

Example of Likelihood Function

- There are only two outputs → Bernoulli distribution
- Bernoulli distribution: the probability distribution of a random variable which takes the value 1 with success probability of p and the value 0 with failure probability of $q = 1 - p$
 - For random variable following Bernoulli distribution,
$$p(X = 1) = 1 - p(X = 0) = p = 1 - q$$
 - Probability mass function over possible outcomes y

$$f(y; p) = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{if } y = 0 \end{cases}$$

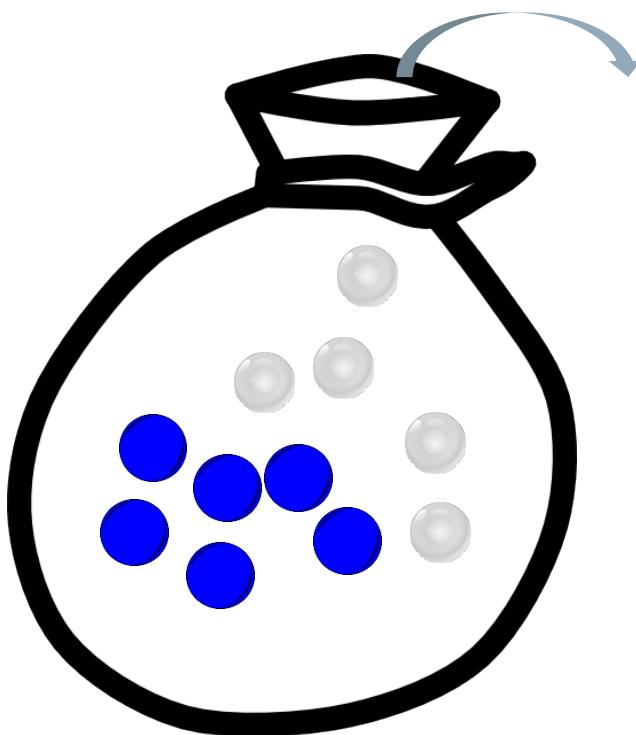
- This can also be expressed as

$$f(y; p) = p^y(1 - p)^{1-y} \quad \text{for } y \in \{1, 0\}$$

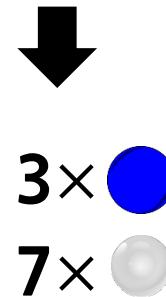
- For Bernoulli distribution, p is θ
 - In this example, assume that blue ball is 1

$$\begin{aligned} p &= p_{blue} \\ 1 - p &= p_{white} \end{aligned}$$

Example of Likelihood Function



Sampling with replacement



- Likelihood function

- If blue ball, $f(1; p) = p$
 - If white ball, $f(0; p) = 1 - p$

$$\mathcal{L} = \prod_{i=1}^{10} f(y_i; p) = p^3(1-p)^7$$

- Maximize \mathcal{L} with respect to p

$$\arg \max_y p^3(1-p)^7$$

$$\Rightarrow \arg \max_y \ln(p^3(1-p)^7) \\ = 3\ln p + 7\ln(1-p)$$

$$\frac{d \ln \mathcal{L}}{dp} = 0, \quad \hat{p} = \frac{3}{10}$$

Example of Likelihood Function

- 1D data samples from Gaussian distribution with $\sigma = 1$

$$f(x; \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$$

	1	2	3	4	5
x	2.61	3.73	2.80	4.29	3.12

- Likelihood function is function of parameter θ

$$\mathcal{L}(\theta; \mathbf{x}) = \prod_{i=1}^5 \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}}$$

- If $\theta = 2$, $\mathcal{L}(2) \approx 0.33 \times 0.09 \times 0.29 \times 0.03 \times 0.21 = 0.0000542619$

	1	2	3	4	5
x	2.61	3.73	2.80	4.29	3.12
$f(x; \theta)$	0.33	0.09	0.29	0.03	0.21

- Maximum likelihood estimation is method to find parameter to maximize likelihood function with given data samples

Maximum Likelihood Estimation

- Compare likelihood with different parameters
 - ▣ If $\theta = 2, \mathcal{L}(2) \approx 0.33 \times 0.09 \times 0.29 \times 0.03 \times 0.21 = 0.0000542619$

	1	2	3	4	5
x	2.61	3.73	2.80	4.29	3.12
$f(x; \theta)$	0.33	0.09	0.29	0.03	0.21

- ▣ If $\theta = 3, \mathcal{L}(3) \approx 0.37 \times 0.31 \times 0.39 \times 0.17 \times 0.40 = 0.003041844$

	1	2	3	4	5
x	2.61	3.73	2.80	4.29	3.12
$f(x; \theta)$	0.37	0.31	0.39	0.17	0.40

- ▣ If $\theta = 4, \mathcal{L}(4) \approx 0.15 \times 0.38 \times 0.19 \times 0.38 \times 0.27 = 0.001111158$

	1	2	3	4	5
x	2.61	3.73	2.80	4.29	3.12
$f(x; \theta)$	0.15	0.38	0.19	0.38	0.27

Solve Optimization Problem

- Likelihood function

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{\sum_{i=1}^n (\theta^2 - 2x_i\theta + x_i^2)}{2}\right) \\ &\propto \exp\left(-\sum_{i=1}^n (\theta^2 - 2x_i\theta + x_i^2)\right)\end{aligned}$$

- When $\sum_{i=1}^n (\theta^2 - 2x_i\theta + x_i^2)$ is minimum, $\mathcal{L}(\theta; \mathbf{x})$ is maximized

$$\frac{d}{d\theta} n\theta^2 - 2(\sum_{i=1}^n x_i)\theta + \sum_{i=1}^n x_i^2 = 2n\theta - 2\sum x_i = 0. \quad \theta = \frac{\sum x_i}{n} = \bar{x}$$

- Second order equation of $\theta \rightarrow$ There is a solution to minimize equation
- Example
 - <https://www.geogebra.org/m/zOmGcvXq>

※ Gaussian (Normal) Distribution

- The Gaussian distribution is a continuous probability distribution
 - ▣ probability density function

$$\mathcal{N}(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- μ : mean or expectation of the distribution
- σ : standard deviation

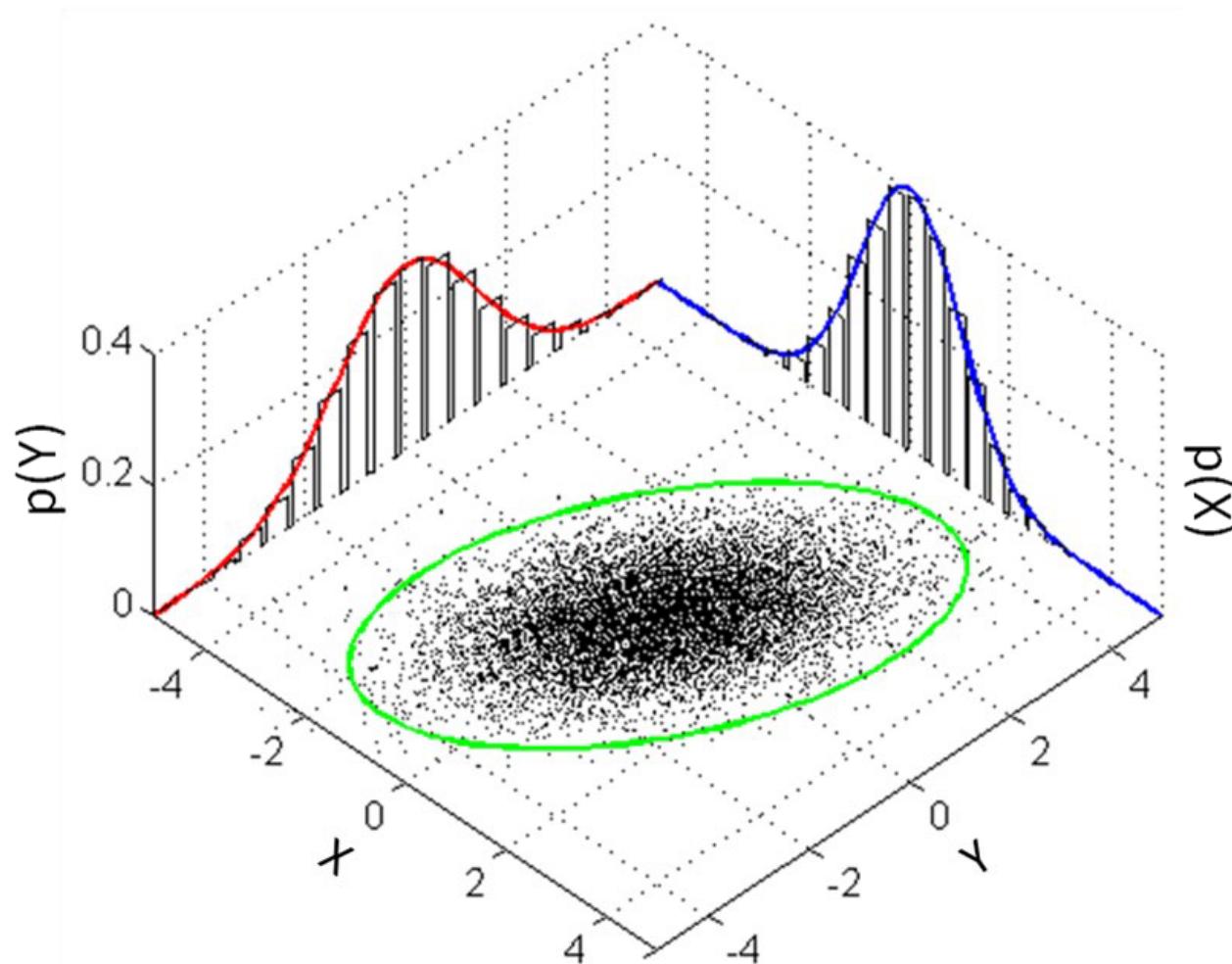
- ▣ When $\mu = 0$ and $\sigma = 1$, the distribution is called the standard normal distribution
- Multivariate normal distribution is a generalization of the 1D normal distribution
 - ▣ probability density function

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\frac{1}{(2\pi)^p |\boldsymbol{\Sigma}|} \right)^{1/2} e^{-\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- p : dimensionality
- $\boldsymbol{\mu}$: mean vector
- $\boldsymbol{\Sigma}$: covariance matrix

※ Gaussian (Normal) Distribution

- Two dimensional normal distribution



How to Find Parameters for Logistic Regression?

- Output is 0 or 1 → Output follows Bernoulli distribution with parameter p
- Each sample has different p depending on input
 $y_i \sim \text{Bernoulli}(P_i)$
 - ▣ P_i is the probability that output value is 1
 - ▣ Assume that all samples from the same Bernoulli distribution

$$f(y_i) = P\{Y = y_i\} = P_i^{y_i} (1 - P_i)^{1-y_i}, \quad y_i \in \{0,1\}$$

- Likelihood function of logistic regression model

$$\mathcal{L} = \prod_{i=1}^n f(y_i) = \prod_{i=1}^n P_i^{y_i} (1 - P_i)^{1-y_i}$$

$$\square P_i = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p}}$$

Issues with Using the Likelihood Function Directly

1. Numerical Stability
 - ▣ The likelihood function involves a product of probabilities, which can become extremely small for large datasets
 - ▣ Floating-point precision issues can arise, leading to numerical underflow
2. Computational Simplicity
 - ▣ The product operation in $\mathcal{L}(\theta)$ makes differentiation complex when optimizing θ .

Log-Likelihood Function

- Likelihood function

$$\mathcal{L}(\theta; x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta)$$

- Log-likelihood function

$$\log \mathcal{L}(\theta) = \log \prod_{i=1}^n f(x_i | \theta) = \sum_{i=1}^n \log f(x_i | \theta)$$

- Log-likelihood function for logistic regression

$$\log \mathcal{L} = \sum_{i=1}^n y_i \log P_i + \sum_{i=1}^n (1 - y_i) \log(1 - P_i)$$

- Find parameters $\beta_0, \beta_1, \dots, \beta_p$ to maximize $\log \mathcal{L}$

Advantages of Log-Likelihood

1. **Avoids numerical underflow**

- Logarithm converts products into sums, keeping values within a manageable range.

2. **Easier differentiation**

$$\frac{d}{dt} f(t)$$

- Differentiating a sum is simpler than differentiating a product, making gradient-based optimization easier.

3. **Log-convexity and optimization benefits**

- Many likelihood functions become convex in log-space, simplifying optimization.

Odds and Odds Ratio

- Odds reflect the likelihood that the event will take place
 - In gambling, odds represent the ratio between the amounts staked by parties to a wager or bet
 - In logistic regression, odds represent the ratio between $P(y = 1)$ and $P(y = 0)$

$$\frac{P(Wins)}{P(Losses)}$$

$$Odd = \frac{P}{1-P}$$

$\cancel{P \rightarrow 1} \quad Odd = \infty, \quad \cancel{P \rightarrow 0} \quad Odd = 0$

$$\text{odds} = \frac{P(y = 1)}{P(y = 0)} = \frac{\frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p}}}{1 - \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p}}} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)$$

- Odds ratio is the ratio between odds when unit increment of a variable

$$\text{odds ratio} = \frac{\text{odds when input is } x_1 = x + 1}{\text{odds when input is } x_1 = x} = \frac{\exp(\beta_0 + \beta_1(x + 1) + \beta_2 x_2 + \dots + \beta_p x_p)}{\exp(\beta_0 + \beta_1 x + \beta_2 x_2 + \dots + \beta_p x_p)} = e^{\beta_1}$$

- Odd increases e^{β_1} times for every 1-unit increase in x_1

Logistic Regression: Odds

- A logistic model is one where the log-odds of the probability of an event is a linear combination of independent or predictor variables (binary case)
- Logistic model

$$\ln(\text{odds}) = \ln\left(\frac{P(y = 1)}{P(y = 0)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- Let $P = P(y = 1)$

$$\frac{P(y = 1)}{P(y = 0)} = \frac{P}{1 - P}$$

$$g(P) = \ln\left(\frac{P}{1 - P}\right) = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}_{\text{linear predictor}}$$

Link function

Logistic Regression

- Link function
 - ▣ A link function connects the linear predictor (i.e., a linear combination of input variables) to the expected value of the dependent variable
 - ▣ The choice of the link function determines how we transform probabilities into a form suitable for regression modeling

Logistic Regression: Link Function

□ Logit function

- ▣ The logit function is the **most commonly used link function** in logistic regression
- ▣ It transforms probabilities $P(P = P(Y = 1|X))$ into log-odds:

$$g(P) = \log\left(\frac{P}{1 - P}\right)$$

- ▣ It ensures symmetry and interpretability in terms of odds
- ▣ Logistic regression with this function estimates the log-odds as a linear function of predictors

Logistic Regression: Link Function

symmetric ✓

□ Probit function

input $[0,1]$ output $(-\infty, \infty)$

- The probit function uses the inverse cumulative distribution function (CDF) of the standard normal distribution

$$g(P) = \Phi^{-1}(P)$$

where Φ^{-1} is the inverse of the standard normal CDF

- It is similar to the logit function but assumes a normal distribution of the underlying latent variable
- Probit regression is used in **econometrics and psychometrics** when normality assumptions are more appropriate

Logistic Regression: Link Function

ASYMMETRY

- Complementary Log-Log (cloglog) function
 - ▣ The cloglog function is asymmetric and models extreme event probabilities
$$g(P) = \log(-\log(1 - P))$$
 - ▣ It is useful when **probabilities near 1 are more common than those near 0** (e.g., survival analysis, infectious disease modeling)
 - ▣ Unlike the logit and probit functions, it does not have symmetry around $P = 0.5$
- Log-Log and Negative Log-Log functions
 - ▣ Log-log function
$$g(P) = -\log(-\log P)$$
 - Used when small probabilities dominate the data
 - ▣ Negative log-log function
$$g(P) = \log(-\log P)$$
 - Similar to cloglog but in the opposite direction

Logistic Regression for Multi-class Classification

- Logistic function only can be used in binary classification
- For K classes, $P(y_i = k)$ is the probability that i th data point belong to class k ($k \in \{1,2,3, \dots, K\}$)
 - It is reasonable to select class k whose probability is the highest

Multinomial Logistic Regression

- Multinomial logistic regression assumes that log ratio between probabilities of two different classes is linear
 - ▣ Log linear model

$$\ln p(y_i = 1) = \boldsymbol{\beta}_1 \cdot \mathbf{x}_i - \ln Z$$

$$\ln p(y_i = 2) = \boldsymbol{\beta}_2 \cdot \mathbf{x}_i - \ln Z$$

⋮

$$\ln p(y_i = K) = \boldsymbol{\beta}_K \cdot \mathbf{x}_i - \ln Z$$

- $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})$

- $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \beta_{k2}, \dots, \beta_{kp})$

- $\boldsymbol{\beta}_k \cdot \mathbf{x}_i = \beta_{k0} + \beta_{k1}x_{i1} + \dots + \beta_{kp}x_{ip}$



$$\mathcal{Z} \left(p(y_i = k) = \frac{1}{Z} e^{\boldsymbol{\beta}_k \cdot \mathbf{x}_i} \right) = 1.$$

$$Z = \sum_{k=1}^K e^{\boldsymbol{\beta}_k \cdot \mathbf{x}_i}$$

※ Multinomial distribution

- Multinomial distribution is a generalization of the binomial distribution

- Binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments with success probability p

$$p(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = {}^n C_k p^k (1-p)^{n-k}$$

- Example of binomial distribution is the distribution of the number of head when flipping a coin n times (in this case, $p = 0.5$)

- Probability that k times head occur among n trials

$$p(k) = \frac{n!}{k!(n-k)!} 0.5^k 0.5^{n-k} = \frac{n!}{k!(n-k)!} 0.5^n$$

- In multinomial distribution, possible outcome is more than two and each outcome has its own probability to occur, (p_1, \dots, p_d)

- $p_1 + \dots + p_d = 1$
 - d is the number of possible outcomes
 - $n_x = \sum_{i=1}^d x_i$

$$p(\mathbf{x} = (x_1, x_2, \dots, x_d)) = \frac{n_{\mathbf{x}}!}{x_1! \cdots x_d!} p_1^{x_1} \cdots p_d^{x_d}$$

Likelihood Function

- Likelihood function

$$\mathcal{L} = \prod_{i=1}^n \prod_{k=1}^K P_{ik}^{v_{ik}} , \quad v_{ik} = \begin{cases} 1, & y_i = k \\ 0, & y_i \neq k \end{cases}$$

- $P_{ik} = p(y_i = k)$

- Log likelihood function

$$\log \mathcal{L} = \sum_{i=1}^n \sum_{k=1}^K v_{ik} \log P_{ik}$$

- Through maximum likelihood estimation, determine β_k as the same as in binary logistic regression

Multiclass Classification Using Binary Classifiers

- There are other ways to get multiclass classifiers by combining binary classifiers
 - For multiclass classification commonly used approach is to construct K separate binary classifiers $K > H \text{ class} \Rightarrow K \text{ 개 } \text{이진 분류기 } \text{생성}$
 - Each model is trained using the data from class C_k as the positive examples and the data from the remaining $K - 1$ classes as the negative examples

$$y(\mathbf{x}) = \max_k y_k(\mathbf{x})$$

→ One-versus-the rest approach

각 클래스에 대한 분류기

각각 데이터에 대해

각각 클래스에 대한 학습

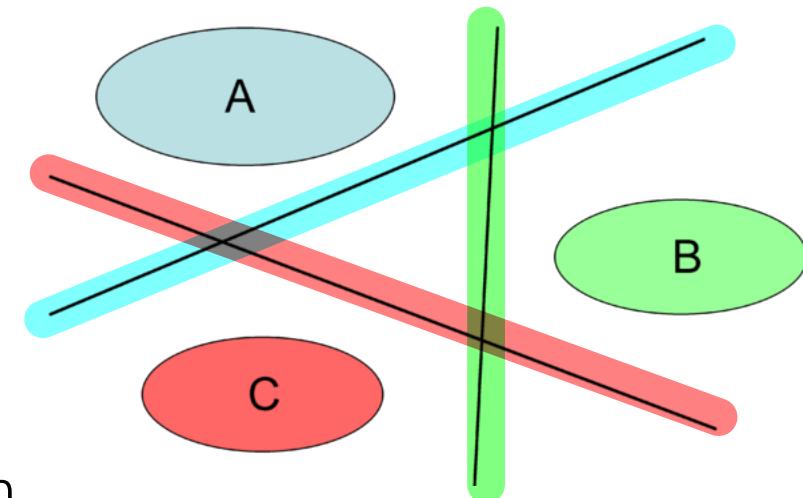
각각 확률(가능성)을 계산

결국은 분류기의 확률로 표시

(A) vs (B C)

(B) vs (A C)

(C) vs (A B)



- Problems of one-versus-the rest approach

- Because each classifier was trained on different task, there is no guarantee that the real-values quantities $y_k(\mathbf{x})$ will have appropriate scales
- Imbalance of data on training

한 클래스는 적고 나머지는 많아서 모델 편향

각 분류기 고유 기준으로 학습. 출력값 스케일 차이
최적화 간 상관 없고 예측

Multiclass Classification Using Binary Classifiers

- Another approach is to train $K(K - 1)/2$ different 2-class classifiers on all possible pairs of classes

kC_2

A-B A-C B-C

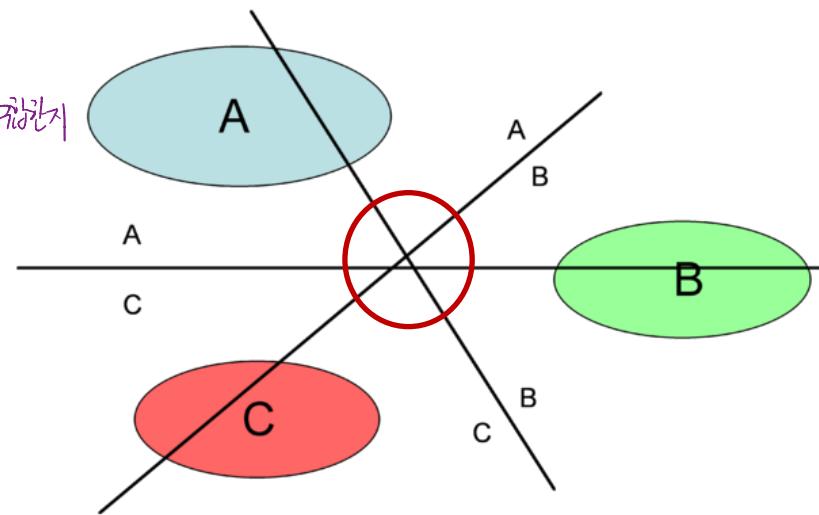
- Classify test points according to which class has the highest number of votes

→ one-versus-one approach

$$\begin{array}{ccc} A - B & A - C & B - C \\ \downarrow & \downarrow & \downarrow \\ ? & ? & ? \end{array}$$

vote

두 클래스 중 어떤 클래스에 더 적합한지
판정하여 그 결과를 반환.
가장 많은 득표 클래스로 판정



득표 풍선 혹은 순환 표
반복 \Rightarrow 오차점.

- Problems of one-versus-one approach

- It can lead to ambiguities in the resulting classification
- For large K , it requires significantly more training time

NAÏVE BAYES CLASSIFIER

Week07

Naïve Bayes Classifier

Review: Types of Classifiers

- A classifier is a function that assigns to a sample, \mathbf{x} a class label \hat{y}
$$\hat{y} = f(\mathbf{x})$$
- A probabilistic classifier obtains conditional distributions $\Pr(Y|\mathbf{x})$, meaning that for a given $\mathbf{x} \in \mathcal{X}$, they assign probabilities to all $y \in Y$
 - Hard classification
$$\hat{y} = \arg \max_y \Pr(Y = y | \mathbf{x})$$
- Logistic regression

$$f(x) = P(Y = 1 | \mathbf{x}) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p}}$$

Any other ways to model $P(Y|X)$?

Naïve Bayes Classifier

- Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumption between features

- **Bayes' theorem**

$$P(A|B) = \frac{\text{prior} \ P(B) \ \text{likelihood}}{P(A)P(B|A)}$$

- A and B are events
- $P(A)$ and $P(B)$ are the probabilities of A and B without regard to each other
- $P(A|B)$, a conditional probability, is the probability of A given that B is true
- $P(B|A)$, is the probability of B given that A is true



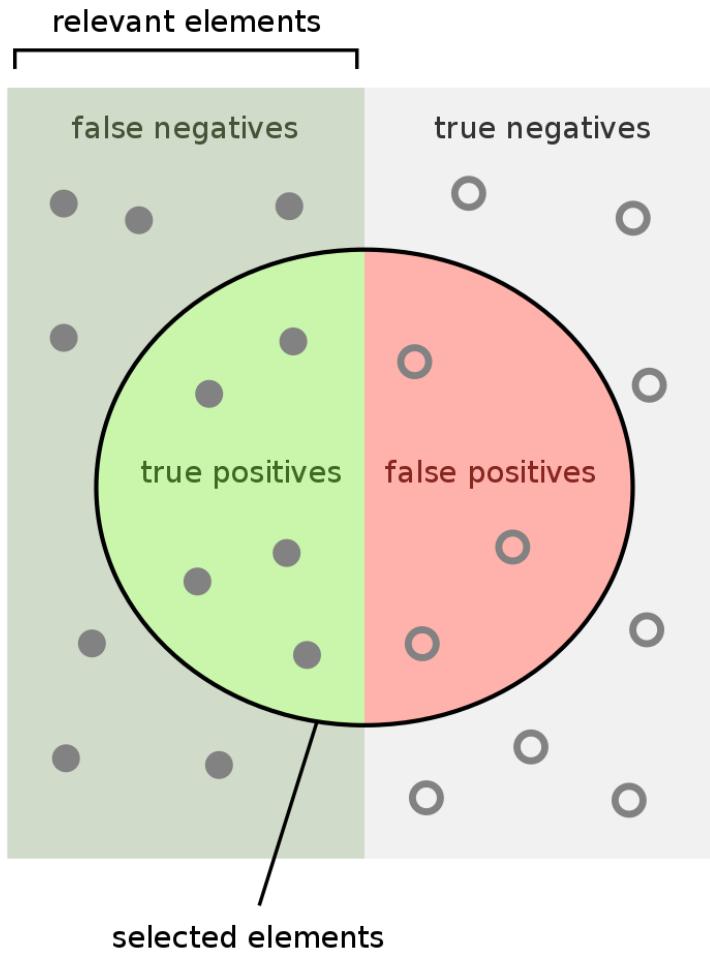
Example of Bayes' Theorem

- Suppose a drug test is 99% sensitive and 99% specific
 - ▣ 99% sensitive=99% true positive over real positive $P(\text{positive} | \text{User}) = 0.99$
 - ▣ 99% specific=99% true negative over real negative $P(\text{negative} | \text{Not-user}) = 0.99$

Real Decision	Positive	Negative
Positive	True positive	False positive (Type I error)
Negative	False negative (Type II error)	True negative

- Suppose that 0.5% of people are users of the drug

* Sensitivity and Specificity



How many relevant items are selected?
e.g. How many sick people are correctly identified as having the condition.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

실제 양성 중
양성이 맞았는가.

How many negative selected elements are truly negative?
e.g. How many healthy people are identified as not having the condition.

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Negative}}$$

실제 음성 중
음성이 맞았는가.

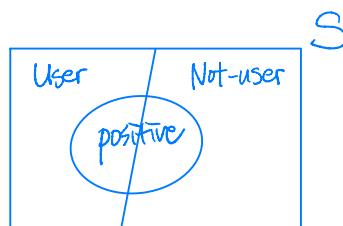
Example of Bayes' Theorem

- If a randomly selected individual tests positive, what is the probability he or she is a user of drug?

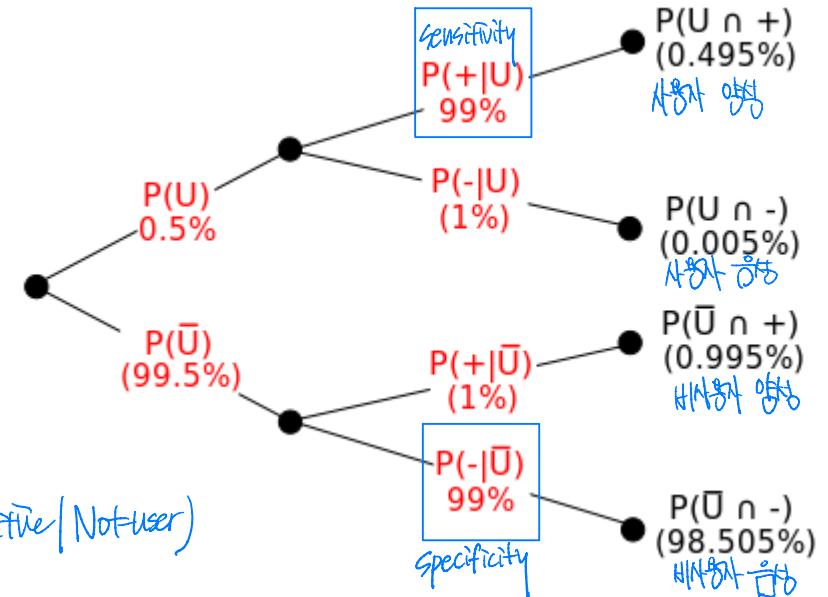
- This problem is to calculate $P(U|+)$

+ means positive drug test

U represents user, \bar{U} represents non-user



$$P(\text{positive}) = P(\text{User}) P(\text{positive} | \text{User}) + P(\text{Not-user}) P(\text{positive} | \text{Not-user})$$



$$\begin{aligned}
 P(U|+) &= \frac{P(U)P(+|U)}{P(+)} = \frac{P(U)P(+|U)}{P(U)P(+|U) + P(\bar{U})P(+|\bar{U})} \\
 &= \frac{0.005 \times 0.99}{0.005 \times 0.99 + 0.995 \times 0.01} \approx 33.2\%
 \end{aligned}$$

law of total probability

$$\therefore P(A) = \sum_{b \in B} P(b)P(A|b)$$

Example of Bayes' Theorem

Other example :

95% sensitive
90% specific
1% user

$$P(\text{User} | \text{positive}) = \frac{P(\text{User}) P(\text{positive} | \text{user})}{P(\text{positive})}$$

$$= \frac{0.01 \times 0.95}{0.01 \times 0.95 + 0.99 \times 0.10} \approx 0.0876, 8.76\%$$

양성반응자 중 8.76%만이 실제 사용자

$P(\text{User})$
 $= 0.01$

$P(\text{Positive} | \text{user})$
 $= 0.95$

$P(\text{Negative} | \text{user})$
 $= 0.05$

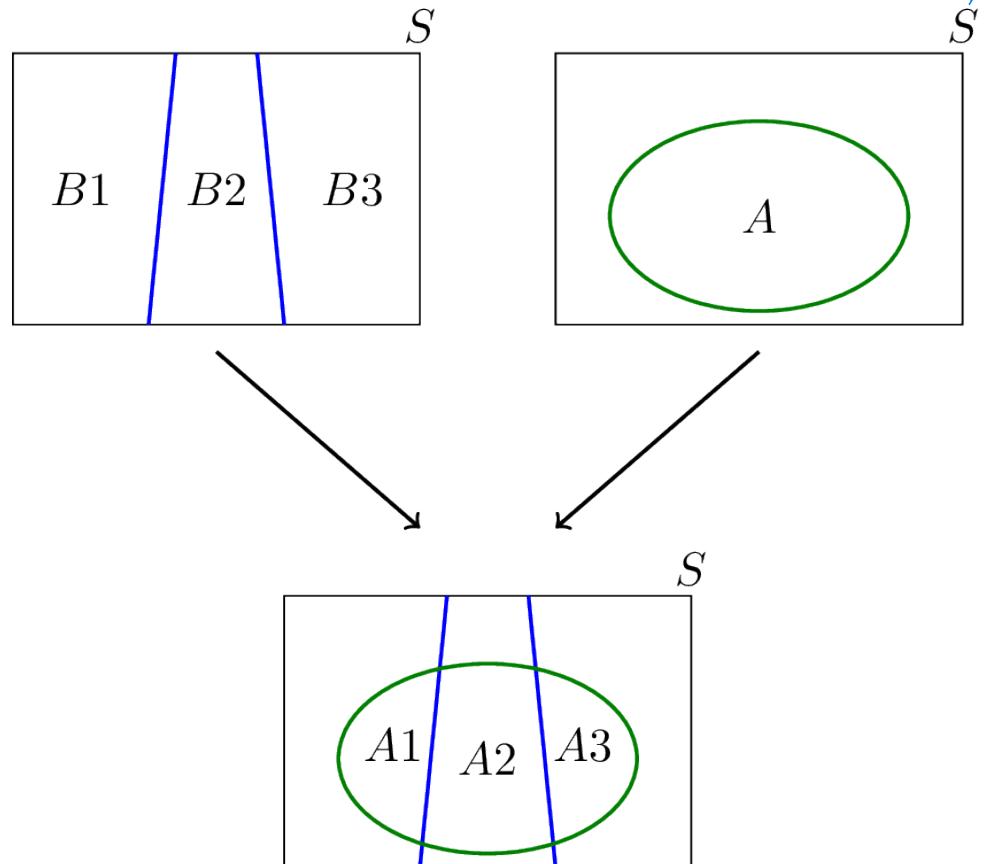
$P(\text{Non-user})$
 $= 0.99$

$P(\text{positive} | \text{Non-user})$
 $= 0.10$

$P(\text{Negative} | \text{Non-user})$
 $= 0.90$

※ Law of Total Probability

$$P(A) = \sum_{b \in B} P(b)P(A|b) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3)$$



Naïve Bayes Classifier

- Conditional probability model for Naïve Bayes classifier
 - ▣ Naïve Bayes classifier calculates following probability for every class

$$p(C_k | x_1, \dots, x_p) = p(C_k | \mathbf{x})$$

- x_i represents each feature (independent variable)
- k represents k -th class and classifier assigns output class with the maximum probability

- ▣ Re-formulation using Bayes' theorem

$$p(C_k | \mathbf{x}) = \frac{p(C_k) p(\mathbf{x} | C_k)}{p(\mathbf{x})}$$

Handwritten annotations in blue:

- Top right: C_k 의 확률
- Top right: C_k 의 확률
- Middle left: C_k 의 확률
- Middle left: \mathbf{x} 의 확률
- Bottom left: C_k 의 확률
- Bottom left: \mathbf{x} 의 확률
- Middle right: $p(C_k)$ prior
- Middle right: $p(\mathbf{x} | C_k)$ likelihood
- Bottom right: $p(\mathbf{x})$ evidence

- This equation is also written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Naïve Bayes Classifier

- Naïve Bayes Classifier

$$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

- ▣ Denominator $p(\mathbf{x})$ does not depend on class 가장 학점이 높은 출판사가!

$$\operatorname{argmax}_k p(C_k|\mathbf{x}) = \operatorname{argmax}_k p(C_k)p(\mathbf{x}|C_k)$$

- $p(C_k)p(\mathbf{x}|C_k)$ is equivalent to the joint probability $p(C_k, x_1, \dots, x_p)$

- ▣ Using chain rule $p(C_k, x_1, \dots, x_p)$ can be written as follows

$$\begin{aligned} p(C_k, x_1, \dots, x_p) &= p(C_k)p(x_1, \dots, x_p|C_k) = p(C_k)p(x_1|C_k)p(x_2, \dots, x_p|C_k, x_1) \\ &= p(C_k)p(x_1|C_k)p(x_2|C_k, x_1) \cdots p(x_p|C_k, x_1, \dots, x_{p-1}) \end{aligned}$$

- ▣ Naïve Bayes classifier assumes conditional independence of each feature

$$\begin{aligned} p(x_i|C_k, x_j) &= p(x_i|C_k) \\ p(x_i|C_k, x_j, x_l) &= p(x_i|C_k) \end{aligned}$$

※ Chain Rule

- Chain rule permits the calculation of any member of the joint distribution of a set of random variables using only conditional probabilities

$$P(A_n, \dots, A_1) = P(A_n | A_{n-1}, \dots, A_1) \cdot P(A_{n-1}, \dots, A_1)$$

- Repeating this process with each final term creates the product

$$P\left(\bigcap_{k=1}^n A_k\right) = \prod_{k=1}^n P\left(A_k \mid \bigcap_{j=1}^{k-1} A_j\right)$$

Naïve Bayes Classifier

□ Naïve Bayes Classifier

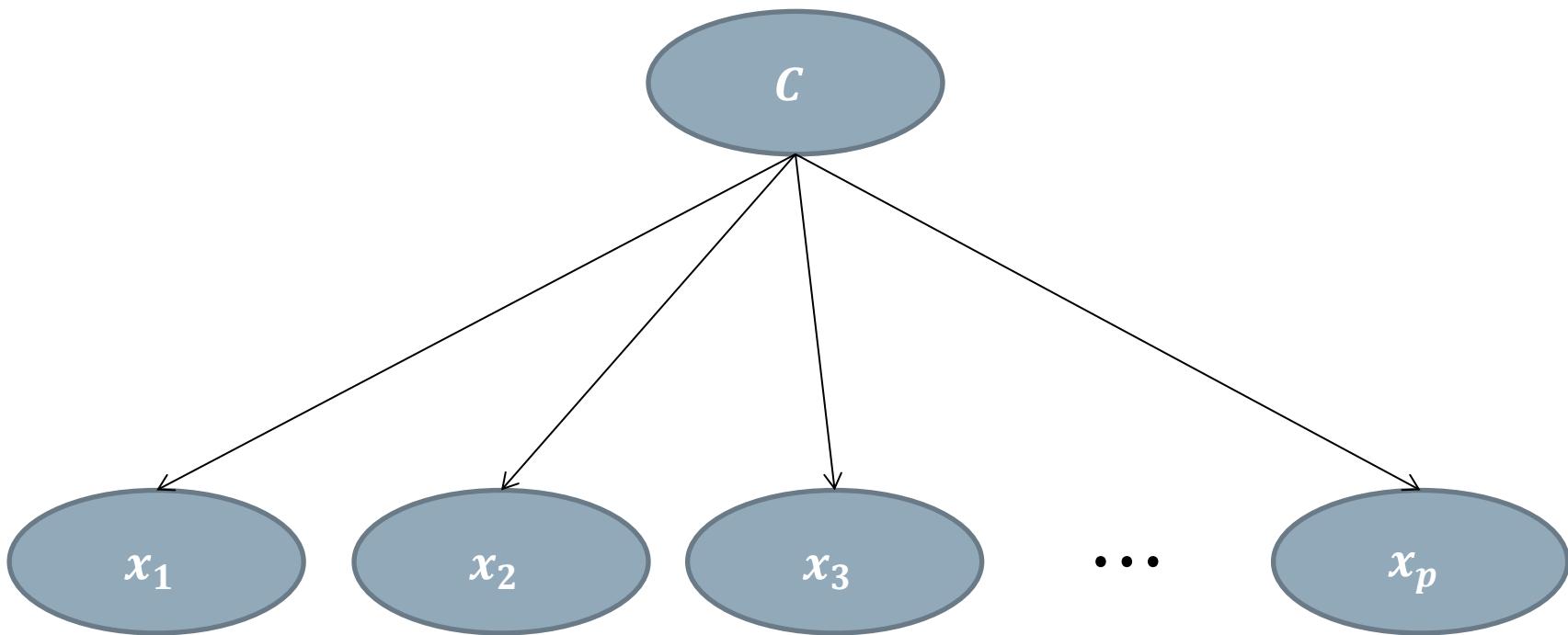
$$\begin{aligned} & p(C_k)p(x_1|C_k)p(x_2|C_k, x_1) \cdots p(x_p|C_k, x_1, \dots, x_{p-1}) \\ &= p(C_k)p(x_1|C_k)p(x_2|C_k) \cdots p(x_p|C_k) = p(C_k) \prod_{i=1}^p p(x_i|C_k) \\ &\therefore p(C_k)p(\mathbf{x}|C_k) = p(C_k) \prod_{i=1}^p p(x_i|C_k) \end{aligned}$$

□ Decision function of Naïve Bayes classifier

$$\hat{y} = \operatorname{argmax}_{k \in \{1, \dots, K\}} p(C_k) \prod_{i=1}^p p(x_i|C_k)$$

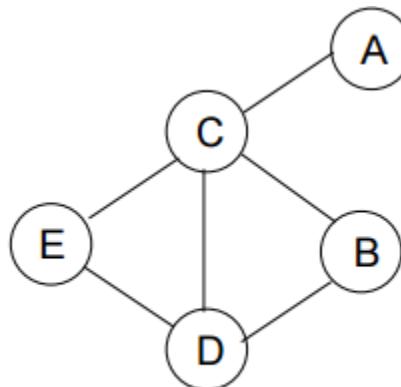
Naïve Bayes Classifier

- Graphical model representation of Naïve Bayes

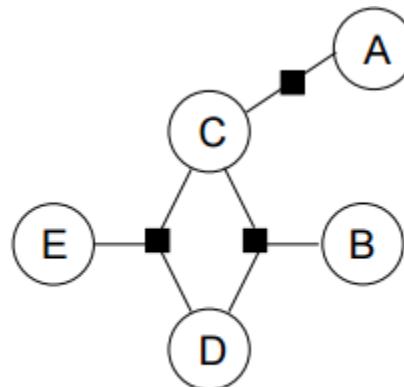


* Graphical Model

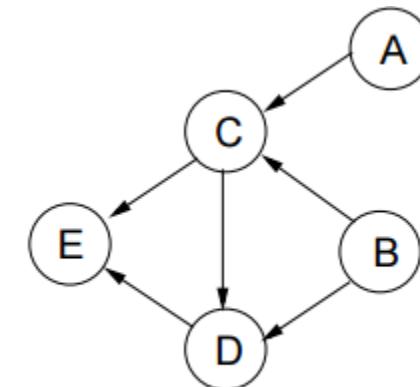
- In graphical model, each node represents a random variable and the edge express probabilistic relationships between these variables



Undirected Graph



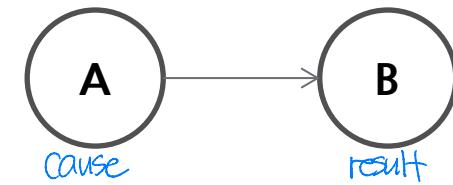
Factor Graph



Bayesian Network

- Bayesian network is a specific form of graphical model
 - Directed acyclic graphs
 - If two nodes A and B are connected by edge and direction of edge heads to B, it means that the state of A affects on probability of B

$$p(B, A) \neq p(B)p(A)$$
$$p(B|A) \neq p(B)$$



$A \& B \Rightarrow$ not independent

Naïve Bayes Classifier: Advantages and Limitation

- Advantages of Naïve Bayes
 - ▣ Simple and fast to train
 - ▣ Works well with small datasets
 - ▣ Handles missing data well
 - ▣ Requires less training data compared to other classifiers
- Limitations of Naïve Bayes
 - ▣ Independence assumption is often unrealistic
 - ▣ Struggles with correlated features
 - ▣ Zero probability problem (solved using Laplace smoothing)

Naïve Bayes Classifier: Parameter Estimation

- Parameter estimation and event models
 - ▣ A class' prior setting: by assuming equiprobable classes ($p(C_k) = 1/K$) or by calculating an estimate for the class probability from the training set ($p(C_k) = n_k/n$)
 - ▣ Select appropriate probability distribution for $p(x_i|C_k)$
 - For continuous variables, Gaussian distribution is the common choice
 - For discrete variables, multinomial distribution is the common choice
 - ▣ After setting, probabilistic model for naïve Bayes classifier, parameters of distributions are estimated using training data
 - Calculate $\tilde{p}(C_{y_j}|\mathbf{x}) = (C_{y_j}) \prod_{i=1}^p p(x_i|C_{y_j})$ for j -th sample
 - Calculate $\tilde{p}(\mathbf{C}|\mathbf{X}) = \prod_{j=1}^n \tilde{p}(C_{y_j}|\mathbf{x})$ and maximize this probability

Naïve Bayes Classifier: Parameter Estimation

Outlook	Temp	Humidity
Sunny	Hot	N
Cloud	H	N
Rain	Low	Y
	Mid	Y
	H	N
C	M	Y
S	L	Y

$$P(\text{Rain} = Y | \text{R}, \text{H}) = P(Y) \cdot P(R|Y) \cdot P(H|Y) = \frac{5}{8} \times \frac{2}{5} \times \frac{1}{5} = \frac{1}{20} = 0.05$$
$$P(\text{Rain} = N | \text{R}, \text{H}) = P(N) \cdot P(R|N) \cdot P(H|N) = \frac{3}{8} \times \frac{1}{3} \times \frac{2}{3} = \frac{1}{8} = 0.125$$

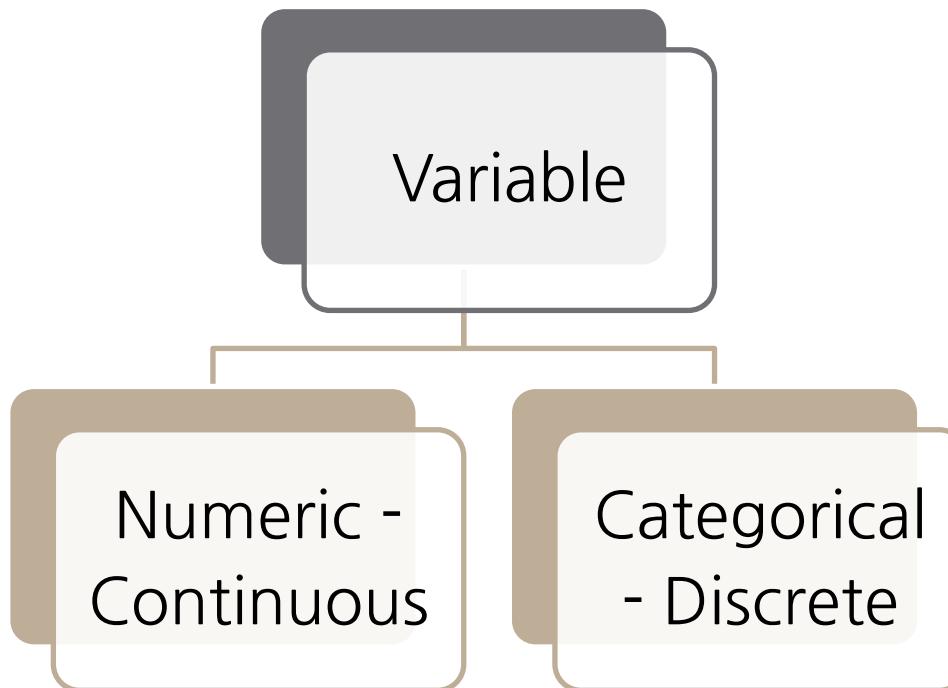
$$P(Y) = \frac{5}{8} \quad P(R|Y) = \frac{2}{5} \quad P(H|Y) = \frac{1}{5}$$
$$P(N) = \frac{3}{8} \quad P(R|N) = \frac{1}{3} \quad P(H|N) = \frac{2}{3}$$

$$P(Y | R, H) = 0.05 < P(N | R, H) = 0.125$$

∴ 雨 高温 高湿 X

Naïve Bayes Classifier

- Determine $p(x_i|C_k)$
 - ▣ The probability functions depend on the type of variables
 - Probability distributions for discrete random variables: Binomial, Multinomial, Geometric, ...
 - Probability distributions for continuous random variables: Gaussian(Normal), χ^2 , beta, F , t , ...





Bernoulli Naïve Bayes

- Bernoulli naïve Bayes
 - ▣ In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs
 - Ex) Each variable only takes values 0 or 1
 - ▣ Each x_i is a boolean expressing the occurrence of event

$$p(\mathbf{x}|C_k) = \prod_{i=1}^d p_{ki}^{x_i} (1 - p_{ki})^{1-x_i}$$

- d is the number of input features $x_i \in \{0, 1\}$
- p_{ki} is the probability that x_i is 1(true) for class k

※ Bernoulli distribution

- The probability distribution of a random variable which takes the value 1 with success probability of p and the value 0 with failure probability of $q = 1 - p$

- For random variable following Bernoulli distribution,

$$p(X = 1) = 1 - p(X = 0) = p = 1 - q$$

- Probability mass function over possible outcomes y

$$f(y; p) = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{if } y = 0 \end{cases}$$

- This can also be expressed as

$$f(y; p) = p^y(1 - p)^{1-y} \quad \text{for } y \in \{1, 0\}$$

- Expected value of a Bernoulli random variable X

$$\mathbb{E}[X] = p$$

- Variance of a Bernoulli random variable

$$\text{Var}[X] = p(1 - p)$$

Bernoulli Naïve Bayes: Estimation of Parameters

- Maximum likelihood estimates for Bernoulli Naïve Bayes model
 - ▣ Each input variable can take values 0 or 1
 - ▣ There exist n samples with d features
 - ▣ Prior probability $p(C_k)$

$$p(C_k) = \frac{n_k}{n}$$

- n_k is the number of samples that y_i belongs to class k
- ▣ Likelihood $p(\mathbf{x}|C_k)$

$$p(\mathbf{x}|C_k) = \prod_{i=1}^d p_{ki}^{x_i} (1 - p_{ki})^{1-x_i}$$

- ▣ Posterior probability $p(C_k|\mathbf{x})$

$$p(C_k|\mathbf{x}) \propto p(C_k)p(\mathbf{x}|C_k)$$

$$\frac{n_k}{n} \times \prod_{i=1}^d p_{ki}^{x_i} (1 - p_{ki})^{1-x_i}$$

Bernoulli Naïve Bayes: Estimation of Parameters

- Maximum likelihood estimates for Bernoulli Naïve Bayes model

$$L = \prod_{j=1}^n p(C_{y_j}) p(\mathbf{x}_j | C_{y_j}) = \prod_{j=1}^n \frac{n_{y_j}}{n} \left(\prod_{i=1}^d p_{y_j i}^{x_{ji}} (1 - p_{y_j i})^{1-x_{ji}} \right)$$

$$\log L = \sum_{j=1}^n \log \frac{n_{y_j}}{n} + \sum_{j=1}^n \sum_{i=1}^d (x_{ji} \log p_{y_j i} + (1 - x_{ji}) \log (1 - p_{y_j i}))$$

- n is the total number of data samples
- n_{y_j} is the number of data samples belong to class y_j ($y_j \in \{1, 2, \dots, k\}$)
- x_{ji} is the i -th input variable's value for j -th data sample

- Parameters to be estimates

- For each class k , probability to occur 1 for each feature i , p_{ki}

Bernoulli Naïve Bayes: Estimation of Parameters

- Maximum likelihood estimates for Bernoulli Naïve Bayes model
 - ▣ To obtain optimal p_{ki} , set $\frac{\partial \log L}{\partial p_{ki}} = 0$

$$\begin{aligned}\frac{\partial \log L}{\partial p_{ki}} &= \sum_{j \in \{m: y_m=k\}} \left\{ \frac{x_{ji}}{p_{ki}} - \frac{1 - x_{ji}}{1 - p_{ki}} \right\} \\ &= \frac{n_{k1}}{p_{ki}} - \frac{n_{k0}}{1 - p_{ki}} = 0\end{aligned}$$

- $n_{k1} = |\{m: x_{mi} = 1, y_m = k\}|$ is the number of data samples in set of $\{m: x_{mi} = 1, y_m = k\}$
- $n_{k0} = |\{m: x_{mi} = 0, y_m = k\}|$ is the number of data samples in set of $\{m: x_{mi} = 0, y_m = k\}$
- $\{m: x_{mi} = 1, y_m = k\}$ is a set that contains every sample with $x_i = 1$ in class k
- $\{m: x_{mi} = 0, y_m = k\}$ is a set that contains every sample with $x_i = 0$ in class k
- $|\{m: x_{mi} = 1, y_m = k\}| + |\{m: x_{mi} = 0, y_m = k\}| = n_k$

$$p_{ki} = \frac{n_{ki}}{n_k}$$

이미지에서
1가지 이상 비율
} 전부 확률과
동일하다.

Bernoulli Naïve Bayes: Estimation of Parameters

x	y
1	0
1	0
1	0
0	0
0	1
0	1
1	1



$$p(x = 0|y = 0) = p_{00} = \frac{1}{4}$$

$$p(x = 1|y = 0) = p_{01} = \frac{3}{4}$$



$$p(x = 0|y = 1) = p_{10} = \frac{2}{3}$$

$$p(x = 1|y = 1) = p_{11} = \frac{1}{3}$$

Categorical Naïve Bayes

- Discrete random variables with more than two outcomes

$$P(\mathbf{x} = (x_1, x_2, \dots, x_m)) = \prod_{j=1}^m p_j^{x_j}$$

- m outcomes
- x_j is a binary indicator variable: 1 when j -th outcome is observed; otherwise 0
- p_j is probability of j -th outcome ($\sum_{j=1}^m p_m = 1$)

$$p(\mathbf{X}|C_k) = \prod_{i=1}^n \prod_{j=1}^m p_{kj}^{x_j}$$

$$\log p(\mathbf{X}|C_k) = \sum_{i=1}^n \sum_{j=1}^m x_j \log p_{kj}$$
$$p_{ki} = \frac{n_{kj}}{n_k}$$

Categorical Naïve Bayes

- Discrete random variables with more than two outcomes

x	y
High	0
Mid	0
High	0
Low	0
High	0
Low	0
High	0
High	0

$$p(x = \text{High}|y = 0) = p_{0,\text{High}} = \frac{5}{8}$$

$$p(x = \text{Mid}|y = 0) = p_{0,\text{Mid}} = \frac{1}{8}$$

$$p(x = \text{Low}|y = 0) = p_{0,\text{Low}} = \frac{2}{8}$$

Parameter Smoothing

- Smoothing techniques for parameter estimation
 - ▣ Smoothing techniques helps prevent zero probabilities when a category is not observed in the training data
 - ▣ Laplace smoothing (additive smoothing)
$$p_{kj} = \frac{n_{kj} + \alpha}{n_k + \alpha m}$$
- α : smoothing parameter, commonly $\alpha = 1$

Multinomial Naïve Bayes

- Multinomial naïve Bayes
 - ▣ With a multinomial event model, samples represent the frequencies with which certain events have been generated by a multinomial (p_1, \dots, p_d)
 - p_i is the probability that event i occurs
 - m is the number of features in input data

$$p(\mathbf{x}|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_{ki}^{x_i}$$

- $\mathbf{x} = (x_1, x_2, \dots, x_d)$ represents each sample and it can be seen as a histogram with x_i counting the number of times event i was observed in a particular instance
- ▣ This is the event model typically used for document classification
 - With events representing the occurrence of a word in a single document
→ bag of words representation



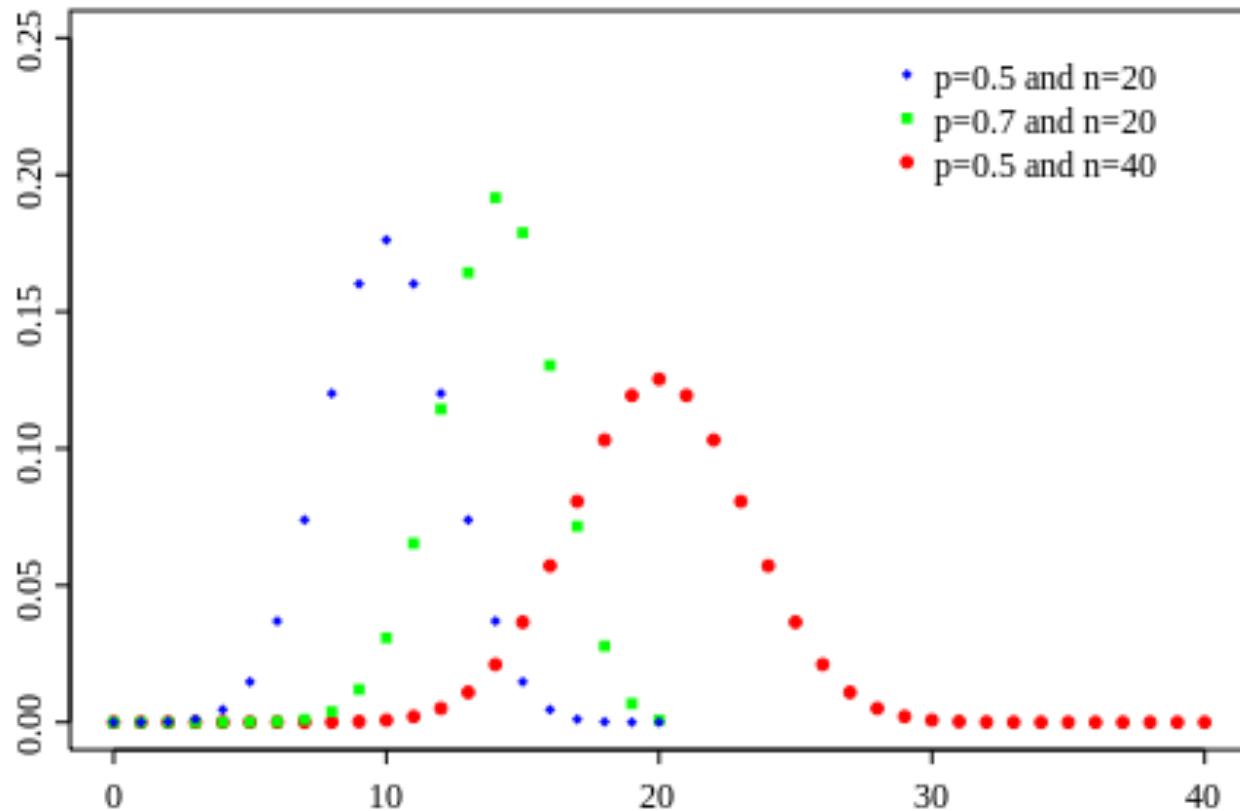
※ Multinomial distribution

- Multinomial distribution is a generalization of the binomial distribution
 - Binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments with success probability p
$$p(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$
 - Example of binomial distribution is the distribution of the number of head when flipping a coin n times (in this case, $p = 0.5$)
 - Probability that k times head occur among n trials
$$p(k) = \frac{n!}{k!(n-k)!} 0.5^k 0.5^{n-k} = \frac{n!}{k!(n-k)!} 0.5^n$$
 - In multinomial distribution, possible outcome is more than two and each outcome has its own probability to occur, (p_1, \dots, p_d)
 - $p_1 + \dots + p_d = 1$
 - d is the number of possible outcomes
 - $n_x = \sum_{i=1}^d x_i$

$$p(\mathbf{x} = (x_1, x_2, \dots, x_d)) = \frac{n_{\mathbf{x}}!}{x_1! \cdots x_d!} p_1^{x_1} \cdots p_d^{x_d}$$

※ Multinomial distribution

□ Binomial distribution



Multinomial Naïve Bayes: Application

- Relies on very simple representation of document
 - ▣ Bag of words

I simply loved this film but was shocked by the bad reviews that people gave it. To this I say to them: You seriously misunderstood the meaning of it. Although I won't reveal any real details about the meaning because I think that you should try and understand it yourself. The movie was terrific and simply breathless the whole time. I felt awestruck about how the life of one man could be so changed after an experience that Hanks went through. I say that every element of the film was perfect. And for those of you who hate Wilson, you have to understand about how human he really was to Chuck. I was amazed on how well this movie was made and think that everybody should have an experience that should cause you to take stock of your life

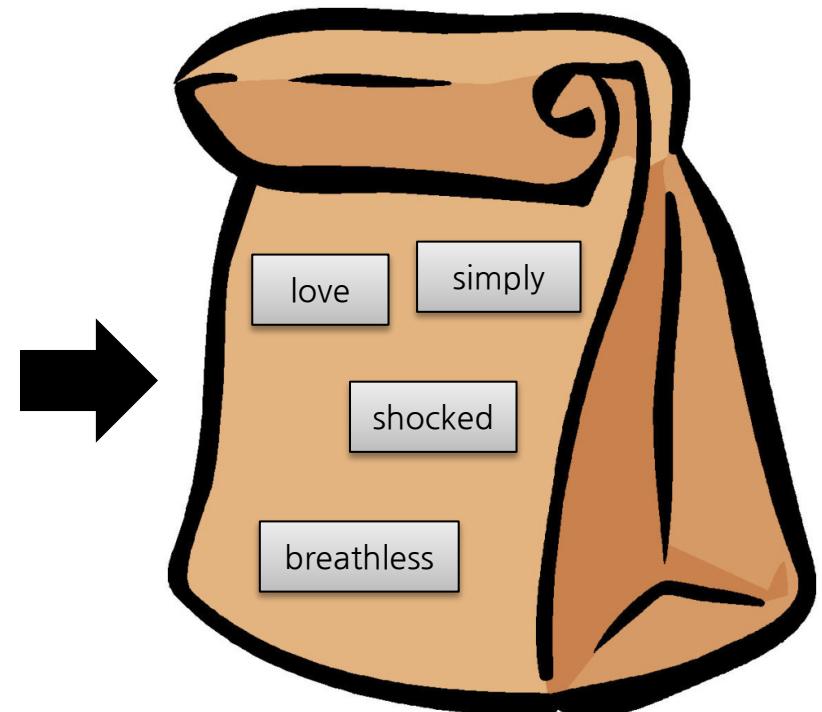
$f($

) = C

Multinomial Naïve Bayes: Application

- The bag of words representation

I simply loved this film but was shocked by the bad reviews that people gave it. To this I say to them: You seriously misunderstood the meaning of it. Although I won't reveal any real details about the meaning because I think that you should try and understand it yourself. The movie was terrific and simply breathless the whole time. I felt awestruck about how the life of one man could be so changed after an experience that Hanks went through. I say that every element of the film was perfect. And for those of you who hate Wilson, you have to understand about how human he really was to Chuck. I was amazed on how well this movie was made and think that everybody should have an experience that should cause you to take stock of your life



Multinomial Naïve Bayes: Application

- The bag of words representation
 - ▣ It is possible to select a subset of words

Word	Frequency
love	1
great	2
recommend	1
simply	1
happy	1
bad	2
:	:

$f() = C$

Multinomial Naïve Bayes: Application

- The bag of words representation
 - ▣ It is possible to select a subset of words

Word	Frequency
love	1
great	2
recommend	1
old	0
simply	1
happy	1
bad	2
dog	0
:	:

$f($

$) = C$

Multinomial Naïve Bayes: Application

- Revisit naïve Bayes

$$\hat{y} = \operatorname{argmax}_{k \in \{1, \dots, K\}} p(C_k) \prod_{i=1}^p p(x_i | C_k)$$

- Each i th column represent the specific word
- C_k is the possible output class
 - Ex.) Spam filter: spam or non-spam
- The basic idea that uses Naïve Bayes for text classification
 - Each class has the different distribution of words
 - Spam mails usually contain “click below”, “free dvd”, “great offer” and etc.

Multinomial Naïve Bayes: Estimation of Parameters

- Maximum likelihood estimator with Laplace smoothing

$$p_{ki} = \frac{\sum_{j \in C_k} x_{ji} + \alpha}{\sum_{j \in C_k} \sum_{l=1}^d x_{jl} + \alpha d}$$

- p_{ki} : The probability of occurrence of i -th variable for class k
- d : The number of input variables
- x_{ji} : The occurrence of i -th variable for sample j

Spam	free	click	w̄in
A	2	1	0
B	1	0	1
C	0	1	1
	3	2	2

$$d=3$$

$$\alpha = 1$$

$$P(\text{free} | \text{Spam}) = \frac{3}{7} \Rightarrow \frac{3+1}{7+3 \times 1} = 0.4$$

$$P(\text{w̄in} | \text{Spam}) = \frac{2}{7} \Rightarrow \frac{2+1}{7+3 \times 1} = 0.3$$

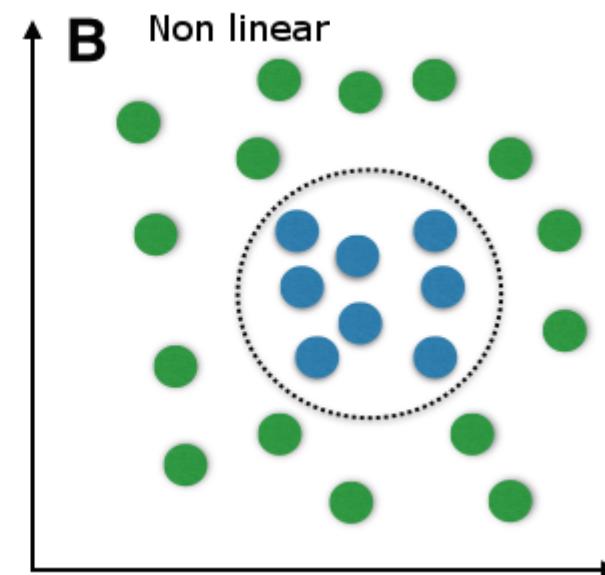
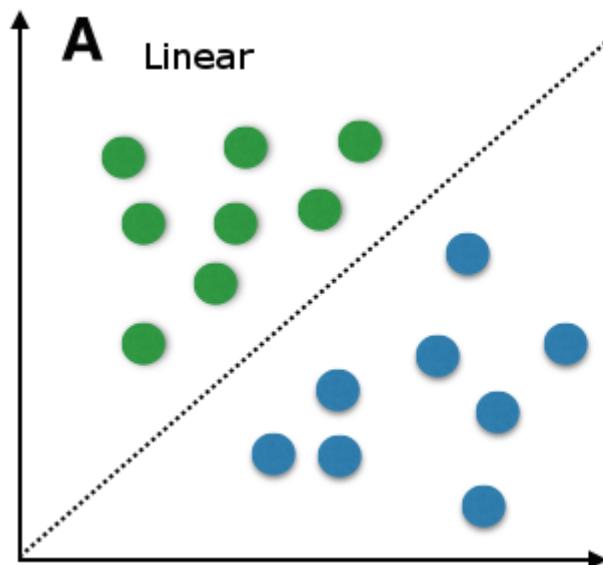
Multinomial Naïve Bayes

- Multinomial naïve Bayes
 - ▣ The multinomial naive Bayes classifier becomes a linear classifier when expressed in log-space

$$\log p(C_k | \mathbf{x}) \propto \log p(C_k) \prod_i p_{k_i}^{x_i} = \log p(C_k) + \sum_{i=1}^n x_i \log p_{ki} = b + \mathbf{w}_k^T \mathbf{x}$$

- $b = \log p(C_k)$
- $w_{ki} = \log p_{ki}$
- $\frac{(\sum_i x_i)!}{\prod_i x_i!}$ term only depends on \mathbf{x} and does not depend on class

* Decision Boundary



Complement Naïve Bayes

- Complement naïve Bayes
 - ▣ Complement naïve Bayes is an adaptation of the standard multinomial naïve Bayes algorithm that is particularly suited for imbalanced data sets
 - ▣ Unlike multinomial naïve Bayes, it uses word occurrence frequencies from the complement of each class

$$p_{ki} = \frac{\sum_{j \notin C_k} x_{ji} + \alpha}{\sum_{j \notin C_k} \sum_{l=1}^d x_{jl} + \alpha d}$$

Spam vs Ham

⇒ Spam 예측 시 Ham 흑자 사용
Ham 예측 시 Spam 흑자 사용

Gaussian Naïve Bayes

- Gaussian naïve Bayes
 - ▣ When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution

$$p(x = v | C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-u_k)^2}{2\sigma_k^2}}$$

$$p(C_k) \prod_{i=1}^p p(x_i | C_k) = p(C_k) \prod_{i=1}^p \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} e^{-\frac{(v_i-u_{ki})^2}{2\sigma_{ki}^2}}$$

여러개의 변수 \Rightarrow 각 변수에서 평균과 분산이 다른데 고려

Naïve Bayes Classifier: Variations

- Semi-naïve Bayes classifier
 - ▣ A type of Bayesian classifier that relaxes the strict independence assumption of the Naive Bayes classifier, allowing for limited dependencies between features, while still maintaining computational efficiency and simplicity
 - It assumes that correlations exist only within disjoint subsets of features, meaning that features within a subset can depend on each other, but not across subsets
$$\begin{aligned}P(\mathbf{x}|C) &= P(x_1, x_2, \dots, x_n|C) \\&= P(x_1|C)P(x_2|C) \cdot \dots \cdot P(x_k|C) \cdot P(x_{k+1}, x_{k+2}, \dots, x_n|C)\end{aligned}$$
 - x_1, x_2, \dots, x_k are the features assumed to be conditionally independent
 - $x_{k+1}, x_{k+2}, \dots, x_n$ are the features that are allowed to have dependencies

011 521
27 718

일부 특성 X, 다른!

Naïve Bayes Classifier: Variations

- Hidden naïve Bayes classifier
 - ▣ The Hidden naïve Bayes classifier is an extension of the standard naïve Bayes classifier, which introduces the concept of latent or hidden variables into the model
 - Hidden variables are assumed to influence the observed features but are not directly observed or measured

$$P(C|\mathbf{x}) = \sum_{h \in H} P(C|h) \prod_i P(x_i|C,h)$$

where h represents the hidden variables

- ▣ This model is especially useful when we suspect that there are unobserved or hidden factors influencing the outcome, and these factors should be taken into account for better classification performance

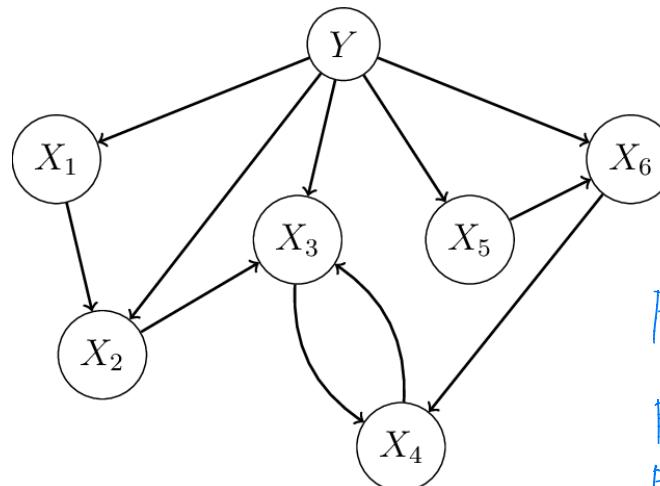
간접 feature 와 숨겨진 변수 존재 가능.

Naïve Bayes Classifier: Variations

- Bayesian network classifier
 - ▣ A Bayesian network classifier a probabilistic classifier that uses a **Bayesian network** to model the relationships between variables
 - ▣ It allows dependencies between features by using a probabilistic graphical model

$$P(\mathbf{x}) = \prod_{i=1} P(x_i | pa_i)$$

- pa_i represents the set of parent nodes for x_i
- ▣ It requires structure learning which can be computationally expensive, especially for large datasets with many features



Study sleep
 \ /
 intelligence
 |
 pass

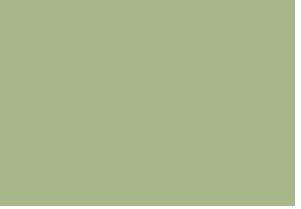
$$P(i = \text{high} \mid \text{study} = \text{yes}, \text{sleep} = \text{no.}) = 0.7$$

$$P(\text{pass} = \text{yes} \mid i = \text{high}) = 0.95$$

$$P(\text{pass} = \text{yes} \mid \text{study} = \text{yes}, \text{sleep} = \text{no.}) = 0.7 \times 0.95$$

NEAREST NEIGHBORS METHODS/ MODEL EVALUATION

Week09



k -NN

Review: Types of Classifiers

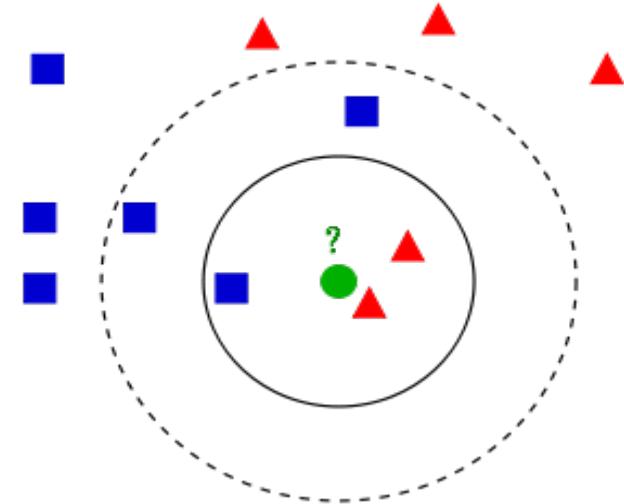
- A classifier is a function that assigns to a sample, \mathbf{x} a class label \hat{y}
$$\hat{y} = f(\mathbf{x})$$
- A probabilistic classifier obtains conditional distributions $\Pr(Y|\mathbf{x})$, meaning that for a given $\mathbf{x} \in \mathcal{X}$, they assign probabilities to all $y \in Y$
 - Hard classification

$$\hat{y} = \arg \max_y \Pr(Y = y | \mathbf{x})$$

Any other classifiers not belonging to a probabilistic approach?

k -Nearest Neighbors(k NN)

- Nonparametric method used for classification and regression
- For classification
 - Output class of data sample is determined by output class of its k -nearest neighbors
 - Majority vote
 - assign the output class to the most common class among k -nearest neighbors
- For regression
 - Output value of data sample is determined by output value of its k -nearest neighbors of the data sample
 - Output value is the average value of k -nearest neighbors
 - There are several different ways to calculate average



※ Parametric Methods

273 453

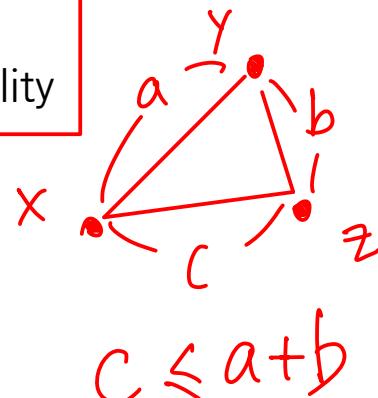
- Parametric method
 - ▣ Parametric methods are statistical and machine learning techniques that assumes a fixed number of parameters for the model (e.g., linear regression, logistic regression)
 - ▣ They requires assumptions about the data distribution
- Key characteristics
 - ▣ Fixed Number of Parameters: Parametric models assume the data follows a specific distribution (e.g., normal distribution) and uses a fixed set of parameters to describe that distribution
 - ▣ Simplicity and Efficiency: The assumption of a fixed form simplifies the learning process, allowing for faster training and predictions
 - ▣ Interpretability: Parametric models are often easier to interpret because their parameters have a clear meaning within the assumed distribution.
 - ▣ Limited Flexibility: This fixed structure and the reliance on specific assumptions mean that parametric models can be less flexible and may not perform well when the data's characteristics don't match the model's assumptions

- Nonparametric method
 - ▣ Nonparametric methods are statistical and machine learning techniques that do not assume a specific form for the underlying data distribution
 - ▣ These methods are flexible and adapt to the structure of the data without relying on predefined parametric models
- Key characteristics
 - ▣ **No Fixed Parameters:** Unlike parametric models, they do not assume a fixed number of parameters.
 - ▣ **Data-Driven Models:** The structure of the model is determined by the data rather than predefined equations.
 - ▣ **Flexible and Adaptive:** Can capture complex relationships without assuming a functional form.
 - ▣ **Higher Computational Cost:** Since they rely on the data directly, they may require more computation compared to parametric methods

Distance Metrics

- Distance metrics are mathematical functions that quantify the "distance" or dissimilarity between two points or data objects
- Distance measure should hold the following

- $d(x, y) \geq 0$
 - Non-negativity
- $d(x, y) = 0 \Leftrightarrow x = y$
 - Identity of indiscernibles
- $d(x, y) = d(y, x)$
 - symmetry
- $d(x, z) \leq d(x, y) + d(y, z)$
 - Subadditivity or triangle inequality



Distance Metrics: Numerical Data

Euclidean / Manhattan / Minkowski
Mahalanobis

regression

□ Euclidean distance *scale에 따른 거리 기준도 불안정. 벡터 간 상대적 중요도 있다.*

- Calculates the straight-line distance between two points in a multi-dimensional space

- Euclidean distance of two dimensional data points, $(x_1, y_1), (x_2, y_2)$

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

- In general, Euclidean distance of two data points, $(x_1, x_2, \dots, x_n), (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

□ Manhattan distance

- Calculates the sum of the absolute differences between two points

$$\sum_i^n |x_i - y_i|$$

- Less sensitive to outliers than Euclidean distance

Distance Metrics: Numerical Data regression

Minkowski distance

- Generalizes both Euclidean and Manhattan distances

$$\left(\sum_i^n |x_i - y_i|^p \right)^{1/p}$$

- $p=2 \rightarrow$ Euclidean distance
- $p=1 \rightarrow$ Manhattan distance

Mahalanobis distance : different scaling / correlation between parameters \Rightarrow useful

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T S^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}$$

($p \times p$ $p \times p$ $p \times 1$ \Rightarrow $|X|$ scalar)

- S is sample covariance matrix
- If covariance matrix is diagonal (no correlation), the resulting distance measure is as the same as the normalized distance

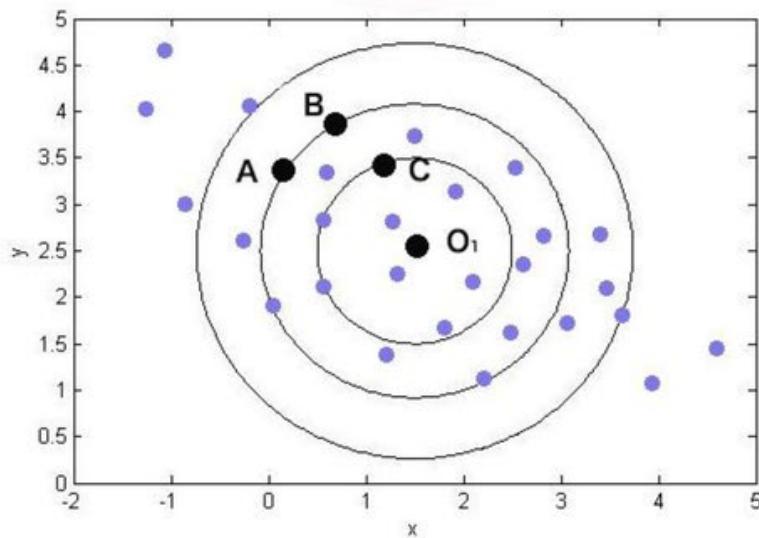
$$\xrightarrow{\text{d}} x'_i = \frac{x_i - \mu}{\sigma}, \quad y'_i = \frac{y_i - \mu}{\sigma}$$

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^p \frac{(x_{1i} - x_{2i})^2}{s_i^2}}$$

$$\begin{aligned} d(x, y) &= \sqrt{\sum (x'_i - y'_i)^2} \\ &= \sqrt{\sum \left(\frac{x_i - y_i}{\sigma} \right)^2} \end{aligned}$$

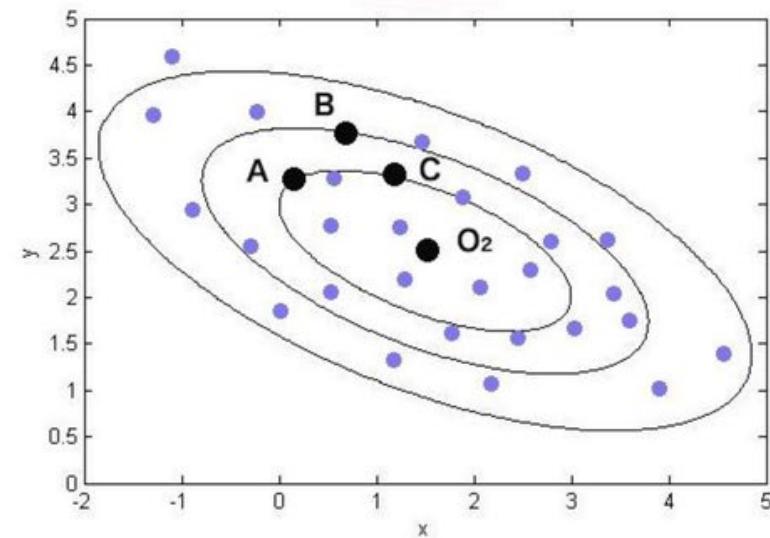
Distance Metrics: Numerical Data regression

- Comparison between Euclidean distance and Mahalanobis distance



(a)

Euclidean distance



(b)

Mahalanobis distance

Distance Metrics: Categorical Data

Jaccard / Hamming

classification.

- Jaccard distance 둘 중 어떤 원소가 (b,c)가 같은지 Jaccard 카운트.
 - ▣ Used to calculate the distance between binary vectors

	x	y	B
A	0	0	1
x	0	a	b
	1	c	d

	A	B
apple	1	1
banana	1	0
melon	0	1
orange	0	0
kiwi	1	0

- a : the total number of attributes where x and y both have a value of 0
- b : the total number of attributes where the attribute of x is 0 and the attribute of y is 1
- c : the total number of attributes where the attribute of x is 1 and the attribute of y is 0
- d : the total number of attributes where x and y both have a value of 1

$$d(x, y) = \frac{b + c}{b + c + d} \quad \frac{1+2}{1+1+2} = \frac{3}{4}$$

두 가지 경우를 가정해
⇒ a는 어떤

Distance Metrics: Categorical Data

□ Hamming distance

$$d(x, y) = \frac{\sum_i I(x_i \neq y_i)}{\dim(x)}$$

- $I(x_i \neq y_i)$ is 1 if and only if $x_i \neq y_i \Rightarrow x_i \neq y_i$ 가 참인 경우는 1.
- $\dim(x)$ is the dimension of x

두 데이터가 몇 개의 항목에서 서로 다른 값을 가지는가.
 \Rightarrow 서로 다른 항목의 개수의 합.

	A	B
apple	1	1
banana	1	0
melon	0	1
orange	0	0
kiwi	1	0

Hamming distance, $d(A, B) = \frac{1+1+1}{5} = \frac{3}{5}$

Question

- Find k -nearest neighbors based on given data points

1) Find k -nearest neighbors of 5th objects when $k=3$ using Euclidean distance

index	x	y
① 1^2+3^2	1 $ +3 \textcircled{3}$	1
② 0^2+1^2	2 $0+1 \textcircled{1}$	3
③ 2^2+2^2	3 $2+2 \textcircled{2}$	6
④ 1^2+1^2	4 $ +1 \textcircled{1}$	1
5	2	4
⑤ 2^2+4^2	6 $2+4 \textcircled{4}$	0
⑥ 5^2+1^2	7 $5+1 \textcircled{1}$	5
⑦ 4^2+2^2	8 $4+2 \textcircled{2}$	2

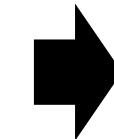
2) Find k -nearest neighbors of 5th objects when $k=3$ using Manhattan distance

tie

Feature Scaling

- Scale of variable affects on determination of nearest neighbors
- Which sample is the nearest neighbor of data sample 1?

i	x_1	x_2	x_3	x_4	y
1	9	30	100	0.5	1
2	9	25	250	0.1	0
3	9	44	220	0.7	0
4	7.5	75	170	1.2	1
...



i	Distance from p_1
1	-
2	150.0838
3	120.8141
4	83.23305
...	...

- Scale of variable x_3 dominates over other variables
- The nearest neighbor is strongly dependent on x_3

It is unfair!

Normalization

- Normalization is to adjust values of variables with different scales to common scale
 - ▣ There are several different ways for normalization
- Commonly used normalization method

$$x \rightarrow \frac{x - \mu}{\sigma}$$

- ▣ μ =mean value of the variable
- ▣ σ =standard deviation of the variable
- ▣ μ and σ are computed by sample data points

$$x \rightarrow \frac{x - x_{min}}{x_{max} - x_{min}}$$

- ▣ x_{max} is the maximum value of variable x and x_{min} is the minimum value of variable x
- ▣ Normalized value is within [0, 1]

Normalization

- Normalization based on normal distribution ($x \rightarrow \frac{x-\mu}{\sigma}$) assumes that the sample points are distributed about the center of mass in a spherical manner
 - In real data, variables are correlated with other variables

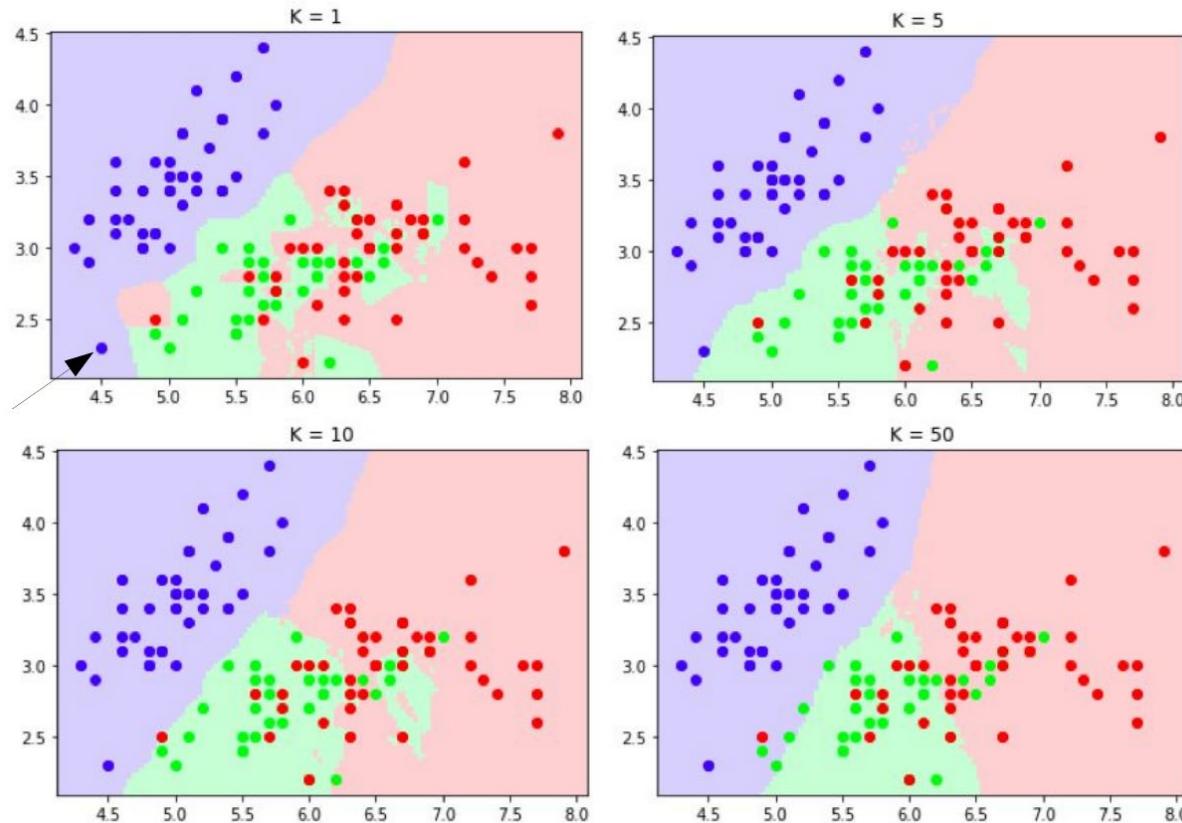
Need to consider scale (level of spread along axis) and correlation to measure distance



Mahalonobis distance

Choosing the Value of k

- Small k leads to high variance (overfitting)
- Large k smooths decision boundaries (underfitting)
- Common approach: use cross-validation to find optimal k



Procedure of k NN

Decide the number of nearest neighbors k and distance measure

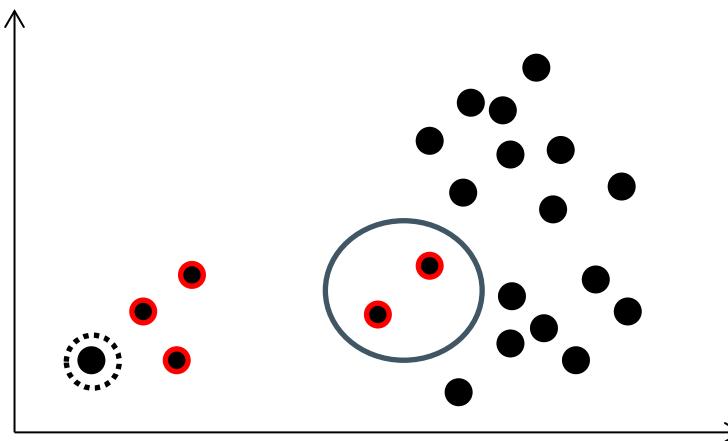
For all data point in test set, find k nearest neighbors

Obtain output value based on output values of neighbors

Fixed-radius Near Neighbors

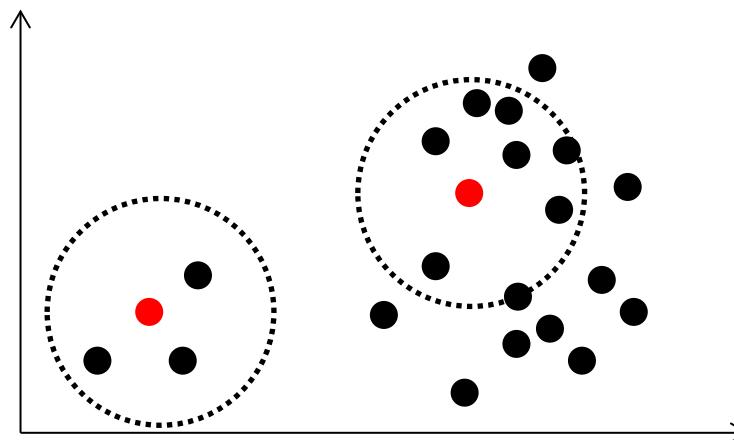
Problem of Fixed-Number of Nearest Neighbors

- When distribution of data set is not homogenous, samples not similar to data point x can be obtained in the nearest neighbors
 (?))
 - $k = 5$



Fixed-Radius Near Neighbors

- Fixed-radius near neighbors are neighbors within fixed range from data point x
 - Because of that, the number of neighbors can be differ from position



Choosing the Value of r

- Small r may lead to insufficient data points
- Large r may include noisy or irrelevant data points
- Adjust based on dataset density

Fixed-Radius Near Neighbors Methods

- The only difference of fixed-radius NN from k NN is the method to find the nearest neighbors
 - ▣ Remained steps of classification and regression are the same

Decide radius of range from data point and distance measure



For all data point in test set, find fixed-radius near neighbors



Obtain output value based on output values of neighbors

Pros and Cons of Nearest Neighbors

KNN & Fixed-Radius NN

□ Pros

- Simple and easy to implement
- No need for explicit training
- Works well with well-separated data

□ Cons

- Computationally expensive for large datasets
- Requires a meaningful distance metric
- Sensitive to irrelevant or redundant features.

Evaluation of Classifiers

Model Evaluation

- In linear regression, several tests are utilized to evaluate regression models and validity of linear regression

How about classification?

- For classification, model evaluation step is required to check validity of classification algorithms
 - Is the model well trained?
 - Is performance of the model enough?

Model Evaluation

- Confusion matrix
 - ▣ A confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data

		Real	
		Positive	Negative
Model	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

- Positive can be used as 1
- Negative can be used as 0

Model Evaluation

Metrics related with confusion matrix

		Real	
		Positive	Negative
Model	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

- Accuracy = $(TP + TN) / (TP + FP + FN + TN)$ 전체 인스턴스 중 모델이 올바르게 예측한 비율
- Misclassification rate = $(FP + FN) / (TP + FP + FN + TN)$ 모델이 잘못 예측한 비율
- True positive rate = $TP / (TP + FN)$ 재현율. 실제 긍정 종에서 모델이 긍정이라고 예측한 비율
 - Of all the actual positive instances, how many did the model correctly identify?
 - Also known as sensitivity or recall
- False positive rate = $FP / (TN + FP)$
 - When it's actually negative, how often does it predict true?
- Precision = $TP / (TP + FP)$
 - Of all the items the model labeled as positive, how many were actually positive?
- True negative rate (Specificity) = $TN / (TN + FP)$ 실제 부정 인스턴스 중 모델이 부정으로 예측한 비율

Model Evaluation

정밀도와 재현율

F-score

$$\frac{TP}{TP+FP}$$

$$\frac{TP}{TP+FN}$$
 재현율 중 정밀도

- Consider both the precision and recall

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (\beta=1)$$

- Harmonic mean of precision and recall

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

$$\begin{array}{c|l} \beta > 1 & \text{recall } \uparrow \\ 0 < \beta < 1 & \text{precision } \uparrow \end{array}$$

- The formula in terms of Type I and type II errors

$$F_\beta = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2FN + FP}$$

Type I : precision \approx FP

Type II : recall \approx FN

Model Validation

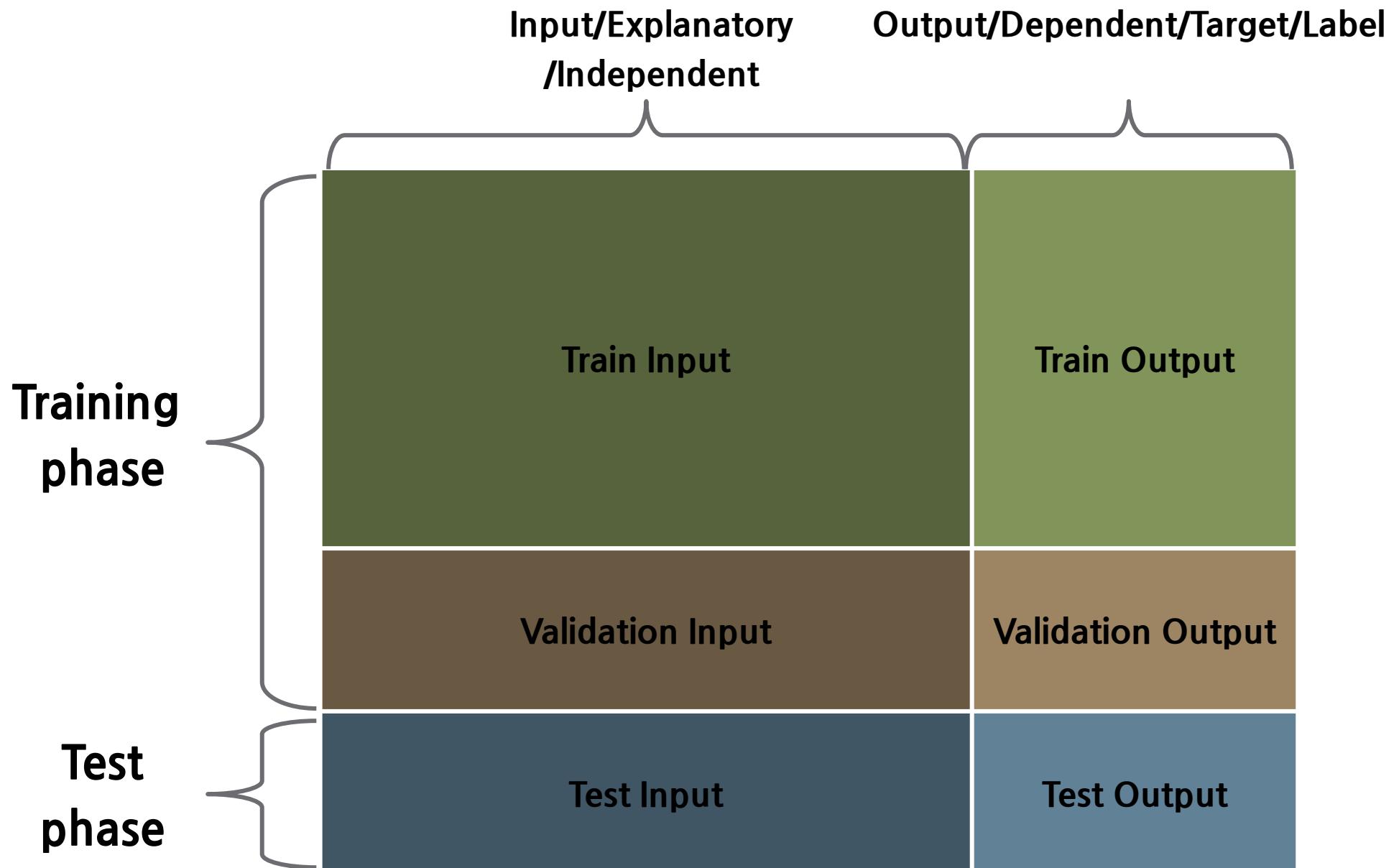
Model Evaluation

Is it right way to evaluate the model?

- The same samples used for training are used to calculate accuracy
 - When learning a model, we know correct answers

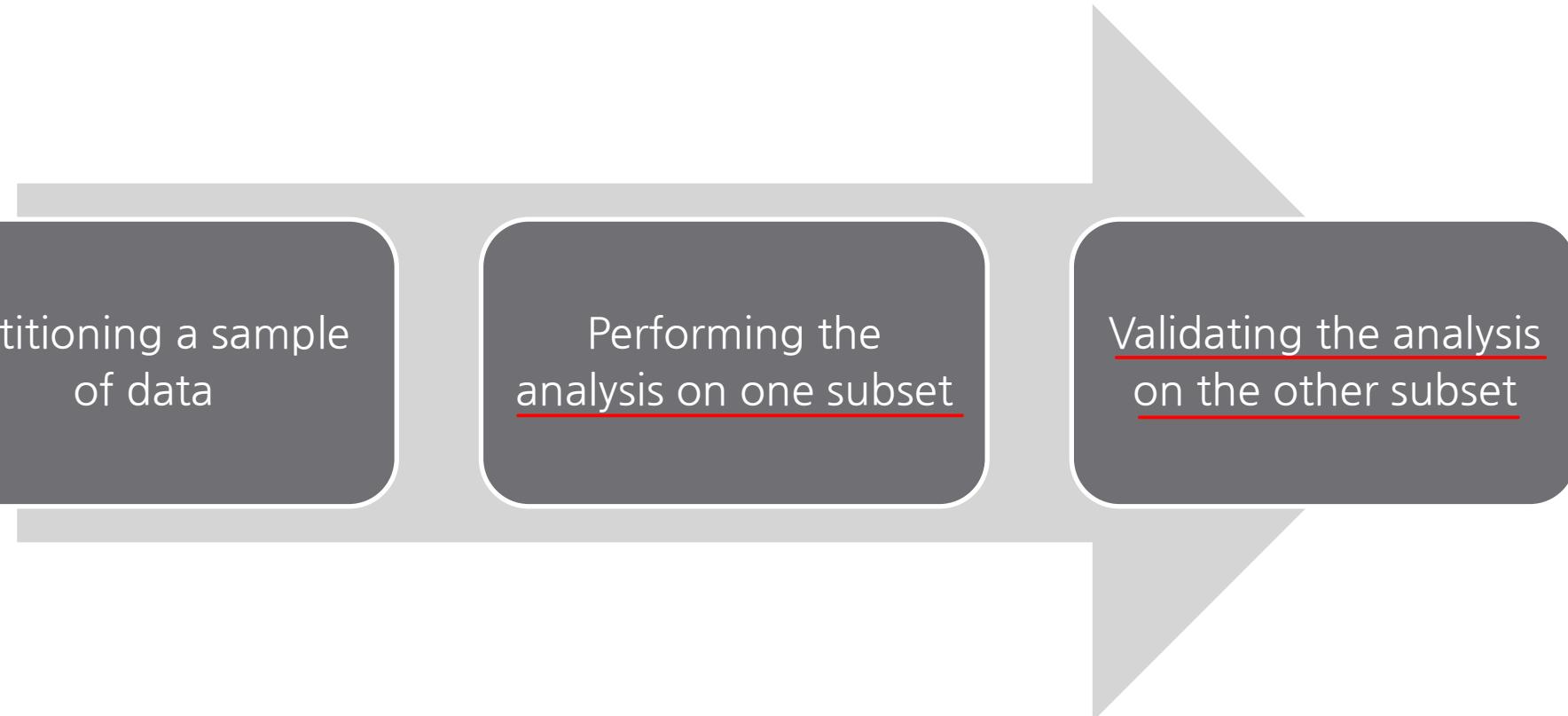
It is not fair way!

Data Partition



Cross-validation

- A model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set
 - Estimate how accurately a predictive model will perform in practice
 - It is also used to determine the best set of parameters



Partitioning a sample
of data

Performing the
analysis on one subset

Validating the analysis
on the other subset

***k*-fold Cross-validation**

- In k -fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples
 - ▣ Of the k subsamples, a single subsample is retained as the validation data
 - ▣ The remaining $k - 1$ subsamples are used as training data
 - ▣ The cross-validation process is then repeated k times with each of the k samples used exactly once as the validation data



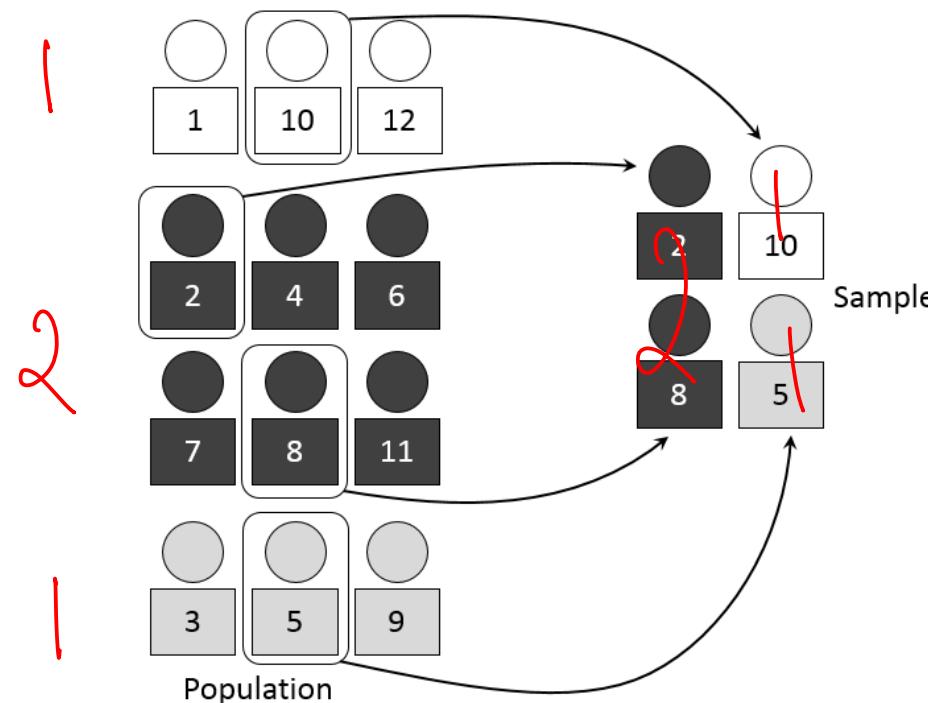
Stratified k -fold Cross-validation

계층적

- A variation of k -fold cross validation that preserves the proportion of classes in each fold
 - ▣ Ensures that class distribution is approximately the same across all folds
 - ▣ Useful for imbalanced classification problems
- Advantages
 - ▣ Better handling of class imbalances
 - ▣ More representative test sets
 - ▣ Suitable for classification tasks with skewed data distributions

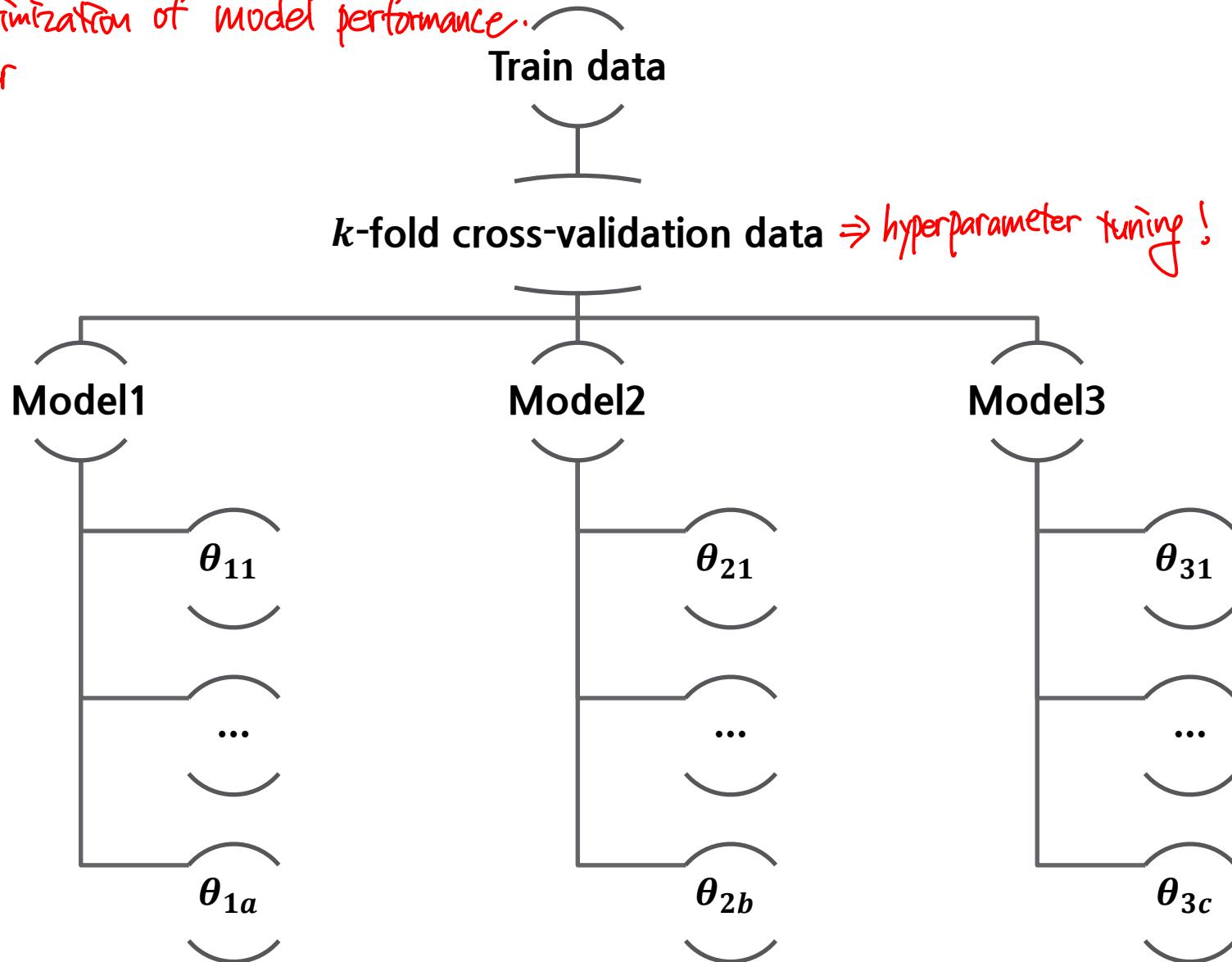
* Stratified Sampling

- Stratified sampling is a method of sampling from a population which can be partitioned into subpopulations
 - For classification analysis, stratified sampling aims at splitting one data set so that each split are similar with respect to class distribution
 - To ensure that the train and test sets have approximately the same percentage of samples of each target class as the complete set



Hyperparameter Tuning & Model Selection

value for optimization of model performance.
Get by user



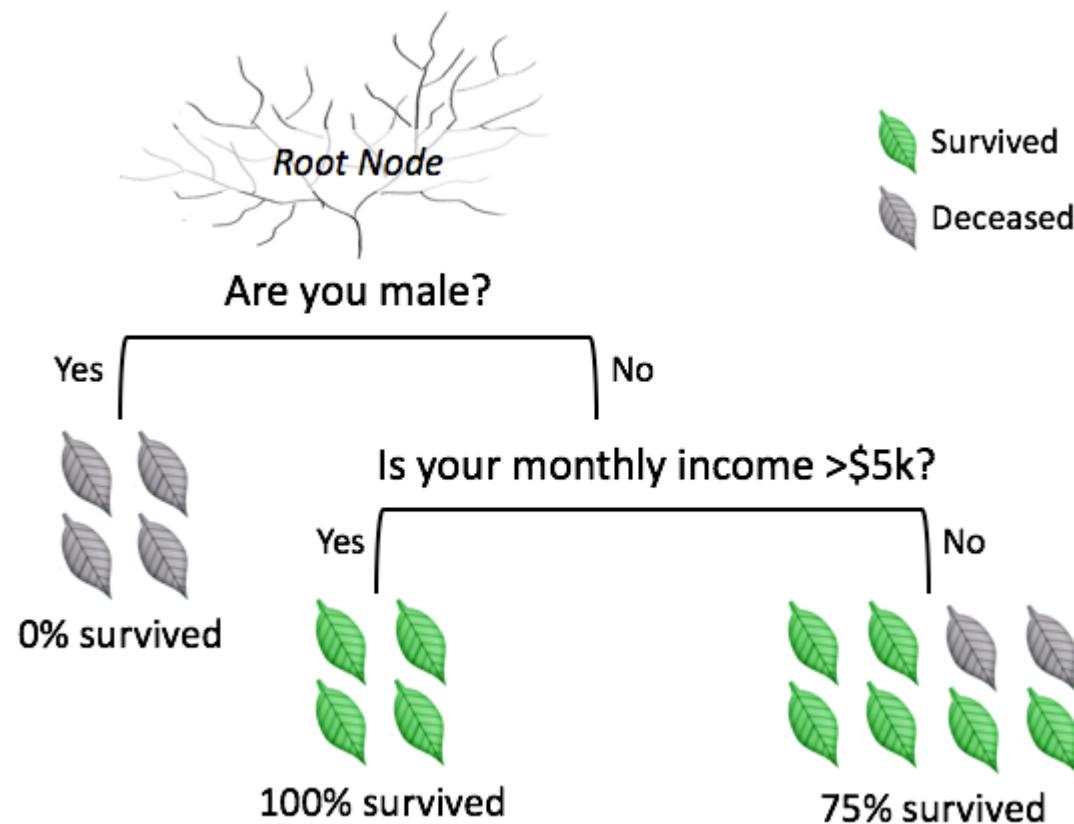
DECISION TREE

Week10

Decision Tree

Introduction to Decision Trees

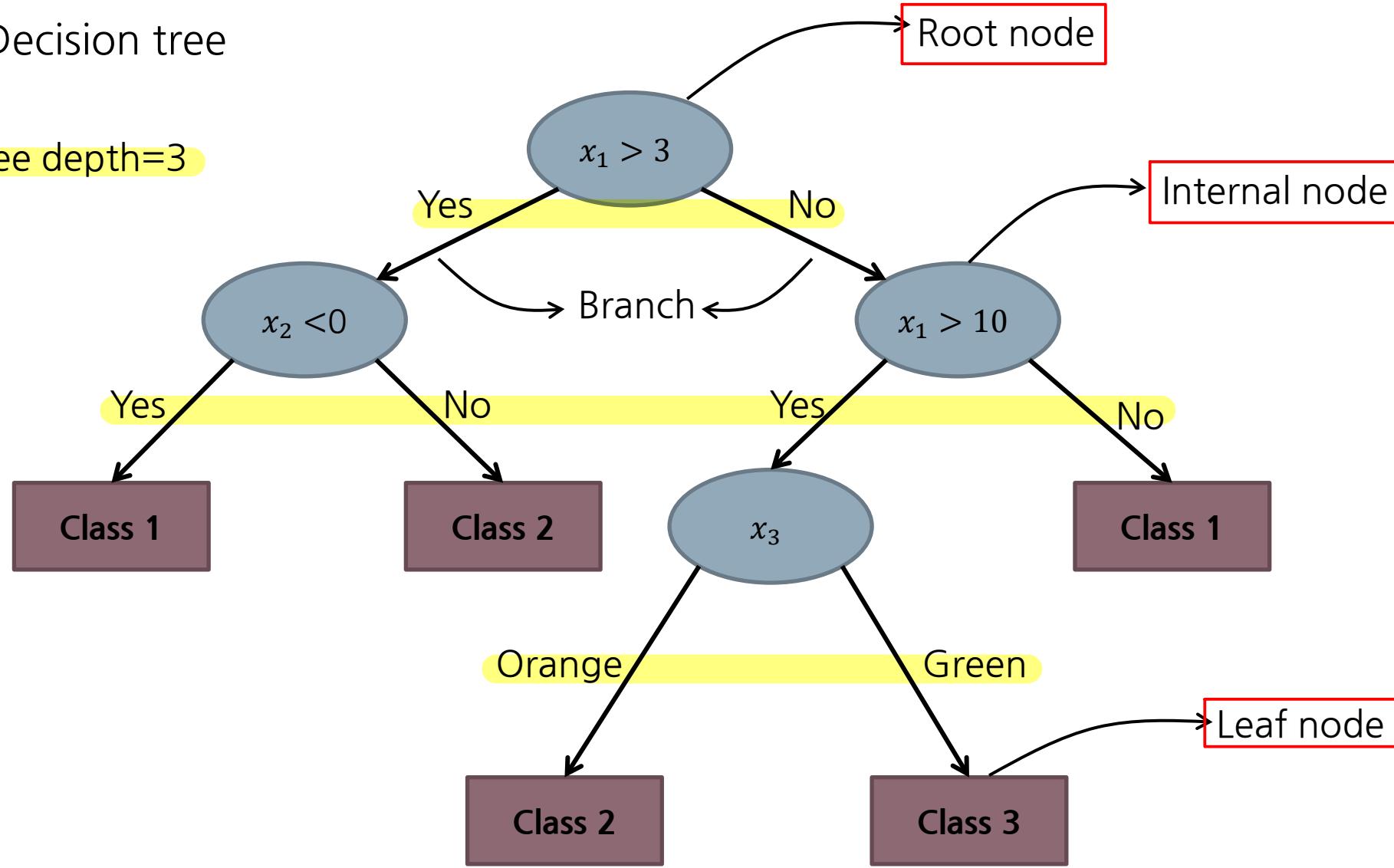
- Decision tree
 - A decision tree is a hierarchical model used for classification and regression
 - It consists of nodes representing decisions, branches representing choices, and leaves representing final outcomes
 - Commonly used due to its interpretability and ease of implementation



Decision Tree

□ Decision tree

Tree depth=3



Decision Tree

- Each root node and internal node represent a specific input variable
 - ▣ Root and internal node tests each attribute
 - $x_1 > 1$
 - x_3 is orange
- Each branch corresponds to the result of the test of node
 - ▣ Yes/No
 - ▣ Values of attribute
 - Orang/Green
 - Long/Short
- Each leaf node assigns a class

**In each node, how to choose attribute?
how to split branches?**

Decision Tree Algorithm

How to determine which one is the most effective?



Need some criteria for measure of effectiveness

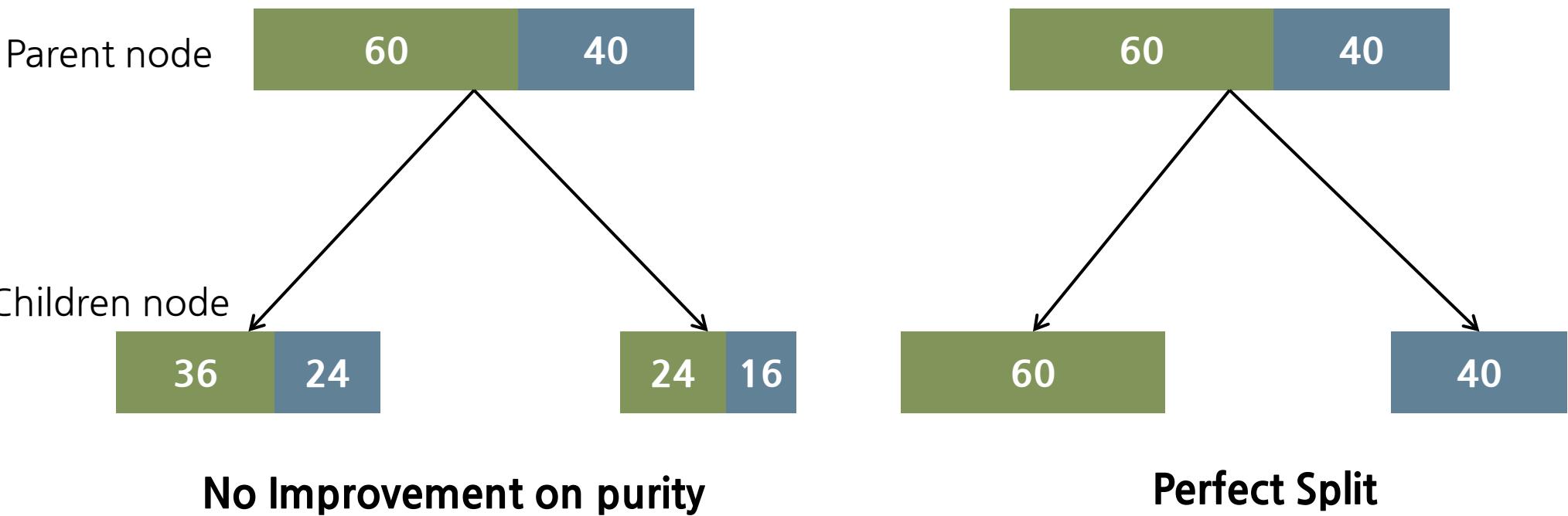
Splitting Criteria

- Categorical target - Classification
 - Entropy
 - Gini impurity
- Continuous target - Regression
 - MSE
 - Friedman MSE
 - MAE

Splitting Criteria for Classification: Purity

특정 노드의 샘플이 얼마나 순수한지
클래스에 정착되어 있는지 측정

- Select each split of a node so that in each of the child nodes are purer or less impure than that in the parent node



- Entropy and Gini impurity are measures of impurity
 - Split a node toward decreasing impurity → Maximize reduction in impurity

Splitting Criteria for Classification: Entropy ↓ purity ↑

□ Expected value of the information

- The less likely an event is, the more information it provides when it occurs.
- Information \leftrightarrow uncertainty

□ Entropy H of event X

$$H(X) = E[I(X)] = E[-\ln P(X)]$$

↑ ↑
Expectation value Probability function

Information content of X

□ For a finite sample

$$H(X) = \sum_i P(x_i)I(x_i) = -\sum_i P(x_i) \log_b P(x_i)$$

- X with possible values of $\{x_1, x_2, \dots, x_n\}$
- Commonly b is 2 ($10, e$ are also used)

Example: Entropy

high entropy: high uncertainty, low purity

- When you flip one coin (X)
 - ▣ Possible output of X : (H), (T)
 - ▣ $P(H) = 0.5, P(T) = 0.5$

$$\begin{aligned} H(X) &= -P(H) \log_2 P(H) - P(T) \log_2 P(T) \\ &= -0.5 \log_2 0.5 - 0.5 \log_2 0.5 \\ &= 1 \end{aligned}$$

- When you flip two coins (X)
 - ▣ Possible output of X : (H,H), (H,T), (T,H), (T,T)
 - ▣ $P(H,H) = P(H,T) = P(T,H) = P(T,T) = 0.25$

$$\begin{aligned} H(X) &= -P(H,H) \log_2 P(H,H) - P(H,T) \log_2 P(H,T) - P(T,H) \log_2 P(T,H) - P(T,T) \log_2 P(T,T) \\ &= -4 \times 0.25 \log_2 0.25 \\ &= 2 \end{aligned}$$

How to Define Effectiveness of Split

- If split is effective, information gain is large
 - Information gain=reduction of uncertainty

Information gain
with the split on attribute a

$$IG(T, a) = H(T) - H(T|a)$$

↑ ↑ ↑
Entropy of original set Entropy of new state after split

Entropy of new state
after split

- Entropy of new state after split=normalized sum of entropy of split sets

$$H(T|a) = \sum_i^n \frac{|T'_i|}{|T|} \cdot H(T'_i)$$

- T is split to T'_1, T'_2, \dots, T'_n

Example: Calculate Information Gain through Entropy

- The node is split by age to predict profit of company

Age(x)	old	old	old	mid	mid	mid	mid	new	new	new
Profit(y)	down	down	down	down	down	up	up	up	up	up

$$H(T) = -P(\text{down}) \log_2 P(\text{down}) - P(\text{up}) \log_2 P(\text{up})$$

$$= -2 \times 0.5 \log_2 0.5 = 1$$

$P(\text{down}) = P(\text{up}) = \frac{1}{2}$

Is age old?

Yes

$P(\text{down}) = 1$

No

$P(\text{down}) = \frac{2}{7}, P(\text{up}) = \frac{5}{7}$

Age(x)	old	old	old
Profit(y)	down	down	down

Age(x)	mid	mid	mid	mid	new	new	new
Profit(y)	down	down	up	up	up	up	up

$$H(T'_1) = -1 \log_2 1 = 0$$

$$H(T'_2) = -\frac{2}{7} \log_2 \frac{2}{7} - \frac{5}{7} \log_2 \frac{5}{7} \approx 0.86$$

$$IG = H(\text{before}) - H(\text{after}) = 1 - 0.86 \times \frac{7}{10} = 1 - 0.602 = 0.398$$

$1 - \left(\frac{3}{10} \times 0 + \frac{7}{10} \times 0.86 \right) = 0.398$

Splitting Criteria for Classification: Gini Impurity $G \downarrow$ purity \uparrow

- Gini Impurity: measure of impurity

$$G(T) = \sum_{i \neq j} P(i|T)P(j|T) = 1 - \sum_j P(j|T)^2 = 1 - \sum_j \left(\frac{n_j(T)}{n(T)} \right)^2$$

- $P(j|t)$ is the probability of output j in node T 노드 T 에 속한 샘플 중 클래스 j 해당하는 확률.
- $n(t)$ is the total number of samples in node T 노드 T 속한 샘플 총 개수.
- $n_j(t)$ is the number of samples with output j in node T 노드 T 속한 j 클래스 개수.



$$\begin{aligned} G &= 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{2}{7}\right)^2 - \left(\frac{2}{7}\right)^2 \\ &= \frac{32}{49} \end{aligned}$$



$$\begin{aligned} G &= 1 - \left(\frac{1}{7}\right)^2 - \left(\frac{5}{7}\right)^2 - \left(\frac{1}{7}\right)^2 \\ &= \frac{22}{49} \Rightarrow \text{more pure} \end{aligned}$$

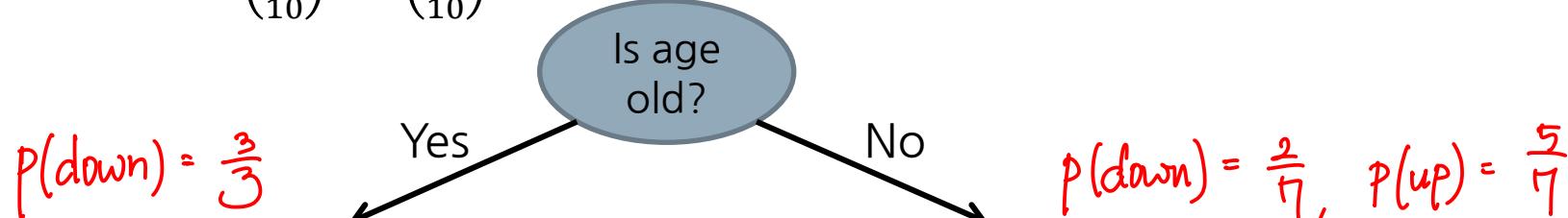
Example: Calculate Information Gain through Gini Impurity

- The node is split by age to predict profit of company

Age(x)	old	old	old	mid	mid	mid	mid	new	new	new
Profit(y)	down	down	down	down	down	up	up	up	up	up

$$G(T) = 1 - P^2(\text{down}) - P^2(\text{up})$$

$$= 1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2 = 0.5$$



Age(x)	old	old	old
Profit(y)	down	down	down

Age(x)	mid	mid	mid	mid	new	new	new
Profit(y)	down	down	up	up	up	up	up

$$G(T'_1) = 1 - \left(\frac{3}{3}\right)^2 = 0$$

$$G(T'_2) = 1 - \left(\frac{2}{7}\right)^2 - \left(\frac{5}{7}\right)^2 \approx 0.41$$

$$IG = G(\text{before}) - G(\text{after}) = 0.5 - 0.41 \times \frac{7}{10} = 0.5 - 0.287 = 0.213$$

$$0.5 - \left(\frac{3}{10} \times 0 + \frac{7}{10} \times 0.41\right) = 0.213$$

Question

- Predict profit of companies based on age of company, type of company, and competition status

$$\begin{aligned} \text{IG} &= H(\text{before}) - H(\text{after}) \\ &= G(\text{before}) - G(\text{after}) \end{aligned}$$

$$H(X) = -\sum_i P(x_i) \log_2 P(x_i)$$

$$G(T) = 1 - \sum_i \left(\frac{n_i(T)}{n(T)} \right)^2$$

Age	Competition	Type	Profit
old	yes	S/W	down
old	no	S/W	down
old	no	H/W	down
mid	yes	S/W	down
mid	yes	H/W	down
mid	no	H/W	up
mid	no	S/W	up
new	yes	S/W	up
new	no	H/W	up
new	no	S/W	up

$$P(\text{down}) = 0.5 \quad P(\text{up}) = 0.5$$

$$\begin{aligned} H(X) &= -0.5 \log_2 0.5 - 0.5 \log_2 0.5 \\ &= -2 \times \frac{1}{2} \times \log_2 2^{-1} = 1 \end{aligned}$$

$$\begin{aligned} \text{yes} \Rightarrow P(\text{down}) &= \frac{3}{4}, \quad P(\text{up}) = \frac{1}{4} \\ -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} &\approx 0.8113 \end{aligned}$$

$$\begin{aligned} \text{no} \Rightarrow P(\text{down}) &= \frac{2}{6}, \quad P(\text{up}) = \frac{4}{6} \\ -\frac{2}{6} \log_2 \frac{2}{6} - \frac{4}{6} \log_2 \frac{4}{6} &\approx 0.9183 \end{aligned}$$

- How much information gain based on Entropy is obtained with the splitting on competition?

$$\text{IG} = 1 - \left(\frac{4}{10} \times 0.8113 + \frac{6}{10} \times 0.9183 \right) = 0.1245$$

- How much information gain based on Gini impurity is obtained with the splitting on type of company?

Question

- Predict profit of companies based on age of company, type of company, and competition status

$$IG = H(\text{before}) - H(\text{after})$$

$$= G(\text{before}) - G(\text{after})$$

$$H(X) = - \sum_i P(x_i) \log_2 P(x_i)$$

$$G(T) = 1 - \sum_i \left(\frac{n_i(T)}{n(T)} \right)^2$$

Age	Competition	Type	Profit
old	yes	S/W	down
old	no	S/W	down
old	no	H/W	down
mid	yes	S/W	down
mid	yes	H/W	down
mid	no	H/W	up
mid	no	S/W	up
new	yes	S/W	up
new	no	H/W	up
new	no	S/W	up

$$G(T) = 1 - \left(\frac{1}{2} \right)^2 - \left(\frac{1}{2} \right)^2 = 0.5$$

$$SW \Rightarrow P(\text{down}) = \frac{3}{6} \quad P(\text{up}) = \frac{3}{6}$$

$$1 - \left(\frac{3}{6} \right)^2 - \left(\frac{3}{6} \right)^2 = \frac{1}{2}$$

$$HW \Rightarrow P(\text{down}) = \frac{2}{4} \quad P(\text{up}) = \frac{2}{4}$$

$$1 - \left(\frac{2}{4} \right)^2 - \left(\frac{2}{4} \right)^2 = \frac{1}{2}$$

- How much information gain based on Entropy is obtained with the splitting on competition?

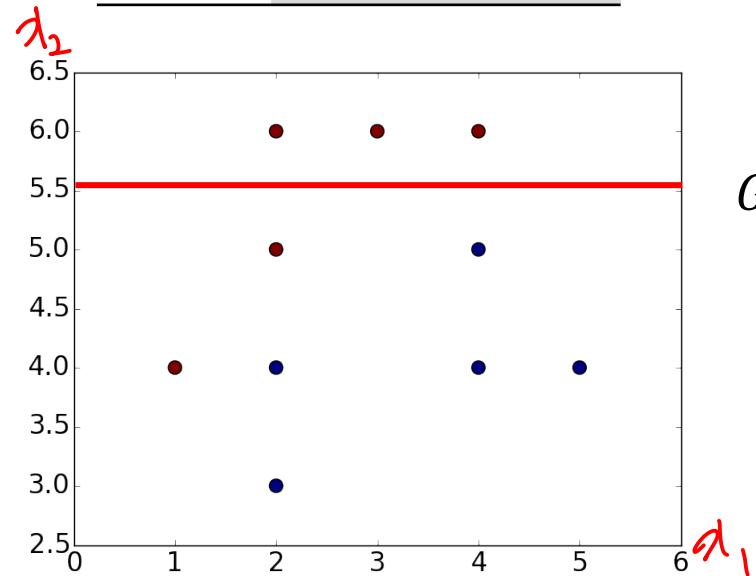
0.1245

- How much information gain based on Gini impurity is obtained with the splitting on type of company?

$$IG = 0.5 - \left(\frac{6}{10} \times \frac{1}{2} + \frac{4}{10} \times \frac{1}{2} \right) = 0$$

Simple Example for Tree

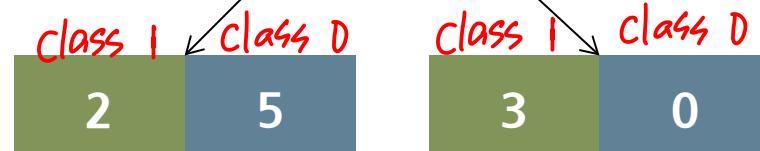
Class	x_1	x_2
1	1	4
1	2	6
1	2	5
0	2	4
0	2	3
1	3	6
1	4	6
0	4	5
0	4	4
0	5	4



$$G = 1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2 = 0.5$$



Yes $x_2 < 5.5$ No



$$G = 1 - \left(\frac{2}{7}\right)^2 - \left(\frac{5}{7}\right)^2 \approx 0.4083$$

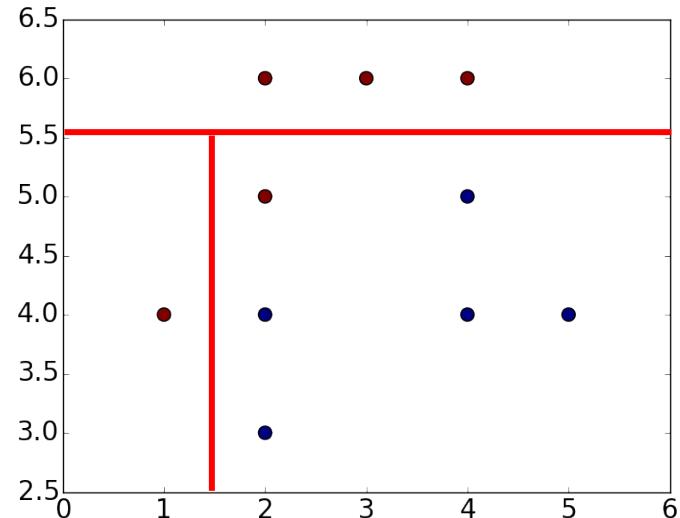
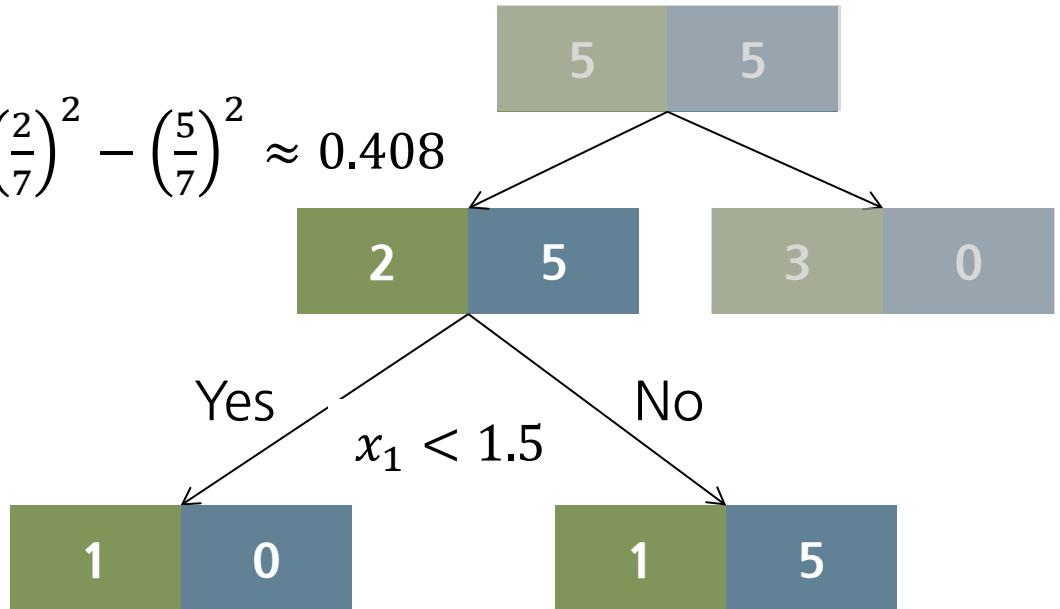
$$G = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$$IG = 0.5 - 0.408 \times \frac{7}{10} - 0 \times \frac{3}{10} = 0.2144$$

Simple Example for Tree

Class	x_1	x_2
1	1	4
1	2	6
1	2	5
0	2	4
0	2	3
1	3	6
1	4	6
0	4	5
0	4	4
0	5	4

$$G = 1 - \left(\frac{2}{7}\right)^2 - \left(\frac{5}{7}\right)^2 \approx 0.408$$



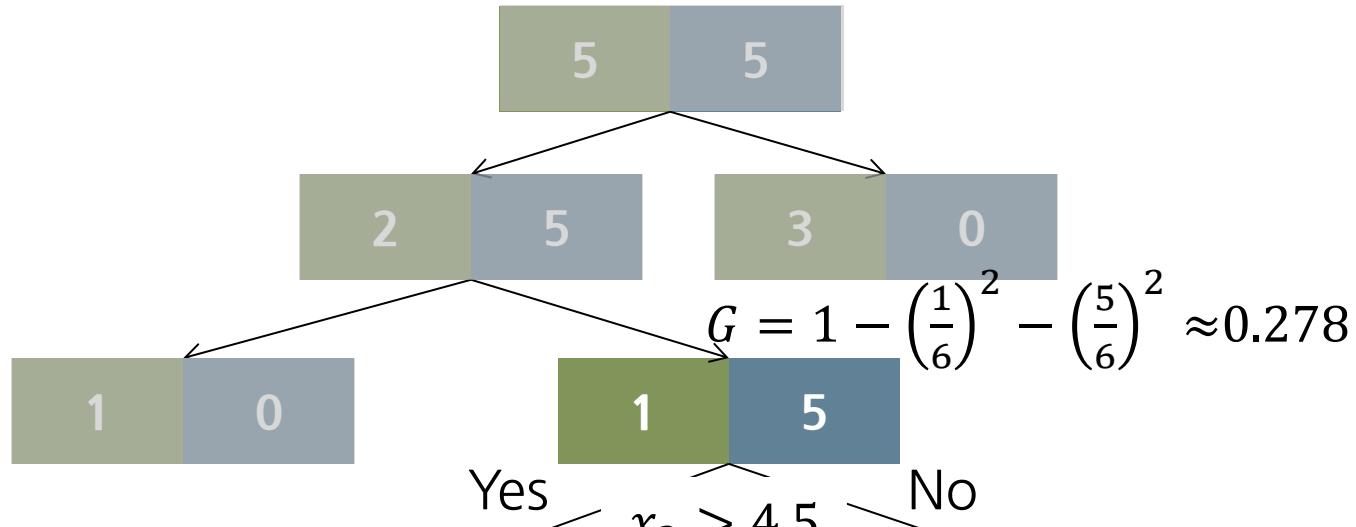
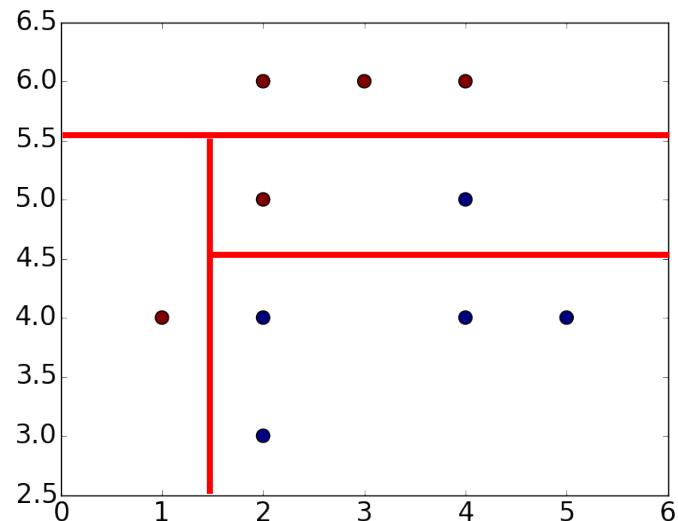
$$G = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2 = 0$$

$$G = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 \approx 0.278$$

$$IG = \left(0.408 - 0 \times \frac{1}{7} - 0.278 \times \frac{6}{7}\right) \times \frac{7}{10} \approx 0.11$$

Simple Example for Tree

Class	x_1	x_2
1	1	4
1	2	6
1	2	5
0	2	4
0	2	3
1	3	6
1	4	6
0	4	5
0	4	4
0	5	4



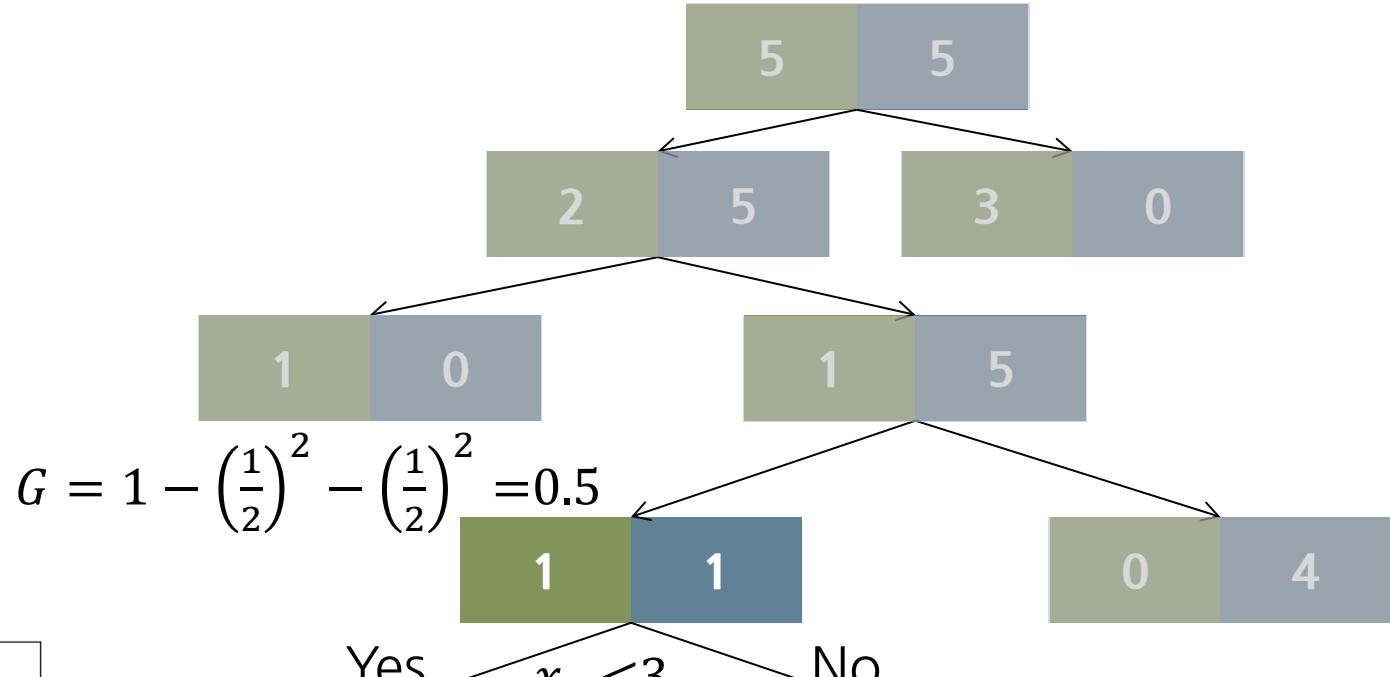
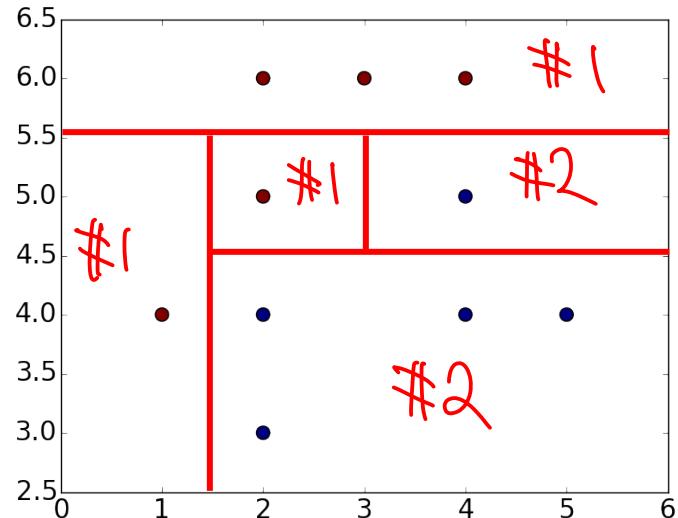
$$G = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$G = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 = 0$$

$$IG = \left(0.278 - 0.5 \times \frac{2}{6} - 0 \times \frac{4}{6}\right) \times \frac{6}{10} \approx 0.067$$

Simple Example for Tree

Class	x_1	x_2
1	1	4
1	2	6
1	2	5
0	2	4
0	2	3
1	3	6
1	4	6
0	4	5
0	4	4
0	5	4

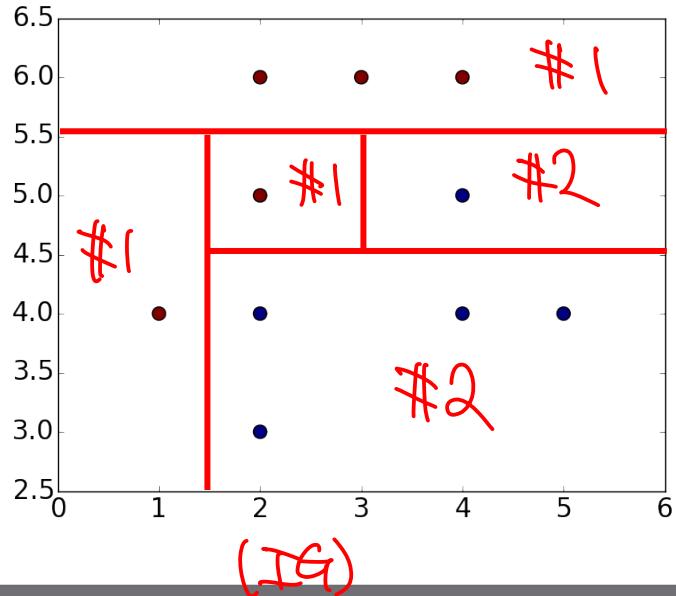


$$G = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$G = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2 = 0 \quad G = 1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2 = 0$$

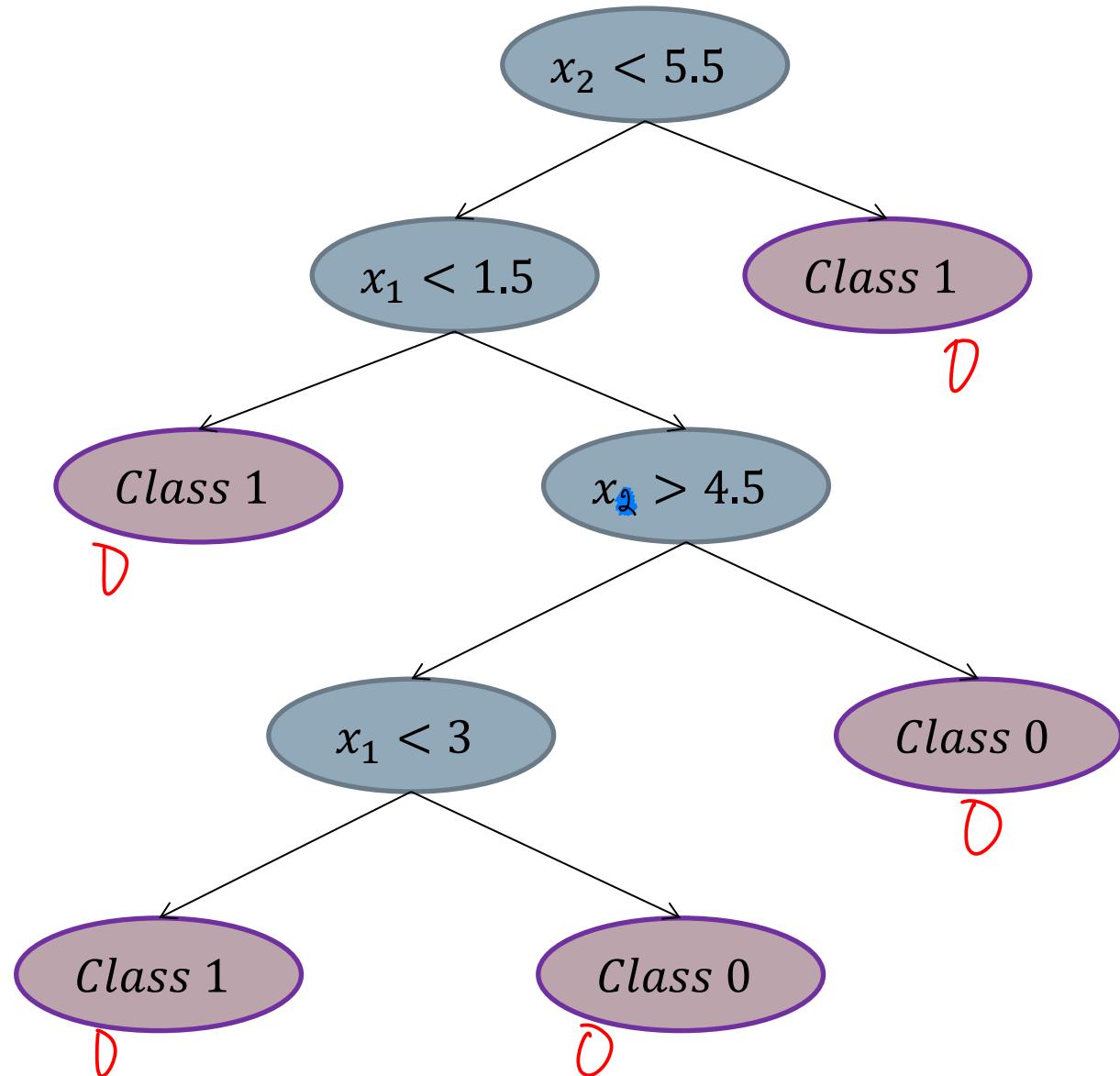
$$IG = \left(0.5 - 0 \times \frac{1}{2} - 0 \times \frac{1}{2}\right) \times \frac{2}{10} = 0.1$$

Simple Example for Tree



(IG)

Step	Impurity Change	Total impurity
0	0.000	0.500
1	0.214	0.286
2	0.119	0.167
3	0.067	0.100
4	0.100	0.000



Pros and Cons of Decision Tree

- Pros
 - ▣ Easy interpretation
 - ▣ Non-parametric approach
 - ▣ Inherently non-linear
 - ▣ Easy to handle categorical variables
 - ▣ Implicitly perform feature selection

- Cons
 - ▣ Large computing cost
 - ▣ Lack of linearity or main effects
 - ▣ Each node only considers single variable
 - Many algorithms has been proposed to overcome this problem

When Does Tree Stop Growing?

- Growing full-size tree can cause overfitting
 - ▣ Low classification accuracy on test set
- Introduce pruning step after growing tree
 - ▣ Pruning simplifies the tree by trimming some branches of the fully grown tree
 - ▣ Generate several pruned trees and select best tree
- Pruning techniques
 - ▣ Pre-Pruning (Early Stopping): Stop growing the tree when certain conditions are met (e.g., max depth, min samples per leaf)
 - ▣ Post-Pruning: Grow the tree fully, then remove branches that do not improve accuracy significantly

Cost Complexity Pruning

- Cost complexity pruning
 - ▣ A post-pruning method that reduces overfitting by controlling tree complexity
 - ▣ Uses a penalty term to balance model complexity and accuracy
- Cost complexity pruning algorithm
 1. Grow a fully expanded tree
 2. Compute the cost complexity for each subtree
 3. Remove nodes that do not significantly improve performance
 4. Select the optimal subtree using cross-validation.

Cost Complexity Pruning

- Cost complexity measure

$$R_\alpha(T) = R(T) + \alpha|T|$$

Complexity parameter

- $R(T)$ is misclassification cost of T
- $|T|$ is tree complexity=the number of terminal nodes

Cost Complexity Pruning

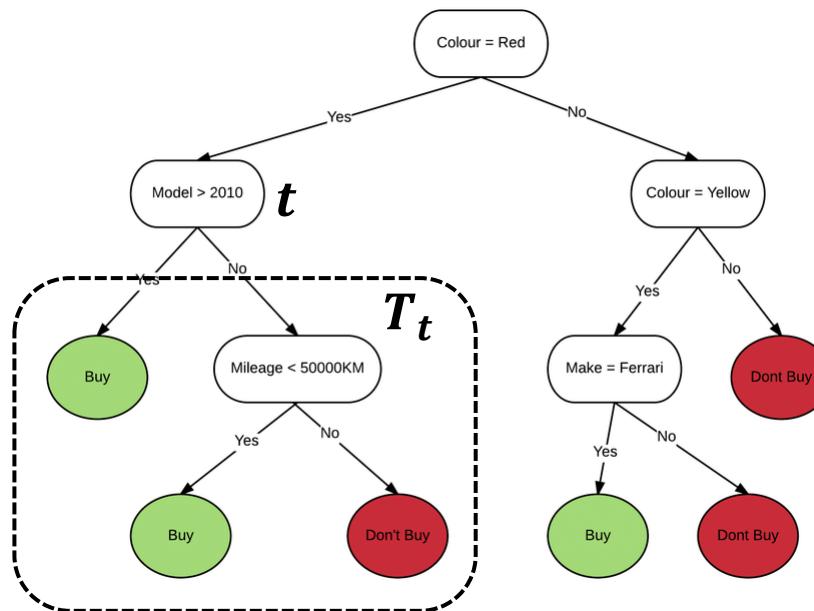
- Calculate values of cost complexity measure for internal node t before and after removing subtree T_t

- For a subtree at t *subtree 티지*

$$R_\alpha(T_t) = R(T_t) + \alpha|T_t|$$

- For a terminal node t *subtree 가지지 않고 단일 leaf로 만든다*

$$R_\alpha(t) = R(t) + \alpha$$



Cost Complexity Pruning

□ Misclassification cost at node t

Class 뷰포가 $\{0.2, 0.7, 0.1\}$ 이라고
 $\max_k p(k|t) = 0.7, r(t) = 1 - 0.7 = 0.3$

$$r(t) = \min_i \sum_{k=1}^K C(i|k)p(k|t) \rightarrow r(t) = 1 - \boxed{\max_k p(k|t)}$$

- $C(i|k) = \begin{cases} 1, & \text{if } i \neq k \\ 0, & \text{if } i = k \end{cases}$
- $p(k|t)$ is probability that class of data point is k given that it is in node t
- Misclassification cost of tree T

$$R(T) = \sum_{t \in \tilde{T}} r(t)p(t) = \sum_{t \in \tilde{T}} R(t)$$

- \tilde{T} is set of terminal nodes of tree T
- $p(t)$ is probability that data point is in node t 상수 비례
- Set $R(t) = r(t)p(t)$

Cost Complexity Pruning

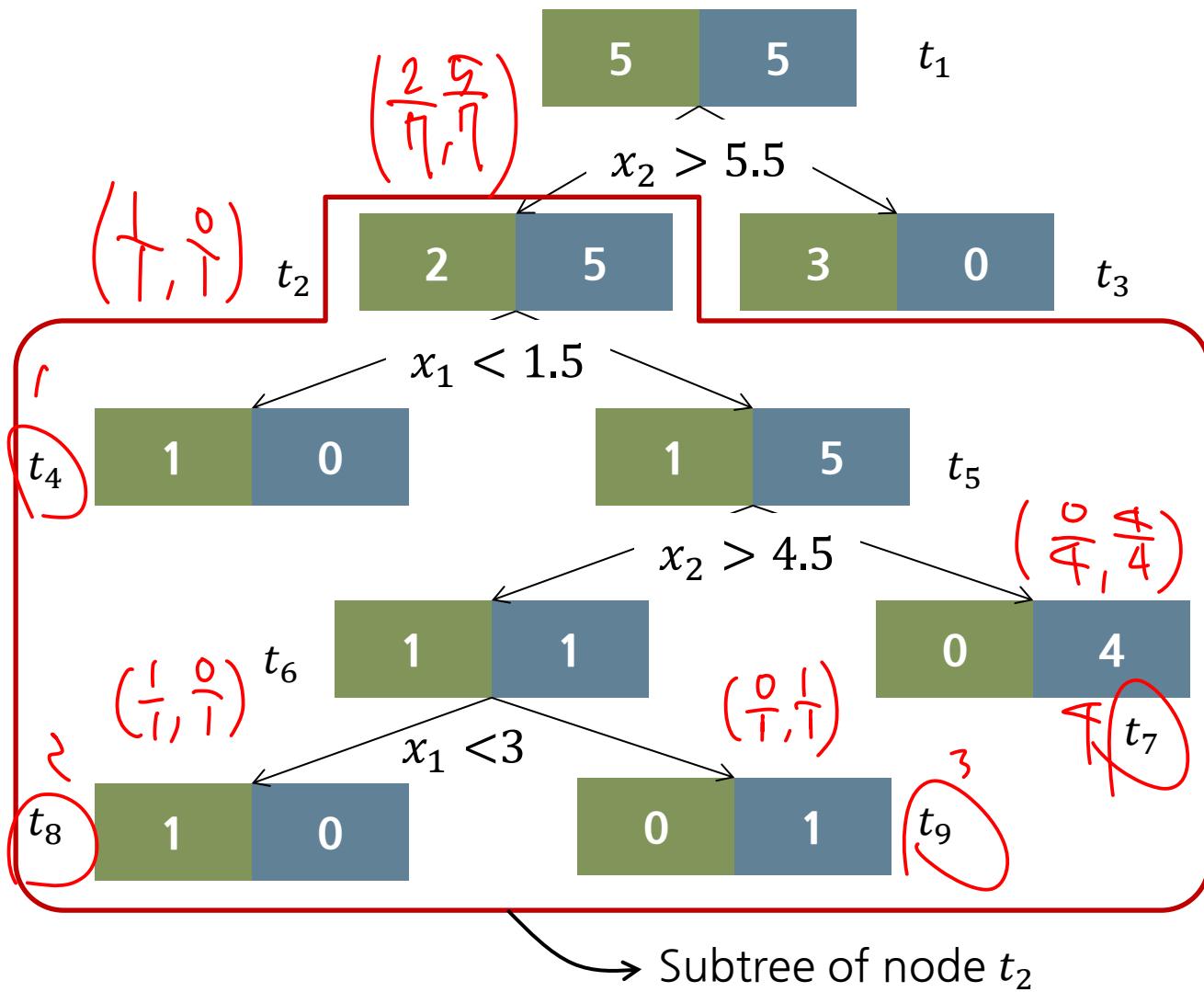
- Cost complexity pruning prunes subtree at t by comparing $R_\alpha(T_t)$ and $R_\alpha(t)$

$$\frac{\text{Subtree 유지}}{R(T_t) + \alpha|T_t|} = \frac{\text{단일노드 대체}}{R(t) + \alpha}$$

$$\alpha(t) = \frac{R(t) - R(T_t)}{|T_t| - 1} \quad \text{복잡도 감소위해 감소하는 만큼의 비중}$$

- A small alpha value at a node means that removing (pruning) the subtree rooted at that node results in only a small increase in the error relative to the reduction in tree complexity
 - This suggests the subtree is not very important and can be pruned early with little cost in accuracy.
- A large alpha value at a node indicates that pruning the subtree would significantly increase the error compared to the gain from simplifying the tree
 - This means the subtree contains useful structure and should be retained longer during pruning.

Cost Complexity Pruning



At node t_2

$$\begin{aligned}
 R(t_2) &= r(t_2)p(t_2) \\
 &= \left(1 - \frac{5}{7}\right) \times \frac{7}{10} = 0.2
 \end{aligned}$$

$\max(\frac{2}{7}, \frac{5}{7})$

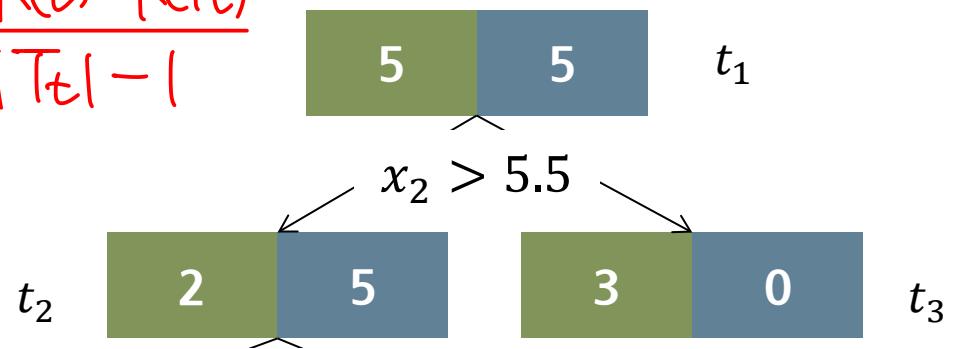
$$|T_{t_2}| = 4$$

$$\begin{aligned}
 R(T_{t_2}) &= (1 - 1) \times \frac{1}{10} + (1 - 1) \times \frac{1}{10} \\
 &\quad + (1 - 1) \times \frac{1}{10} + (1 - 1) \times \frac{4}{4} \\
 &= 0
 \end{aligned}$$

$$\alpha(t_2) = \frac{0.2 - 0}{4 - 1} \approx 0.067$$

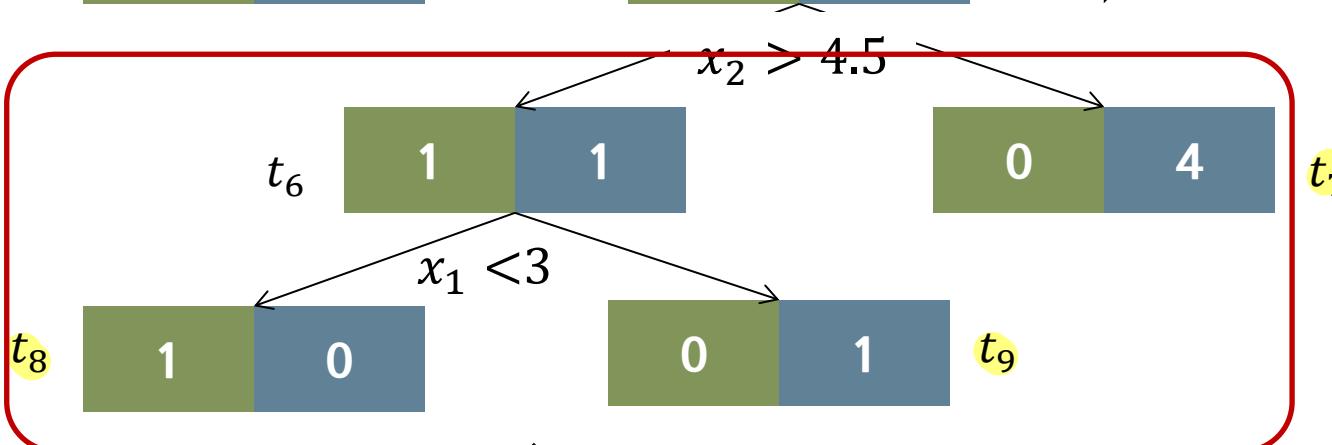
Cost Complexity Pruning

$$\alpha(t) = \frac{R(t) - R(T_t)}{|T_t| - 1}$$



node	$R(t)$	$R(T_t)$	$ T_t $	$\alpha(t)$
t_2	0.2	0	4	0.067
t_5	0.1	0	3	0.050
t_6	0.1	0	2	0.100

Change t_5 to leaf node to classify class 0



$$R(t_5) = R(t_5) P(t_5) = \left(1 - \frac{5}{6}\right) \times \frac{6}{10} = \frac{1}{10}$$

$$|T_{t_5}| = 3$$

$$R(T_{t_5}) = \left(1 - \frac{1}{4}\right) \frac{1}{10} + \left(1 - \frac{1}{4}\right) \frac{1}{10} + \left(1 - \frac{4}{4}\right) \frac{4}{10} = 0$$

Regression Tree: Splitting Criteria

Split 층의

□ Splitting criteria for regression problems

IG↑ 예측오차↓ 산수

- Entropy and Gini impurity are not appropriate split measure for regression analysis
- MSE (Mean Squared Error)
 - The split that most decreases the MSE is selected

$$\hat{y}_i = \frac{\sum_{j \in t_i} y_j}{|t_i|}$$

N_t : 노드 t 생길 수

$$R(t_i) = \frac{1}{N_{t_i}} \sum_{j \in t_i} (y_j - \hat{y}_i)^2$$

$$IG = \underbrace{p(t_p)R(t_p)}_{\text{parent}} - \underbrace{p(t_r)R(t_r)}_{\text{right child}} - \underbrace{p(t_l)R(t_l)}_{\text{left child}}$$

□ Friedman MSE

$$\hat{y}_i = \frac{\sum_{j \in t_i} y_j}{|t_i|}$$

t_r: right node
t_l: left node

$$IG = \frac{N_{t_r} N_{t_l}}{N_{t_r} + N_{t_l}} (\hat{y}_{t_r} - \hat{y}_{t_l})^2$$

Regression Tree

- Splitting criteria for regression problems
 - ▣ MAE (Mean Absolute Error)
 - The split that minimizes the L1 loss using the median of each terminal node is selected

$\hat{y}_i = \text{the median of each terminal node}$

$$R(t_i) = \frac{1}{N_{t_i}} \sum_{j \in t_i} |y_j - \hat{y}_i|$$

$$IG = p(t_p)R(t_p) - p(t_r)R(t_r) - p(t_l)R(t_l)$$

Regression Tree

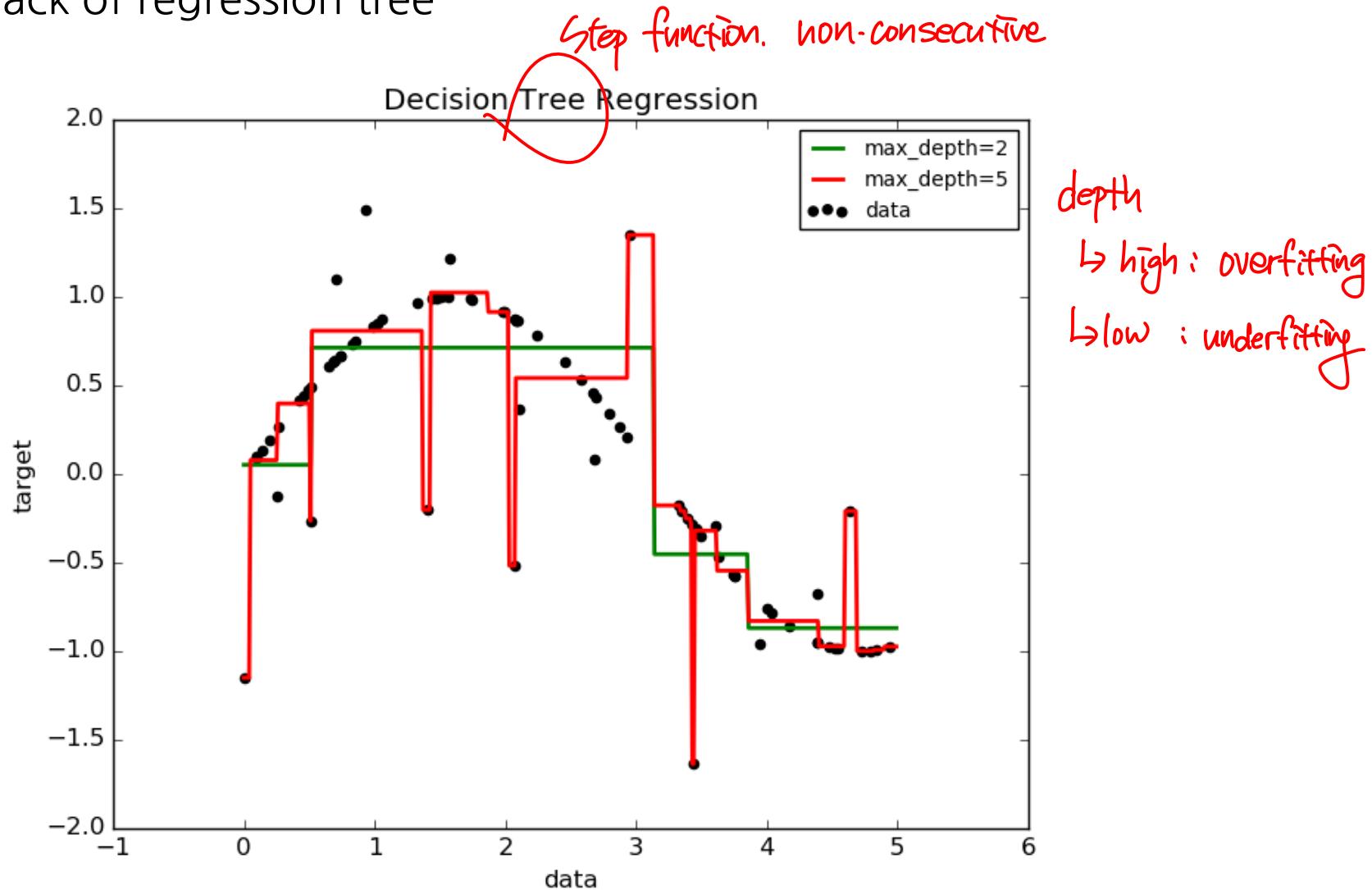
- Predicting a target value
 - ▣ The predicted value for each leaf node is the mean of the target values in that node

$$\hat{y}_t = \frac{1}{N_t} \sum_{i \in node_t} y_i$$

- $node_t$: t -th leaf node
- N_t : the number of training samples assigned to $node_t$

Regression Tree

- Drawback of regression tree



Random Forest

- Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve the predictive performance and reduce overfitting
 - ▣ Combination of bagging idea and the random selection of features
 - ▣ T number of trees are training from T bootstrap samples
 - For regression trees, output is obtained by averaging the predictions from all trees

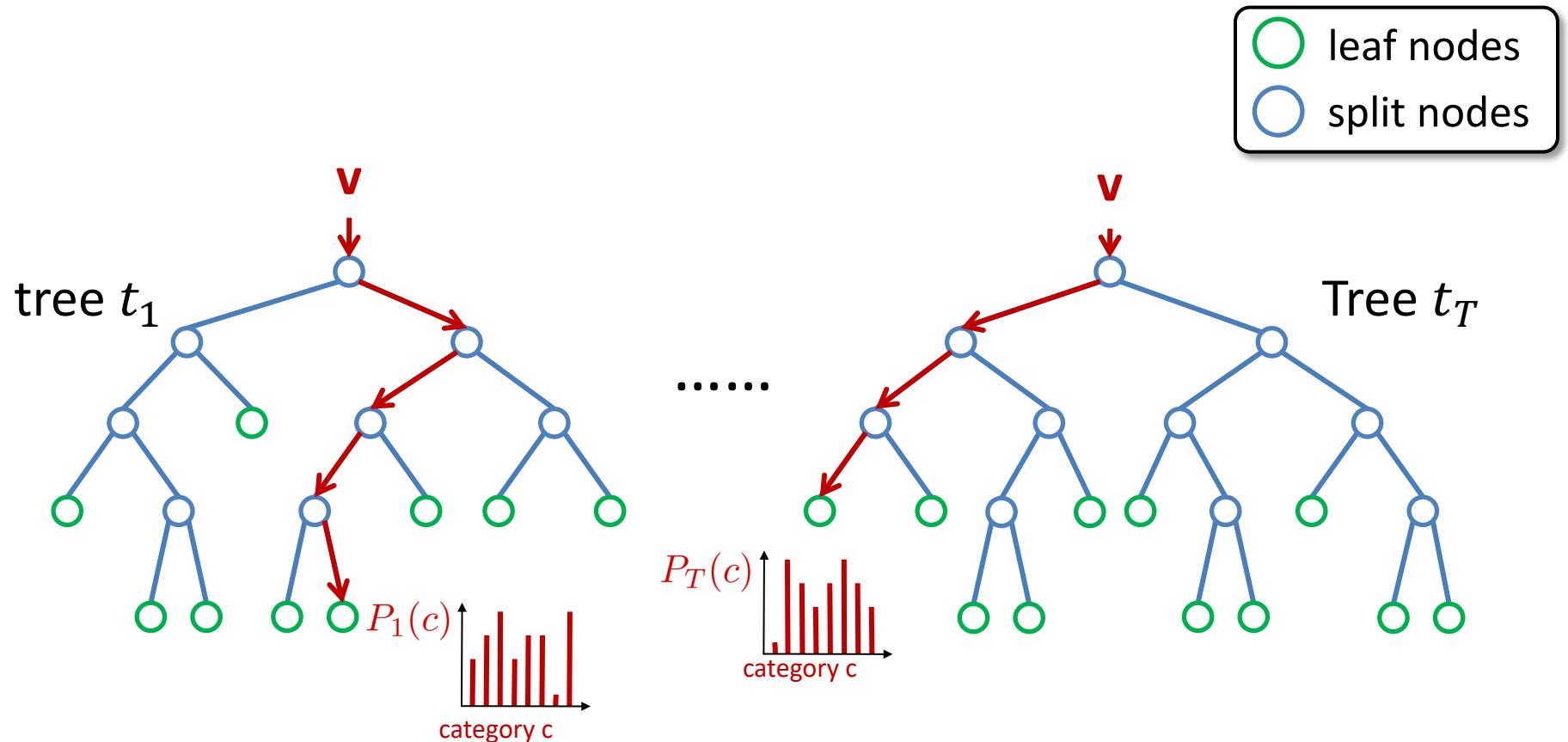
$$\hat{f}(\mathbf{x}) = \frac{1}{T} \sum_{i=1}^T f_i(\mathbf{x})$$

- For classification tree, output is determined by taking the majority vote in the case of decision trees
- Through bootstrapping, model performance increases over single tree (single tree is highly sensitive to noise in its training set)

원본 훈련 데이터에서 복원 추출로 여러 데이터 샘플 생성

Random Forest

- Random forest create T different trees on each bootstrap samples

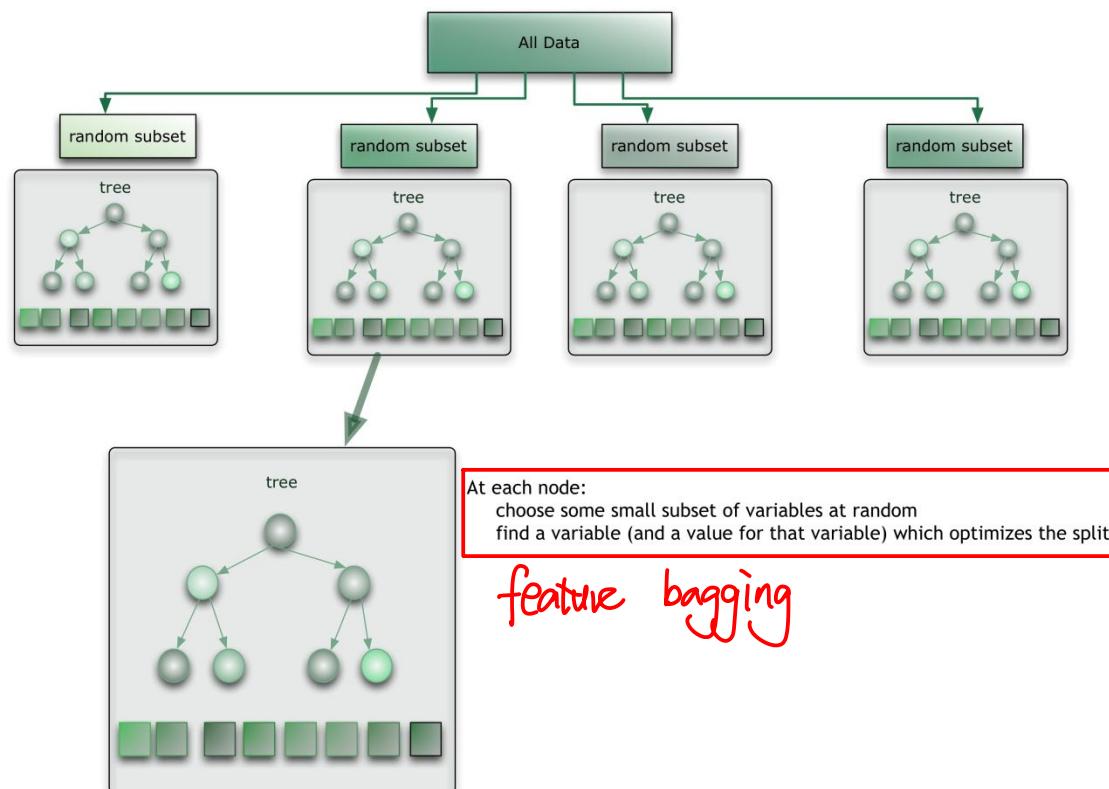


Random Forest

- Random forest differs from bagging in the way that this algorithm **selects a random subset of the features at each candidate split**
 - ▣ It is sometimes called **'feature bagging'**
 - ▣ The reason for doing this is to **reduce the correlation of the trees**
 - If one or a few features are very strong predictors for the response variable, these features will be selected in many trees → Trees become correlated
 - Detailed splitting process for each tree
 - At each node
 - ▣ For some number m , **m predictor variables are selected at random from all the predictor variables**
 - ▣ **The predictor variable that provides the best split**, according to some objective function is used to do a binary split on that node
 - ▣ At the **next node, choose another m variables at random** from all predictor variables and do the same
- 전체 feature에서 랜덤한 feature 추출
subset에서 가장 좋은 split criteria 선정

Random Forest

- Three different types of ensemble methods on tree
 - Random splitter selection: $m = 1$
 - Breiman's bagger: $m = \text{total number of predictor variables}$
 - **Random forest:** $m \ll \text{number of predictor variables}$
 - Literature suggests three possible values for m : $\frac{1}{2}\sqrt{m}$, \sqrt{m} and $2\sqrt{m}$



Random Forest

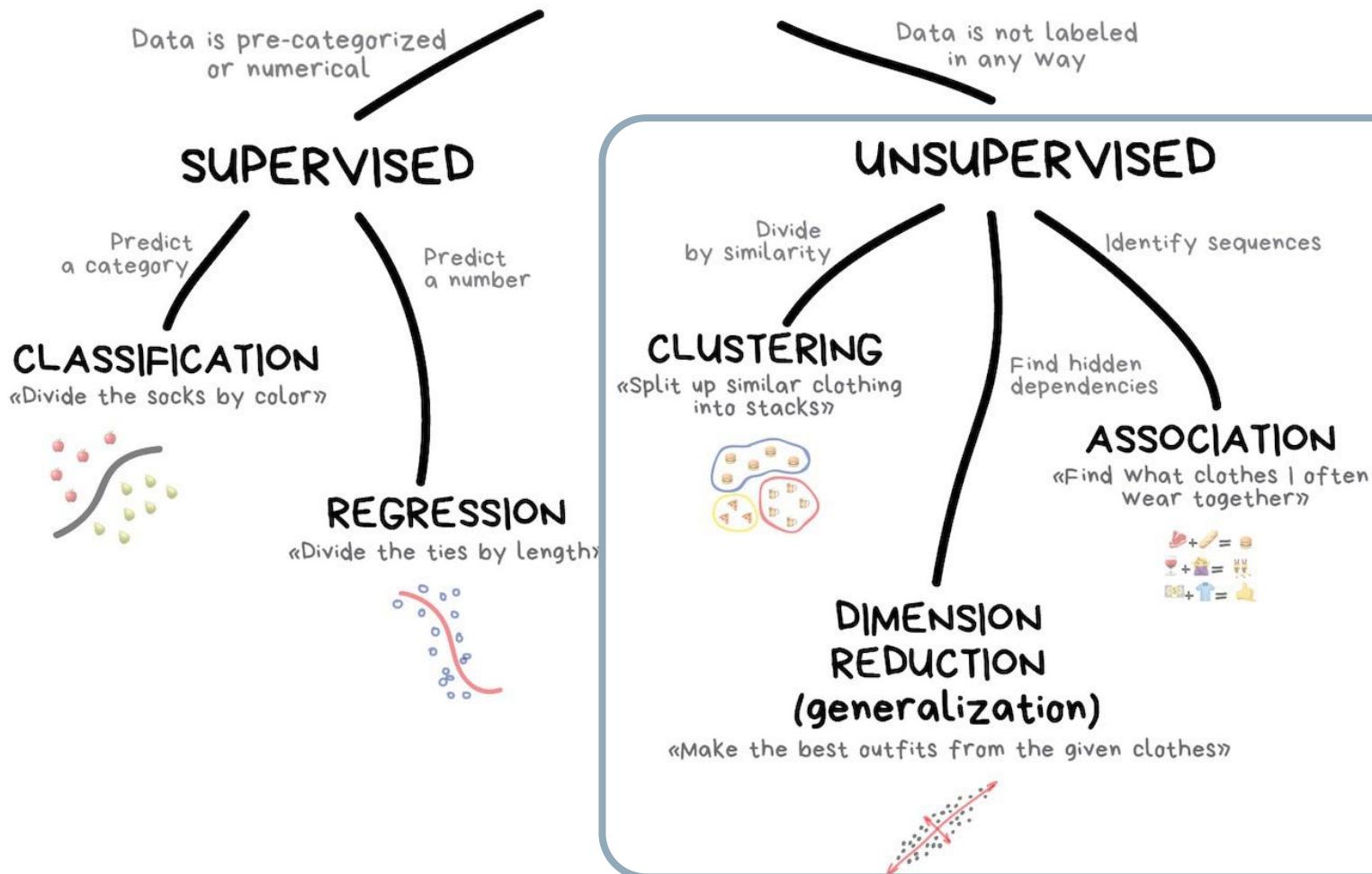
- Pros
 - ▣ **High Accuracy:** Random Forest generally provides high accuracy in both classification and regression tasks
 - ▣ **Robust to Overfitting:** Since it uses multiple trees and random subsets of features and data, it is less likely to overfit than a single decision tree
 - ▣ **Feature Importance:** Random Forest can be used to determine the importance of each feature in making predictions
- Cons
 - ▣ **Computationally Intensive:** Since it builds multiple decision trees, Random Forest can be slow to train and require more memory, especially on large datasets
 - ▣ **Less Interpretable:** While individual decision trees are interpretable, the Random Forest model as a whole is more of a “black box,” making it harder to interpret the relationships between features and predictions
 - ▣ **Predictions Are Slower:** Predicting with a large number of trees can be slower compared to a single decision tree, especially in real-time applications

CLUSTERING

Week 11

Type of Learning

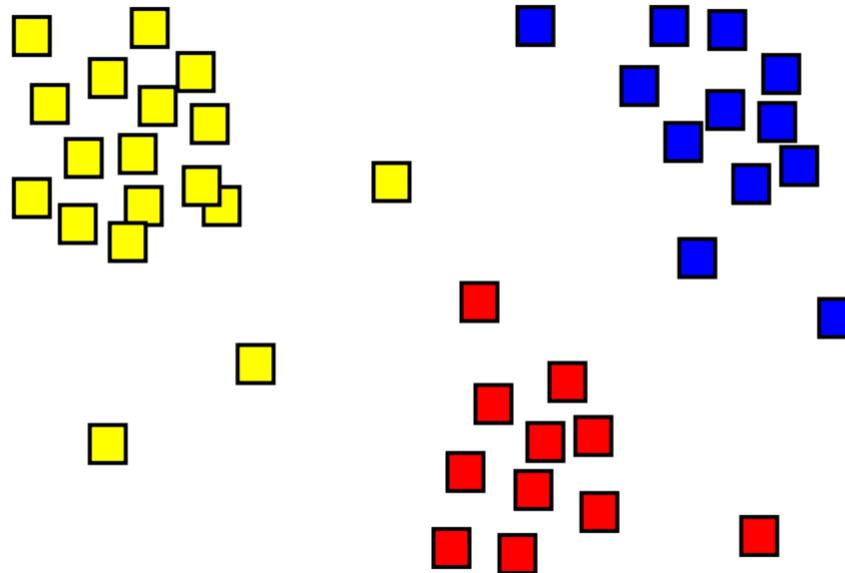
CLASSICAL MACHINE LEARNING



Clustering

Unsupervised Learning: Clustering

- [Remind] Unsupervised learning is learning with unlabeled data
 - No certain output to be estimated
- Clustering is to group a set of data points to satisfy following conditions as much as possible
 - Data points in the same group are more similar to each other than to data points in other groups



Clustering

- Purpose
 - Discover patterns in data
 - Data compression and summarization
 - Anomaly detection
 - Preprocessing for other machine learning tasks

Clustering

- Types of clustering
 - ▣ Hard Clustering: Each data point belongs to only one cluster
 - ▣ Soft Clustering (Fuzzy Clustering): Each data point has a probability of belonging to multiple clusters
- Types of clustering methods
 - ▣ Partition-based Clustering (e.g., K-Means): Partition-based clustering assigns data points into k clusters by minimizing intra-cluster variance.
 - ▣ Hierarchical Clustering: Builds a hierarchy of clusters using a tree-like structure (dendrogram)
 - ▣ Density-based Clustering (e.g., DBSCAN): Groups points that are closely packed together based on density
 - ▣ Model-based Clustering (e.g., Gaussian Mixture Models): Assumes that data is generated from a mixture of probability distributions

Clustering

- Data points in the same group are more similar to each other than to data points in other groups

**1. How to know
some points are more similar than others?**



Using distance measure

2. How to group?



Determine certain rule to group

k -means Clustering

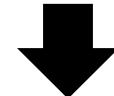
k-means Clustering

- Objective function of clustering

$$\sum_i \min_j \|x_i - \mu_j\|^2$$

X_i가 j번 클러스터에 속하는 경우 μ_j까지
Euclidean distance square
between X_i & μ_j

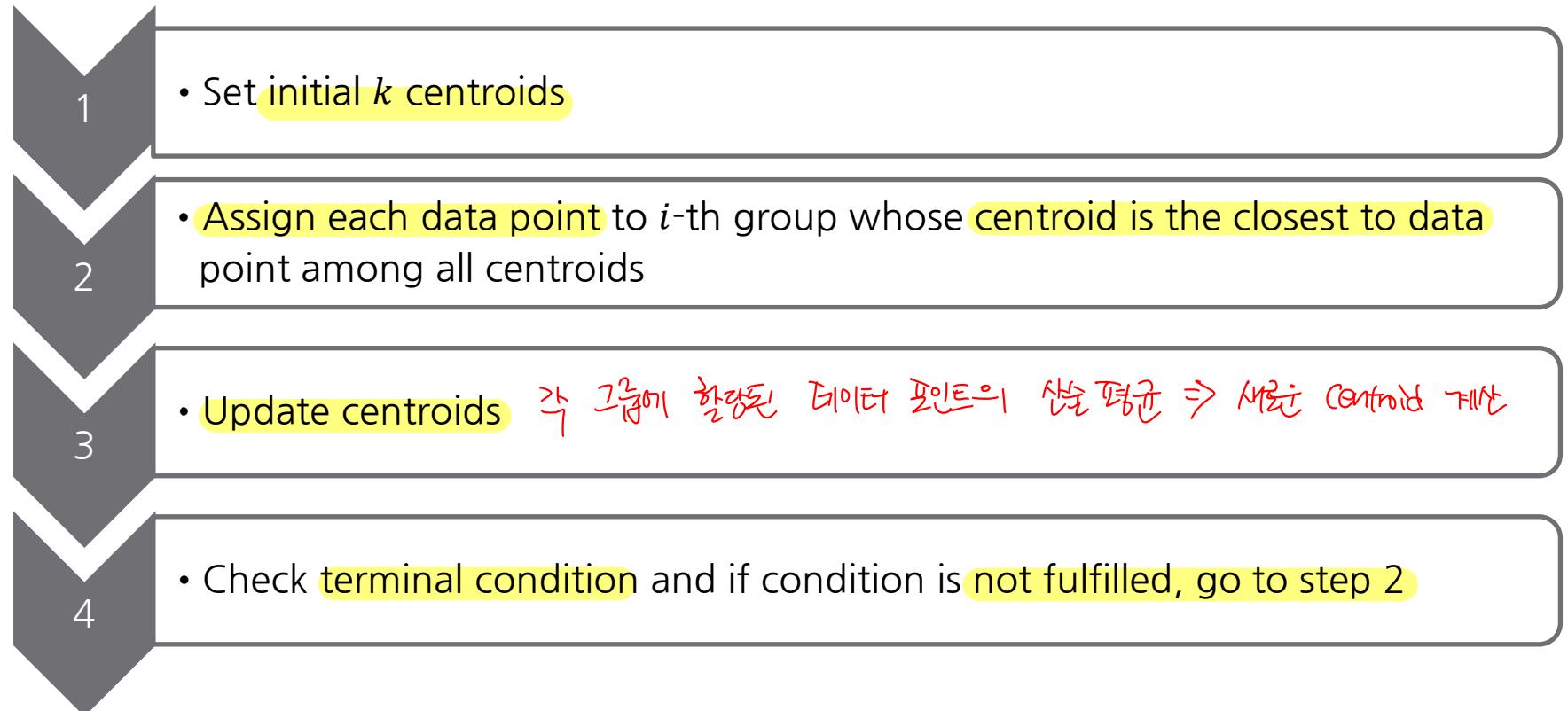
- $j \in [1, 2, \dots, k]$
- μ_j is the centroid of j-th cluster



Combinatorial Optimization Problem

k-means Clustering

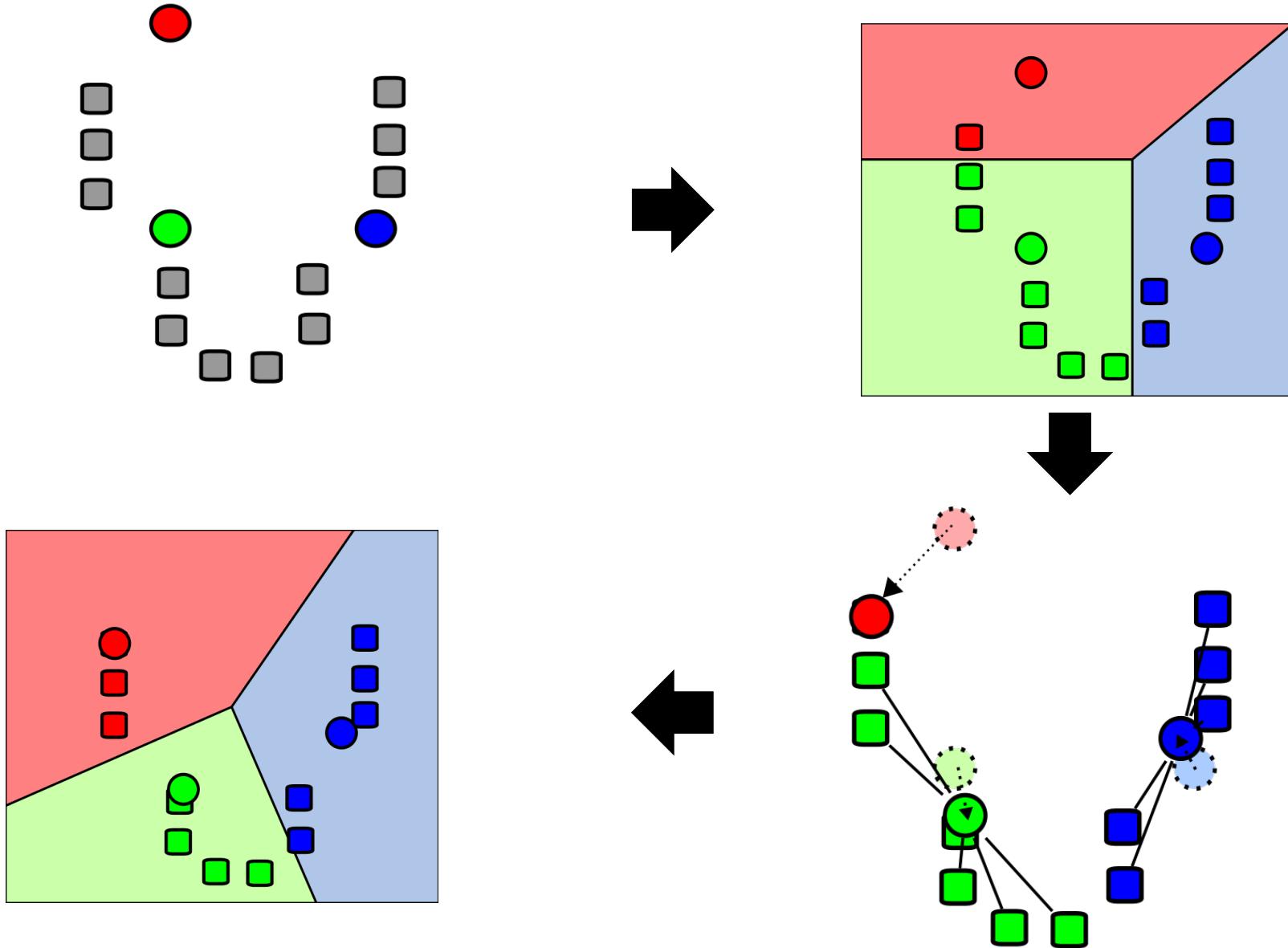
□ Procedure of *k*-means clustering



Terminal conditions

[No change in centroids or the number of iteration is over the pre-specified threshold]

k-means Clustering



How to Update Centroids

- Arithmetic mean

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

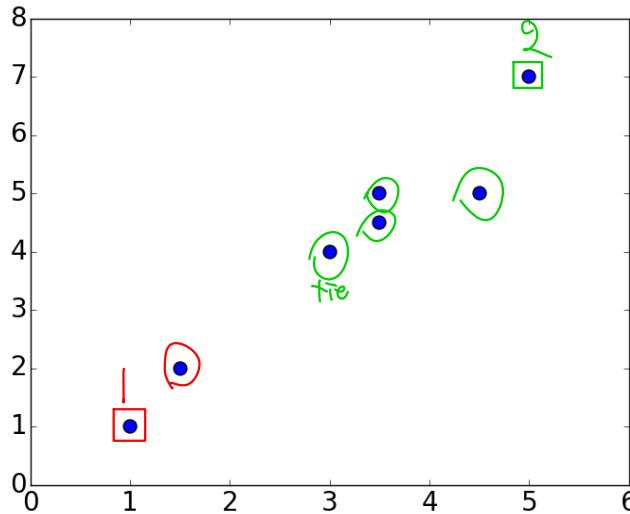
- t is iteration
- m_i is i -th group centroid
- S_i is a set of i -th group and $|S_i|$ is size of S_i

- Example
 - If $(3,1), (2,2), (4,6)$ belong to group, updated centroid is

$$\left(\frac{3+2+4}{3}, \frac{1+2+6}{3} \right) = (3, 2)$$

Question

- Clustering for 2D data set



	1	2	3	4	5	6	7
x	1.0	1.5	3.0	5.0	3.5	4.5	3.5
y	1.0	2.0	4.0	7.0	5.0	5.0	4.5

- 1) When $k = 2$ and initial centroids are $(1.0, 1.0)$ and $(5.0, 7.0)$, determine group of each data point
- 2) What are new centroids of two groups?

$$\left(\frac{1+1.5}{2}, \frac{1+2}{2} \right) = \left(\frac{5}{4}, \frac{3}{2} \right) \quad \left(\frac{3+5+3.5+4.5+3.5}{5}, \frac{9+7+5+5+4.5}{5} \right) = \left(3.9, 5.1 \right)$$

k-means Clustering: Pros and Cons

- Pros
 - Simple and efficient
 - Works well with large datasets
- Cons
 - Requires specifying k
 - Sensitive to initial centroid selection

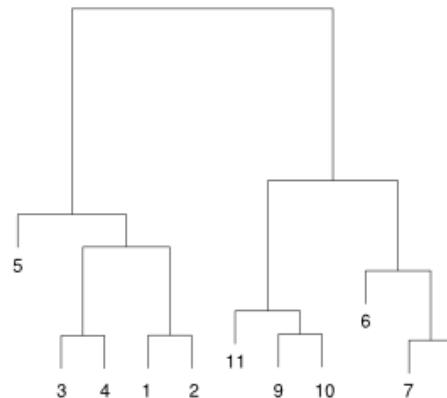
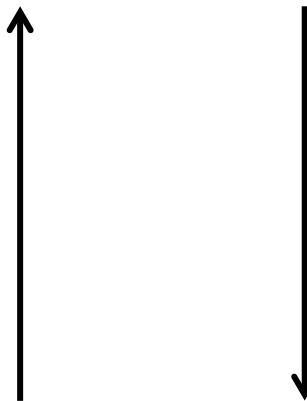
Hierarchical Clustering

Hierarchical Clustering

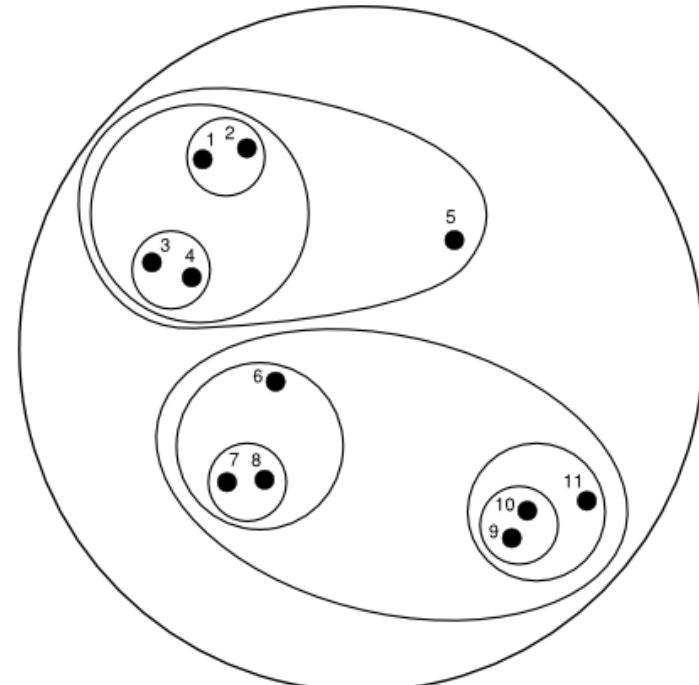
- Hierarchical clustering builds a hierarchy of clusters
 - ▣ Agglomerative: Bottom up approach, each data point starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy
 - ▣ Divisive: Top down approach, all data points start in one cluster and splits are performed recursively as one moves down the hierarchy

top down -

Divisive



Agglomerative
bottom up



Linkage Criteria for Agglomerative Clustering

bottom up

- Way to calculate similarity between two clusters A, B 클러스터 간 거리 정의하는 방법

Type	Formula
Complete-linkage	$\max\{d(a, b) : a \in A, b \in B\}$ 두 클러스터 중 가장 먼 데이터 포인트 사이의 거리
Single-linkage	$\min\{d(a, b) : a \in A, b \in B\}$ 가장 가까운 데이터 포인트 사이의 거리
Mean linkage	$\frac{1}{ A B } \sum_{a \in A} \sum_{b \in B} d(a, b)$ 두 클러스터 간 모든 데이터 쌍의 평균 거리
Centroid linkage	$d(c_A, c_B)$ centroid 간 거리
Ward linkage	$\text{Var}(A \cup B) - \text{Var}(A) - \text{Var}(B)$ 두 클러스터 병합시 발생하는 총간의 내분산의 증가량.

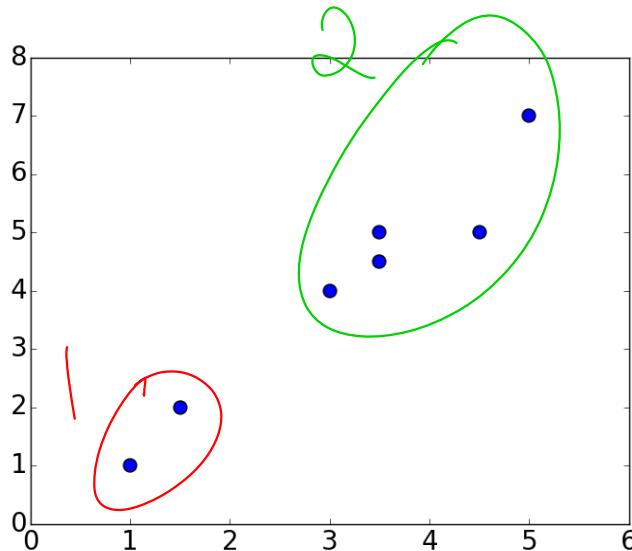
- a belongs to A , b belongs to B
- $\text{Var}(X)$ is within-cluster variance (variance of cluster X)

$$\text{Var}(X) = \frac{1}{n_A} \sum_{i \in A} \|\mathbf{x}_i - \mu_A\|^2$$

- $d(a, b)$ is distance between two data points a and b

Question

- Clustering for 2D data set



	1	2	3	4	5	6	7
x	1.0	1.5	3.0	5.0	3.5	4.5	3.5
y	1.0	2.0	4.0	7.0	5.0	5.0	4.5
c	1	1	2	2	2	2	2

- 1) Using complete-linkage, calculate linkage criterion of cluster 1 and 2

$$(1,1) \quad (5,7) \Rightarrow \sqrt{f^2 + f^2} = \sqrt{2} \approx 1.41$$

- 2) Using centroid-linkage, calculate linkage criterion of cluster 1 and 2

$$\text{Centroid of 1 : } (1.25, 1.5) \Rightarrow \sqrt{2b_1^2 + 3b_2^2} \approx 1.41$$

$$\text{Centroid of 2 : } (3.9, 5.1)$$

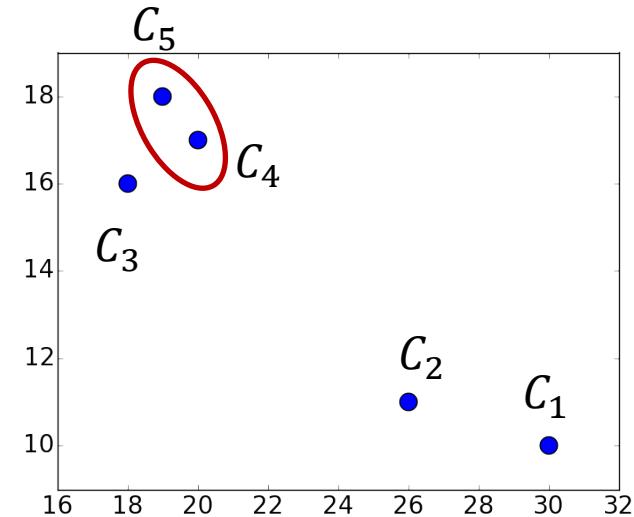
Example: Single Linkage Clustering

- Find clusters through single linkage hierarchy clustering
 - ▣ Start each data as own cluster
 - ▣ Distance measure between two points:
Euclidean distance

Distance $d(A, B) = \min\{d(a, b) : a \in A, b \in B\}$

	1	2	3	4	5
1	0				
2	4.12	0			
3	15.23	11.18	0		
4	12.21	8.48	4.12	0	
5	13.60	9.90	3.61	1.41	0

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18



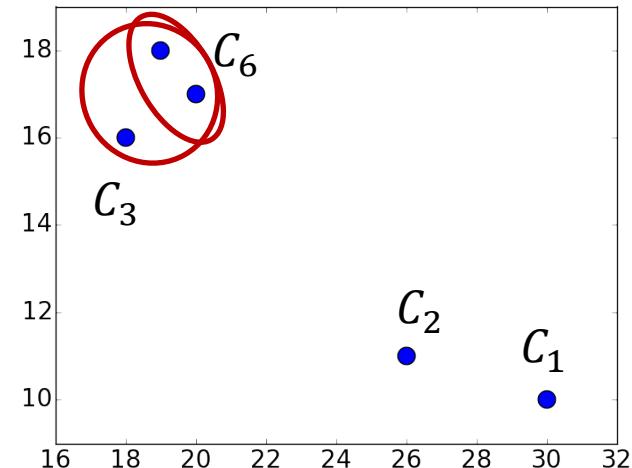
Example: Single Linkage Clustering

- Find clusters through single linkage hierarchy clustering
 - ▣ Merge cluster 4 and 5 to create new cluster
 $\Rightarrow \text{cluster } 6$

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18

Distance $d(A, B) = \min\{d(a, b) : a \in A, b \in B\}$

	1	2	3	6
1	0			
2	4.12	0		
3	15.23	11.18	0	
6	12.21	8.48	3.61	0



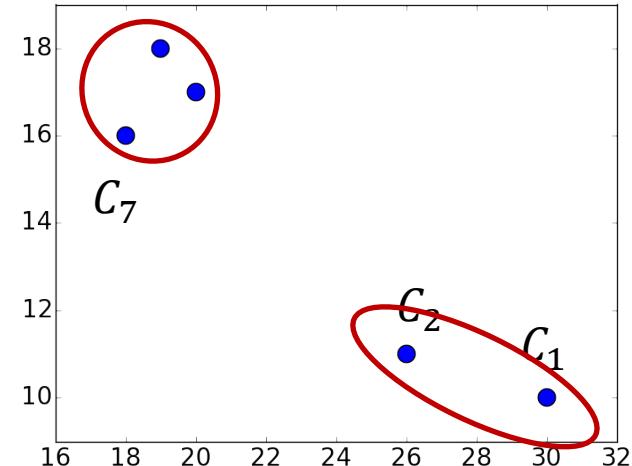
Example: Single Linkage Clustering

- Find clusters through single linkage hierarchy clustering
 - ▣ Merge cluster 3 and 6 to create new cluster
 $\Rightarrow \text{cluster } 7$

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18

Distance $d(A, B) = \min\{d(a, b) : a \in A, b \in B\}$

	1	2	7
1	0		
2	4.12	0	
7	12.21	8.48	0



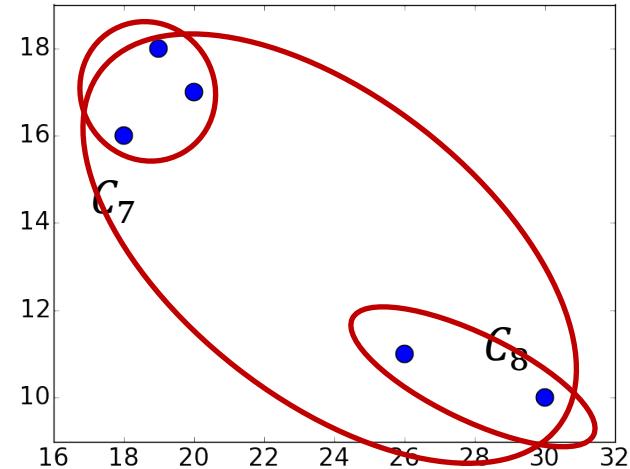
Example: Single Linkage Clustering

- Find clusters through single linkage hierarchy clustering
 - ▣ Merge cluster 1 and 2 to create new cluster
⇒ *cluster 8*

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18

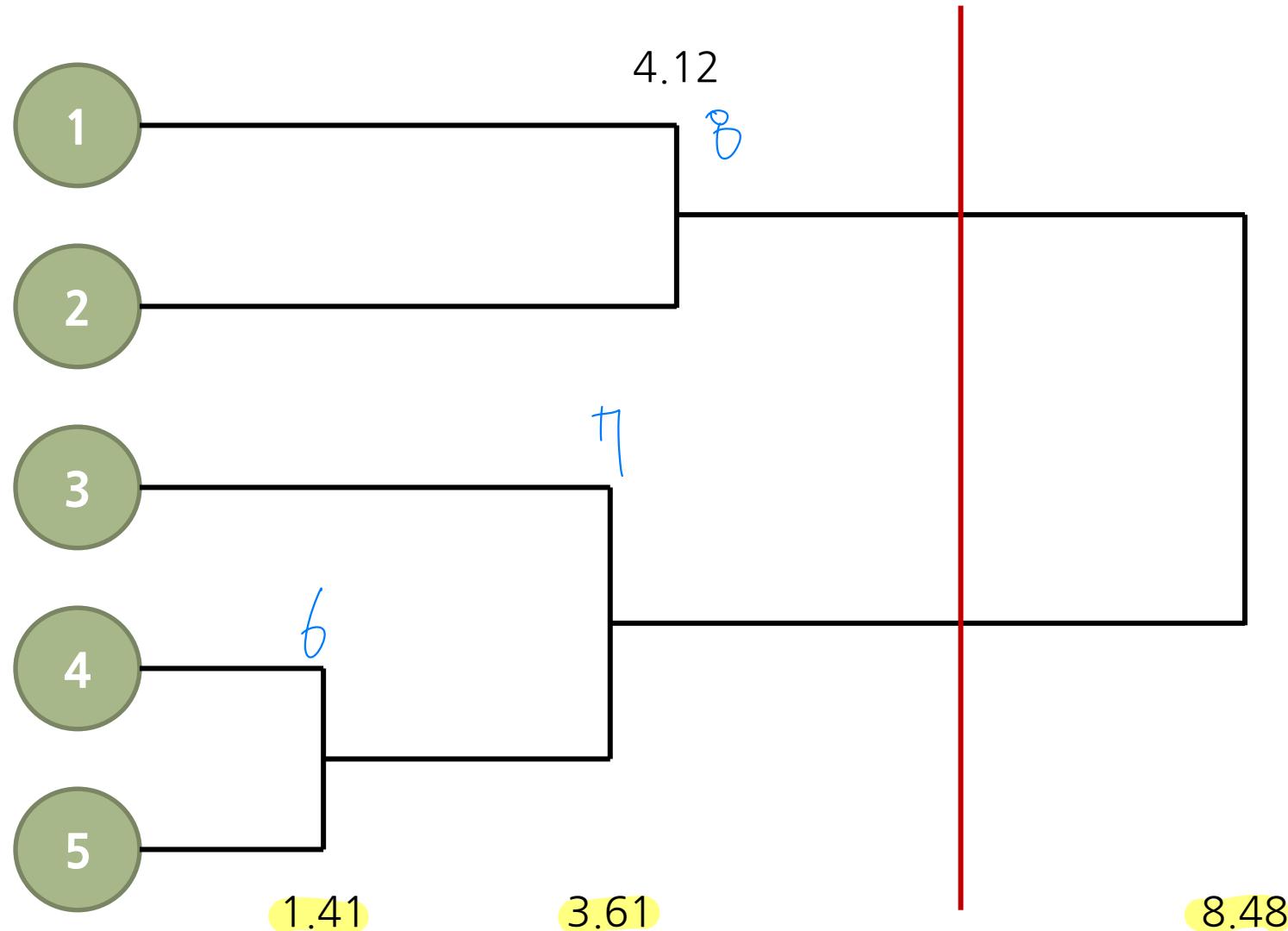
Distance $d(A, B) = \min\{d(a, b) : a \in A, b \in B\}$

	7	8
7	0	
8	8.48	0



Example: Single Linkage Clustering

- Dendrogram



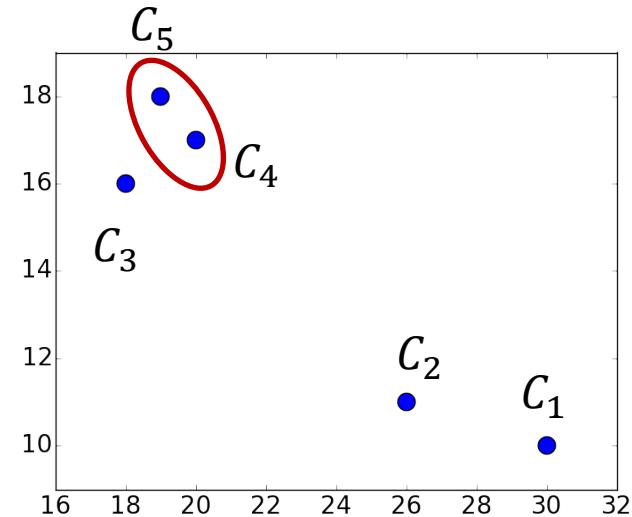
Example: Complete Linkage Clustering

- Find clusters through complete linkage hierarchy clustering
 - ▣ Start each data as own cluster

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18

Distance $d(A, B) = \max\{d(a, b) : a \in A, b \in B\}$

	1	2	3	4	5
1	0				
2	4.12	0			
3	15.23	11.18	0		
4	12.21	8.48	4.12	0	
5	13.60	9.90	3.61	1.41	0



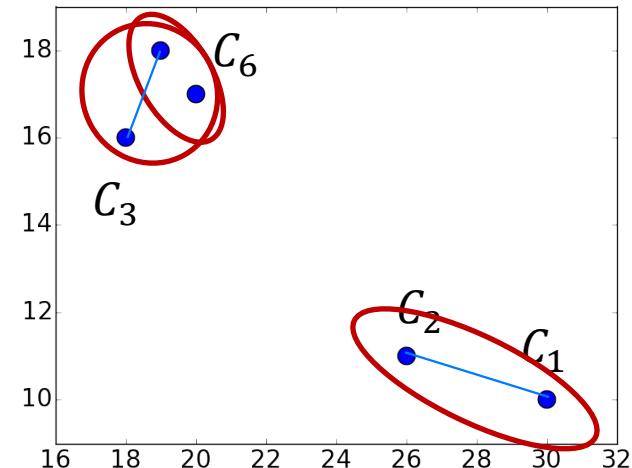
Example: Complete Linkage Clustering

- Find clusters through complete linkage hierarchy clustering
 - ▣ Merge cluster 4 and 5 to create new cluster
 $\Rightarrow \text{cluster 6}$

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18

Distance $d(A, B) = \max\{d(a, b) : a \in A, b \in B\}$

	1	2	3	6
1	0			
2	4.12	0		
3	15.23	11.18	0	
6	13.60	9.90	4.12	0



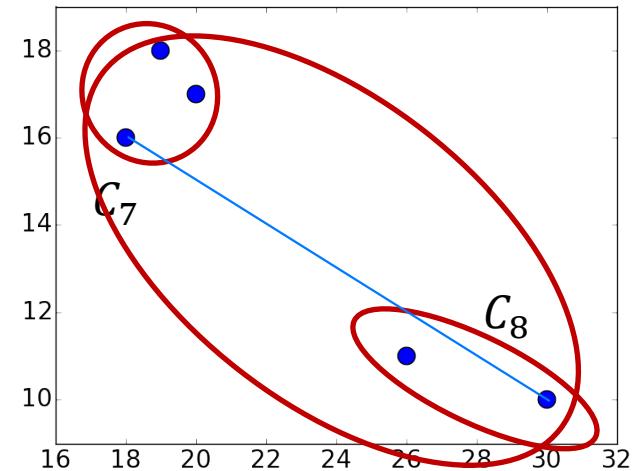
Example: Complete Linkage Clustering

- Find clusters through complete linkage hierarchy clustering
 - ▣ Merge cluster 1 and 2 to create new cluster 8
 - ▣ Merge cluster 3 and 6 to create new cluster 7

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18

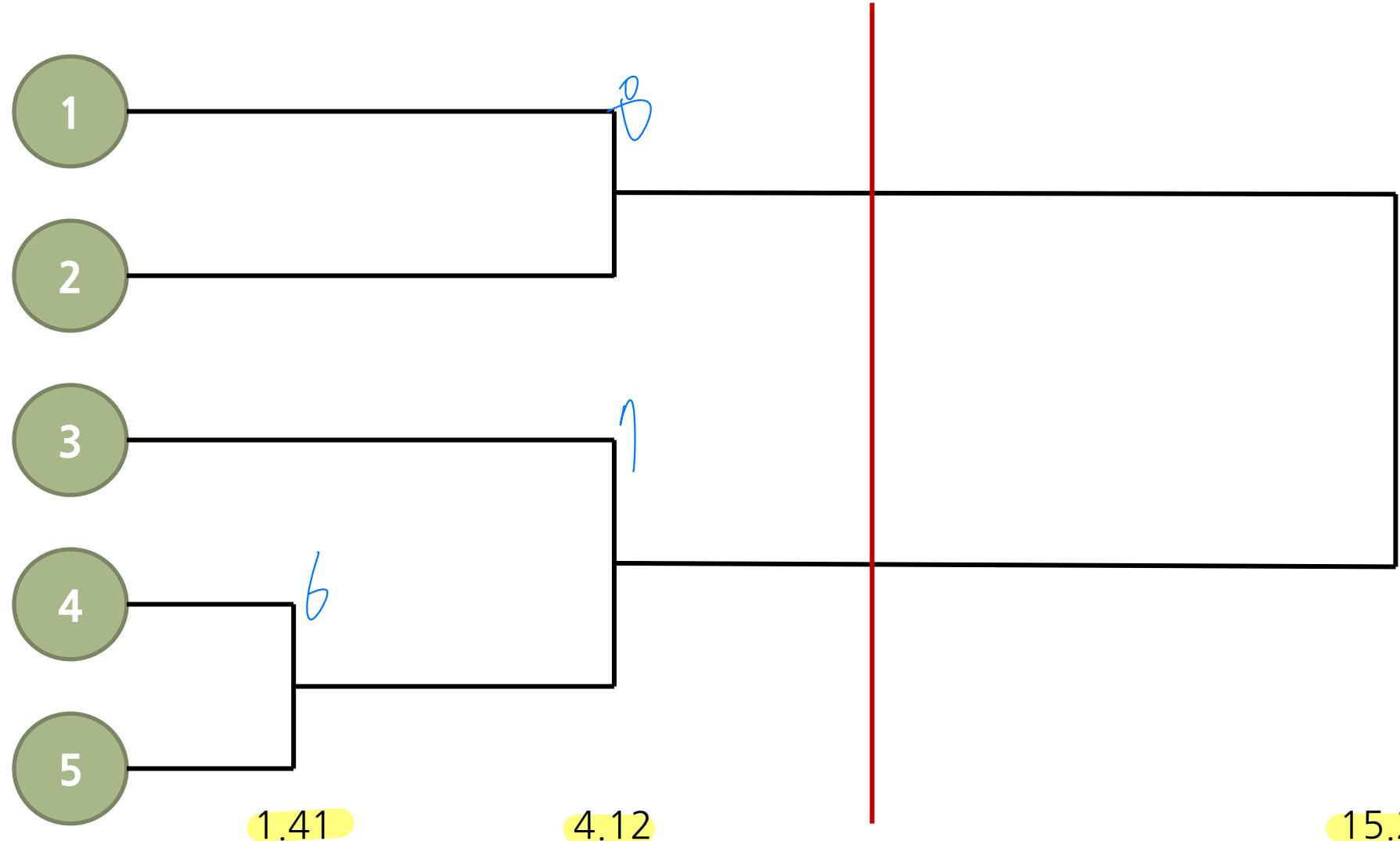
Distance $d(A, B) = \max\{d(a, b) : a \in A, b \in B\}$

	7	8
7	0	
8	15.23	0



Example: Complete Linkage Clustering

- Dendrogram



Divisive Clustering - DIANA

top down

- Divisive method starts with one cluster including all samples
 - At each step, divide cluster into two sub clusters until every cluster consists of one data point
 - This algorithm is based on the average distance between one object and the others

점과 클러스터 C 사이의 평균 거리.

$$\bar{d}(i, C) = \begin{cases} \frac{1}{|C|-1} \sum_{j \in C, j \neq i} d(i, j), & \text{if } i \in C \\ \frac{1}{|C|} \sum_{j \in C} d(i, j), & \text{if } i \notin C \end{cases}$$

자기 자신 제외한 다른 점들과의 거리 평균

i represent i -th object

DIANA Algorithm

$$e(i) = \bar{d}(i, C) - \bar{d}(i, C') \Rightarrow \text{값이 커록 } C' \text{에 속하는게 나은}$$

- 1 • Consider all samples as one cluster
- 2 • Select the cluster C containing two objects with the longest distance
- 3 • Divide cluster C into two as follows (At first, C' is empty set(\emptyset))
 - Find object i with maximum $\bar{d}(i, C)$ \Rightarrow 다른 점들와의 평균 거리가 가장 먼 i는 선택
 - $C \leftarrow C - \{i\}, C' \leftarrow C' \cup \{i\}$ \Rightarrow i는 C에서 제외 C'에 추가
 - If there exist the objects j in C whose $e(j) = \bar{d}(j, C) - \bar{d}(j, C') > 0$, select one of them with maximum $e(j)$, remove j from C and add j into C'
 - If $e(j) \leq 0$ for all objects in C , finish this step
- 4 • Repeat step 2 and 3 until the number of clusters is the same as the number of samples

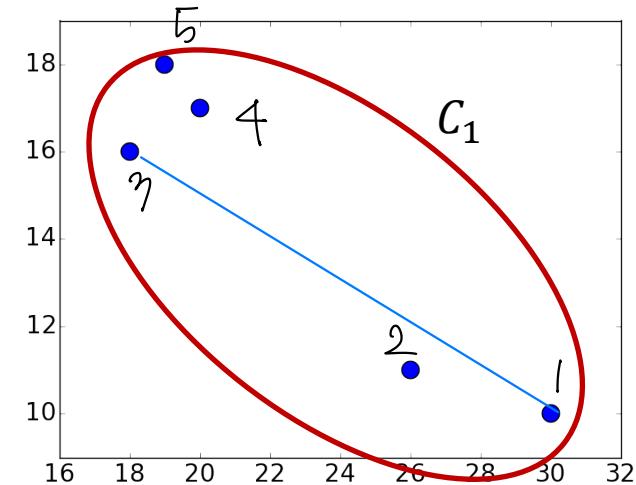
Example: DIANA

- Find clusters through DIANA
 - Start with a cluster consisting of all objects
 - $C_1 = \{1,2,3,4,5\}$
 - $C_2 = \{\}$

Step 2: Find pair of objects wit the longest distance

$d(i,j)$	1	2	3	4	5
1	0				
2	4.12	0			
3	15.23	11.18	0		
4	12.21	8.48	4.12	0	
5	13.60	9.90	3.61	1.41	0

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18



Example: DIANA

- Find clusters through DIANA
 - ▣ C_1 is the selected cluster

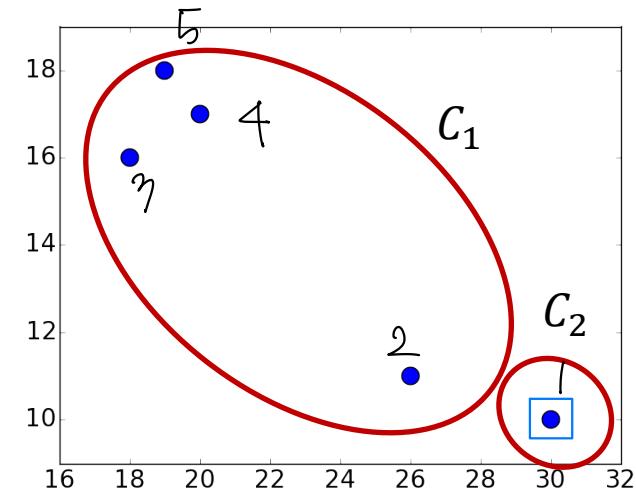
Step 3

$d(i, j)$	1	2	3	4	5
1	0	4.12	15.23	12.21	13.60
2	4.12	0	11.18	8.48	9.90
3	15.23	11.18	0	4.12	3.61
4	12.21	8.48	4.12	0	1.41
5	13.60	9.90	3.61	1.41	0

↓ Average except 0

	1	2	3	4	5
$\bar{d}(i, C_1)$	11.29	8.42	8.54	6.56	7.13

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18



Example: DIANA

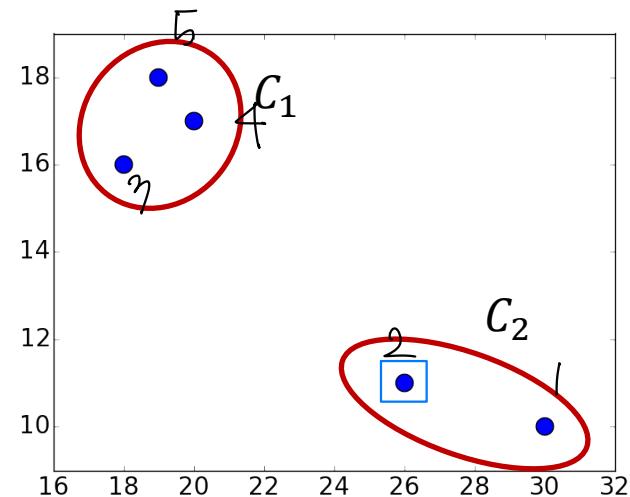
- Find clusters through DIANA
 - ▣ $C_1 = \{2,3,4,5\}, C_2 = \{1\}$

$$e(i) = \bar{d}(i, c) - \bar{d}(i, c')$$

Step 3

	2	3	4	5
$\bar{d}(i, C_1)$	9.85	6.30	4.67	4.97
$\bar{d}(i, C_2)$	4.12	15.2	12.2	13.6
$e(i)$	5.73	-8.9	-7.53	-8.63

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18



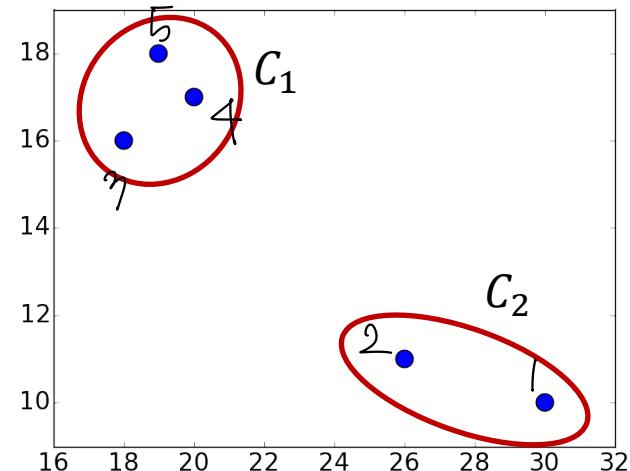
Example: DIANA

- Find clusters through DIANA
 - ▣ $C_1 = \{3, 4, 5\}, C_2 = \{1, 2\}$

Step 3

	3	4	5
$\bar{d}(i, C_1)$	3.87	2.77	2.51
$\bar{d}(i, C_2)$	13.21	10.35	11.75
$e(i)$	-9.34	-7.58	-9.24

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18



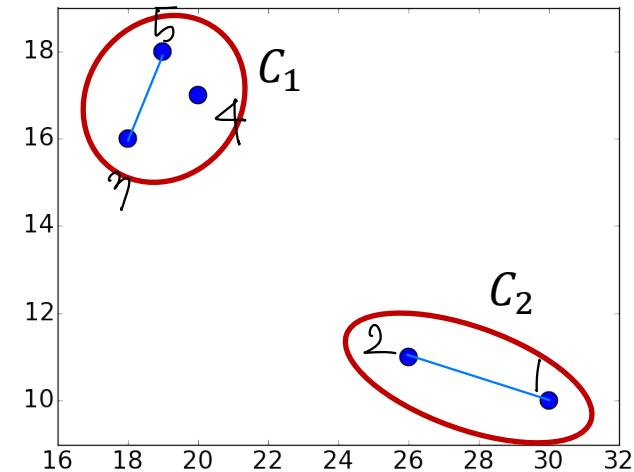
Example: DIANA

- Find clusters through DIANA
 - ▣ $C_1 = \{3,4,5\}, C_2 = \{1,2\}$
 - ▣ Find pair of objects wit the longest distance

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18

Step 2: Find pair of objects wit the longest distance

$d(i,j)$	1	2	3	4	5
1	0				
2	4.12	0			
3			0		
4			4.12	0	
5			3.61	1.41	0



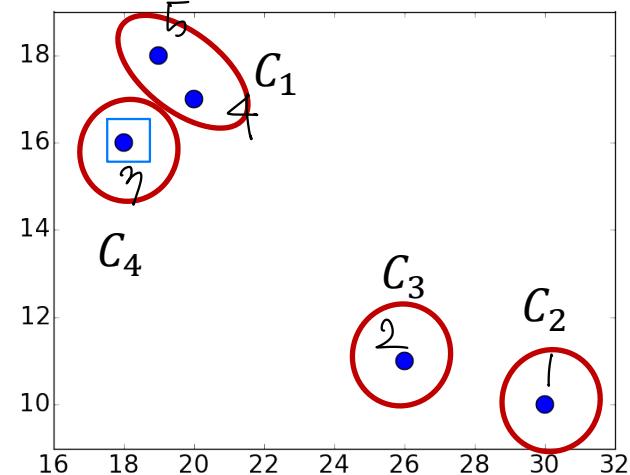
Example: DIANA

- Find clusters through DIANA
 - ▣ $C_1 = \{4, 5\}, C_4 = \{3\}$

Step 3

	4	5
$\bar{d}(i, C_1)$	1.41	1.41
$\bar{d}(i, C_4)$	4.12	3.61
$e(i)$	-2.71	-2.20

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18



Example: DIANA

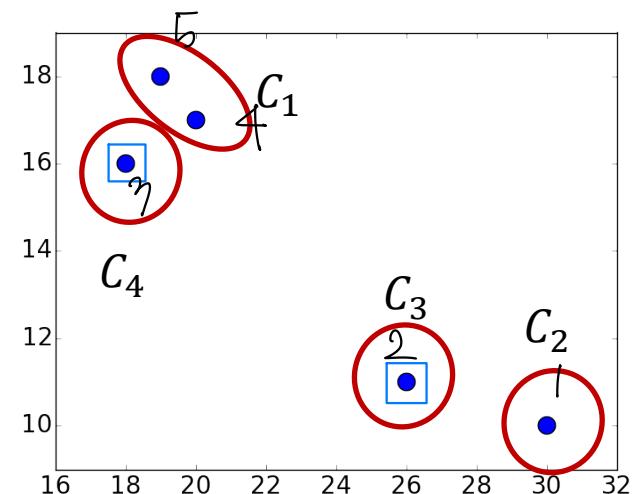
- Find clusters through DIANA
 - ▣ Select C_2
 - ▣ $C_1 = \{1,2\}, C_3 = \{\}$
 - ▣ C_2 contains only two object, so divide C_2 into two clusters directly: $C_2 = \{1\}, C_3 = \{2\}$

- ▣ Select C_1
- ▣ $C_1 = \{3,4,5\}, C_4 = \{\}$

Step 3

	3	4	5
$\bar{d}(i, C_1)$	3.87	2.77	2.51

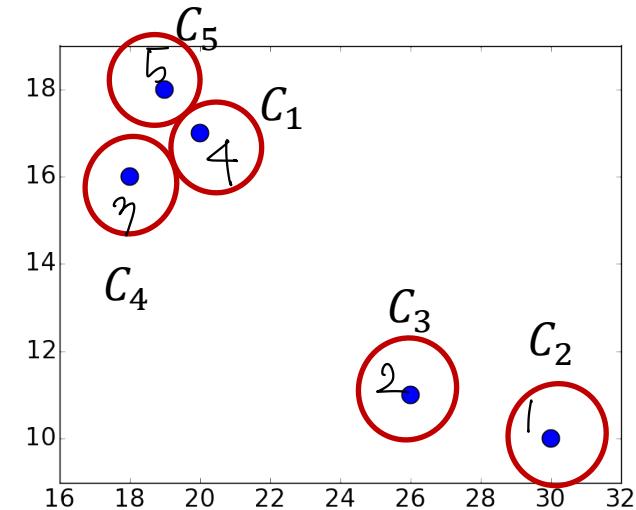
	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18



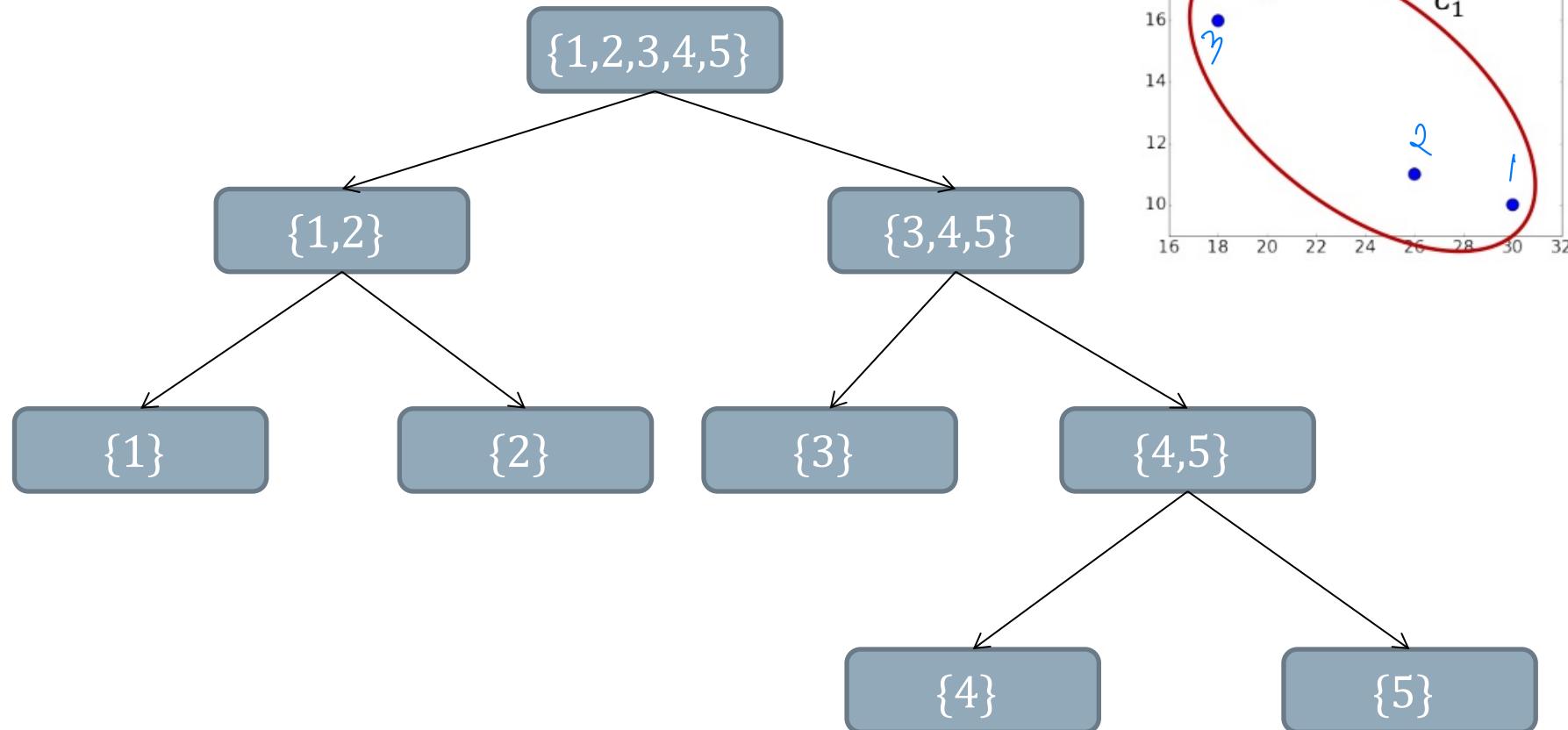
Example: DIANA

- Find clusters through DIANA
 - ▣ Select C_1
 - ▣ $C_1 = \{4,5\}, C_5 = \{\}$
 - ▣ C_1 contains only two object, so divide C_1 into two clusters directly: $C_1 = \{4\}, C_5 = \{5\}$

	1	2	3	4	5
x	30	26	16	20	19
y	10	11	16	17	18

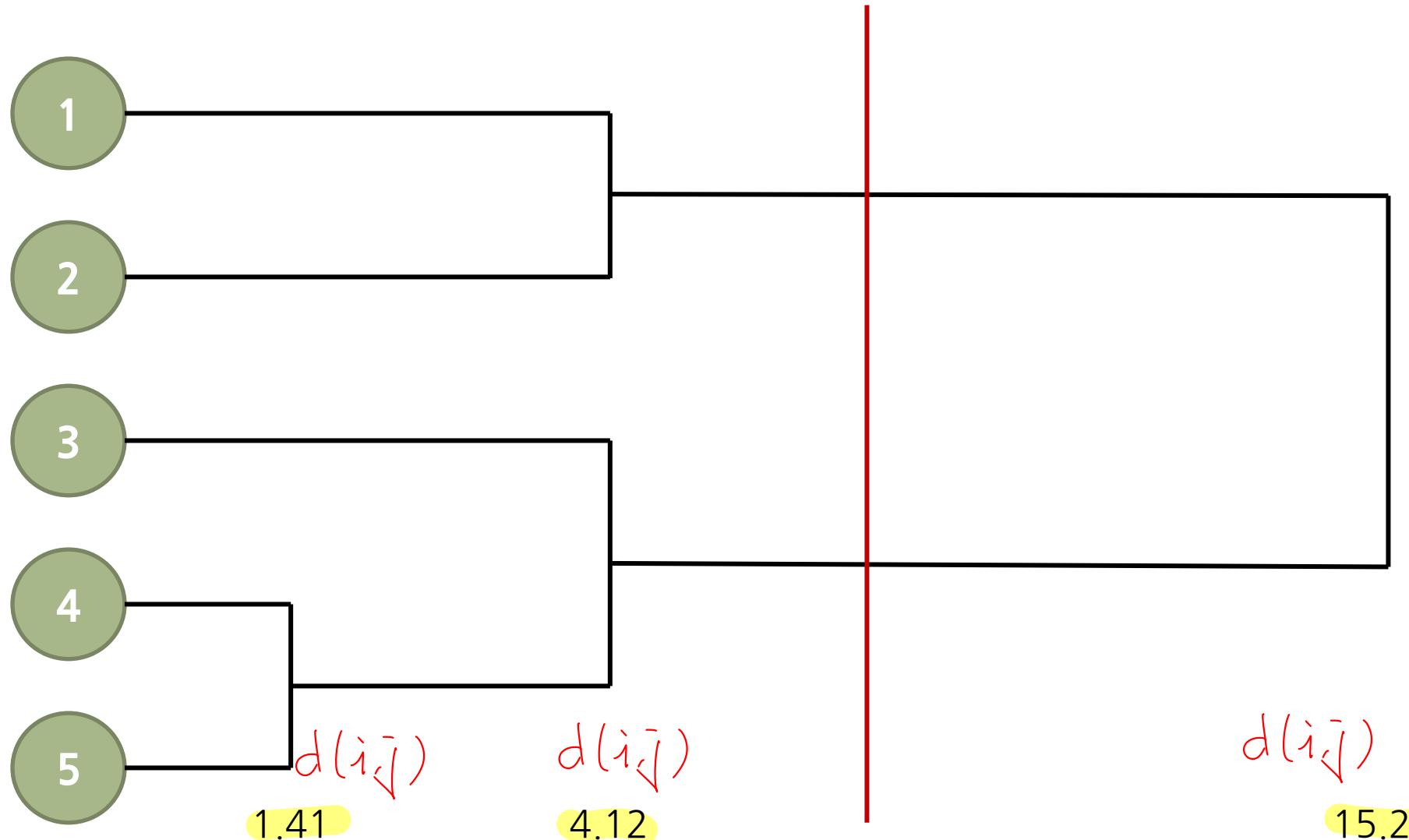


Example: DIANA



Example: DIANA

- Dendrogram



Hierarchical Clustering: Pros and Cons

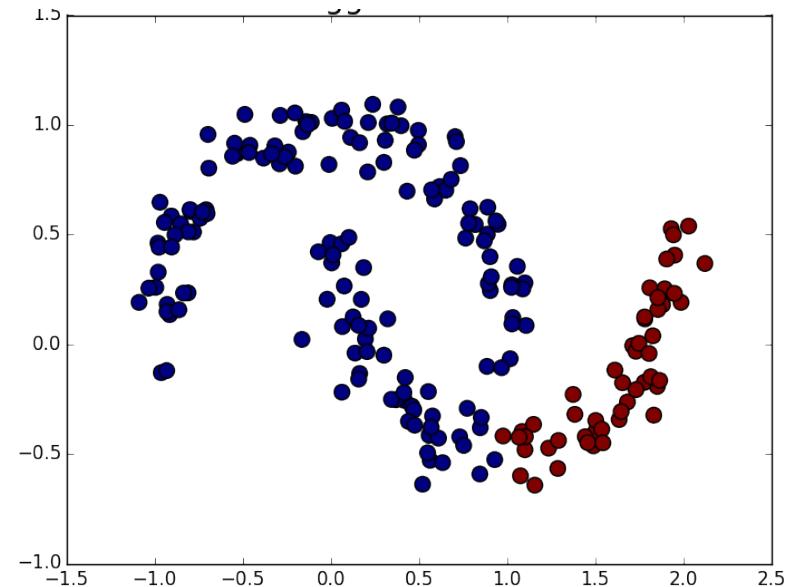
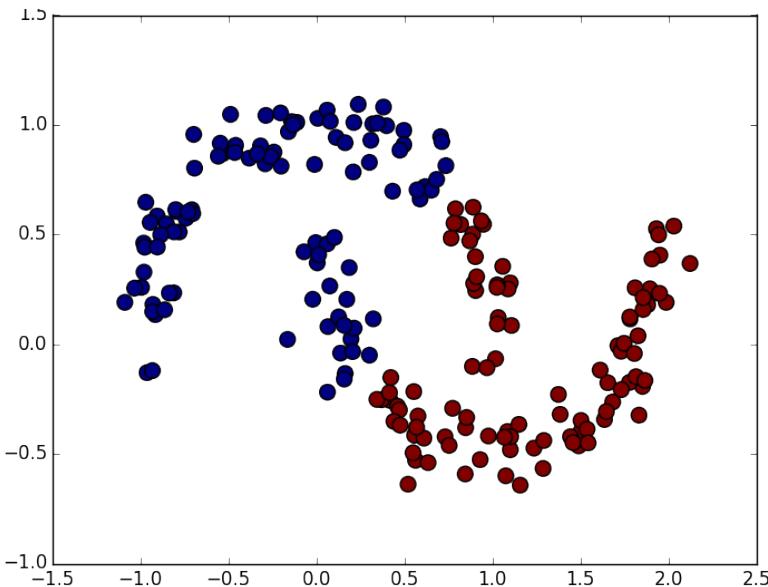
- Pros
 - No need to specify k in advance
 - Produces a hierarchical structure
- Cons
 - Computationally expensive for large datasets
 - Sensitive to noise

Evaluation Measures

How to Measure Clustering Quality

- Clustering problem is unsupervised problem
 - No explicit answer for learning
 - We need to define a method to measure quality of clustering

Which one is better?



How to Measure Clustering Quality

- Measures that do not require ground truth labels

- Inertia

- Within-cluster sum-of-squares

$$\sum_{i=0}^n \min_{\mu_j \in C} \|x_j - \mu_i\|^2$$

k-mean clustering 목적 함수 \Rightarrow WCSS 최소화

각 점과 클러스터 중심 간 거리 제곱합,
작음수록 밀도 ↑

- Silhouette Coefficient

- $s(i)$: Silhouette coefficient of i -th sample
 - $a(i)$: The mean distance between a sample and all other points in the same class
 - $b(i)$: The mean distance between a sample and all other points in the next nearest cluster

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$
$$-1 \leq s(i) \leq 1$$

- Overall clustering quality can be obtained by averaging $s(i)$ for all samples

$s(i) \approx 1$ | 잘 분류됨

$s(i) \approx 0$ | 약간 균형

$s(i) \approx -1$ | 잘못된 클러스터에 속함

How to Measure Clustering Quality

□ Clustering performance **evaluation measure** External Evaluation → 외부 평가지표

■ **Homogeneity**: each cluster contains only members of a single class

하나의 클러스터에 하나의 클래스만
존재하는 정도

$$h = 1 - \frac{H(C|K)}{H(C)}$$

External Evaluation → 외부 평가지표

$K \Rightarrow$ cluster

■ $H(C)$ is the entropy of the classes

$$H(C) = - \sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log\left(\frac{n_c}{n}\right)$$

$C \Rightarrow$ class

■ $H(C|K)$ is the conditional entropy of the classes given the cluster assignments

$$H(C|K) = - \sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{n_{c,k}}{n} \log\left(\frac{n_{c,k}}{n_k}\right)$$

- n is the total number of samples, n_c and n_k are the number of samples respectively belonging to class c and cluster k

- $n_{c,k}$ is the number of samples from class c assigned to cluster k

■ **Completeness**: all members of a given class are assigned to the same cluster

같은 클래스 속한 샘플들이
하나의 클러스터에 모인 정도

$$c = 1 - \frac{H(K|C)}{H(K)}$$

How to Measure Clustering Quality

- Clustering performance evaluation measure 우연의 일치 보정
 - ▣ Adjusted Rand Index(ARI) 실제 클래스 할당과 클러스터링 알고리즘 할당 간 유사도 측정
 - Given the knowledge of the ground truth class assignments and our clustering algorithm assignments of the same samples, the adjusted Rand index is a function that measures the similarity of the two assignments

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]}$$

- C is a ground truth class assignment, K is the clustering
- a is the number of pairs of elements that are in the same set in C and in the same set in K C 에서도, K 에서도 동일 그룹에 속하는 경우
- b is the number of pairs of elements that are in different sets in C and in different sets in K C 에서도, K 에서도 다른 그룹에 속하는 경우.
- Raw Rand index $\text{RI} = \frac{a+b}{C_2^n}$ (C_2^n is the total number of possible pairs in the dataset)

$$C_2^n = \frac{n!}{2!(n-2)!} = n C_2$$

$\text{ARI} = 1$	클러스터링 결과와 실제 클래스 할당 완벽 일치
$\text{ARI} = 0$	클러스터링이 우연 무작위로 할당된 것과 다르지 않음.
$\text{ARI} < 0$	우연에 의한 결과보다 못한 클러스터링 결과.

How to Measure Clustering Quality

□ Contingency table

	K_1	K_2	...	K_s	$sums$
C_1	n_{11}	n_{12}	...	n_{1s}	a_1
C_2	n_{21}	n_{22}	...	n_{2s}	a_2
:	:	:	..	:	:
C_r	n_{r1}	n_{r2}	...	n_{rs}	a_r
$sums$	b_1	b_2	...	b_s	n

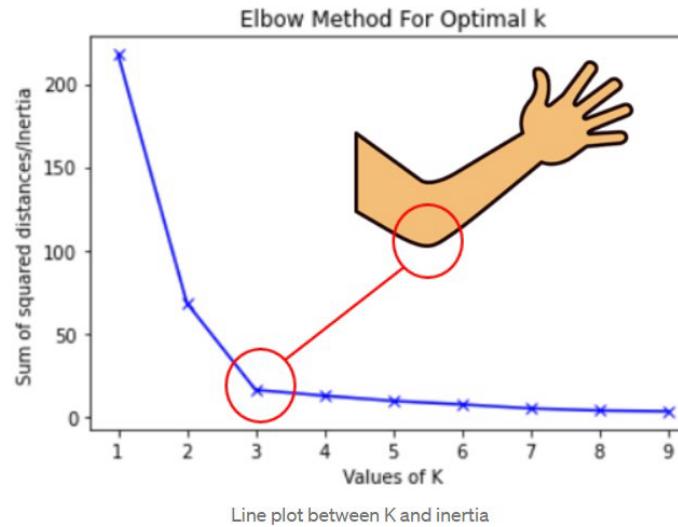
$$ARI = \frac{RI - \mathbb{E}[RI]}{\max(RI) - \mathbb{E}[RI]} = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}$$

n_{ij} = 실제 클래스 i 속한 데이터 중
클래스 j에 속한 데이터 수 $\binom{n}{k} = nC_k$

a_i = C_i , K_i 에서 모두 같은 그룹에 속한 데이터 수 b_i = C_i , K_i 에서 모두 다른 그룹에 속한 데이터 수

Determining the Optimal Number of Clusters

- Elbow method *WCSS*
 - ▣ Plot the total within-cluster sum of squares against the number of clusters
 - ▣ Look for an “elbow” point where the rate of decrease sharply slows



- Silhouette Score
 - ▣ Measures how similar an object is to its own cluster compared to other clusters
 - ▣ Value ranges from -1 to 1 and a higher value indicates better clustering

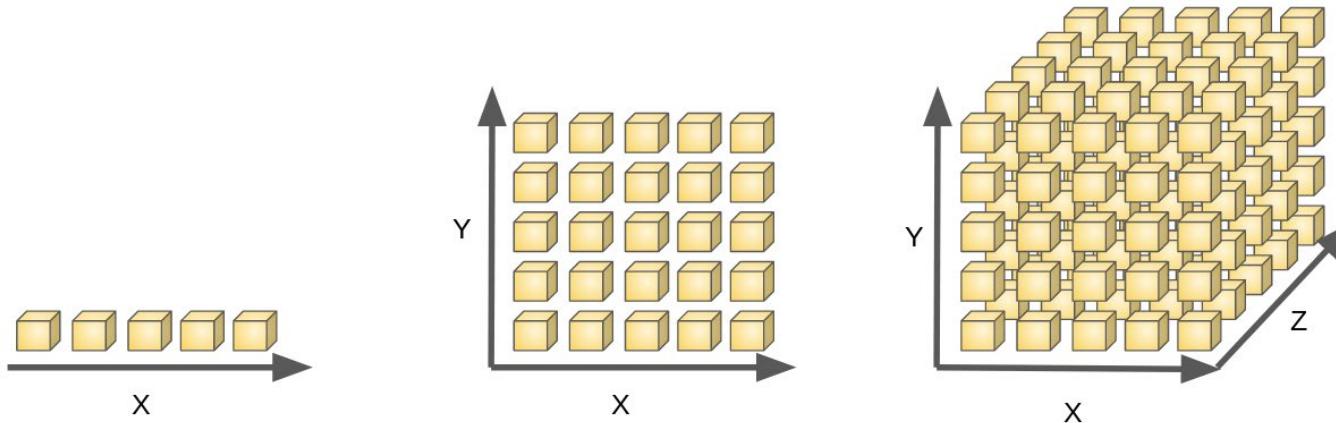
DIMENSIONALITY REDUCTION: PCA

Week 12

Dimensionality Reduction

Why is Dimensionality Reduction Needed?

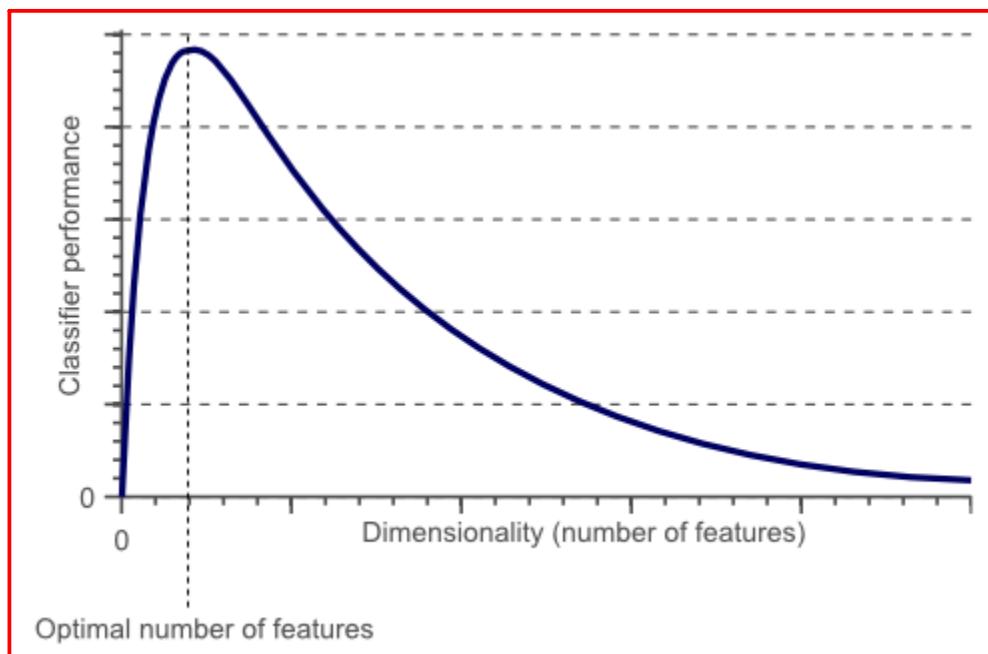
- Avoid the curse of dimensionality
 - ▣ Curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional space



- As the dimensionality increases, we need more data to fill the space

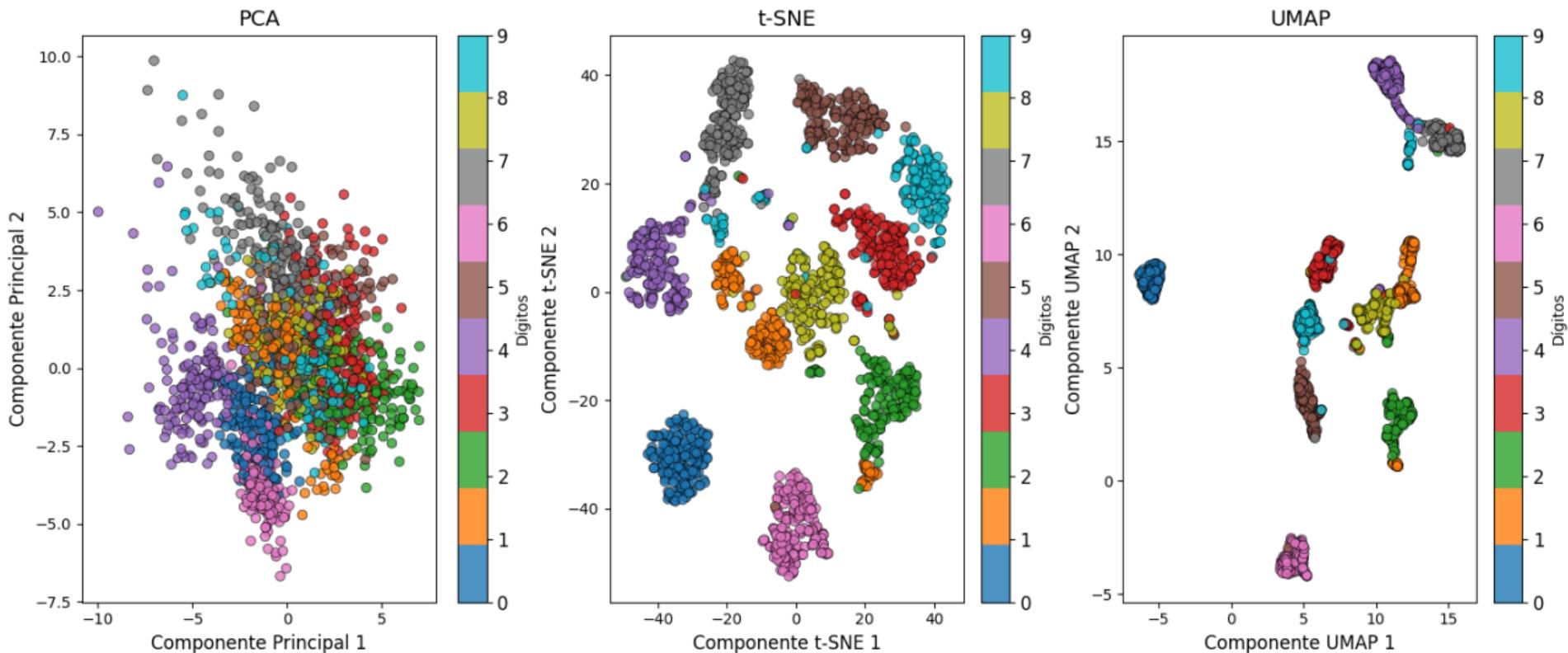
Why is Dimensionality Reduction Needed?

- With a fixed number of training samples, the predictive power of a classifier or regressor first increases as number of dimensions/features used is increased but then decreases



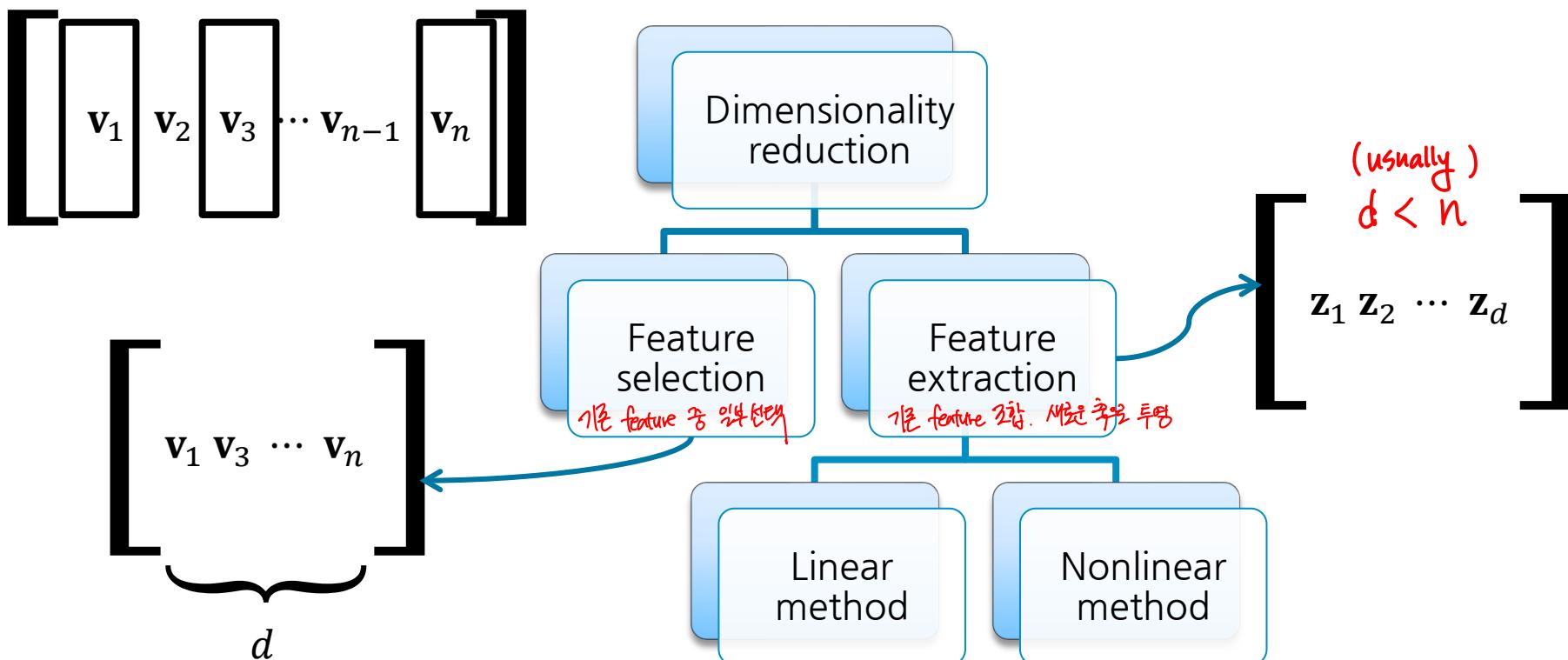
Why is Dimensionality Reduction Needed?

- Dimensionality reduction helps in understanding data by reducing it to 2D or 3D representations



Dimensionality Reduction

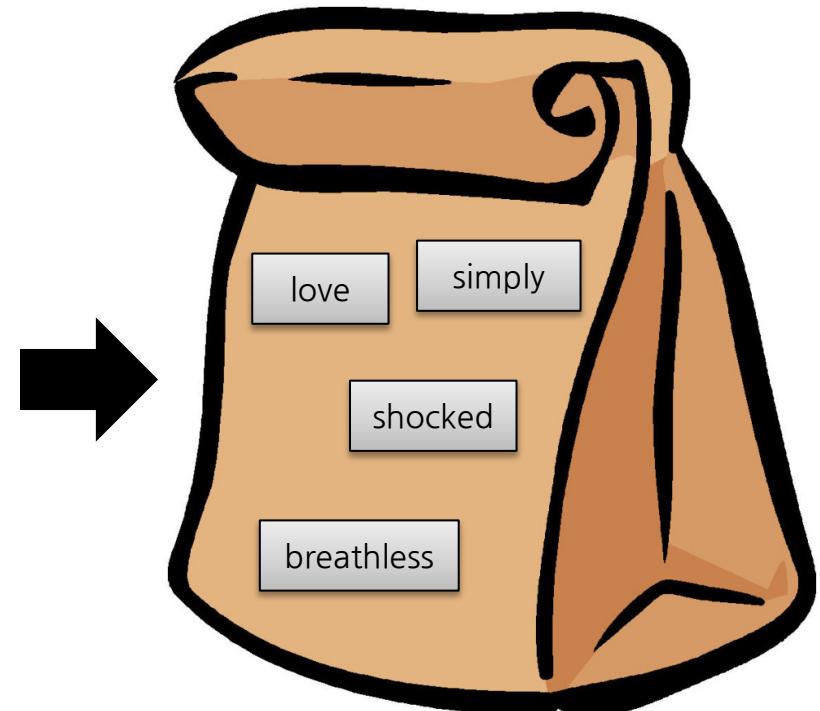
- Dimensionality reduction
 - ▣ The process of reducing the number of variables
- Hierarchy of dimensionality reduction



Example: Feature Selection

- Do you remember the bag of words representation for text data?

I simply loved this film but was shocked by the bad reviews that people gave it. To this I say to them: You seriously misunderstood the meaning of it. Although I won't reveal any real details about the meaning because I think that you should try and understand it yourself. The movie was terrific and simply breathless the whole time. I felt awestruck about how the life of one man could be so changed after an experience that Hanks went through. I say that every element of the film was perfect. And for those of you who hate Wilson, you have to understand about how human he really was to Chuck. I was amazed on how well this movie was made and think that everybody should have an experience that should cause you to take stock of your life



Example: Feature Selection

Do not stand at my grave and weep.
I am not there. I do not sleep.



Term	Frequency
do	2
not	2
stand	1
at	1
my	1
grave	1
and	1
weep	1
I	2
am	1
there	1
sleep	1

Example: Feature Selection

□ Text Categorization

Politics

TRACKING TRUMP'S AGENDA | VIDEO | THE UPSHOT



Trump Is 'Not Happy' With Border Deal, but Doesn't Say if He Will Sign It

The president, who said he would have to study the deal, all but ruled out another government shutdown and emphasized that he would find "other methods" to finance a

Business | Tech | Econ | Media | Money

MARKET
SNAPSHOT

11:58
PM

S&P 500 ↗ 2744.73 +1.29%

DOW
INDUSTRIALS

25425.76
+1.49%

Smaller Tax Refunds Surprise Those Expecting More Relief

President Trump's tax plan promised benefits, but as returns are being filed, some frustrated people are getting smaller refunds, or even writing checks.

4h ago · By TARA SIEGEL BERNARD



CHRISTIE HEMM KLOK

Technology

DEALBOOK | MARKETS | ECONOMY | ENERGY | MEDIA | TECHNOLOGY | PERSONAL TECH | ENTREPRENEURSHIP | YO



JENAH MOON FOR THE NEW YORK TIMES

T-Mobile-Sprint Deal Gets New Scrutiny From the Left

Democratic lawmakers, empowered by their new House majority, have amplified their criticism of the deal, and two hearings are set this week.

4h ago · By CECILIA KANG

Example: Feature Selection

- Titles of the news articles

Trump Is ‘Not Happy’ With Border Deal, but Doesn’t
Say if He Will Sign It

Video

Smaller Tax Refunds Surprise Those Expecting More
Relief

T-Mobile-Sprint Deal Gets New Scrutiny From the
Left

Are there any irrelevant or redundant words?

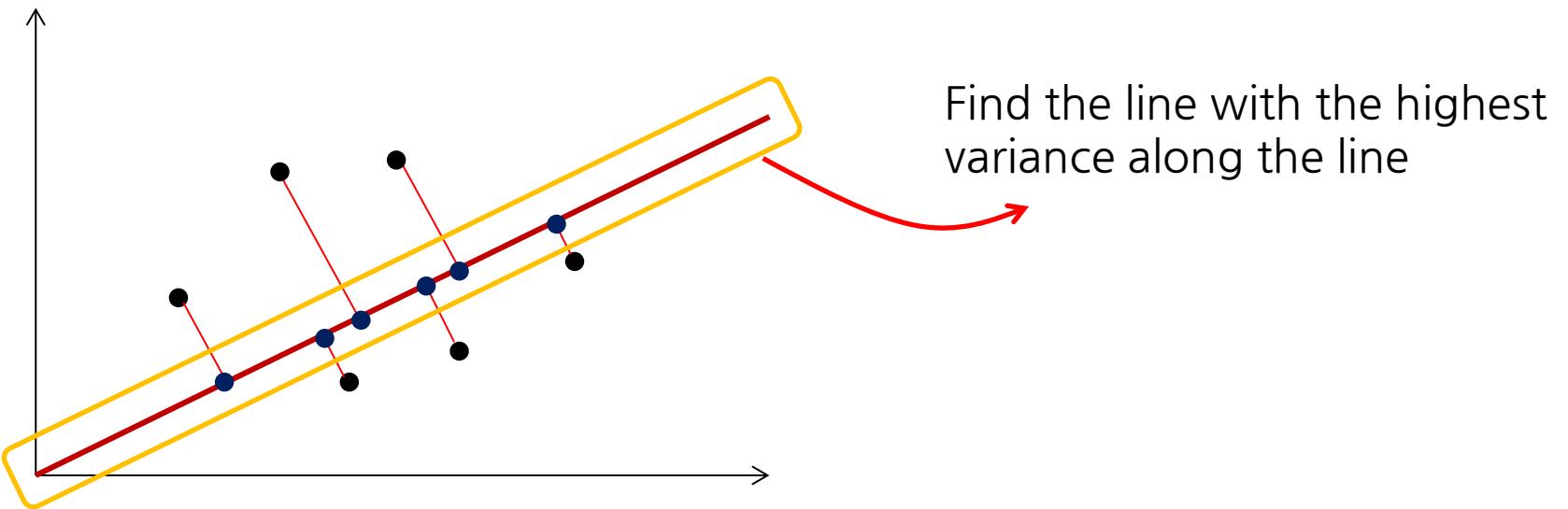
Principal Component Analysis

Principal Component Analysis (PCA)

- PCA is the **orthogonal transformation** to possibly correlated variables into a set of values into linearly uncorrelated variables
 - Uncorrelated variables are called principal components
principal components < original variables
- The number of principal components is less than or equal to the number of original variables
 - Even though the number of principal components are equal to the number of original variables, we select smaller number of components and then use for analysis

Principal Component Analysis (PCA)

- Criterion to find principal component is to achieve the highest variance
 - Variance of projected data samples on principal component



- ✗ □ First principal component has the highest possible variance
- Each succeeding component in turn has the highest variance possible and it should be orthogonal to the preceding components

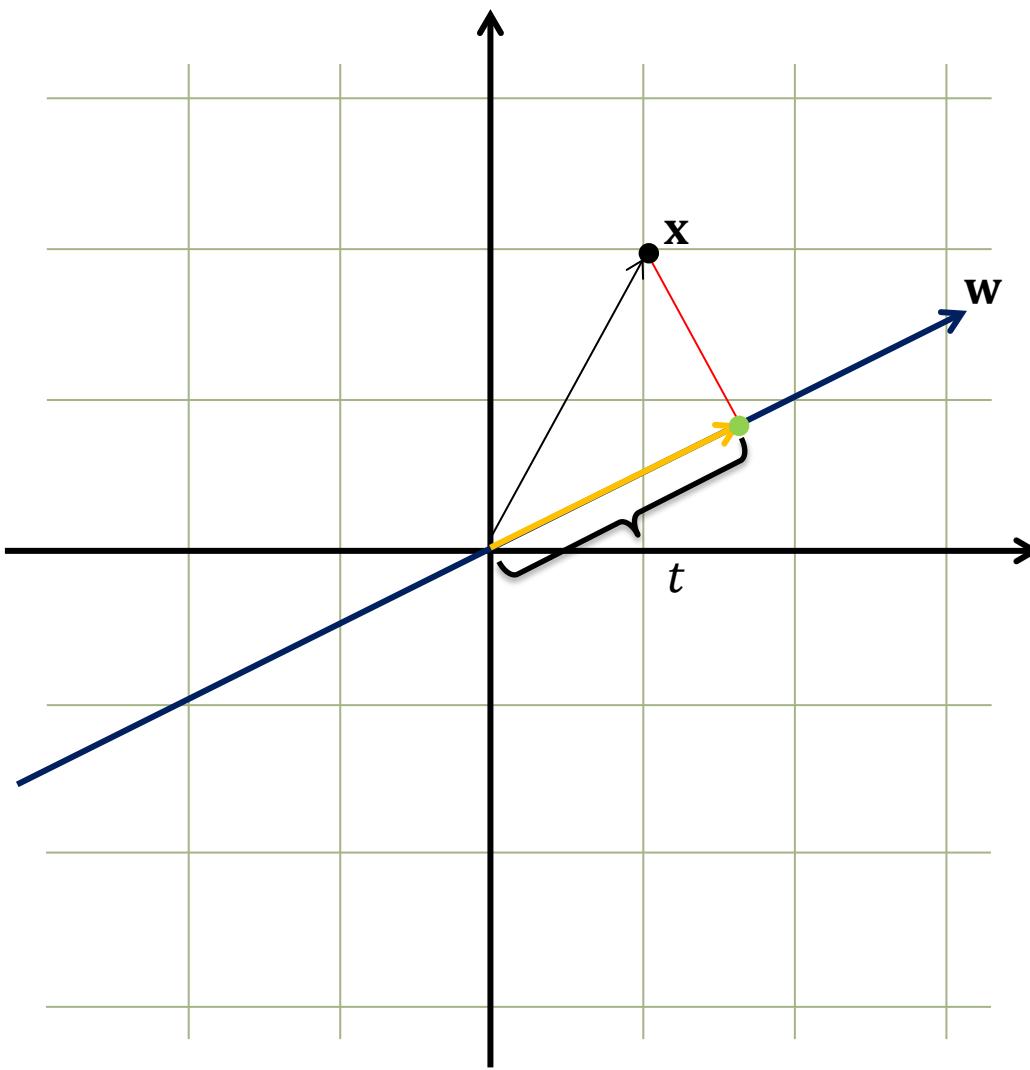
Principal Component Analysis (PCA)

- Which feature is the most helpful to distinguish one house from another?

ID	Value	Area	Floors	Household
1	148	72	4	20
2	156	76	4	22
3	160	86	4	22
4	165	79	4	24
5	169	88	5	30
6	184	90	5	35

- It's a good thing to have features with **high variance**, since they will be more **informative** and more important
 - Maximize variance
- It's a bad thing to have **highly correlated features**, or **high covariance**, since they can be deduced from one another with little loss in information, and thus keeping them **together** is redundant
 - Obtain orthogonal features

* Projection on the Line



Projected point on the line of data point \mathbf{x}

$$t = \mathbf{w} \cdot \mathbf{x}$$

Direction of line is defined as \mathbf{w} and \mathbf{w} is unit length vector

$$\mathbf{w} = \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}} \right)$$

$$\mathbf{x} = (1, 2)$$

$$t = \mathbf{w} \cdot \mathbf{x} = \mathbf{w}^T \mathbf{x} = \frac{2}{\sqrt{5}} + \frac{2}{\sqrt{5}} = \frac{4}{\sqrt{5}}$$

Principal Component Analysis (PCA)

- Find first component, \mathbf{w}_1 for data set that each dimension has zero mean

$$\mathbf{w}_1 = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \sum_i (t_{1i})^2 = \boxed{\underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \sum_i (\mathbf{x}_i \cdot \mathbf{w})^2}$$

- t_{1i} is the score(projected point on the first component) of i -th data point

- Define data matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}^T = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Dot product
 $\mathbf{x}_1 \cdot \mathbf{w}$

$$\mathbf{X}\mathbf{w} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}^T \mathbf{w} = \begin{bmatrix} \mathbf{x}_1 \cdot \mathbf{w} \\ \mathbf{x}_2 \cdot \mathbf{w} \\ \vdots \\ \mathbf{x}_n \cdot \mathbf{w} \end{bmatrix}$$

- Rewrite \mathbf{w}_1

$$\mathbf{w}_1 = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \|\mathbf{X}\mathbf{w}\|^2 = \boxed{\underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}$$

Principal Component Analysis (PCA)

- Since unit vector constraint

$$\mathbf{w}_1 = \arg \max \left(\frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right)$$

- The larger $\|\mathbf{w}\|$ is, the larger $\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}$ is
- $\mathbf{w}^T \mathbf{w}$ is the penalty term on $\|\mathbf{w}\|$ ($\|\mathbf{w}\| = \sqrt{\mathbf{w}^T \mathbf{w}}$)

$$\mathbf{x} = (1, 2)$$

$$\mathbf{w}_1 = (1, 2), \mathbf{w}_2 = (2, 4)$$

$$t_1 = \mathbf{w}_1 \cdot \mathbf{x} = \mathbf{w}_1^T \mathbf{x} = 1 + 4 = 5$$

$$t_2 = \mathbf{w}_2 \cdot \mathbf{x} = \mathbf{w}_2^T \mathbf{x} = 2 + 8 = 10$$

$$\therefore t_1^2 < t_2^2$$

- Solution of optimization problem

- \mathbf{w}_1 = eigenvector of $\mathbf{X}^T \mathbf{X}$ with the largest eigenvalue
- $\mathbf{X}^T \mathbf{X}$ is proportional to covariance matrix of data when mean of each dimension is zero \Rightarrow 特수한 covariance matrix ($\mathbf{X}^T \mathbf{X}$)의 eigenvectors가 연관.

※ Covariance

- Variance of a random variable X is the expected value of the squared deviation from the mean ($\mu = \mathbb{E}[X]$)

$$Var(X) = \mathbb{E}[(X - \mu)^2]$$

- Sample variance is calculated by

$$Var(x) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n - 1)}$$

- Covariance is a measure of how much two random variables change together

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

- Variance is the covariance of a random variable with itself

$$Var(X) = Cov(X, X)$$

- Sample covariance is calculated by

$$cov(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

※ Covariance

□ Covariance matrix, \mathbf{C}

- Matrix whose elements correspond to possible covariance values between all the different dimensions

$$\mathbf{C} = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_p) \\ cov(x_2, x_1) & cov(x_2, x_2) & \cdots & cov(x_2, x_p) \\ \vdots & \vdots & & \vdots \\ cov(x_p, x_1) & cov(x_p, x_2) & \cdots & cov(x_p, x_p) \end{bmatrix}$$

- If mean of each dimension is zero in data matrix, \mathbf{X} ,

$$\mathbf{C} \propto \mathbf{X}^T \mathbf{X}$$

※ Eigenvector and Eigenvalue

- For some matrix \mathbf{A} , the vector \mathbf{x} satisfying following relation is eigenvector of matrix \mathbf{A}

$$\mathbf{Ax} = \lambda \mathbf{x}$$

- λ is the eigenvalue of eigenvector \mathbf{x}
 - The number of eigenvectors depends on matrix
-
- Example

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$$

- For vector $\mathbf{x} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$

$$\mathbf{Ax} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 11 \\ 5 \end{bmatrix}$$

- For vector $\mathbf{y} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$

$$\mathbf{Ay} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 8 \end{bmatrix} = 4 \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

eigenvector
eigenvalue

※ Eigenvector and Eigenvalue

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc = 0$$

- How to get eigenvector and eigenvalue?

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$$

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = 0$$

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

- Eigenvector and eigenvalue should satisfy $\mathbf{Ax} = \lambda \mathbf{x}$

$$\begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \lambda \begin{bmatrix} x \\ y \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} - \lambda \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 - \lambda & 3 \\ 2 & 1 - \lambda \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$$

- If there exist nontrivial solution (trivial solution = $\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$), determinant of $\begin{bmatrix} 2 - \lambda & 3 \\ 2 & 1 - \lambda \end{bmatrix}$ should be 0

$$(2 - \lambda)(1 - \lambda) - 6 = 0 \rightarrow \lambda^2 - 3\lambda + 4 = 0$$

eigenvalue

$$\lambda = 4 \text{ or } -1$$

- When $\lambda = 4$, $\mathbf{x} = [3 \quad 2]^T$
- When $\lambda = -1$, $\mathbf{x} = [1 \quad -1]^T$

eigenvector

Principal Component Analysis (PCA)

- Succeeding process
 - ▣ Subtracting preceding components from $\mathbf{X} \rightarrow$ Create new data matrix

$$\widehat{\mathbf{X}}_k = \mathbf{X} - \sum_{i=1}^{k-1} \mathbf{X} \mathbf{w}_i \mathbf{w}_i^T$$

($k-1$)개의 주성분을 제거

- ▣ Find the principal component that extracts the maximum variance from new data matrix

$$\mathbf{w}_k = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \|\widehat{\mathbf{X}}_k \mathbf{w}\|^2 = \operatorname{argmax} \left(\frac{\mathbf{w}^T \widehat{\mathbf{X}}_k^T \widehat{\mathbf{X}}_k \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right) \quad k\text{번째 주성분 계산}$$

- Solve above equation is the also same as calculated the remaining eigenvectors of $\mathbf{X}^T \mathbf{X}$
- \mathbf{w}_2 =eigenvector of $\mathbf{X}^T \mathbf{X}$ with the second largest eigenvalue

$\mathbf{w}_k = \mathbf{X}^T \mathbf{X}$ 에서 k 번째 큰 eigenvalue 가진 eigenvector

\mathbf{X} 에서 이미 설명된 방향 제거 후, 남은 부분에서 가장 분산이 큰 방향 찾기

PC \Rightarrow 분산 최대 / 서로 orthogonal / 충복 없는 정보 분할

※ Eigendecomposition

- Eigendecomposition
 - ▣ Eigendecomposition is the factorization of a matrix into a canonical form, whereby the matrix is represented in terms of its eigenvalues and eigenvectors
 - ▣ Let A be a square $n \times n$ matrix with n linearly independent eigenvectors q_i (where $i = 1, \dots, n$). Then A can be factored as

$$A = Q\Lambda Q^{-1}$$

- Q : $n \times n$ matrix whose i th column is the eigenvector q_i of A
 - Λ : diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\Lambda_{ii} = \lambda_i$

Principal Component Analysis (PCA)

- Finally,

$$\mathbf{T} = \textcolor{red}{\cancel{\mathbf{X}}} \mathbf{W}$$

- \mathbf{W} is p -by- p matrix whose columns are the eigenvectors of $\mathbf{X}^T \mathbf{X}$ and it is called loading matrix

$$\mathbf{W} = [\mathbf{w}_1 \quad \mathbf{w}_2 \quad \cdots \quad \mathbf{w}_p]$$

- λ_i : eigenvalue of $\mathbf{w}_i \in \mathbb{R}^{p \times 1}$
- $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$
- Principal component is linear combinations of original features and transformation by loading matrix is linear transformation

Principal Component Analysis (PCA)

- Dimensionality reduction by PCA
 - ▣ Keeping only the first l principal components (where $p > l$)

$$\mathbf{W}_l = [\mathbf{w}_1 \quad \mathbf{w}_2 \quad \cdots \quad \mathbf{w}_l]$$

- \mathbf{W}_l is $p \times l$ matrix

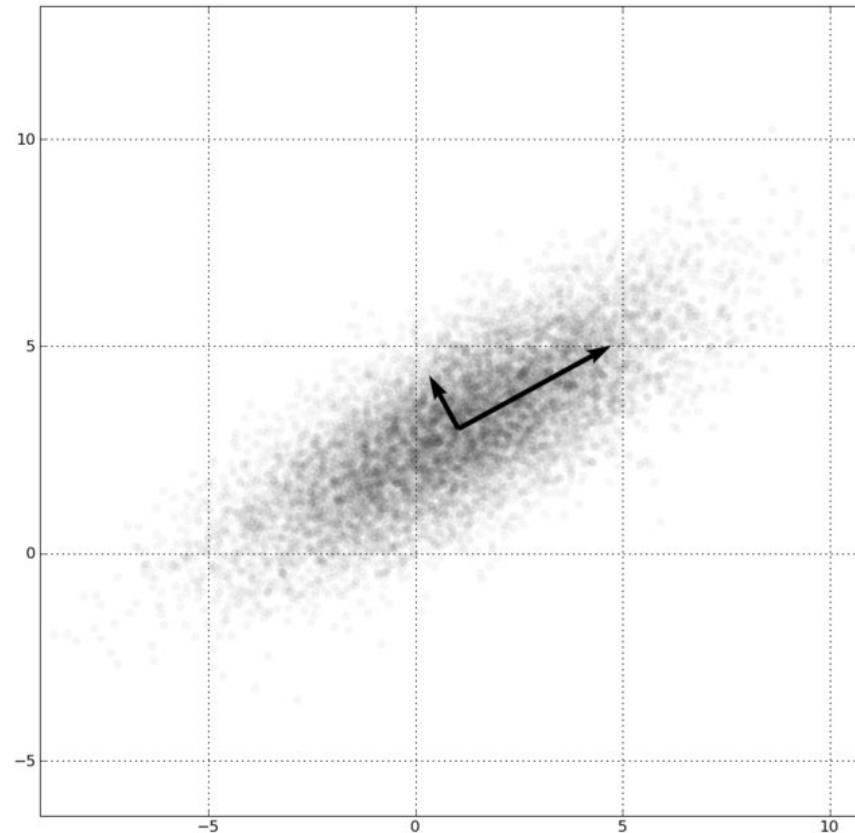
$$\mathbf{w}_i \in \mathbb{R}^{p \times 1}$$

- Dimension-reduced data set is obtained by truncated transformation

$$\mathbf{T}_l = \mathbf{X}\mathbf{W}_l$$

Principal Component Analysis (PCA)

- 2-dimensional data set and its principal components



- Web applet
 - <https://setosa.io/ev/principal-component-analysis/>

※ Linear transformation

- Linear transformation is a mapping $f: V \rightarrow W$ that preserves the operations of addition and scalar multiplication

Addition: $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$

Scalar multiplication: $f(\alpha\mathbf{x}) = \alpha f(\mathbf{x})$

- Example

- Identity map $f(\mathbf{x}) = \mathbf{x}$ is linear transformation

- $f(\mathbf{x} + \mathbf{y}) = \mathbf{x} + \mathbf{y} = f(\mathbf{x}) + f(\mathbf{y})$
 - $f(\alpha\mathbf{x}) = \alpha\mathbf{x} = \alpha f(\mathbf{x})$

- map $f: x \rightarrow x^2$ is not linear transformation

- $f(\mathbf{x} + \mathbf{y}) = \mathbf{x}^2 + 2\mathbf{x}\mathbf{y} + \mathbf{y}^2 \neq f(\mathbf{x}) + f(\mathbf{y})$
 - $f(\alpha\mathbf{x}) = \alpha^2\mathbf{x}^2 \neq \alpha f(\mathbf{x})$

- map $f: x \rightarrow x + 1$ is not linear transformation

- $f(\mathbf{x} + \mathbf{y}) = \mathbf{x} + \mathbf{y} + 1 \neq f(\mathbf{x}) + f(\mathbf{y})$ ($f(\mathbf{x}) + f(\mathbf{y}) = \mathbf{x} + 1 + \mathbf{y} + 1$)
 - $f(\alpha\mathbf{x}) = \alpha\mathbf{x} + 1 \neq \alpha f(\mathbf{x})$ ($\alpha f(\mathbf{x}) = \alpha\mathbf{x} + \alpha$)

- If \mathbf{A} is $m \times n$ matrix, then map $f: \mathbf{x} \rightarrow \mathbf{Ax}$ is linear transformation

$$f(\alpha\mathbf{x}) = \alpha f(\mathbf{x}) = \alpha\mathbf{Ax}, \quad f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}) = \mathbf{Ax} + \mathbf{Ay} = \mathbf{A}(\mathbf{x} + \mathbf{y})$$

※ Linear Combination

- If $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ are vectors and a_1, a_2, \dots, a_n are scalars, then linear combination of those vectors with those scalars as coefficient is

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n$$

- 3-dimensional vector (a_1, a_2, a_3) is linear combination of $e_1 = (1,0,0)$, $e_2 = (0,1,0)$, $e_3 = (0,0,1)$

$$\begin{aligned}(a_1, a_2, a_3) &= (a_1, 0, 0) + (0, a_2, 0) + (0, 0, a_3) \\ &= a_1(1,0,0) + a_2(0,1,0) + a_3(0,0,1) = a_1e_1 + a_2e_2 + a_3e_3\end{aligned}$$

$$\mathbf{w} = w_1\mathbf{e}_1 + w_2\mathbf{e}_2 + \cdots + w_p\mathbf{e}_p$$

- Principal component $\mathbf{w} = (w_1, w_2, \dots, w_p)$ is linear combination of unit vectors on each dimension representing by each variables

Example: PCA

- Find principal components of given data

x	2.5	0.5	2.2	1.9	3.1	2.3	2.0	1.0	1.5	1.1
y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

- Step 1) subtract from of the data dimensions for each dimension to have zero mean
 - $\bar{x} = 1.81, \bar{y} = 1.91$

X

$x_i - \bar{x}$	0.69	-1.31	0.39	0.09	1.29	0.49	0.19	-0.81	-0.31	-0.71
$y_i - \bar{y}$	0.49	-1.21	0.99	0.29	1.09	0.79	-0.31	-0.81	-0.31	-1.01

- Step 2) Calculate covariance matrix of new data ($\mathbf{X}^T \mathbf{X}$) $2 \text{ feature} \Rightarrow 2 \times 2 \text{ matrix}$

$$\text{Cov}(x,y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$C = \begin{bmatrix} 0.617 & 0.615 \\ 0.615 & 0.717 \end{bmatrix} = \begin{bmatrix} \text{Cov}(x,x) & \text{Cov}(y,x) \\ \text{Cov}(x,y) & \text{Cov}(y,y) \end{bmatrix}$$

Example: PCA

- Find principal components of given data

x	2.5	0.5	2.2	1.9	3.1	2.3	2.0	1.0	1.5	1.1
y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

- Step 3) Calculate the eigenvectors and eigenvalues of the covariance matrix
 - The largest eigenvalue is 1.28 and corresponding eigenvector is

$$C = \begin{bmatrix} 0.617 & 0.615 \\ 0.615 & 0.717 \end{bmatrix}$$

$$\mathbf{w}_1 = \begin{bmatrix} -0.678 \\ -0.735 \end{bmatrix}$$

$$* \begin{bmatrix} -0.927 \\ \sqrt{(-0.927)^2 + (-0.735)^2} & \frac{1}{\sqrt{(-0.927)^2 + (-0.735)^2}} \end{bmatrix}^T = \mathbf{w}_1$$

- The second largest eigenvalue is 0.049 and corresponding eigenvector is

$$\mathbf{w}_2 = \begin{bmatrix} -0.735 \\ 0.678 \end{bmatrix}$$

* Check eigenvector is unit vector!

$$\det(C - \lambda I) = \begin{bmatrix} 0.617 - \lambda & 0.615 \\ 0.615 & 0.717 - \lambda \end{bmatrix} = 0$$

$$\lambda = 1.28 \text{ or } 0.049$$

$$(C - \lambda I) \mathbf{w} = 0$$

$$\text{unit norm} \Rightarrow \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

$$\text{if } \lambda = 1.28, \begin{bmatrix} -0.663 & 0.615 \\ 0.615 & -0.563 \end{bmatrix} \begin{bmatrix} 1 \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$1:y = 0.927 : 1$$

$$\begin{bmatrix} -0.927 \\ -1 \end{bmatrix} \xrightarrow{\text{unit vector}} \mathbf{w}_1 = \begin{bmatrix} -0.678 \\ -0.735 \end{bmatrix}$$

$$\text{if } \lambda = 0.049, \dots \mathbf{w}_2 = \begin{bmatrix} -0.735 \\ 0.678 \end{bmatrix}$$

Example: PCA

- Find principal components of given data

$$\bar{x} = 1.81$$

$$\bar{y} = 1.91$$

x	2.5	0.5	2.2	1.9	3.1	2.3	2.0	1.0	1.5	1.1
y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

- Step 4) Choosing components and forming a loading matrix

- If you choose two principal components both

$$W = \begin{bmatrix} -0.678 & -0.735 \\ -0.735 & 0.678 \end{bmatrix}$$

- If you want to reduce dimensionality

$$W = \begin{bmatrix} -0.678 \\ -0.735 \end{bmatrix}$$

$Q \times 2$

1×2

- Step 5) Derive the new data set

$$\begin{bmatrix} 0.69, 0.49 \end{bmatrix} \begin{bmatrix} -0.678 & -0.735 \\ -0.735 & 0.678 \end{bmatrix} = \begin{bmatrix} 0.69 \times (-0.678) + 0.49 \times (-0.735) \\ 0.69 \times (-0.735) + 0.49 \times 0.678 \end{bmatrix} \xrightarrow{\text{10x2} \cdot 2x2} T = XW$$

$$W = \begin{bmatrix} x_1, y_1 \\ \vdots \\ x_{10}, y_{10} \end{bmatrix} \quad W = \begin{bmatrix} x'_1, y'_1 \\ \vdots \\ x'_{10}, y'_{10} \end{bmatrix}$$

x'	-0.83	1.78	-0.99	-0.27	-1.68	-0.91	0.99	1.14	0.44	1.22
y'	-0.18	0.14	0.38	0.13	-0.21	0.18	-0.35	0.46	0.02	-0.16

Explained Variance $\stackrel{=?}{\lambda}$ eigenvalue

- Explained variance and explained variance ratio
 - ▣ Explained variance: Variance that that is captured by each principal component
 - ▣ Explained variance ratio: The proportion of the dataset's total variance that is captured by each principal component

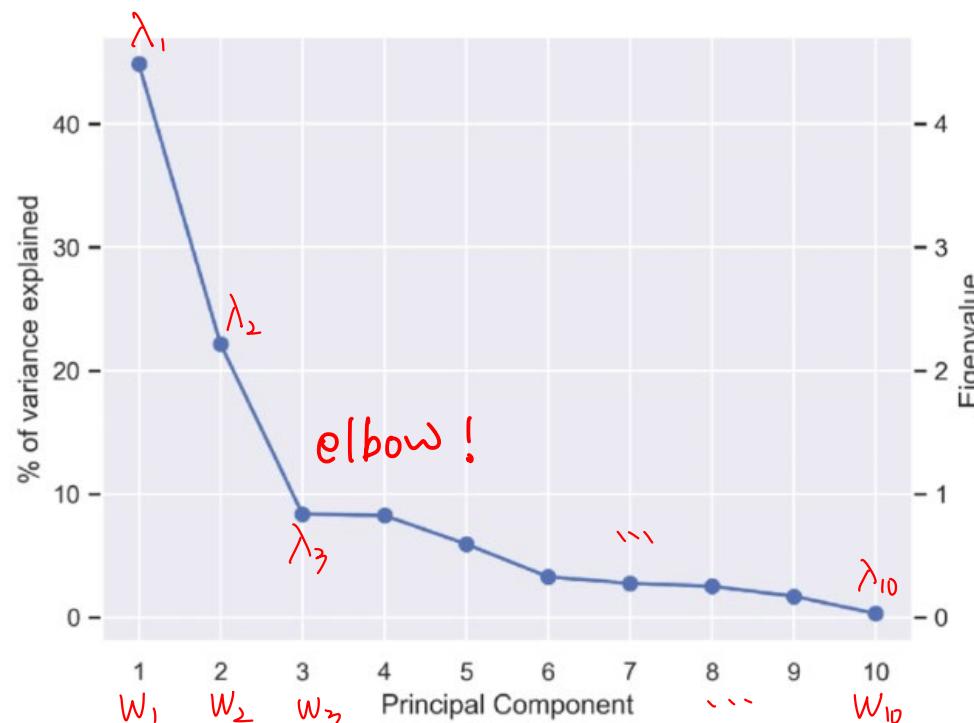
Explained variance of $w_i = \lambda_i$ 각 주성분이 얼마나 많은 분산을 포착하는지

Explained variance ratio of $w_i = \frac{\lambda_i}{\sum_i \lambda_i}$
각 주성분이 데이터셋의 총 분산 중 얼마만큼의 비율을 포착하는지

Scree Plot in PCA

□ Scree plot

- A scree plot is a line plot of the eigenvalues (or explained variance) associated with each principal component
- The components are ordered on the x-axis from the first to the last
- The y-axis shows the eigenvalue or explained variance or explained variance ratio for each component



Scree Plot in PCA

- Purpose of the Scree Plot
 - ▣ Helps to determine the optimal number of principal components to retain
 - ▣ Aims to identify the point after which additional components contribute only marginally to explaining variance \Rightarrow elbow point
- Limitations of the Scree Plot
 - ▣ The “elbow” may not be obvious in some datasets
 - ▣ Interpretation can be subjective
 - ▣ Should ideally be used in combination with
 - Cumulative variance explained
 - Cross-validation

Feature Scaling

- PCA finds principal component to achieve the highest variance
 - ▣ The variable with large scale is dominated on principal component
- ex) When distance measure is change from m to cm, variance increases 10000(100^2) times

Length(m)	Length(cm)
1.5	150
1.7	170
2.3	230
3.3	330
2.7	270
1.9	190

Sample variance(m)=0.46

Sample variance(cm)=4586

$$\frac{x - \mu}{\sigma}$$

- Before apply PCA to data samples, standardization is applied
 - ▣ Transform each dimension to have unit variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Reconstruct to Original Space

- Transformed data by PCA can be reconstructed to original space
 - ▣ Recall the final transformation

$$\mathbf{T} = \mathbf{X}\mathbf{W}$$

- ▣ Old data can be written as

$$\mathbf{X} = \mathbf{T}\mathbf{W}^{-1}$$

- If \mathbf{W} consists of unit vectors which are orthogonal to each other, inverse matrix of \mathbf{W} is the same as the transpose of \mathbf{W} , \mathbf{W}^T

$$\therefore \mathbf{X} = \mathbf{T}\mathbf{W}^T$$

- If you subtract mean of each dimension from original data

$$\mathbf{X} = \mathbf{T}\mathbf{W}^T + \boldsymbol{\mu}$$

- $\boldsymbol{\mu}$ is mean vector of \mathbf{X} ($\boldsymbol{\mu} = [\bar{x}_1 \bar{x}_2 \cdots \bar{x}_p]^T$)
 - Actually, if \mathbf{W} is not square matrix(if you choose the smaller number of principal components than original dimension), \mathbf{W}^{-1} does not exist. However, in this case, reconstruction is performed through \mathbf{W}^T

Reconstruction Error

- Reconstruction error
 - ▣ The reconstruction error measures how well the original data can be approximated from the lower-dimensional representation

$$\mathbf{X}' = \mathbf{T}\mathbf{W}^T$$

$$Error = \|\mathbf{X} - \mathbf{X}'\|_F^2$$

- $\|\cdot\|_F$: Frobenius norm

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$$

- Another PCA optimization objective

$$\begin{aligned} \min_{\mathbf{W}} \|\mathbf{X} - \mathbf{X}'\|_F^2 &= \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^T\|_F^2 \\ s.t \quad \mathbf{W}^T\mathbf{W} &= \mathbf{I} \end{aligned}$$

- ▣ PCA finds an orthogonal projection $\mathbf{W}\mathbf{W}$ that minimizes reconstruction error

Reconstruction Error

- Intuition
 - ▣ Low-dimensional projection means we **keep only the most "important" directions** (those with high variance)
 - ▣ The **reconstruction error comes from ignoring the remaining, less important directions**
 - ▣ The **more principal components we keep, the lower the error**, but the **higher the dimensionality**

Application: PCA

- Extract important features through PCA for face recognition
 - ▣ For image recognition, simple way to represent each image is to vectorization

16×16 image



Each pixel
represents one
dimension

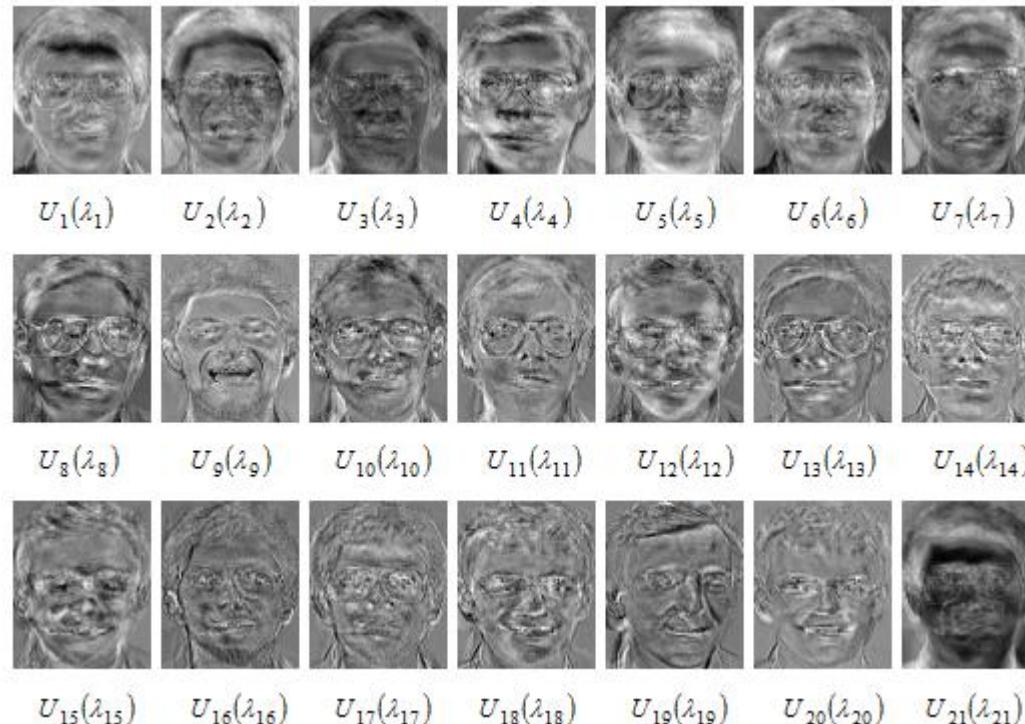
Transform to
vector with 256
dimensions

- ▣ Apply PCA to the set of image vectors and obtain principal components
→ transform image vector to lower dimensional space by loading matrix
 - Usually image is high-dimensional data
 - Through PCA, image can be compressed to low-dimensional data

Application: PCA

□ Eigenfaces

- A set of eigenvectors when they are used in the computer vision problem of human face recognition



- Eigenface can be viewed as a sort of map of the variations between faces
- PCA analysis has identified the statistical patterns in the data

Multidimensional Scaling

Multidimensional Scaling(MDS)

- MDS is a means of visualizing the level of similarity of individual cases of a dataset 고차원 데이터 간 '서로 간 거리'만 알 때, 그 거리를 보존하는 2D-3D 위치 표기 찾기
 - Visualizing requires reduction of dimensionality to 2 or 3
→ dimensionality reduction
- MDS attempts to find an embedding such that distances are preserved
 - Let $d(i,j)$ =distance between i -th and j -th objects
 - Dissimilarity matrix
$$D = \begin{bmatrix} d(1,1) & d(1,2) & \cdots & d(1,n) \\ d(2,1) & d(2,2) & \cdots & d(2,n) \\ \vdots & \vdots & & \vdots \\ d(n,1) & d(n,2) & \cdots & d(n,n) \end{bmatrix}$$
 - The goal of MDS is, given D , to find transformed vector \mathbf{y}_i corresponding to \mathbf{x}_i such that $\|\mathbf{y}_i - \mathbf{y}_j\| \approx d(i,j)$

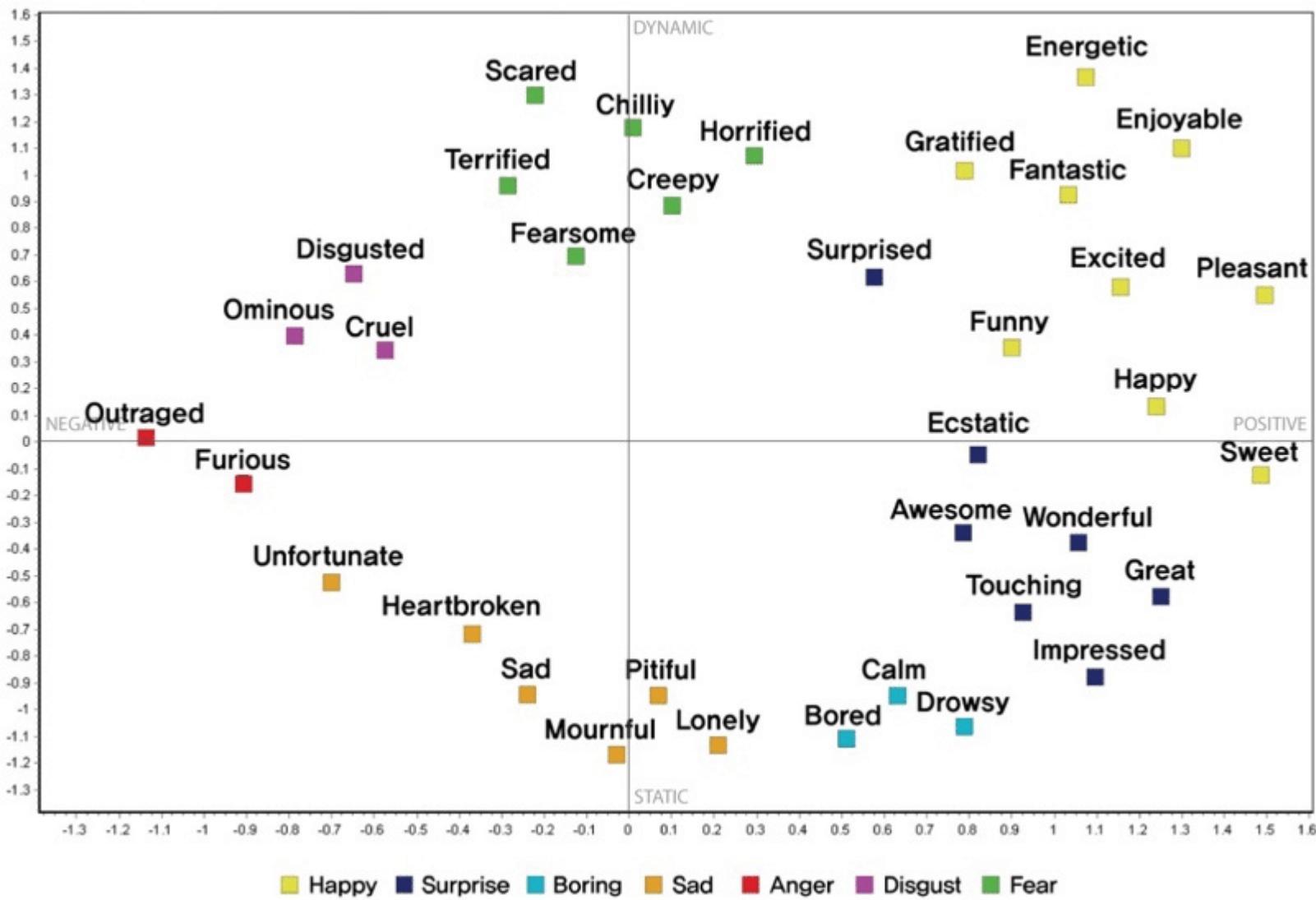
Multidimensional Scaling(MDS)

- In view point of dimensionality reduction, MDS is the method to reduce dimensionality by preserving pair-wise distances in high dimension
 - ▣ In this case MDS uses Euclidean distance for $d(i, j) = \|\mathbf{x}_i - \mathbf{x}_j\|$
→ Find low dimensional vector \mathbf{y}_i which satisfies $\|\mathbf{y}_i - \mathbf{y}_j\| \approx \|\mathbf{x}_i - \mathbf{x}_j\|$
거리 차이를 보존하는 좌표 \mathbf{y}_i 찾기.
- MDS minimizes summation of errors between distances in high dimension and distances in low dimension to obtain \mathbf{y}_i

$$E = \sum_{i,j}^n (\|\mathbf{y}_i - \mathbf{y}_j\| - \|\mathbf{x}_i - \mathbf{x}_j\|)^2$$

거리 차이 제곱합 최소화.

Multidimensional Scaling(MDS)



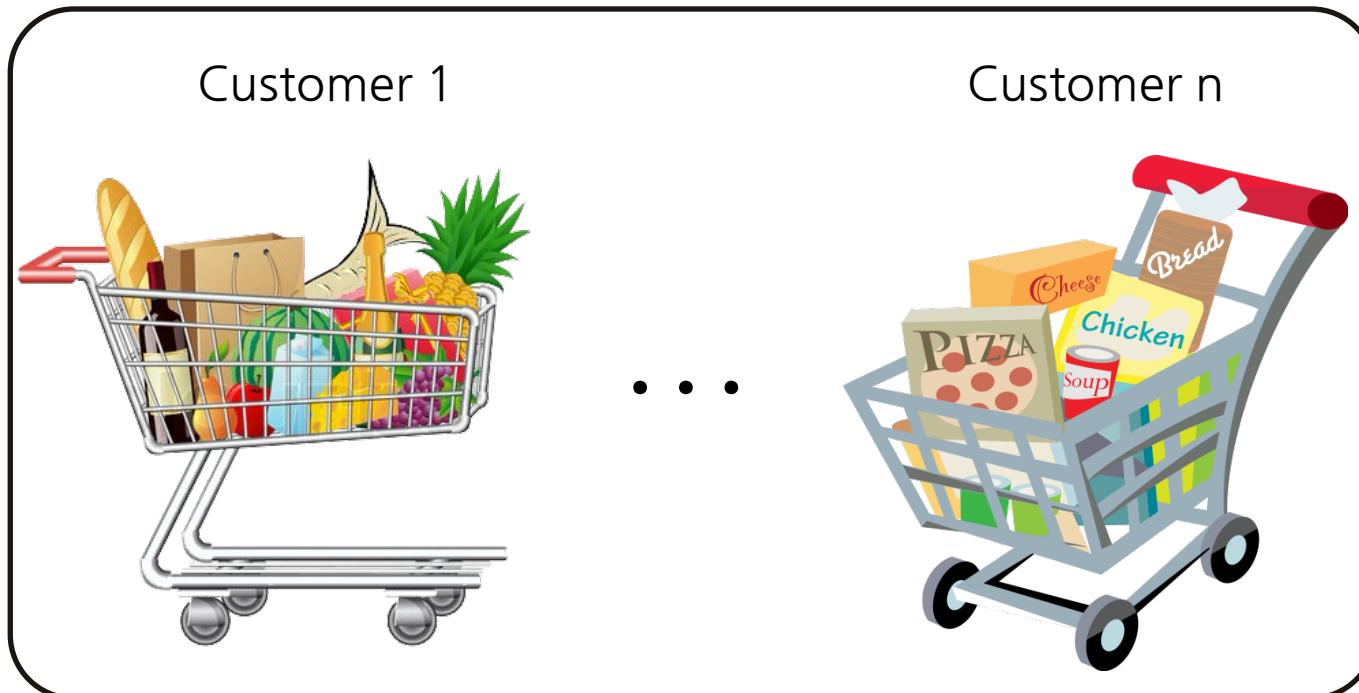
ASSOCIATION RULE MINING

Week13

Association Rule

What is Market Basket Analysis?

- Finding some useful information in ‘market basket’
- What kinds of information?
 - ▣ Who customers are
 - ▣ Which products tend to be purchased together
 - ▣ Why some products tend to be purchased together
- Association rule: Information like “If item A then item B” ($A \Rightarrow B$)



Collecting
transactions &
finding useful
information

Point of Sale Transactions

- Transaction and item

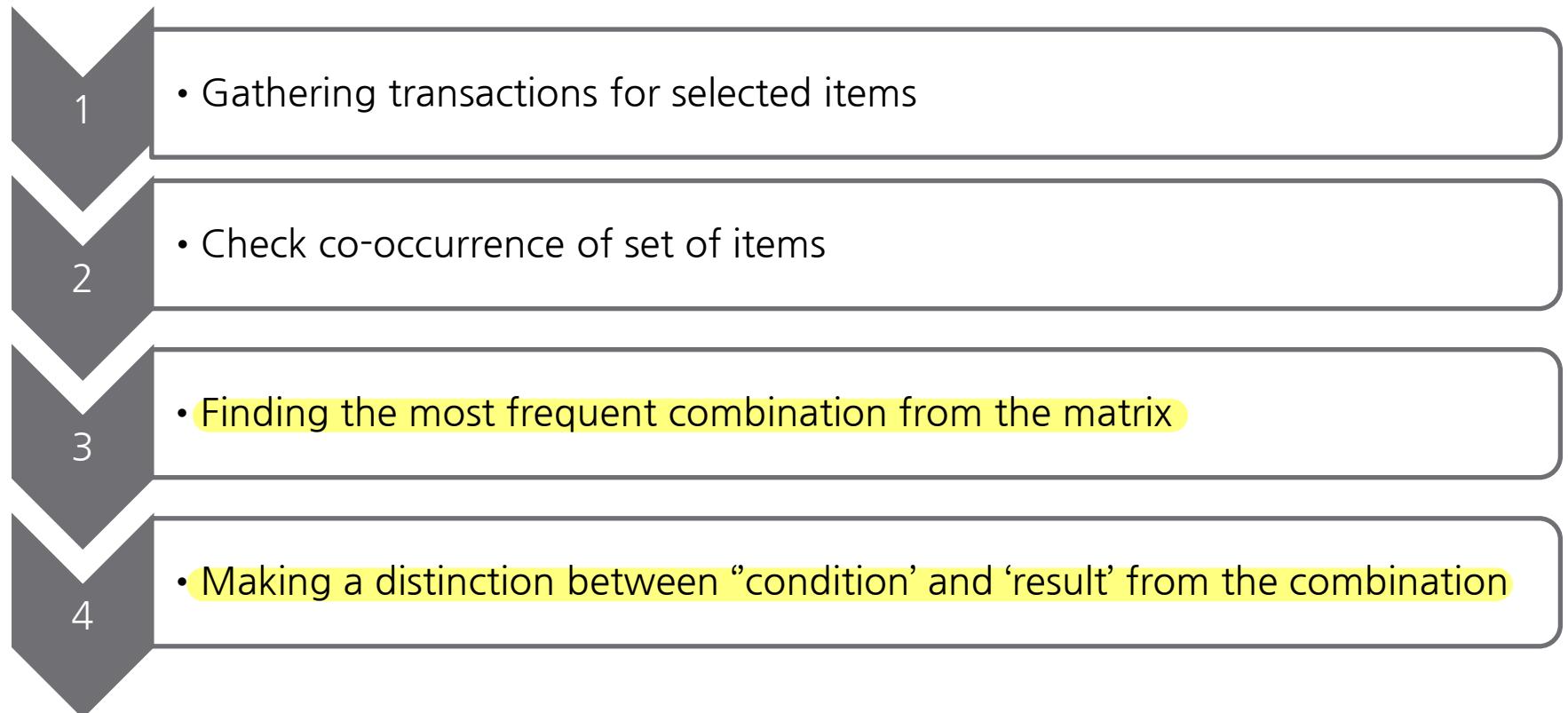
Datetime	Customer	Items
2015-07-15 14:03	1	orange juice, banana
2015-07-15 16:20	2	orange juice, milk
2015-07-16 10:14	3	detergent, banana, orange juice
2015-07-25 19:34	2	milk, bread, soda
2015-07-29 09:41	4	detergent, window cleaner
2015-08-01 20:55	1	bread, milk

- Find pair of items that is more likely to be purchased together based on transactions
 - Banana and orange juice are more likely to be purchased together
 - Milk and bread are more likely to be purchased together

Association Rules

- Association rules obtained from transactions are like
“If item A, then item B”
 - ▣ Rules are defined from co-occurrence of items in the same market basket
- Three types of rules
 - ▣ **Useful:** contains high quality, actionable information
 - On Thursday, customer who purchase diapers are likely to purchase beer
 - ▣ **Trivial:** already known by anyone familiar with the business
 - Customers purchasing paint buy pain brushes
 - ▣ **Inexplicable:** new but no explanation about customer behavior
 - When a new hardware store opens, one of the most commonly sold items is toilet rings

General Process for Finding Rules



Performance Measure for Rules

Rule: If ‘condition’, then ‘result’

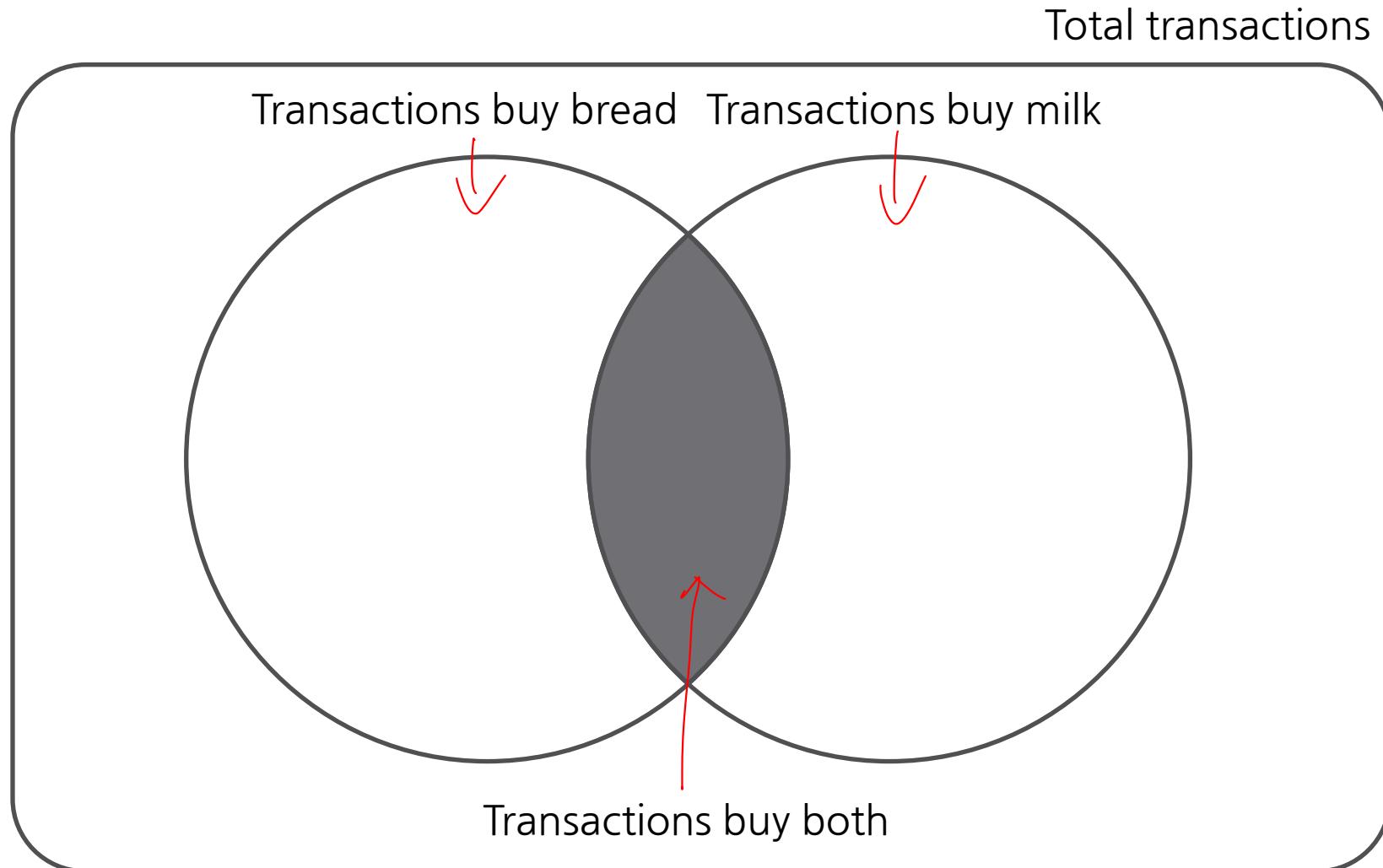
- Support
 - ▣ How many transactions that contain ‘condition(X)’ and ‘result(Y)’ simultaneously

$$\text{Supp}(X \Rightarrow Y) = \text{Supp}(X \cup Y) = \frac{\# \text{ of transactions that include both condition and result}}{\# \text{ of total transactions}}$$

- Confidence
 - ▣ How many transactions that contain ‘condition(X)’ and ‘result(Y)’ among transactions including ‘condition’

$$\text{Conf}(X \Rightarrow Y) = P(Y|X) = \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X)} = \frac{\# \text{ of transactions that include both condition and result}}{\# \text{ of transactions that include condition}}$$

Performance Measure for Rules



Performance Measure for Rules

- Low support
 - ▣ This rule rarely happens → not interesting
- High support, but low confidence *조건 결과 모두 자주 관찰*.
 - ▣ Both 'condition' and 'result' are quite often observed, but comparing with the number of transactions that include condition are much more
 - ▣ The reason that support of the rule is high may be that the number of transactions that include condition is high *조건 포함 거래 중 결과 함께 나타나는 경우가 많다.*
- High support and high confidence
 - ▣ This rule is significant rule
 - ▣ However, high support and high confidence do not guarantee usefulness of the rule

Example: Association Rule

- Example rules from given transactions

TID	Items
1	bread, milk, butter
2	bread, butter
3	bread, juice, butter
4	bread, beer
5	beer, juice

total
5

- If bread, then butter ($\text{bread} \Rightarrow \text{butter}$)

$$\text{Supp}(\text{bread} \cup \text{butter}) = \frac{3}{5}$$
$$\text{support} = \frac{3}{5}, \text{confidence} = \frac{3}{4} = \frac{\frac{3}{5}}{\frac{3}{5}} = \frac{\text{Supp}(\text{bread} \cup \text{butter})}{\text{Supp}(\text{bread})}$$

- If beer, then bread ($\text{beer} \Rightarrow \text{bread}$)

$$\text{support} = \frac{1}{5}, \text{confidence} = \frac{1}{2} = \frac{\frac{1}{5}}{\frac{2}{5}} = \frac{\text{Supp}(\text{beer} \cup \text{bread})}{\text{Supp}(\text{beer})}$$

Performance Measure for Rules

Lift ≥ 0 . non-negative

- Lift or improvement 규칙이 예상한 것보다 얼마나 잘 예측하는가?
 - How much better a rule is at predicting the result than just guessing the result at random

$$\text{lift}(X \Rightarrow Y) = \frac{P(Y|X)}{P(Y)} = \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X) \times \text{Supp}(Y)}$$
$$= \frac{(\# \text{ of transactions that include both condition and result}) \times (\# \text{ of transactions})}{(\# \text{ of transactions that include condition})(\# \text{ of transactions that include result})}$$

Improvement	Interpretation	Example
1	Two items are independent	pepper and cookies
>1	Complementary 상보적.	Bread and butter 침자구역 > 득점구역
<1	Substitutional 대체적.	Butter and margarine 침자구역 < 득점구역

In this case, If A, then NOT B is better than
If A, then B

A는 사면 B는 안될 가능성이↑

Performance Measure for Rules

$$\text{Conv} \geq 0$$

- Conviction $X \Rightarrow Y$ 를 미리 나타낼 확률. $\text{H}(X)$. $P(X), P(Y)$ 모두 사용. Y 뿐만 아니라 $\sim Y$ 도 사용.

- Conviction measures the implication strength of the rule from statistical independence

$$\text{conv}(X \Rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)} = \frac{P(X) \times P(\sim Y)}{P(X \cup \sim Y)} = \frac{\frac{P(\sim Y)}{P(X \cup \sim Y)}}{\frac{P(X)}{P(X \cup \sim Y)}} = \frac{P(\sim Y)}{P(\sim Y | X)}$$

- $P(\sim Y)$ is the probability that Y does not appear in a transaction
 - Conviction compares the probability that X appears without Y if they were dependent with the actual frequency of the appearance of X without Y
 - Unlike confidence, conviction factors in both $P(X)$ and $P(Y)$ and always has a value 1 when the relevant items are completely unrelated.
 - In contrast to lift, conviction is directed measure because it also uses the information of the absence of the consequent

Question

- Calculate performance measures of the rule

Rule1: $a \Rightarrow b$ $\text{Supp} = \frac{2}{5}$ $\text{Conf} = \frac{2}{2}$ $\text{lift} = \frac{2/5}{2/5 \times 3/5}$ conv

Rule2: $e \Rightarrow f$

Rule3: b and $c \Rightarrow g$

TID	Items
1	b, c, g
2	a, b, d, e, f
3	a, b, c, g
4	b, c, e, f
5	b, c, e, f, g

- Calculate support of above three rules

$$\frac{2}{5} \quad \frac{3}{5} \quad \frac{3}{5}$$

- Calculate confidence of above three rules

$$\frac{2}{2} \quad \frac{3/3}{3} \quad \frac{3}{4}$$

- Calculate lift of above three rules

$$\frac{\frac{2}{5}}{\frac{2}{5} \times \frac{3}{5}} \quad \frac{\frac{3}{5}}{\frac{3}{5} \times \frac{3}{5}} \quad \frac{\frac{3}{5}}{\frac{3}{5} \times \frac{4}{5}}$$

- Calculate conviction of above three rules

$$\frac{1 - \text{Supp}(y)}{1 - \text{Conf}(\text{rule})} = \frac{?}{0} = \inf \quad \frac{1 - \frac{3}{5}}{1 - 1} = \inf$$

$$\frac{1 - \frac{3}{5}}{1 - \frac{3}{4}} = \frac{\frac{2}{5}}{\frac{1}{4}}$$

Pros and Cons of Market Basket Analysis

Pros

- Produces understandable and clear results (association rules)
- Handle transactions themselves
- Computational method is simple to implement and understand

Cons

- Require much more computation resource as the problem size grows
- Sometimes require to utilize the taxonomy for mining better rules and reducing complexity
- Discount rare items

Apriori Algorithm

Practical Issues on Market Basket Analysis

- Exponential growth on distinct combinations as the number of items increases
 - ▣ If 100 items are sold in the store, the number of combinations with 3 items

$$C(100,3) = \frac{100!}{3!97!} = \frac{100 \times 99 \times 98}{3 \times 2} = 161,700$$

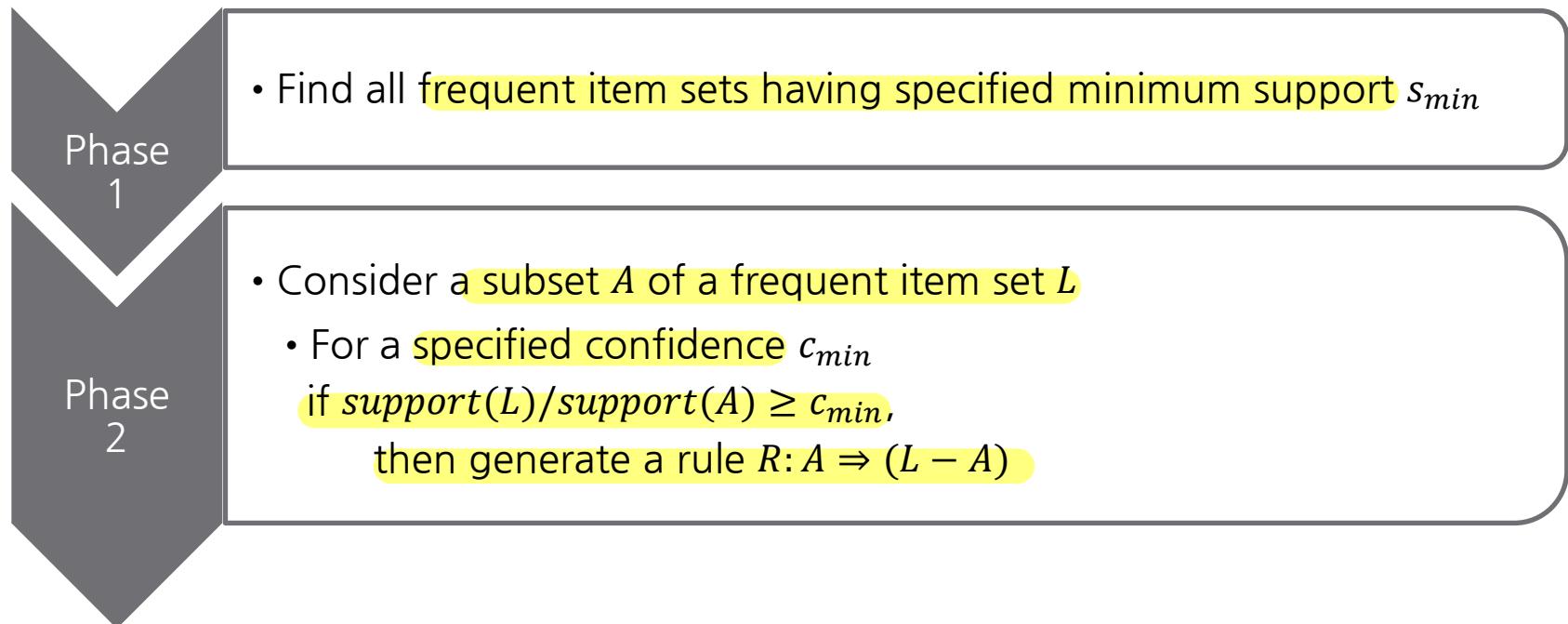
- Methods to solve rapid growth on problem size
 - ▣ Use the taxonomy: generalize items that can meet criterion
 - Vanilla ice cream ∈ Ice cream ∈ Frozen food ∈ Food
 - When there are too many items to handle, use higher level of category instead to reduce combinations
 - ▣ Use pruning: throw out item or combination of items that do not meet criterion
 - Minimum support pruning is the most common method

Apriori Algorithm

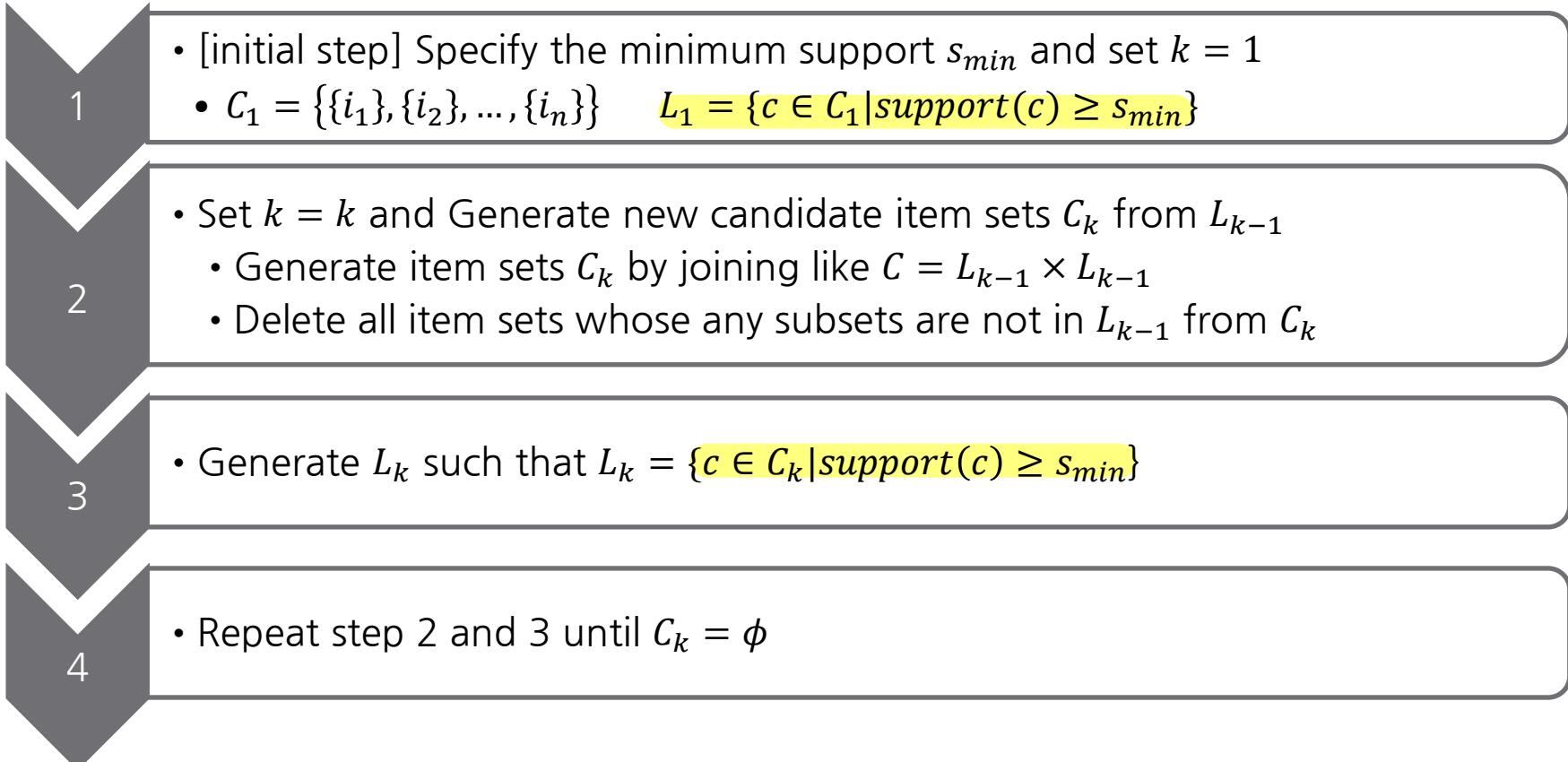
- Apriori is the algorithm to mine rules from transactions
 - ▣ Key idea is that any subsets of a frequent item set are also frequent item sets

→ Satisfy minimum support criteria

{1,2,3} is frequent item set $\Rightarrow \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}$ are frequent item set

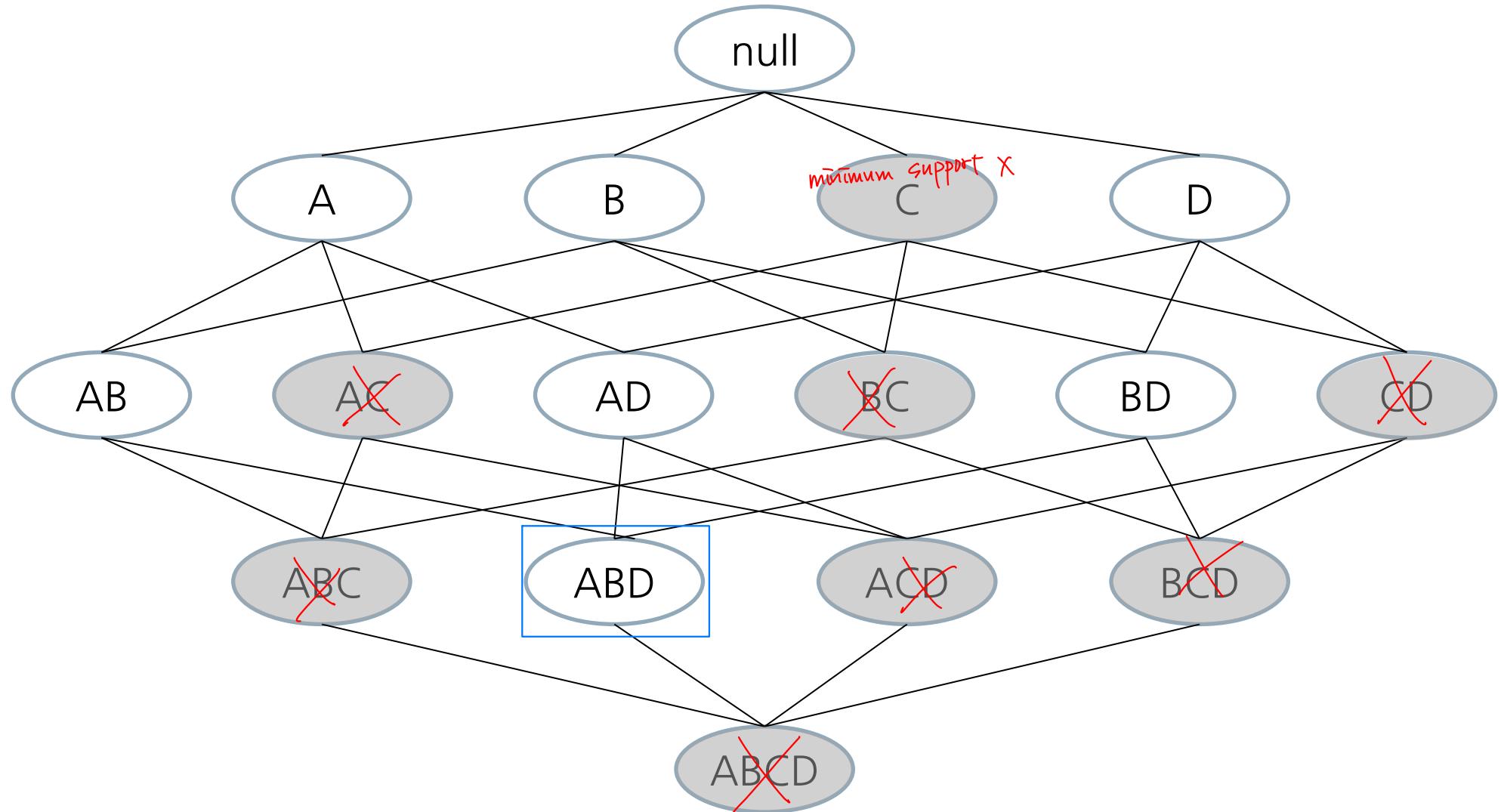


Apriori Algorithm - Phase 1

- 
- 1 • [initial step] Specify the minimum support s_{min} and set $k = 1$
• $C_1 = \{\{i_1\}, \{i_2\}, \dots, \{i_n\}\}$ $L_1 = \{c \in C_1 \mid \text{support}(c) \geq s_{min}\}$
 - 2 • Set $k = k$ and Generate new candidate item sets C_k from L_{k-1}
 - Generate item sets C_k by joining like $C = L_{k-1} \times L_{k-1}$
 - Delete all item sets whose any subsets are not in L_{k-1} from C_k
 - 3 • Generate L_k such that $L_k = \{c \in C_k \mid \text{support}(c) \geq s_{min}\}$
 - 4 • Repeat step 2 and 3 until $C_k = \emptyset$

Apriori Algorithm - Phase 1

- Key idea of phase 1



Example: Apriori Algorithm - Phase 1

- Generate C_k and L_k
 - ▣ Set $s_{min}=0.4$

$C_1 = \{\{a\}, \{b\}, \{c\}, \cancel{\{d\}}, \{e\}, \{f\}, \{g\}\}$ Remove infrequent item sets
 $L_1 = \{\{a\}, \{b\}, \{c\}, \{e\}, \{f\}, \{g\}\}$ Generate item sets by joining
 $C_2 = \{\{a, b\}, \{a, \cancel{c}\}, \{a, \cancel{e}\}, \{a, \cancel{f}\}, \cancel{\{a, g\}}, \{b, c\}, \{b, e\}, \{b, f\}, \{b, g\}, \{c, e\}, \{c, f\}, \{c, g\}, \{e, f\}, \cancel{\{e, g\}}, \cancel{\{f, g\}}\}$
 $L_2 = \{\{a, b\}, \{b, c\}, \{b, e\}, \{b, f\}, \{b, g\}, \{c, e\}, \{c, f\}, \{c, g\}, \{e, f\}\}$ ☆☆
 $C_3 = \{\{b, c, e\}, \{b, c, f\}, \{b, c, g\}, \{b, e, f\}, \{c, e, f\}\}$
 $L_3 = \{\{b, c, e\}, \{b, c, f\}, \{b, c, g\}, \{b, e, f\}, \{c, e, f\}\}$

 $\{a, b, c\}$ is removed from C_3 because
 $\{a, c\}$ does not belong to L_2

 $C_4 = \{\{b, c, e, f\}\}$
 $L_4 = \{\{b, c, e, f\}\}$

TID	Items
1	b, c, g
2	a, b, d, e, f
3	a, b, c, g
4	b, c, e, f
5	b, c, e, f, g

Question

- Generate C_k and L_k
 - Set $s_{min}=0.4$

TID	Items
1	bread, milk, butter
2	bread, butter
3	bread, juice, butter
4	bread, beer
5	beer, juice

- 1) Generate C_1 and L_1

C_1 bread / milk / butter / juice / beer $\leftarrow C_1$
 L_1 bread / / butter / juice / beer

- 2) Generate C_2 and L_2

C_2 bread, butter / bread, juice / bread, beer / $\leftarrow C_2$
 butte, juice / butter, beer / juice, beer
 L_2 bread, butter

Example: Apriori Algorithm - Phase 2

- Rule generation
 - ▣ Candidate frequent item set $L = \{b, c, g\}$
 - ▣ Rules having 1 item in result

$$R_1: \{b, c\} \Rightarrow \{g\}$$

$$R_2: \{b, g\} \Rightarrow \{c\}$$

$$R_3: \{c, g\} \Rightarrow \{b\}$$

TID	Items
1	b, c, g
2	a, b, d, e, f
3	a, b, c, g
4	b, c, e, f
5	b, c, e, f, g

Rule	Support($\{b, c, g\}$)	Support(condition)	Confidence
$R_1: \{b, c\} \Rightarrow \{g\}$	0.6	/	0.6/0.8=0.75
$R_2: \{b, g\} \Rightarrow \{c\}$	0.6	/	0.6/0.6=1
$R_3: \{c, g\} \Rightarrow \{b\}$	0.6	/	0.6/0.6=1

How to Efficiently Generate Rules

$$\text{Supp}(X \cup Y) \leq \text{Supp}(X) \text{ or } \text{Supp}(Y)$$

- Confidence is not anti-monotonic
 - $\text{confidence}(ABC \Rightarrow D)$ can be larger or smaller than $\text{confidence}(AB \Rightarrow D)$
- However, confidence of rules generated from the same item set is anti-monotonic with respect to the number of items in result
 - All conditions should be subsets of the largest condition

$$\text{confidence}(ABC \Rightarrow D) \geq \text{confidence}(AB \Rightarrow CD) \geq \text{confidence}(A \Rightarrow BCD)$$

filter out



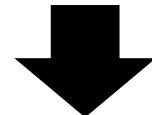
result 항목 수 증가 \Rightarrow anti-monotonic
conf.

If the rule $ABC \Rightarrow D$ has lower confidence than certain value

Then, $BC \Rightarrow AD$, $AC \Rightarrow BD$, $BD \Rightarrow CD$, $C \Rightarrow ABD$, $B \Rightarrow ACD$, $A \Rightarrow BCD$
have lower confidence than certain value

Set Up s_{min}

- If s_{min} is too high, we can miss item sets containing interesting rare items (e.g., expensive products)
- If s_{min} is too low, the number of frequent item sets increases and computational cost becomes expensive



Always, it is really hard to set “appropriate” parameter

It is not effective to use a single s_{min}

Frequent Pattern Growth Algorithm

Frequent Pattern Growth Algorithm

- Why Frequent Pattern Growth (FP-Growth) ?
 - ▣ The Apriori algorithm is widely used for association rule mining, but it has significant drawbacks
 - It generates a large number of candidate itemsets, leading to high computational costs
 - It requires multiple scans of the database, which can be slow for large datasets
- FP-Growth solves these problems by
 - ▣ Using a tree-based data structure (FP-tree) to store frequent items compactly
 - ▣ Avoiding candidate generation, reducing computational complexity

Frequent Pattern Growth Algorithm

- The **FP-Growth algorithm** consists of two main steps
 1. Building the FP-Tree
 2. Mining frequent patterns from the FP-Tree

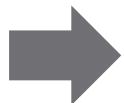
Frequent Pattern Growth Algorithm

- Step 1: Building the FP-Tree
 1. Scan the dataset once to find the frequency of all individual items
 2. Filter out infrequent items based on a minimum support threshold
 3. Sort the frequent items in descending order of support (ties are resolved arbitrarily)
 4. Construct the FP-tree
 - The root is a null node
 - For each transaction
 - Follow the path in the tree corresponding to its frequent items
 - If a node already exists for an item, increment its count
 - Otherwise, create a new node

Frequent Pattern Growth Algorithm

- Step 1-1: Scan the dataset once to find the frequency of all individual items

TID	Items
1	f, a, c, d, g, i, m, p
2	a, b, c, f, l, m, o
3	b, f, h, j, o
4	b, c, k, s, p
5	a, f, c, e, l, p, m, n

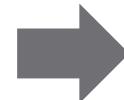


Item	Frequency	Item	Frequency
a	3	j	1
b	3	k	1
c	4	l	2
d	1	m	3
e	1	n	1
f	4	o	2
g	1	p	3
h	1	s	1
i	1		

Frequent Pattern Growth Algorithm

- Step 1-2: Filter out infrequent items based on a minimum support threshold
 - Let $s_{min} = 3$

Item	Frequency	Item	Frequency
a	3	j	1
b	3	k	1
c	4	l	2
d	1	m	3
e	1	n	1
f	4	o	2
g	1	p	3
h	1	s	1
i	1		

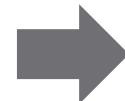


Item	Frequency
a	3
b	3
c	4
f	4
m	3
p	3

Frequent Pattern Growth Algorithm

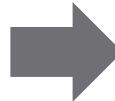
- Step 1-3: Sort the frequent items in descending order of support

Item	Frequency
a	3
b	3
c	4
f	4
m	3
p	3



Item	Frequency
f	4
c	4
a	3
b	3
m	3
p	3

TID	Items
1	f, a, c, d, g, i, m, p
2	a, b, c, f, l, m, o
3	b, f, h, j, o
4	b, c, k, s, p
5	a, f, c, e, l, p, m, n

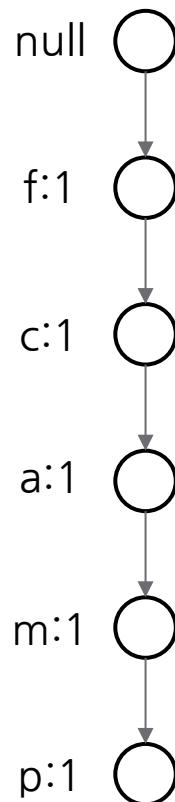


TID	Items
1	f, c, a, m, p
2	f, c, a, b, m
3	f, b
4	c, b, p
5	f, c, a, m, p

*filter out
&
sort*

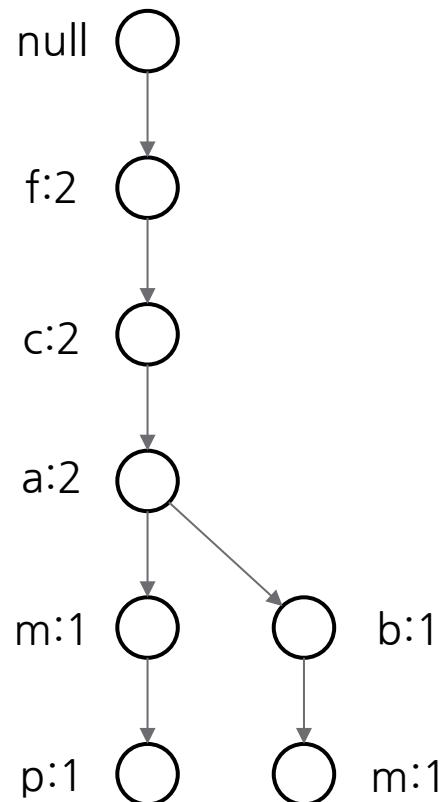
Frequent Pattern Growth Algorithm

- Step 1-4: Construct the FP-tree
 - Reading TID1= {f, c, a, m, p}



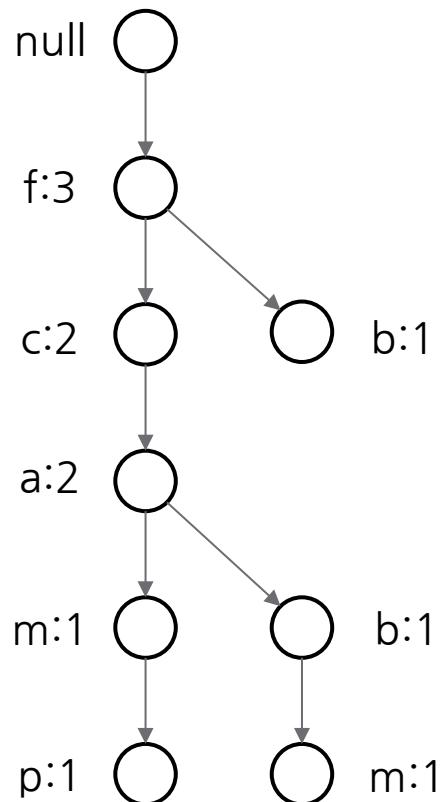
Frequent Pattern Growth Algorithm

- Step 1-4: Construct the FP-tree
 - Reading TID2= {f, c, a, b, m}



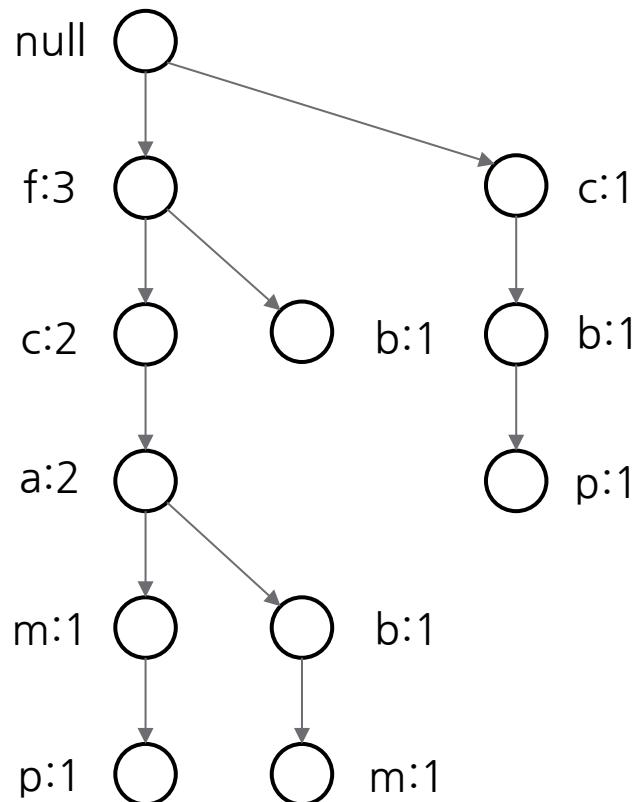
Frequent Pattern Growth Algorithm

- Step 1-4: Construct the FP-tree
 - Reading TID3= {f, b}



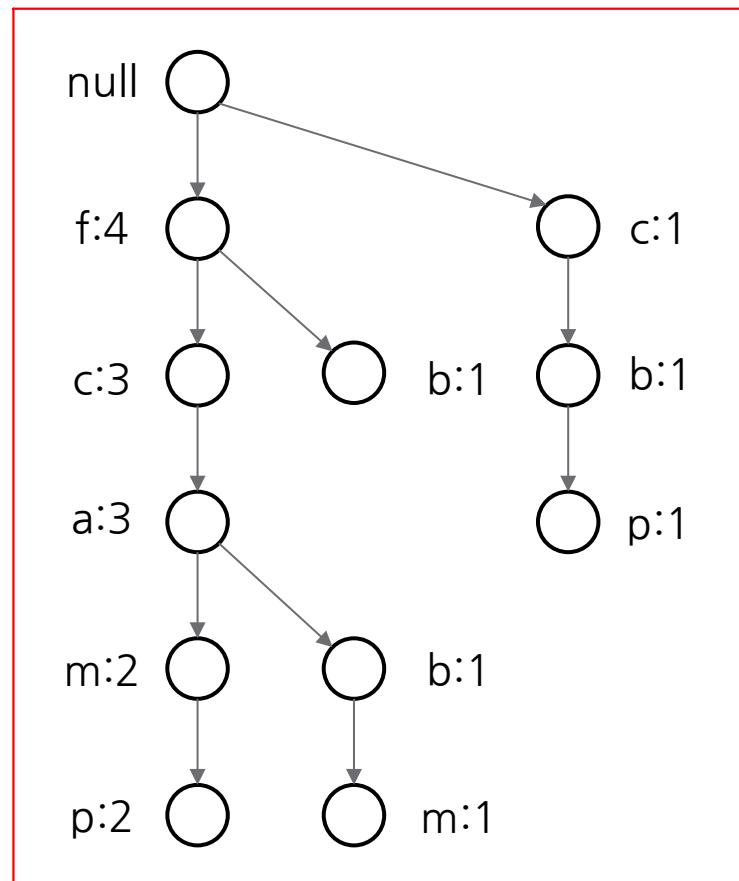
Frequent Pattern Growth Algorithm

- Step 1-4: Construct the FP-tree
 - Reading TID4= {c, b, p}



Frequent Pattern Growth Algorithm

- Step 1-4: Construct the FP-tree
- Reading TID5= {f, c, a, m, p}

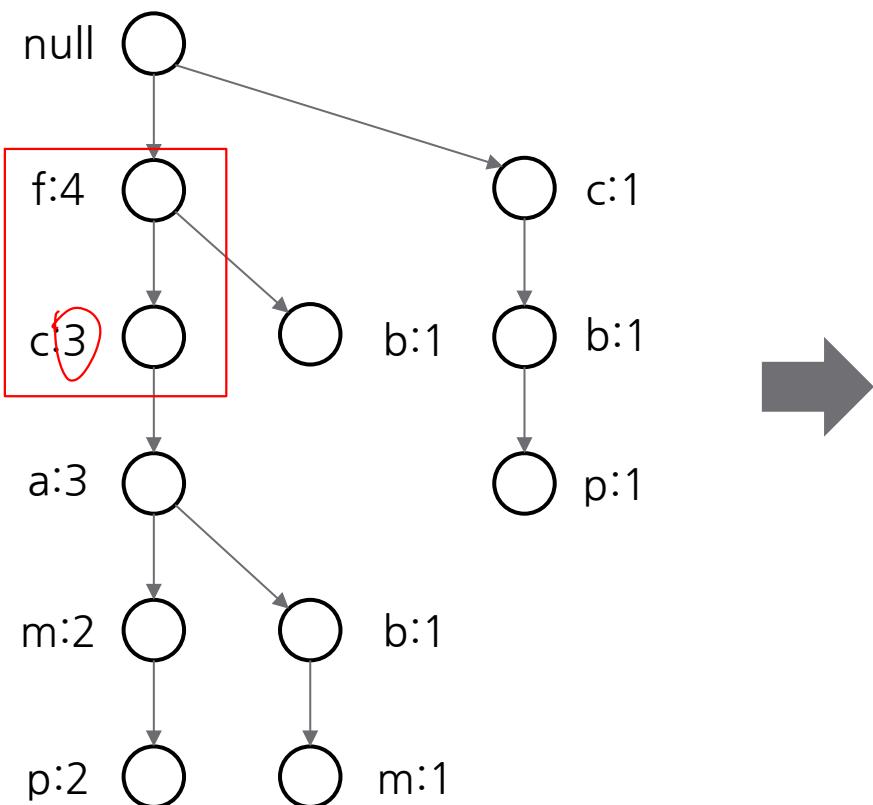


Frequent Pattern Growth Algorithm

- Step 2: Recursive Pattern Growth
 - 1. For each frequent item, compute conditional pattern bases
 - 2. For each frequent item, construct a conditional FP-tree
 - 3. Recursively mine the conditional FP-tree to find frequent patterns

Frequent Pattern Growth Algorithm

- Step 2-1: For each frequent item, compute conditional pattern bases



{ Parents : frequency }

Item	Conditional pattern base
f	{}
c	{f : 1}
a	{f, c : 3}
b	{f, c, a : 1}, {f : 1}, {c : 1}
m	{f, c, a : 2}, {f, c, a, b : 1}
p	{f, c, a, m : 2}, {c, b : 1}

Frequent Pattern Growth Algorithm

- Step 2-2: For each frequent item, construct a conditional FP-tree

$$S_m = ?.$$

Item	Conditional pattern base	Conditional FP-tree
f	{}	{}
c	{{f : 3}}	{f:3}
a	{{f, c : 3}}	{f, c:3}
b	{{f, c, a : 1}, {f : 1}, {c : 1}}	{}
m	{{f, c, a : 2}, {f, c, a, b : 1}}	{f, c, a:3}
p	{{f, c, a, m : 2}, {c, b : 1}}	{c:3}

Frequent Pattern Growth Algorithm

- Step 2-3: Recursively mine the conditional FP-tree to find frequent patterns

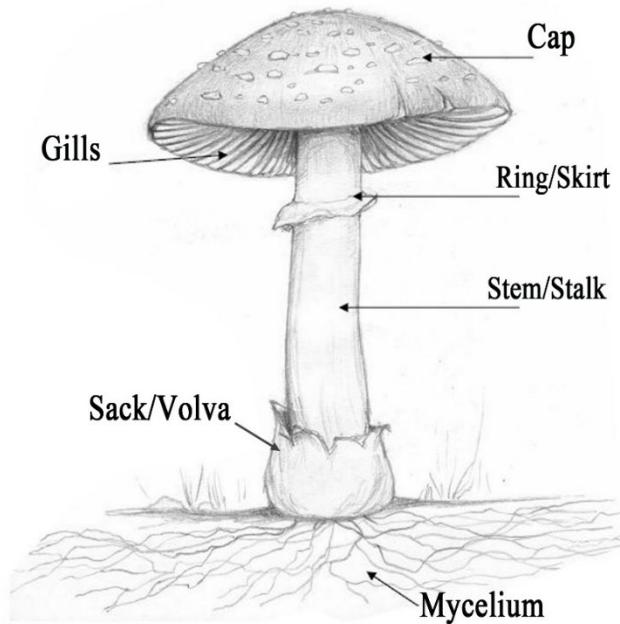
멱집합 - (power set) $2^n - 1$

Item	Conditional pattern base	Conditional FP-tree	Frequent patterns
f	{}	{}	{}
c	{f : 3}	{f:3}	$2^1 - 1$ {<f,c:3>}
a	{f, c : 3}	{f, c:3}	$2^2 - 1$ {<f, a : 3>, <c, a : 3>, <f, c, a:3>}
b	{f, c, a : 1}, {f : 1}, {c : 1}	{}	{}
m	{f, c, a : 2}, {f, c, a, b : 1}	{f, c, a:3}	$2^3 - 1$ {<f, m : 3>, <c, m : 3>, <a, m : 3>, <f, c, m : 3>, <f, a, m : 3>, <c, a, m : 3>, <f, c, a, m:3>}
p	{f, c, a, m : 2}, {c, b : 1}	{c:3}	$2^1 - 1$ {<c,p:3>}

Applications

Applications of Association Rules

- Another example of association rule mining
 - ▣ Ex.) Mushroom data
 - Classify edibility based on the descriptions of mushroom
 - <https://archive.ics.uci.edu/ml/datasets/mushroom>



- ▣ By applying association rule mining, it can be possible to characterize poisoned mushrooms and edible mushrooms

Applications of Association Rules

- Mushroom dataset
 - ▣ Attribute Information: (classes: edible=e, poisonous=p)
 - 1) cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
 - 2) cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s
 - 3) cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
 - 4) bruises?: bruises=t, no=f
 - 5) odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
 - 6) gill-attachment: attached=a, descending=d, free=f, notched=n
 - 7) gill-spacing: close=c, crowded=w, distant=d
 - 8) gill-size: broad=b, narrow=n
 - 9) gill-color: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
 - 10) stalk-shape: enlarging=e, tapering=t
 - 11) stalk-root: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?

Applications of Association Rules

- Mushroom dataset
 - ▣ Attribute Information: (classes: edible=e, poisonous=p)
 - 12) stalk-surface-above-ring: fibrous=f, scaly=y, silky=k, smooth=s
 - 13) stalk-surface-below-ring: fibrous=f, scaly=y, silky=k, smooth=s
 - 14) stalk-color-above-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
 - 15) stalk-color-below-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
 - 16) veil-type: partial=p, universal=u
 - 17) veil-color: brown=n, orange=o, white=w, yellow=y
 - 18) ring-number: none=n, one=o, two=t
 - 19) ring-type: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
 - 20) spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
 - 21) population: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
 - 22) habitat: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d