

9차시 | 빅데이터 기초분석(1)

이긍희 교수



빅데이터 기초분석(1)

1. 데이터분석 개요
2. 데이터분석 도구
3. 데이터 마트
4. 데이터 탐색

1. 데이터분석의 개요

■ 데이터분석의 필요성

“우리는 정보의 홍수에 있지만 지식에 굽주려 있다.” John Naisbitt

- ✓ 빅데이터의 시대 : 데이터 수집과 저장이 용이하고 데이터분석기술의 대중화
 - ▶ 데이터분석의 자동화가 필요 : 머신러닝

- ✓ 주어진 데이터로부터 합리적 의사결정(예측)을 시도 : 통계학, 확률(불확실성)

- ▶ 스몰데이터 분석의 원리와 빅데이터 분석원리는 근본적으로 동일

1. 데이터분석의 개요

■ 데이터분석의 일반적 구분 : 지도(감독) 학습, 비지도(자율) 학습과 강화학습으로 구분

- ✓ **지도학습** : 입력 데이터(설명변수)로부터 출력 데이터(반응변수)을 모델링
 - › 스팸메일 분류, 이미지 인식, 글씨 인식 등
 - › 주가예측, 나이예측, 질병예측 등
- ✓ **비지도학습** : 주어진 입력 데이터로부터 의미 있는 결과를 도출
 - › 인간의 전형적 학습과정
 - › 군집 발견, 추천, 시장바구니 분석
- ✓ **강화학습** : 보상과 벌칙이라는 규칙 하에 학습

1. 데이터분석의 개요

■ 데이터분석(가트너)

- ✓ 서술형 분석 : 주어진 상황 속에서 어떤 일이 벌어졌는지 설명
→ 데이터 시각화, 군집화 등
- ✓ 진단형 분석 : 어떤 이유로 그 현상이 발생했는지 설명
→ 분류, 군집화, 의사결정나무
- ✓ 예측형 분석 : 미래에 어떤 일이 일어날지 설명
→ 회귀분석, 인공신경망, 의사결정나무, 랜덤 포레스트, 시뮬레이션
- ✓ 처방형 분석 : 결과를 최적화하는 방안을 제안
→ 의사결정나무, 선형, 비선형 프로그래밍,
몬테카를로 시뮬레이션, 게임 이론

1. 데이터분석의 개요

■ 데이터분석 프로세스 개요

- ✓ **요건정의** : 데이터분석을 통해 풀기 원하는 사업문제를 찾고 계획
 - › 분석요건 도출 : 비즈니스로부터 사업 문제를 정의
 - › 수행방안 설계 : 탐색적 분석을 통해 어떤 분석을 수행할지 설계
 - › 요건 확정 : 관련 부서와 협의해 최종 요건을 확정
- ✓ **데이터 수집과 정제** : 요건 정의에 따라 데이터를 수집
 - › 마트 설계와 구축
 - ➔ 데이터베이스, 웹 서비스, 외부데이터 등으로부터 자료를 수집 : ETL과정 포함

1. 데이터분석의 개요

■ 데이터분석 프로세스 개요

- ✓ 모델링 : 분석기법을 적용해 모델(수식 또는 알고리즘)을 개발
 - › 탐색적 분석과 유의 변수 도출 : EDA(탐색적 데이터 분석)
→ 누락 데이터와 특이항(outlier)를 찾아서 처리하고 데이터를 일부 변환
 - › 모델링 : 여러 분석방법중 최적의 방법을 선택하여 구체적 모델을 작성
 - › 모델링 성능평가 : 정확도, 정밀도 등으로 판단
 - › 검증 및 테스트 : 분석용 데이터를 훈련용과 검증용으로 분리 후 검증
- ✓ 모델 적용 : 분석결과를 업무 프로세스에 완전히 통합해 운영
 - › 운영시스템 적용과 자동화
 - › 주기적 리모델링 : 시간이 지나면서 모델의 유용성 저하

1. 데이터분석의 개요

■ 데이터분석 방법

- ✓ 탐색적 분석 : 다양한 차원과 값을 조합해가며 특이한 점이나 의미 있는 사실을 도출 → 시각화 등
- ✓ 모델을 이용한 분석
 - › 분류·군집 : 데이터를 여러 그룹으로 분류
 - ➔ 지도학습 : 의사결정나무, 나이브 베이즈, 서포트벡터머신
 - ➔ 자율학습 : K-means 알고리즘
 - › 예측 : 회귀분석, 뉴럴네트워크, 회귀나무
 - › 시뮬레이션 : 선형 프로그래밍, 몬테카롤로 시뮬레이션
 - › 비정형데이터분석 : 텍스트마이닝, 이미지, 비디오 분석
 - › 모델의 결합 : 앙상블 모델

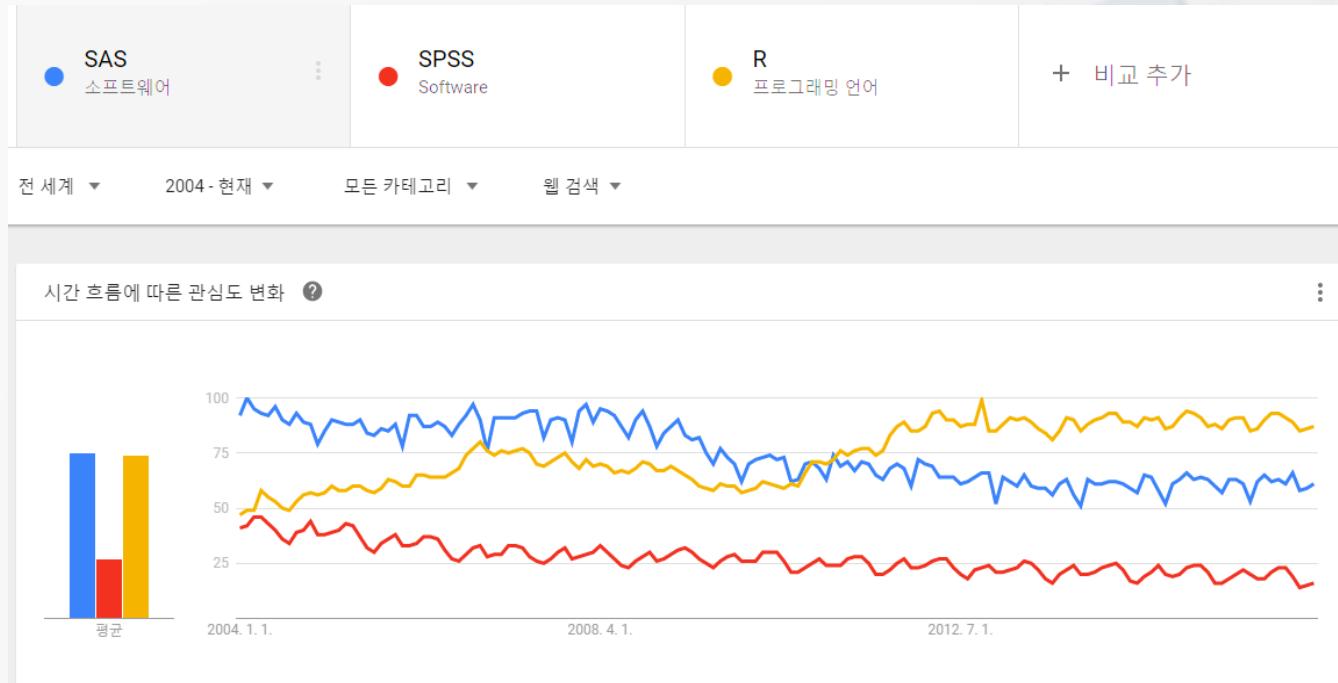
1. 데이터분석의 개요

■ 기초 지식과 소양 : 해당분야의 지식, 통계학과 프로그래밍

- ✓ 주로 이용되는 분석 프로그램 : R, Python
 - 오픈소스, 설치용량이 적고, 기술반영이 빠르고 커뮤니티 활성화
- ✓ 기존 통계 프로그램 : SAS, SPSS
 - 유료, 대용량, 최신기술반영이 느리고 커뮤니티 비활성화

2. 데이터 분석 도구

■ 데이터분석의 도구



2. 데이터 분석 도구

■ R과 Python 비교



[Home]

Download

CRAN

R Project

About R

Logo

Contributors

What's New?

Reporting Bugs

Development Site

Conferences

Search

R Foundation

Foundation

Board

Members

Donors

Donate

Help With R

Getting Help

Documentation

Manuals

FAQs

The R Journal

Books

Certification

Other

Links

Bioconductor

Related Projects

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and Mac OS. To [download R](#), please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to [frequently asked questions](#) before you send an email.

News

- [The R Journal Volume 8/1](#) is available.
- The [useR! 2017](#) conference will take place in Brussels, July 4 - 7, 2017, and details will be appear here in due course.
- [R version 3.3.1 \(Bug in Your Hair\)](#) has been released on Tuesday 2016-06-21.
- [R version 3.2.5 \(Very, Very Secure Dishes\)](#) has been released on 2016-04-14. This is a rebadging of the quick-fix release 3.2.4-revised.
- [Notice XQuartz users \(Mac OS X\)](#) A security issue has been detected with the Sparkle update mechanism used by XQuartz. Avoid updating over insecure channels.
- The [R Logo](#) is available for download in high-resolution PNG or SVG formats.
- [useR! 2016](#), has taken place at Stanford University, CA, USA, June 27 - June 30, 2016.
- [The R Journal Volume 7/2](#) is available.
- [R version 3.2.3 \(Wooden Christmas-Tree\)](#) has been released on 2015-12-10.
- [R version 3.1.3 \(Smooth Sidewalk\)](#) has been released on 2015-03-09.



2. 데이터 분석 도구

■ R과 Python 비교

The screenshot shows the Python.org homepage. At the top, there's a navigation bar with tabs for Python, PSF, Docs, PyPI, Jobs, and Community. Below the header, there's a search bar and links for Socialize and Sign In. The main content area features a Python code snippet for generating a Fibonacci series:

```
# 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 text about Python's extensibility and function definitions. Below the code, there are five numbered buttons (1-5). At the bottom of the main content area, there's a promotional message: "Python is a programming language that lets you work quickly and integrate systems more effectively. [» Learn More](#)".

Get Started
Whether you're new to programming or an experienced developer, it's easy to learn and use Python.
Start with our [Beginner's Guide](#)

Download
Python source code and installers are available for download for all versions! Not sure which version to use? [Check here.](#)
Latest: Python 3.5.2 - Python 2.7.12

Docs
Documentation for Python's standard library, along with tutorials and guides, are available online.
[docs.python.org](#)

Jobs
Looking for work or have a Python related position that you're trying to hire for? Our [relaunched community-run job board](#) is the place to go.
[jobs.python.org](#)

2. 데이터 분석 도구

■ R과 Python 비교

- ✓ R은 통계분석에 유용
- ✓ Python은 데이터 전처리에 유용

3. 데이터 마트

- 데이터마트 : 모델링 또는 분석을 위해 준비된 데이터
 - ✓ 특정 사용자가 관심을 가지는 데이터들을 담은 비교적 작은 데이터웨어하우스

3. 데이터 마트

■ 데이터 reshape

```
> data(airquality)
> head(airquality,10)
```

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10

3. 데이터 마트

■ 데이터 reshape

```
> aqm=melt(airquality, id=c("month","day"),na.rm=TRUE)  
> aqm
```

	month	day	variable	value
1	5	1	ozone	41.0
2	5	2	ozone	36.0
3	5	3	ozone	12.0
4	5	4	ozone	18.0
5	5	6	ozone	28.0
6	5	7	ozone	23.0
7	5	8	ozone	19.0
8	5	9	ozone	8.0
9	5	11	ozone	7.0
10	5	12	ozone	16.0

3. 데이터 마트

■ 데이터 reshape

```
> b<-cast(aqm, month~variable,mean)  
> b
```

	month	ozone	solar.r	wind	temp
1	5	23.61538	181.2963	11.622581	65.54839
2	6	29.44444	190.1667	10.266667	79.10000
3	7	59.11538	216.4839	8.941935	83.90323
4	8	59.96154	171.8571	8.793548	83.96774
5	9	31.44828	167.4333	10.180000	76.90000

3. 데이터 마트

■ SQL을 이용한 데이터 찾기

> head(iris)

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

3. 데이터 마트

■ SQL을 이용한 데이터 찾기

```
> sqldf("select * from iris limit 10")
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

3. 데이터 마트

■ SQL을 이용한 데이터 찾기

```
> sqldf("select count (*) from iris where Species like 'se%'")
```

```
count(*)  
1    50
```

3. 데이터 마트

■ 데이터 결합 : plyr

```
> data.frame(year=rep(2012:2014,each=6), count=round(runif(9, 0, 20)))
```

	year	count
1	2012	5
2	2012	7
3	2012	11
4	2012	18
5	2012	4
6	2012	18
7	2013	19
8	2013	13
9	2013	13
10	2013	5
11	2013	7

3. 데이터 마트

■ 데이터 결합 : plyr

```
> ddply(d,"year", summarise, mean.count=mean(count))
```

	year	mean.count
1	2012	10.50000
2	2013	11.33333
3	2014	14.16667

3. 데이터 마트

■ 데이터 결합 : plyr

```
> ddply(d,"year", transform, total.count=sum(count))
```

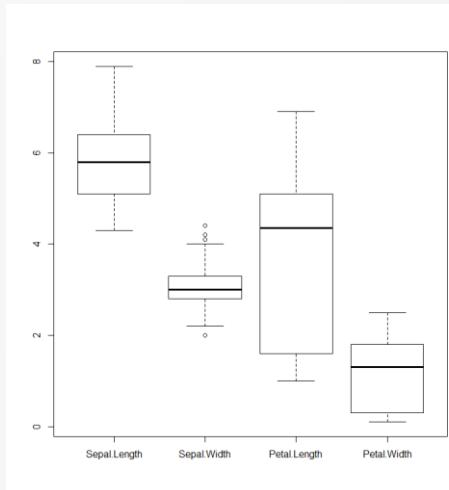
1	2012	5	63
2	2012	7	63
3	2012	11	63
4	2012	18	63
5	2012	4	63
6	2012	18	63

4. 데이터 탐색

■ 요약통계량

> summary(iris)

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
Mean :5.843	Mean :3.057	Mean :4.3758	Mean :1.199	
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	

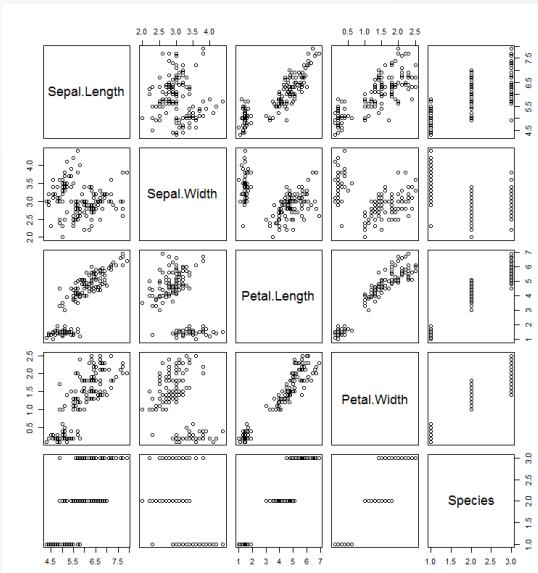


4. 데이터 탐색

■ 상관관계

> `cor(iris[1:4])`

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000



4. 데이터 탐색

- 결측값 처리 : 제거, imputation
- 이상치 탐색



강의를 마쳤습니다

수고하셨습니다.

10차시 | 빅데이터 기초분석(2)

이긍희 교수



빅데이터 기초분석(2)

1. 통계분석의 원리

2. 통계분석

2-1. 회귀분석

2-2. 다변량분석

3. 시계열분석

1. 통계분석의 원리

■ 통계학

✓ 통계학은 데이터를 통해 합리적인 결정을 하게 만드는 과학

- › 데이터를 효과적으로 수집
- › 데이터를 수치 또는 그래프로 요약
- › 데이터로부터 일정한 결론을 도출

✓ 통계학은 분석 과정에서 확률을 이용

1. 통계분석의 원리

■ 모집단과 표본

- › 모집단 : 관심(연구)의 대상이 되는 모든 개체의 집합 또는 집단
- › 표 본 : 모집단에서 추출되어 모집단을 대표하는 집단
 - 모집단에서 표본을 뽑아서 모집단을 추측

1. 통계분석의 원리

■ 데이터 생성방법

✓ 표본추출법

- › 단순랜덤추출법
- › 계통추출법
- › 집락추출법
- › 층화추출법

✓ 실험

✓ 관찰

1. 통계분석의 원리

■ 데이터의 측정

- › 명목척도 : 성, 출생지
- › 순서척도 : 만족도, 학년
- › 구간척도 : 온도, 지수
- › 비율척도 : 키, 몸무게, 시간

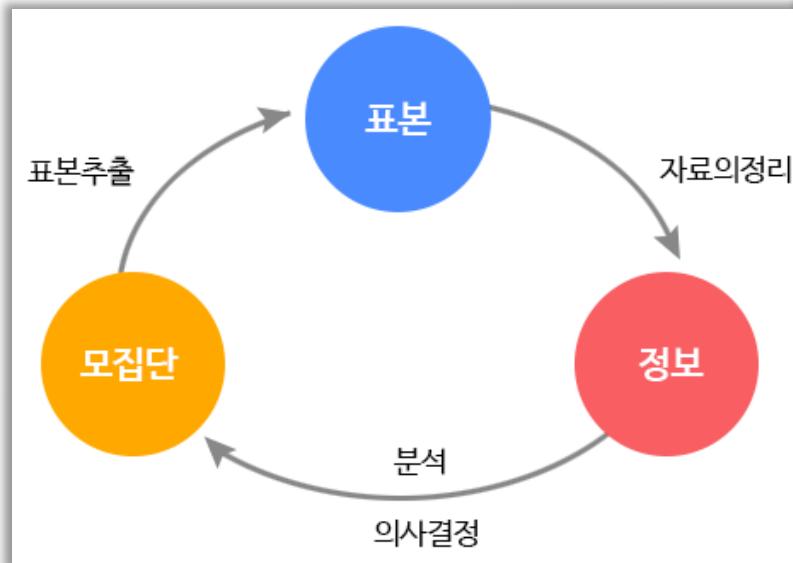
1. 통계분석의 원리

- 통계분석(statistical analysis) : 데이터를 수집하고 이를 적절한 통계 분석방법을 이용하여 요약하고 추론
 - ✓ 관심 대상 전체를 모두 조사하지 않고 일부만 조사 관측하여 전체를 파악하는 것은 통계 분석의 출발 배경

1. 통계분석의 원리

■ 통계분석(statistical analysis)

모집단과 표본의 관계



1. 통계분석의 원리

■ 통계적 추론 : 추정과 검정

- ✓ 추정 : “대상집단의 특성값(모수)이 무엇일까?”에 대한 답을 구하는 것
 - 점추정과 구간추정

1. 통계분석의 원리

■ 통계적 추론 : 추정과 가설검정

- ✓ 가설검정 : 대상집단에 대하여 특정한 가설을 설정한 후에 그 가설의 채택여부를 결정

2. 통계분석

- 회귀분석
- 다변량분석

2-1. 회귀분석

- 회귀분석(regression analysis) : 둘 또는 그 이상의 변수간의 함수관계를 통계자료를 바탕으로 파악하는 통계적 방법
 - ✓ 회귀분석의 주목적 : 변수들간의 관계를 표현하고 동 관계로부터 미래값 등을 추정
 - X : 원인이 되는 변수를 설명변수(explanatory variable) 또는 독립변수 (independent variable)
 - Y : 결과가 되는 변수를 종속변수(dependent variable) 또는 반응변수 (response variable)
 - ✓ 종속변수 Y와 설명변수 X_1, X_2, \dots, X_k 의 관측치를 $(y_i, x_{1i}, x_{2i}, \dots, x_{ki})$, $i=1, 2, \dots, n$ 이라 할 때 설명변수가 k개인 중회귀분석 모형은 다음과 같이 표현

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

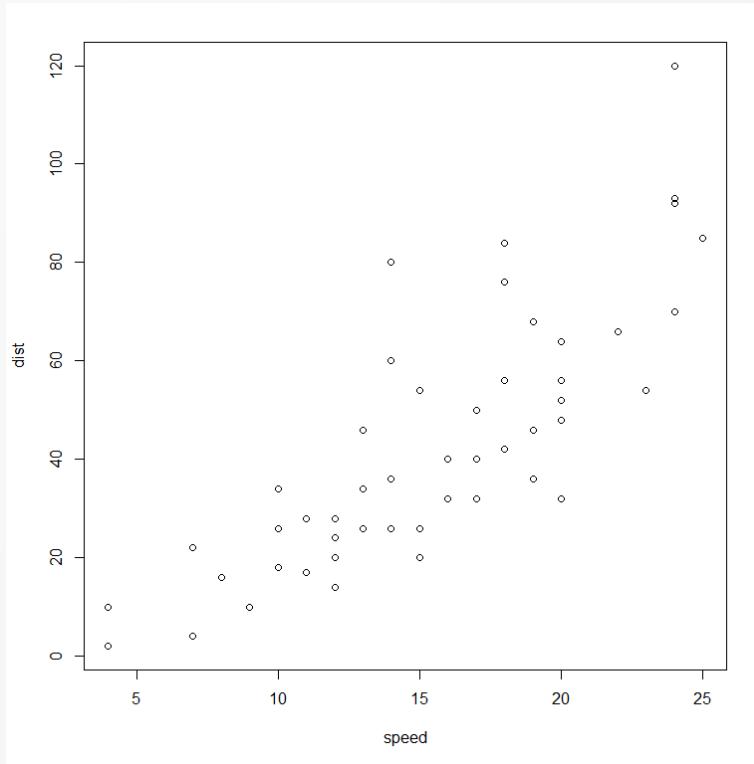
2-1. 회귀분석

■ 회귀분석의 가정

- ✓ 오차항은 서로 독립이며 평균이 0, 분산이 일정
- ✓ 설명변수는 오차항과 무관
- ✓ 설명변수간 선형관계가 존재하지 않음
- ✓ 검정을 하거나 신뢰구간을 구하려면 오차항이 정규분포를 따라야 함

2-1. 회귀분석

■ 회귀모형의 추정 : 최소제곱법, 최대가능도법



2-1. 회귀분석

■ 회귀모형의 추정

```
> summary(lm(dist~speed, data=cars))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-17.5790949	6.7584402	-2.60106	0.012319 *
speed	3.9324088	0.4155128	9.46399	0.00000000000014898 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.37959 on 48 degrees of freedom

Multiple R-squared: 0.6510794, Adjusted R-squared: 0.6438102

F-statistic: 89.56711 on 1 and 48 DF, p-value: 0.0000000000001489836

2-1. 회귀분석

■ 검토사항

- 1) 모형이 통계적으로 유의미한가?
- 2) 회귀계수들이 유의미한가?
- 3) 모형이 얼마나 설명력을 갖는가?
- 4) 모형이 데이터를 잘 적합하고 있는가?
- 5) 오차 가정을 만족시키는가?

2-1. 회귀분석

■ 회귀모형의 추정

```
> summary(lm(dist~speed, data=cars))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-17.5790949	6.7584402	-2.60106	0.012319 *
speed	3.9324088	0.4155128	9.46399	0.00000000000014898 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

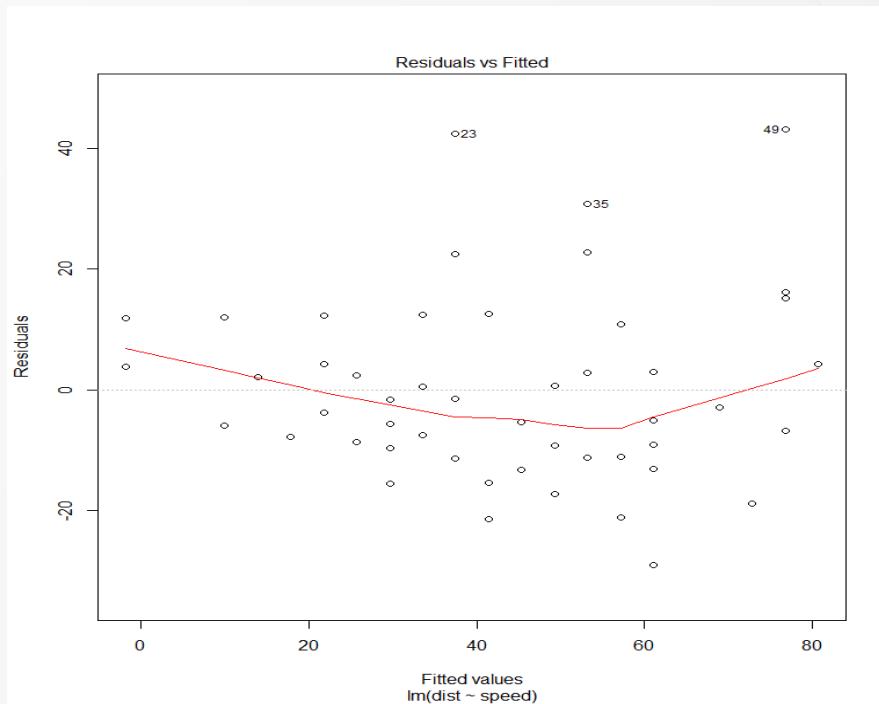
Residual standard error: 15.37959 on 48 degrees of freedom

Multiple R-squared: 0.6510794, Adjusted R-squared: 0.6438102

F-statistic: 89.56711 on 1 and 48 DF, p-value: 0.0000000000001489836

2-1. 회귀분석

■ 회귀모형 잔차의 검토



2-1. 회귀분석

■ 회귀모형의 추정

```
> ss=lm(log(dist)~log(speed), data=cars)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.7296687	0.3758457	-1.94141	0.058092 .
log(speed)	1.6023912	0.1395376	11.48358	0.00000000000000022594 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

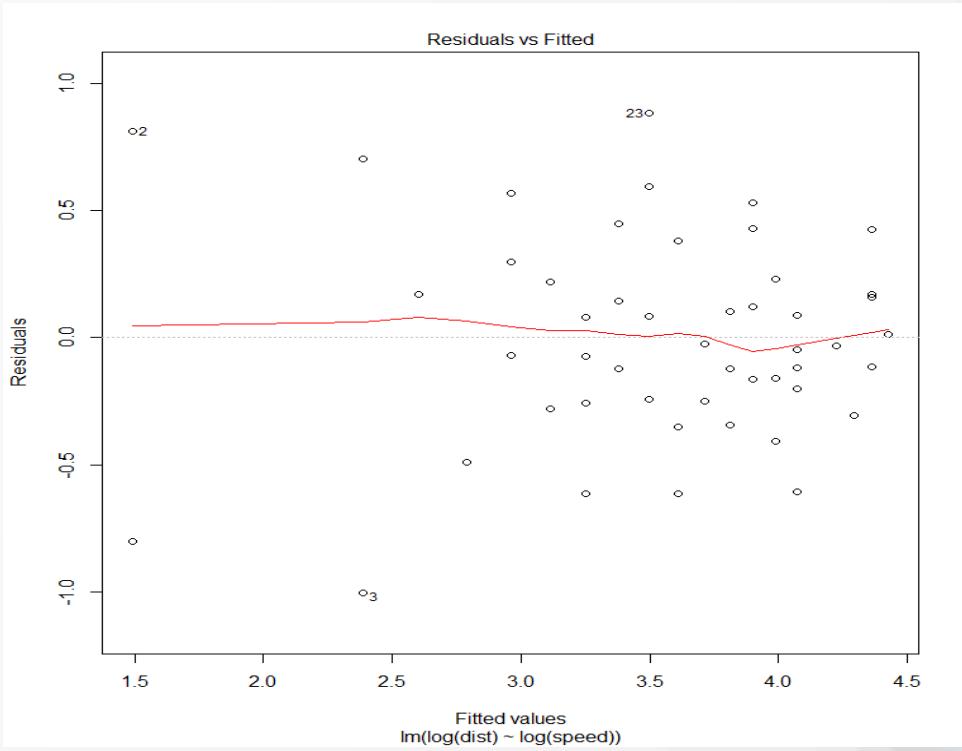
Residual standard error: 0.405267 on 48 degrees of freedom

Multiple R-squared: 0.7331444, Adjusted R-squared: 0.7275849

F-statistic: 131.8726 on 1 and 48 DF, p-value: 0.0000000000000002259356

2-1. 회귀분석

■ 회귀모형 잔차의 검토



2-1. 회귀분석

■ 변수선택법

- 1) 전진선택법
- 2) 후진선택법
- 3) 단계적선택법
- 4) 모든 조합

2-2. 다변량분석

- **다차원 척도법** : 여러 변수로 이루어진 개체에 대한 모든 쌍들간 유사성 (또는 거리)을 이용하여 저차원의 공간에서 개체를 기하학적으로 표현하는 기법

2-2. 다변량분석

■ 주성분 분석: 고차원 자료의 변동을 저차원 자료로 변환

> summary(USArrests)

Murder	Assault	UrbanPop	Rape
Min. : 0.800	Min. : 45.00	Min. :32.00	Min. : 7.300
1st Qu.: 4.075	1st Qu.:109.00	1st Qu.:54.50	1st Qu.:15.075
Median : 7.250	Median :159.00	Median :66.00	Median :20.100
Mean : 7.788	Mean :170.76	Mean :65.54	Mean :21.232
3rd Qu.:11.250	3rd Qu.:249.00	3rd Qu.:77.75	3rd Qu.:26.175
Max. :17.400	Max. :337.00	Max. :91.00	Max. :46.000

> fit <- princomp(USArrests, cor=TRUE)

> summary(fit)

Importance of components:

	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Standard deviation	1.5748782744	0.9948694148	0.59712911550	0.41644938195
Proportion of Variance	0.6200603948	0.2474412881	0.08914079515	0.04335752193
Cumulative Proportion	0.6200603948	0.8675016829	0.95664247807	1.00000000000

3. 시계열분석

■ 시계열 :시간의 흐름에 따라 관찰된 값

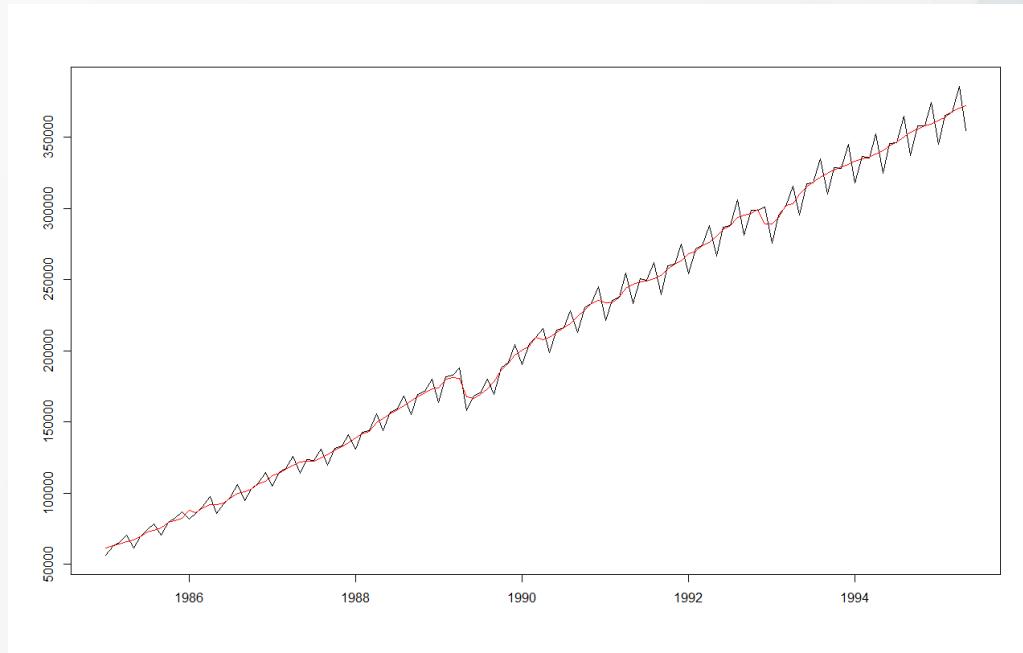
- ✓ 시계열은 주파수 영역(frequency domain) 정보와 시간영역(time domain) 정보를 가지고 있음
 - › 주파수 영역 정보 : 주기적으로 반복되는 정보
 - › 시간영역 정보 : 시간에 따라 전개되는 정보

3. 시계열분석

- 안정 또는 정상(stationary) 시계열 : 구간을 달리하더라도 매 구간별로 그 특성이 동일, 시계열의 평균과 분산 등이 시간의 흐름에 따라 특정한 변화가 없는 시계열

3. 시계열분석

■ GDP와 계절조정 GDP



3. 시계열분석

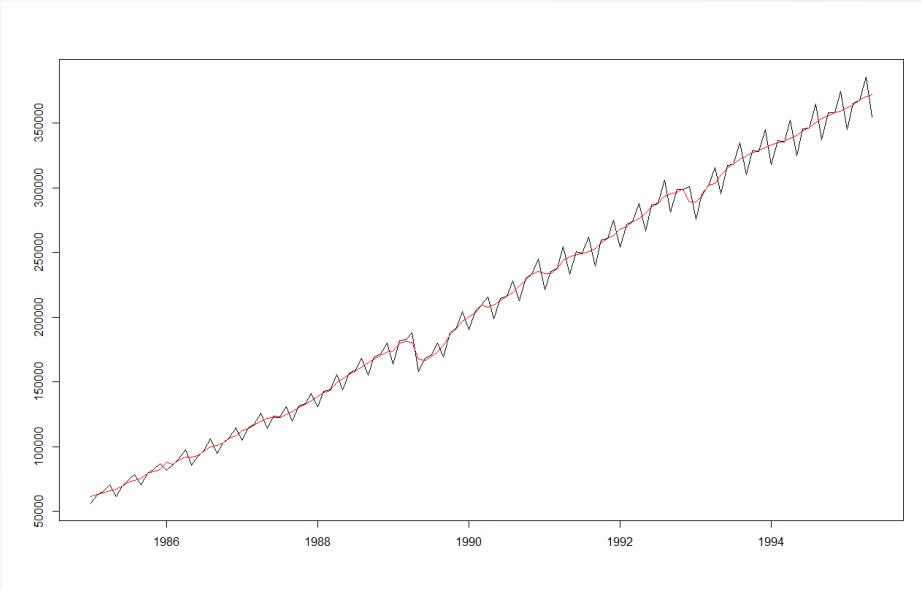
- 경제통계(Y_t)는 일반적으로 그 변동주기에 따라 추세요인(T_t), 순환요인(C_t), 계절요인(S_t), 불규칙요인(I_t)으로 구성된다고 가정
 - ✓ 추세요인(T_t) : 인구증가, 기술변화, 생산성증대 등에 따른 장기변동
 - ✓ 순환요인(C_t) : 경기순환에 따라 반복되는 변동
 - ✓ 계절요인(S_t) : 1년 주기로 반복되는 변동 : 계절변동과 달력변동
 - ✓ 불규칙요인(I_t) : 돌발적, 원인불명의 요인에 의거, 일어나는 변동
(태풍, 지진, 파업 등)

3. 시계열분석

- 시계열(Y_t)은 앞서의 요인들을 이용하여 승법형($Y_t = T_t \cdot C_t \cdot S_t \cdot I_t$),
가법형($Y_t = T_t + C_t + S_t + I_t$) 등으로 결합
 - ✓ 통상적으로 시간에 따라 각 요인이 비례적으로 증가하는 경우 승법형을 선택
- 계절조정계열 : 원계열에서 계절변동을 제거한 계열

3. 시계열분석

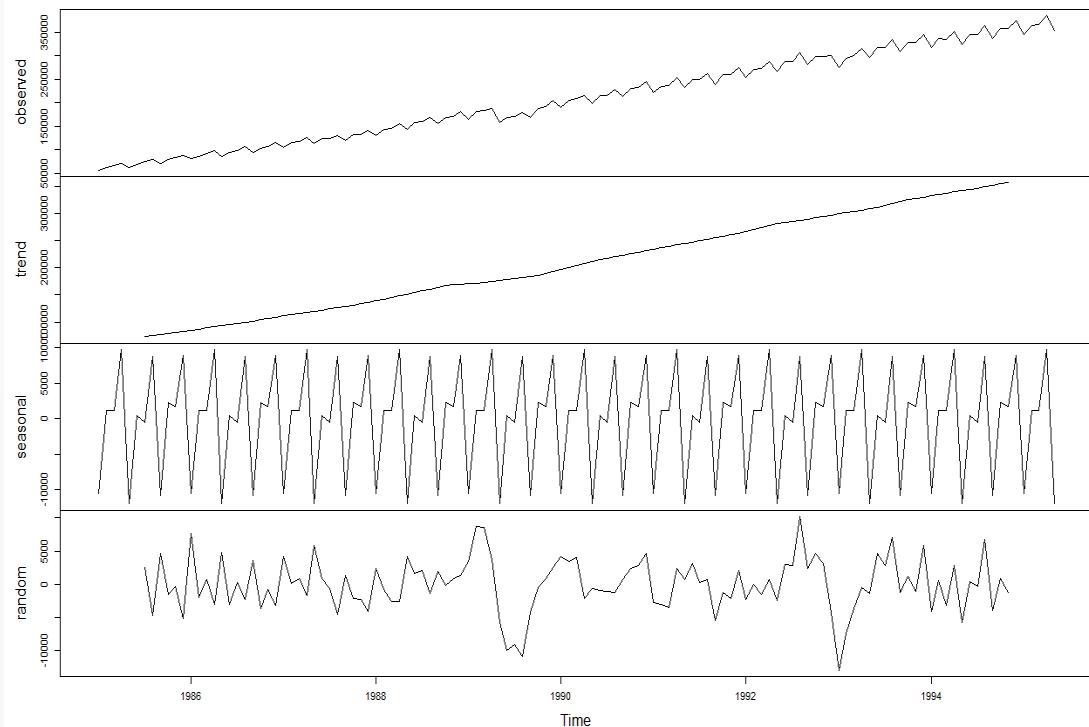
■ GDP와 계절조정 GDP



3. 시계열분석

■ GDP의 분해

Decomposition of additive time series



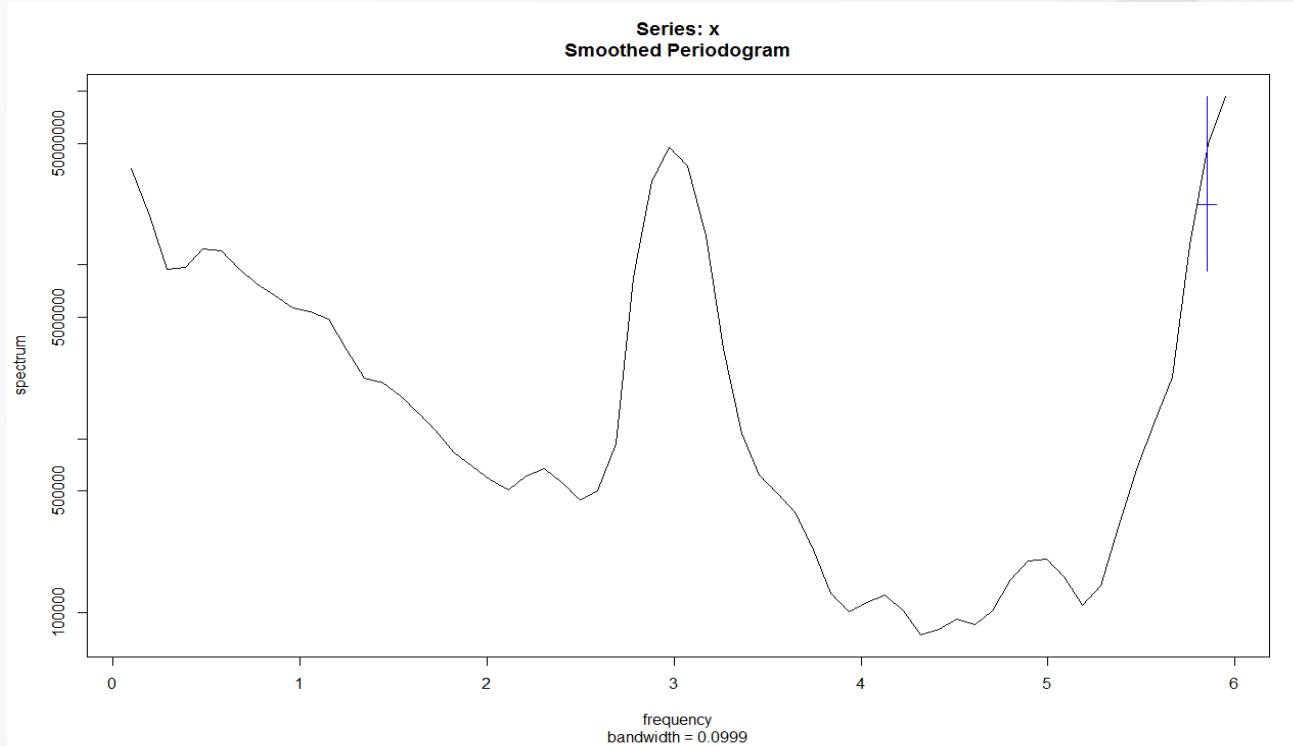
3. 시계열분석

■ 시계열의 주파수 정보 : 어떤 주기의 변동이 존재하는지

- ✓ 주기도(periodogram)은 시계열이 어떤 주기들을 갖고 움직이고 있는지를 나타내주는 도표
- ✓ 주기도에서 특정 주파수에 큰 값이 나타나면 시계열에 해당 주파수(또는 주기)의 변동이 큰 것을 알 수 있음
- ✓ 주기도는 너무 변동성이 커서 평활화 → 스펙트럴 밀도함수(스펙트럼)

3. 시계열분석

■ Log(GDP)의 스펙트럼



3. 시계열분석

- 시계열모형 : 백색잡음과정, AR모형, MA모형, ARMA모형, ARIMA모형
- 백색잡음과정 : 시계열이 과거와 아무런 상관이 없음
 - ✓ 백색잡음(white noise)은 상호독립적이고 같은 분포을 갖는 확률변수로 구성되어 있으며 자기상관함수과 편자기상관함수는 0

3. 시계열분석

■ AR모형 : 시계열이 과거 실제값의 함수로 표현

✓ AR(1)모형 : $Y_t = \phi_0 + \phi_1 Y_{t-1} + \varepsilon_t$

✓ AR(p)모형 : $Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \varepsilon_t$

3. 시계열분석

■ MA모형 : 과거의 충격으로 현재값이 표현

- ✓ MA(1)모형 : $Y_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t$
- ✓ MA(q)모형 : $Y_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t$

3. 시계열분석

■ ARMA모형 : 시계열이 과거의 실제값과 과거에 발생했던 충격으로 동시에 설명되는 모형

✓ ARMA(1.1)모형 : $Y_t = \phi_0 + \phi_1 Y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$

3. 시계열분석

■ 불안정 시계열 : 시간에 따라 평균과 분산이 일정하지 않은 계열

- ✓ 시계열의 추세가 확률적이라면 다음과 같이 차분을 하여 시계열을 안정화

$$\Delta Y_t = Y_t - Y_{t-1}$$

- ✓ 1차 차분하여 ARMA(p,q)모형이 되는 모형 : ARIMA(p,1,q) 모형

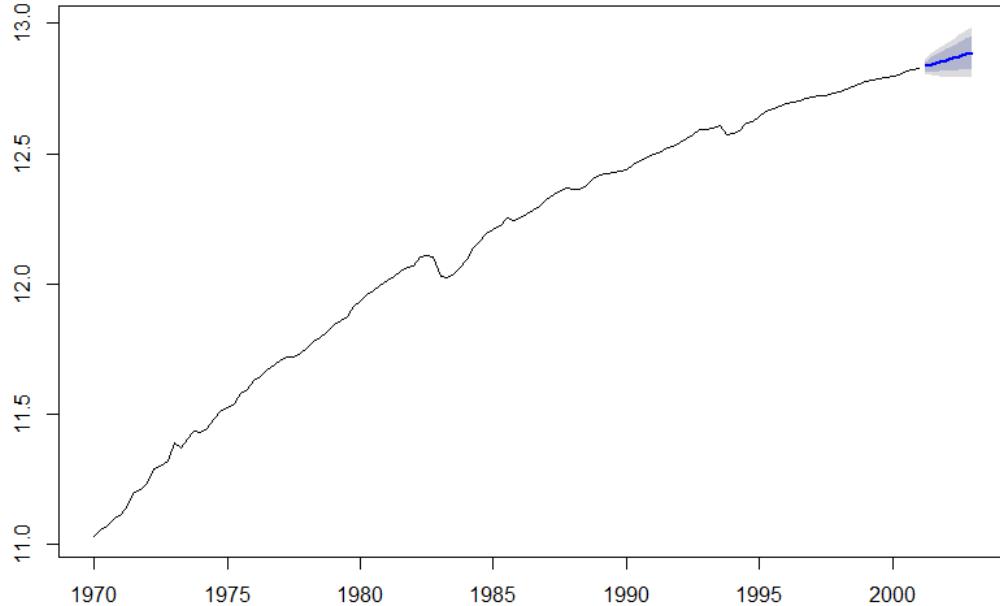
3. 시계열분석

- ARIMA모형의 작성 : 식별 → 추정 → 진단 → 예측
- Log(계절조정 GDP)의 모형 추정

```
ar1          ma1          sar1          sar2  
0.1270732 -0.9449340  0.0666567  0.0171890  
s. e.      0.0876262  0.0265335  0.0891967  0.1023821  
  
sigma^2 estimated as 0.0002127516:  log likelihood=344.49  
AIC=-678.97    AICc= 678.46    BIC= -664.91  
>
```

3. 시계열분석

■ Log(계절조정 GDP)의 예측



강의를 마쳤습니다

수고하셨습니다.

11차시 | 빅데이터 시각화

이긍희 교수



빅데이터 시각화

1. 데이터 시각화의 정의와 필요성
2. 데이터 시각화의 기능
3. 데이터 시각화의 과정
4. 데이터 시각화의 역사
5. 데이터 시각화의 방향
6. R프로그램을 이용한 데이터 시각화

1. 데이터 시각화의 정의와 필요성

■ 데이터 시각화의 정의

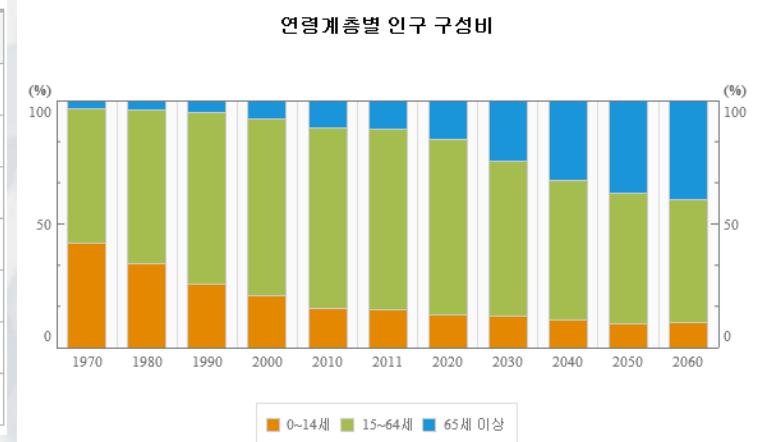
- ✓ 데이터 시각화 : 데이터를 분석하여 데이터를 시각적 형태로 만드는 것
- ✓ 시각화의 목적 : 데이터분석과 의사소통
 - 데이터의 숨은 의미를 발견하고, 설명 또는 이야기하고, 그것을 통해 의사결정을 내리는 통찰력을 확보
 - 데이터 시각화는 다른 사람들과의 소통을 보다 효율적으로 이끌어냄

1. 데이터 시각화의 정의와 필요성

■ 데이터 시각화의 필요성

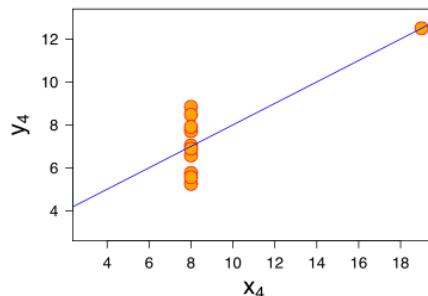
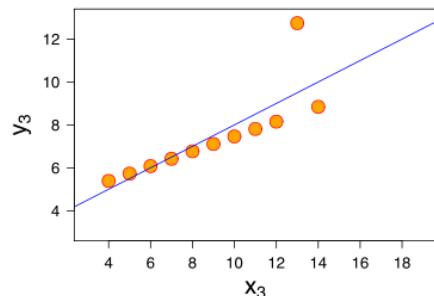
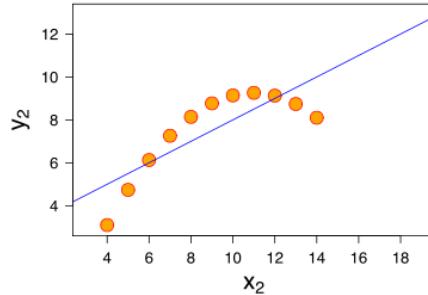
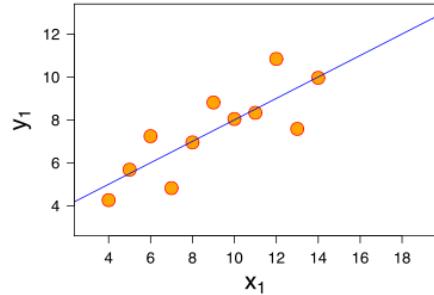
- ✓ 보통 사람의 경우 숫자가 7개 넘어가면 이해하지 못함
 - › 반면 그래프는 100개 이상의 숫자도 그래프로 한눈에 이해
 - › 문자의 경우도 마찬가지 : 트위터

	1970	1980	1990	2000	2010	2012	2020	2030	2040	2050	2060	
인구수	0~14세	13,709	12,951	10,974	9,911	7,975	7,559	6,788	6,575	5,718	4,783	4,473
	15~64세	17,540	23,717	29,701	33,702	35,983	36,556	36,563	32,893	28,873	25,347	21,865
	65세 이상	991	1,456	2,195	3,395	5,452	5,890	8,084	12,691	16,501	17,991	17,622
구성비	0~14세	42.5	34.0	25.6	21.1	16.1	15.1	13.2	12.6	11.2	9.9	10.2
	15~64세	54.4	62.2	69.3	71.7	72.8	73.1	71.1	63.1	56.5	52.7	49.7
	65세 이상	3.1	3.8	5.1	7.2	11.0	11.8	15.7	24.3	32.3	37.4	40.1
	계	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0



1. 데이터 시각화의 정의와 필요성

■ Anscombe's quartet



특성	값
x 의 평균	9
x 의 분산	11
y 의 평균	7.50
y 의 분산	4.122 or 4.127
상관계수	0.816
선형 회귀모형	$y = 3.00 + 0.500x$

출처 : 위키피디아

1. 데이터 시각화의 정의와 필요성

■ 시각과 두뇌

- ✓ 새로운 정보를 받아들이는 비중 : 시각 65%
- ✓ 데이터 시각화를 통해 좌뇌와 우뇌의 기능을 통합

1. 데이터 시각화의 정의와 필요성

■ 데이터 시각화의 필요성

百聞이 不如一見

백 번 듣는 것이 한 번 보는 것만 못하다

«한서(漢書)» 의 <조충국전(趙充國傳)>

2. 데이터 시각화의 기능

■ 데이터 시각화의 위계 : 맥캔들리(McCandless)

- ✓ 데이터(data)
- ✓ 정보(information)
- ✓ 지식(knowledge)
- ✓ 지혜(wisdom)

2. 데이터 시각화의 기능

■ 데이터 시각화의 기능

- ✓ 정보의 기록
- ✓ 패턴의 파악
- ✓ 데이터 분석
- ✓ 스토리텔링

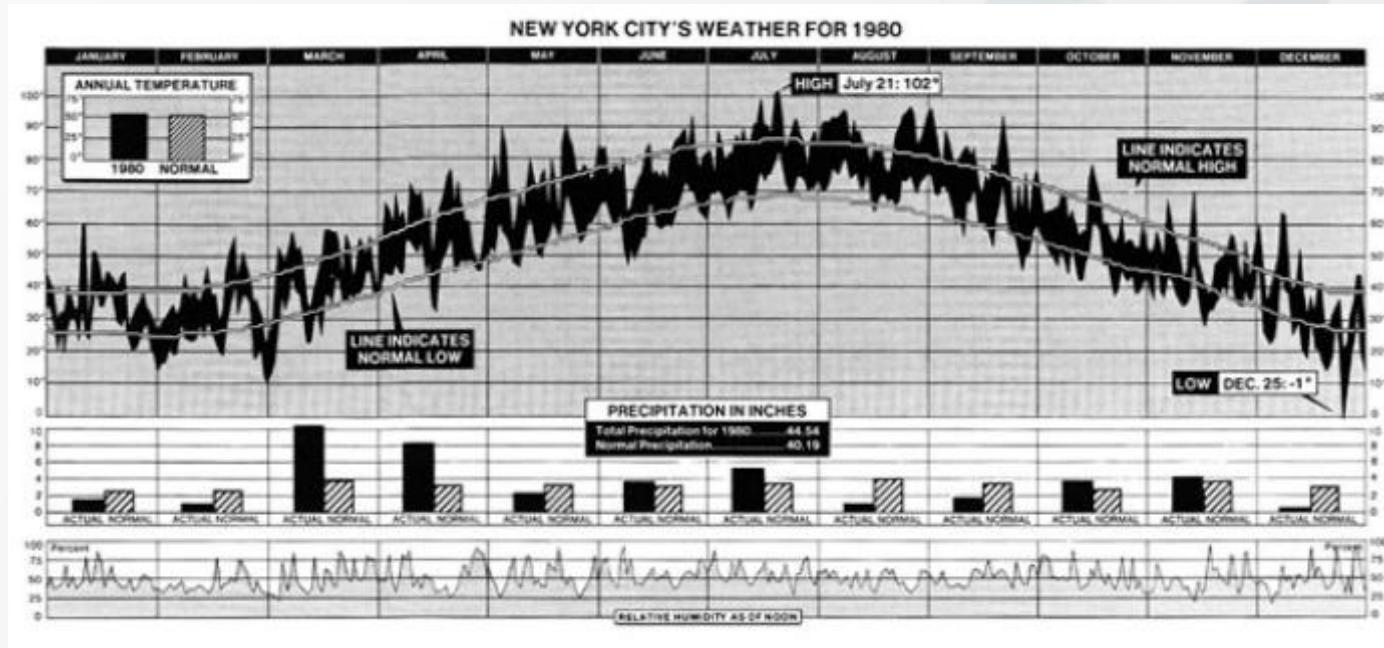
2. 데이터 시각화의 기능

■ 데이터 시각화의 기능 : 정보의 기록



2. 데이터 시각화의 기능

■ 데이터 시각화의 기능 : 패턴의 파악

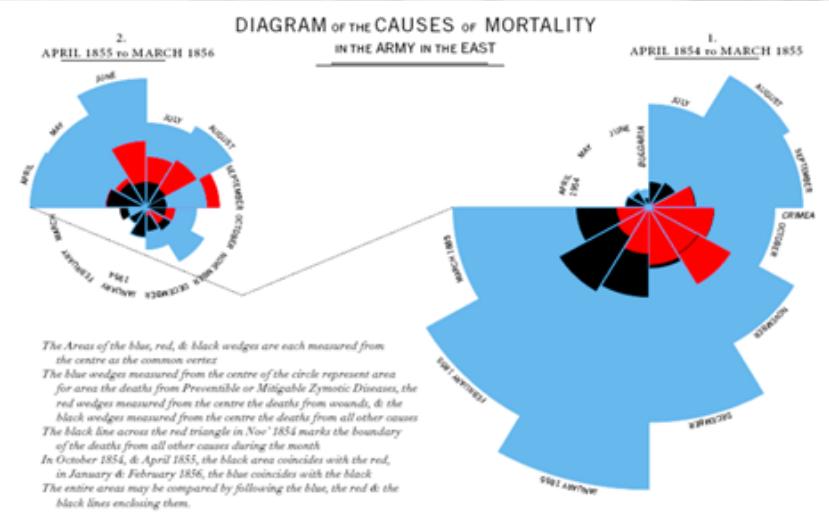


2. 데이터 시각화의 기능

■ 데이터 시각화의 기능 : 데이터의 분석



콜레라 맵, 1855

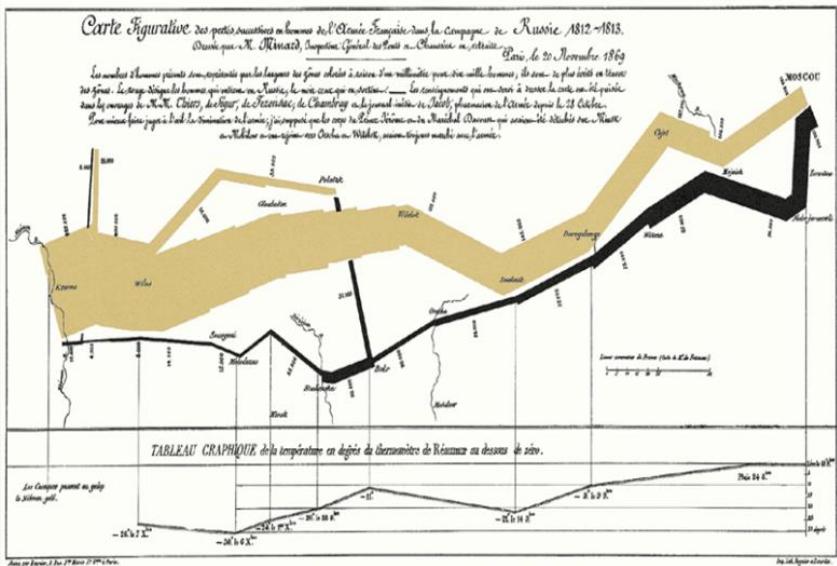


장미도표, 1857

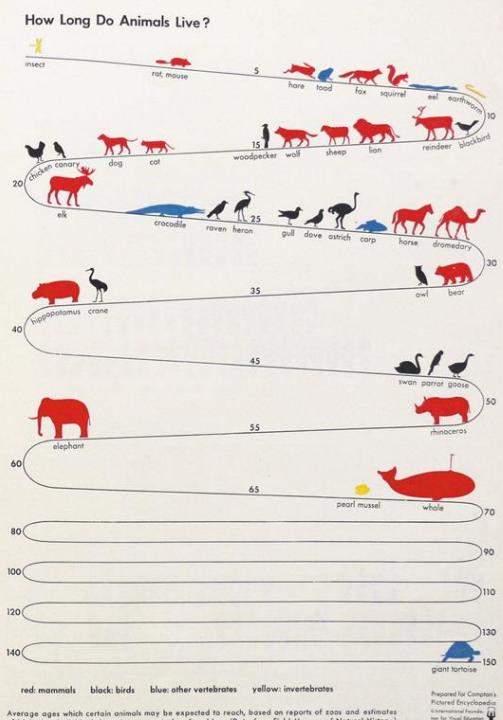
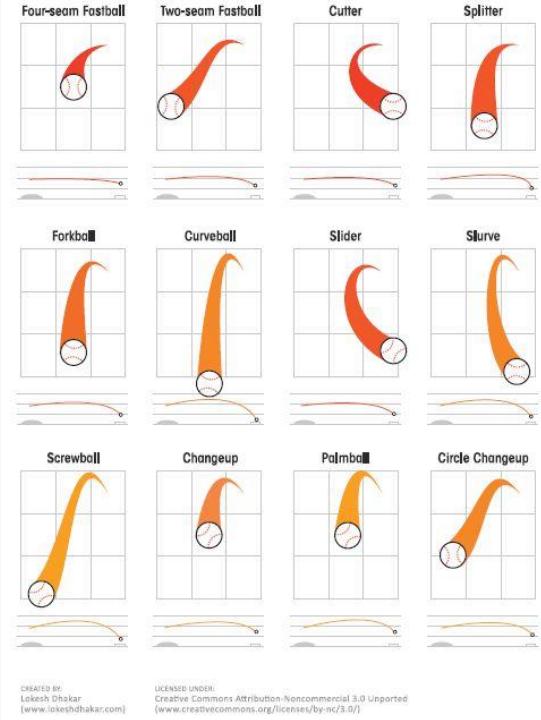
2. 데이터 시각화의 기능

■ 데이터 시각화의 기능 : 스토리텔링

찰스 요셉 미라드 : 나폴레옹의 러시아 침략



2. 데이터 시각화의 기능



출처 : <http://lokeshdhakar.com/baseball-pitches-illustrated/> 출처 : <http://www.informationisbeautiful.net/2011/vintage-infoporn-no-1/>

2. 데이터 시각화의 기능

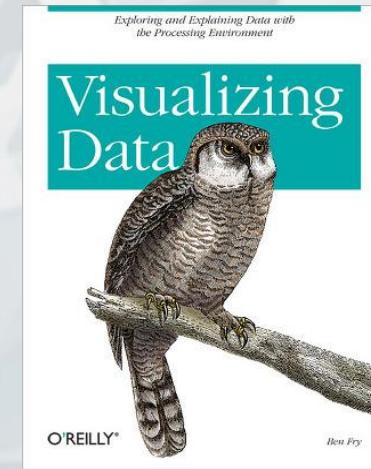
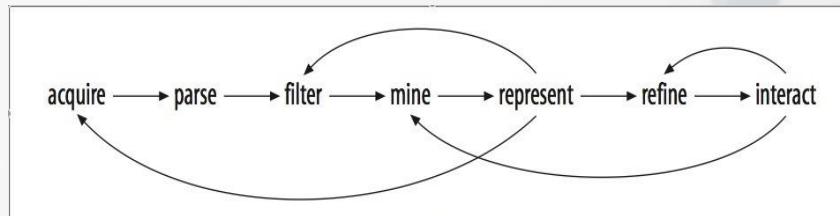
■ 시각화의 장점과 단점

- ✓ 장점 : 많은 양의 데이터를 빠르게 이해하고, 데이터 속의 이야기(통찰) 발견하여 빠르게 의사결정 할 수 있게 함
- ✓ 단점 : 부정확하거나 왜곡된 데이터 시각화는 잘못된 의사결정을 유도

3. 데이터 시각화의 과정

■ 데이터 시각화 과정

- ① Acquire : 데이터 수집
- ② Parse : 데이터 구조화와 분류
- ③ Filter : 관심 데이터 추출
- ④ Mine : 데이터마이닝 기법 적용
- ⑤ Represent : 시각화 표현 방법 선택
- ⑥ Refine : 시각화 개선
- ⑦ Interact : 조건 변화



3. 데이터 시각화의 과정

■ 데이터 시각화 원칙(Tufte)

- › 데이터 그 자체를 보여주는 것이 중요
- › 보는 사람이 화려한 그래픽이나 시각화 방법에 너무 집중하지 않게 함
- › 데이터 자체가 말하고자 하는 바를 왜곡하지 않음
- › 너무 많은 숫자나 문자를 작은 화면에 보여주려고 하지 않음
- › 많은 양의 데이터가 일관성을 가져야 함
- › 서로 다른 데이터를 손쉽게 비교할 수 있게 함
- › 데이터는 몇 가지 단계로 깊이 들어가 자세히 살펴볼 수 있어야 함
- › 통계 결과나 시각화의 설명을 데이터와 함께 보여주어야 함

3. 데이터 시각화의 과정

■ 데이터 시각화 도구

- ✓ EXCEL, Google Chart Tool, Visual.ly
- ✓ Tableau, Manyeyes
- ✓ Leaflet, OpenLayers
- ✓ R, Python, Processing
- ✓ HTML/CSS, Javascript, D3.js

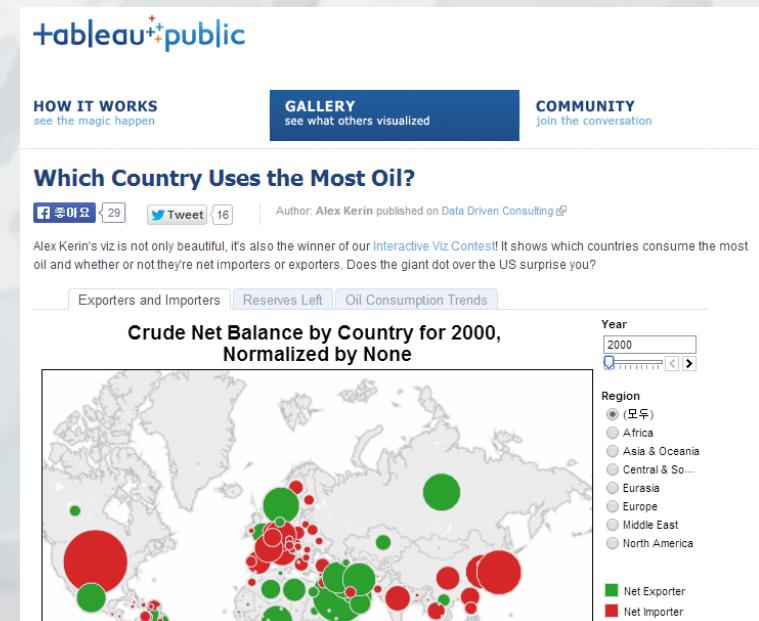
3. 데이터 시각화의 과정

■ 데이터 시각화 도구

<D3.js>

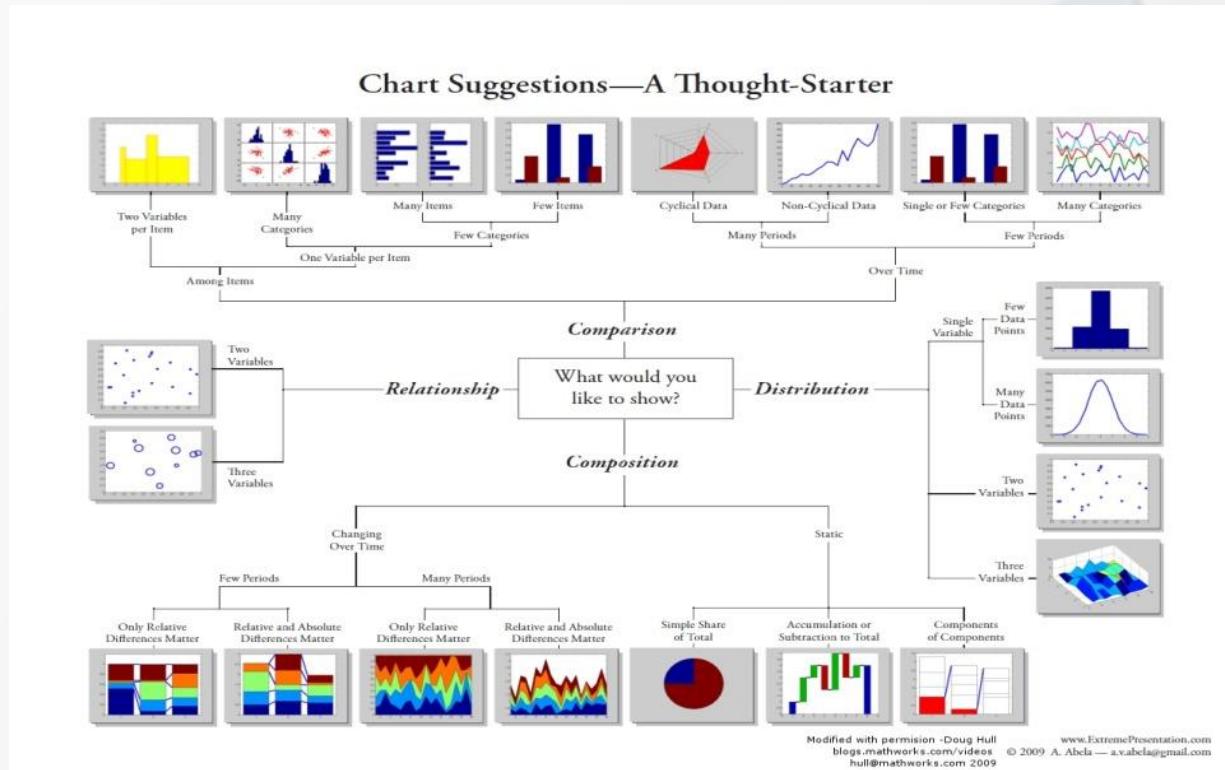


<tableau>



3. 데이터 시각화의 과정

■ 데이터 시각화 방법 개요



3. 데이터 시각화의 과정

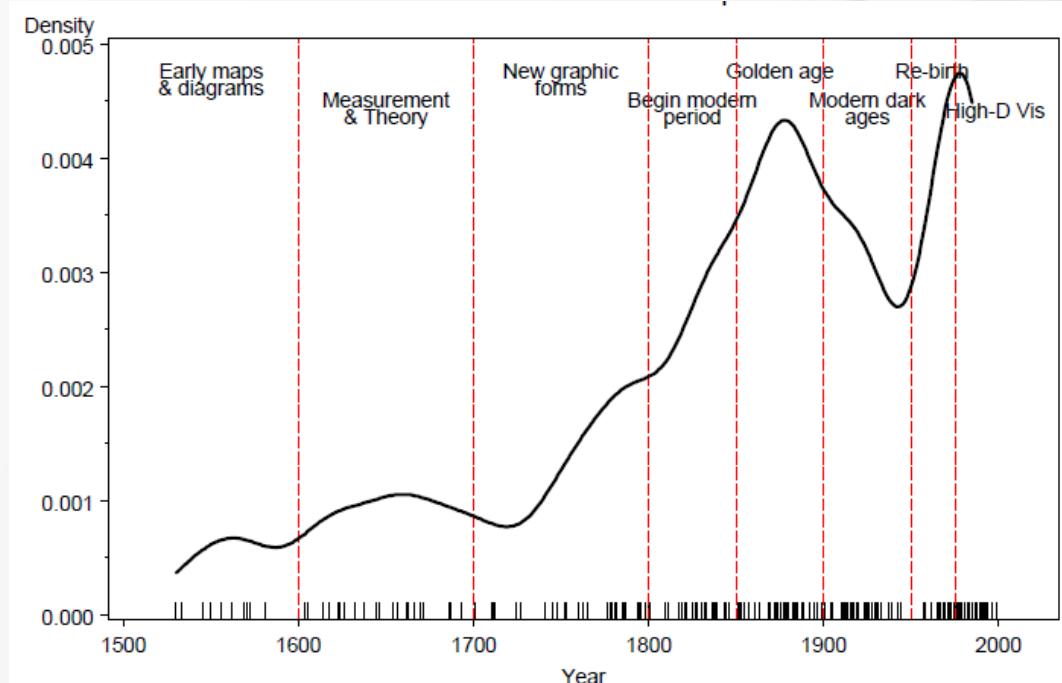
■ 데이터 시각화 방법 개요

시간 시각화	구성 시각화	관계 시각화	비교 시각화	공간 시각화
선그래프 막대 그래프	원그래프 도우넛 그래프 트리맵 누적그래프	산점도 버블차트	히트맵 체르노프 페이스 평행좌표계	지도 이용 그래프

4. 데이터 시각화의 역사

■ 시대적 구분

Friendly(2006) A Brief History of Data Visualization



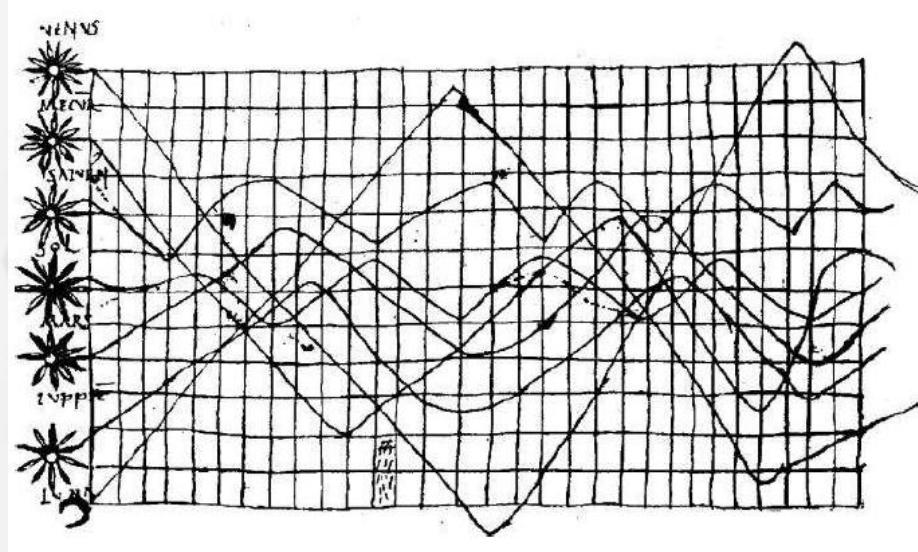
4. 데이터 시각화의 역사

■ 17세기 이전

✓ Maps and Diagrams



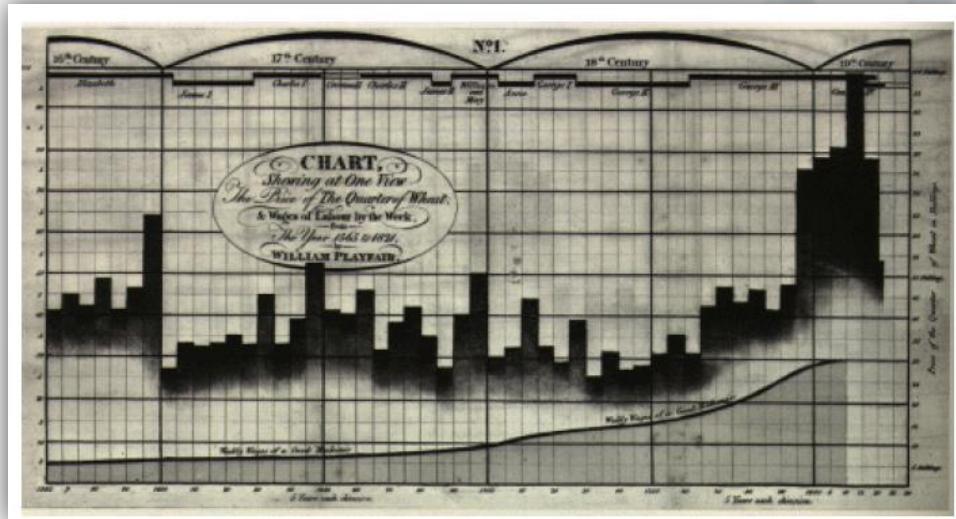
Anaximander's Map of the World



4. 데이터 시각화의 역사

■ 19세기 전반

- ✓ 현대 그래프의 시작



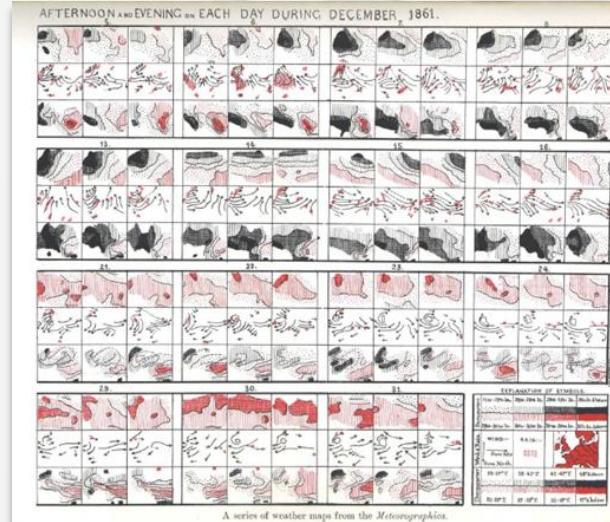
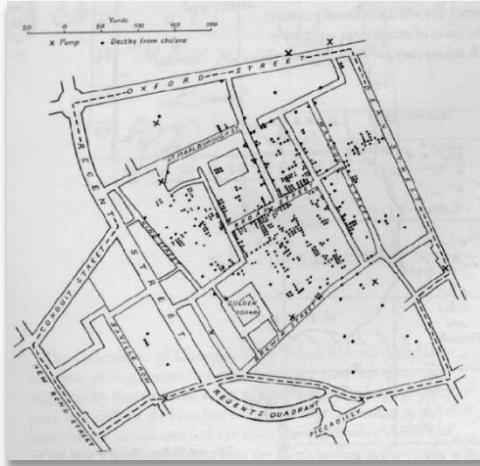
4. 데이터 시각화의 역사

■ 19세기 후반

✓ 통계 그래프의 황금기

› 유럽의 통계청 설립, 통계이론이 확산되면서 시각화가 급격히 발전

<1855년 : John Snow의 콜레라 맵> <1861년 : Francis Galton의 현대식 기상지도>



4. 데이터 시각화의 역사

■ 20세기 후반

✓ 통계 그래프의 재발견

- › 컴퓨터 발전, 통계 이론 발전, 컴퓨터 입력기 발전에 따라 크게 발전
- › 1962년 미국 John W. Tukey "The Future of Data Analysis" 탐색적 자료 분석 "Exploratory Data Analysis" (EDA)을 제안

5. 데이터 시각화의 방향

■ 21세기 들어 인터넷의 확산, 빅데이터 시대 도래, 다양한 시각화 도구 개발 등으로 다양한 데이터 시각화가 시도

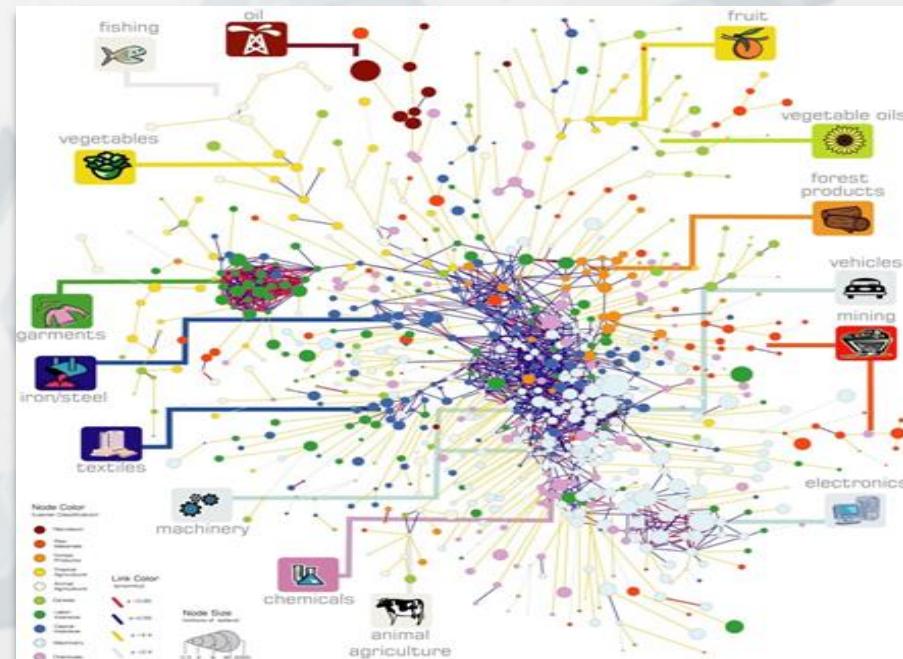
- ✓ 동적 시각화
- ✓ 사용자와의 대화가 가능한 시각화
- ✓ 다양한 정보의 결합 : 지도 등
- ✓ 비정형 데이터 시각화
- ✓ 거시 데이터 중심에서 미시 데이터 중심으로

5. 데이터 시각화의 방향

<GAPMINDER>



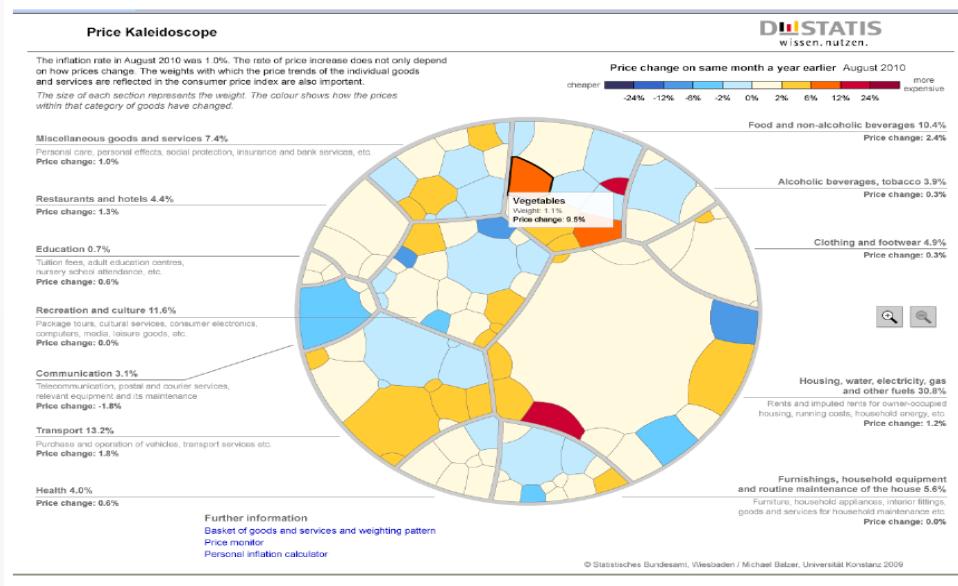
<UN 세계상품교역>



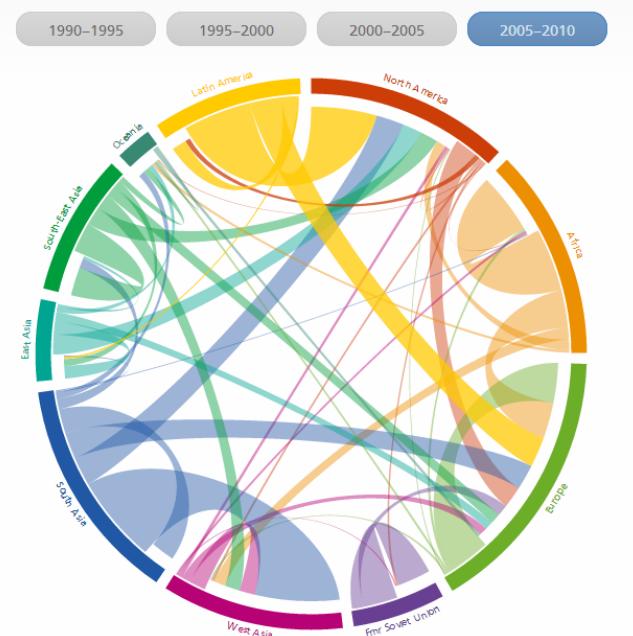
5. 데이터 시각화의 방향

<물가구조 독일통계청>

The German Price Kaleidoscope



<국가별 이민>



5. 데이터 시각화의 방향

<킹목사 연설문>

"I HAVE A DREAM . . ."

(Copyright 1963, MARTIN LUTHER KING, JR.)

Speech by the Rev. MARTIN LUTHER KING
At the "March on Washington"

I am happy to join with you today in what will go down in history as the greatest demonstration for freedom in the history of our nation.

Five score years ago a great American in whose symbolic shadow we stand today signed the Emancipation Proclamation. This momentous decree is a great beacon light of hope to millions of Negro slaves who had been



6. R을 이용한 데이터 시각화

■ ggplot2, plotly

```
library(plotly)
```

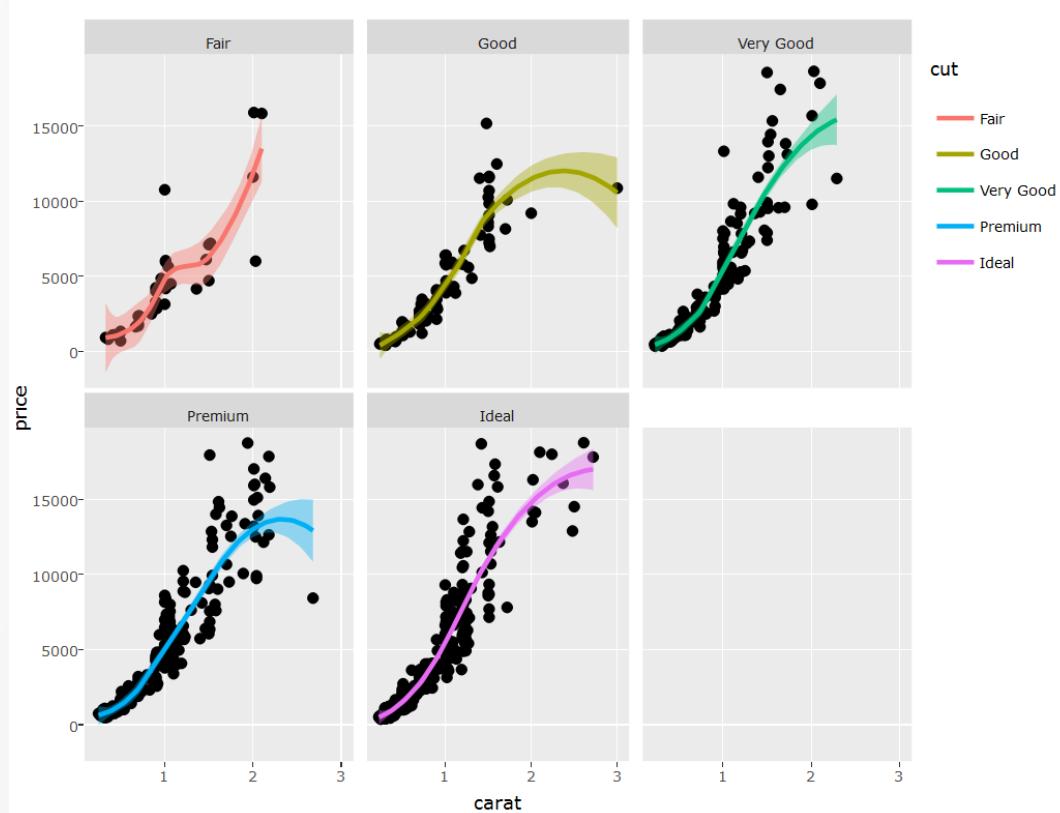
```
library(ggplot2)
```

```
di <- diamonds[sample(nrow(diamonds), 1000), ]
```

```
p1 <- ggplot(data = di, aes(x = carat, y = price)) +  
  geom_point(aes(text = paste("Clarity:", clarity)), size = 2) +  
  geom_smooth(aes(colour = cut, fill = cut)) + facet_wrap(~ cut)
```

```
(g1 <- ggplotly(p1))
```

6. R을 이용한 데이터 시각화



강의를 마쳤습니다

수고하셨습니다.

12차시 | 빅데이터 분석방법1

이긍희 교수



정형 데이터마이닝1

1. 데이터마이닝 개요

2. 분류분석

2-1. 로지스틱 회귀모형

2-2. 의사결정나무모형

2-3. 신경망모형

2-4. 앙상블모형

1. 데이터마이닝 개요

■ 데이터 마이닝의 정의 : 빅데이터에서 드러나지 않는 유용한 정보를
찾아내는 과정

- ✓ 데이터베이스에서의 지식발견
- ✓ 지식추출
- ✓ 정보수학
- ✓ 정보고고학
- ✓ 데이터 패턴 프로세싱

1. 데이터마이닝 개요

■ 데이터마이닝의 기능

✓ 분류(classification)

- › 새로운 현상을 기존의 분류에 배정

✓ 추정(estimation)

- › 주어진 데이터를 바탕으로 미지값을 추정

✓ 예측(prediction)

- › 시간에 따른 미래의 값을 추정하거나 분류

1. 데이터마이닝 개요

■ 데이터마이닝의 기능

- ✓ 연관분석(association analysis)
 - › 장바구니 분석
- ✓ 군집(clustering)
 - › 다른 모집단에서 생성된 데이터를 그룹별로 세분화
- ✓ 기술(description)
 - › 데이터분석 의미를 단순하게 기술

1. 데이터마이닝 개요

■ 데이터 마이닝 추진 단계

✓ 1단계 : 목적 정의

- › 데이터마이닝 도입의 목적을 분명히 하는 단계

✓ 2단계 : 데이터 준비

- › 데이터 마이닝에 필요한 데이터 수집
- › 분석에 적합한 품질을 가지도록 데이터를 정제

✓ 3단계 : 가공

- › 데이터를 데이터 마이닝에 적합한 형식으로 가공

1. 데이터마이닝 개요

■ 데이터 마이닝 추진 단계

- ✓ 4단계 : 데이터 마이닝 기법 적용

- › 데이터 마이닝 기법을 적용하여 데이터로부터 정보를 추출

- ✓ 5단계 : 검증

- › 추출한 정보의 유용성을 검증

2. 분류분석

■ 분류분석 : 데이터를 범주로 예측하는 분석

- ✓ 로지스틱 회귀모형
- ✓ 의사결정나무모형
- ✓ 앙상블모형
- ✓ 신경망모형

2-1. 로지스틱 회귀모형

■ 로지스틱 회귀모형

- ✓ 반응변수가 범주형인 경우에 적용되는 회귀분석 모형
- ✓ 새로운 설명변수의 값이 주어질 때 반응변수의 각 범주(또는 집단)에 속할 확률이 얼마인지 추정

2-1. 로지스틱 회귀모형

■ 로지스틱 회귀모형

```
> data(iris)
> iris_a = subset(iris, Species=="setosa" | Species == "versicolor")
> b <- glm(Species~Sepal.Length, data=iris_a, family=binomial)
> summary(b)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-27.8285	4.8276	-5.765	8.19e-09
Sepal.Length	5.1757	0.8934	5.793	6.90e-09

2-1. 로지스틱 회귀모형

■ 로지스틱 회귀모형

```
> exp(coef(b)[ "Sepal.Length" ])
```

Sepal. Length

176.9201

```
> exp(confint(b, parm="Sepal.Length"))
```

2.5% 97.5%

38.35635 1323.23391

2-1. 로지스틱 회귀모형

■ 로지스틱 회귀모형

```
> fitted(b)[c(1:5, 96:100)]
```

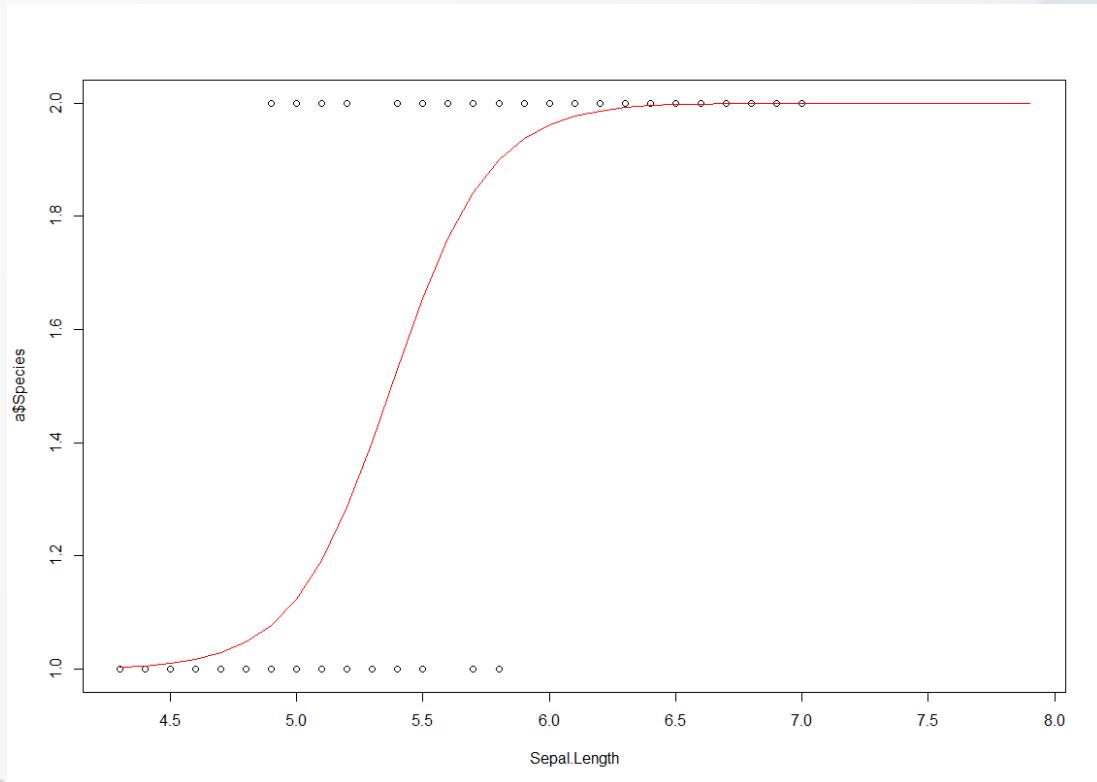
1	2	3	4	5	96	97
98	99	100				
0.19271553	0.07816094	0.02923436	0.01763097	0.12455000	0.84196979	0.84196979
0.98608544	0.19271553	0.84196979				

```
> predict(b, newdata=a [c(1, 50, 51, 100)], type="response")
```

1	50	51	100
0.1927155	0.1245500	0.9997755	0.8419698

2-1. 로지스틱 회귀모형

■ 로지스틱 회귀모형



2-2. 의사결정나무모형

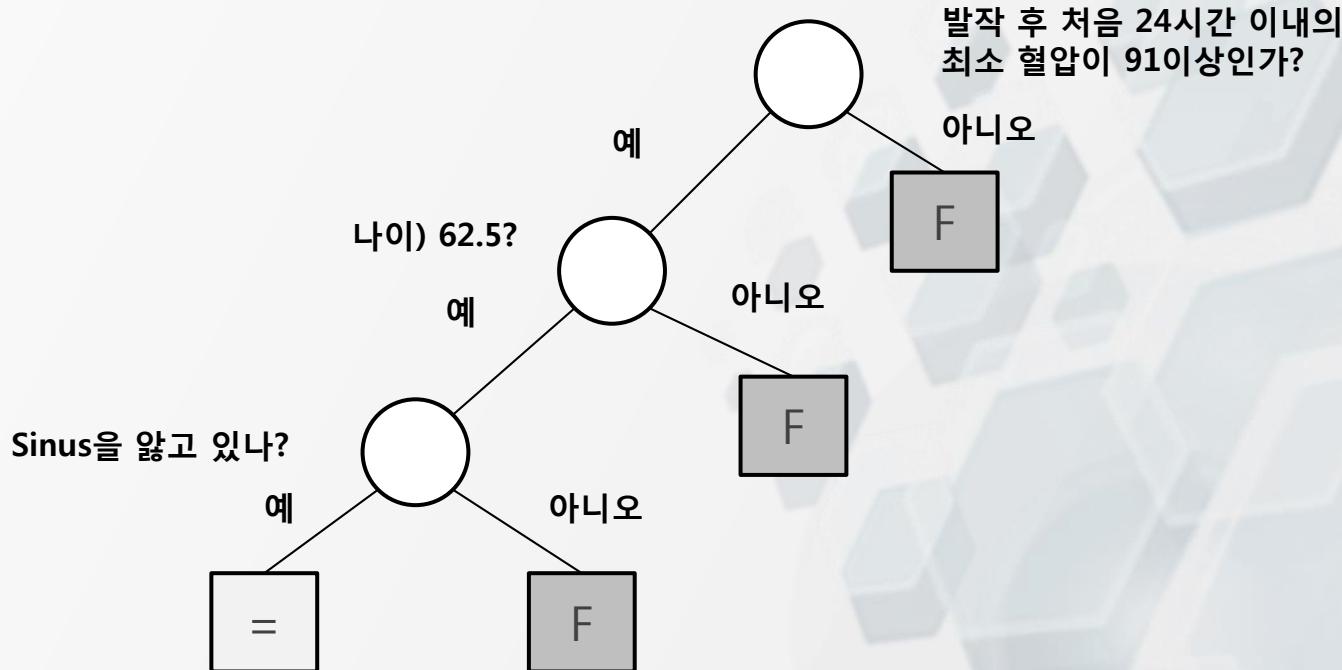
■ 의사결정나무

의사결정 규칙을 나무(tree)구조로 나타내어 데이터를 몇 개의 범주로 분류, 예측하는 방법

- ✓ 나무 구조를 형성하는 분류변수와 분류기준 값 선택이 중요
- ✓ 나무 모형의 크기는 과대적합(또는 과소적합) 되지 않도록 조절

2-2. 의사결정나무모형

■ 심장발작환자 측정에 의한 사망가능성 예측



2-2. 의사결정나무모형

■ 불확실성지표

① Gini 지수

$$\phi(g) = \sum_j \hat{p}_j(g)(1 - \hat{p}_j(g))$$

② Entropy 지수

$$\phi(g) = -\sum_j \hat{p}_j(g) \log \hat{p}_j(g))$$

③ Deviance

$$\phi(g) = -2 \sum_j n_j \log \hat{p}_j(g))$$

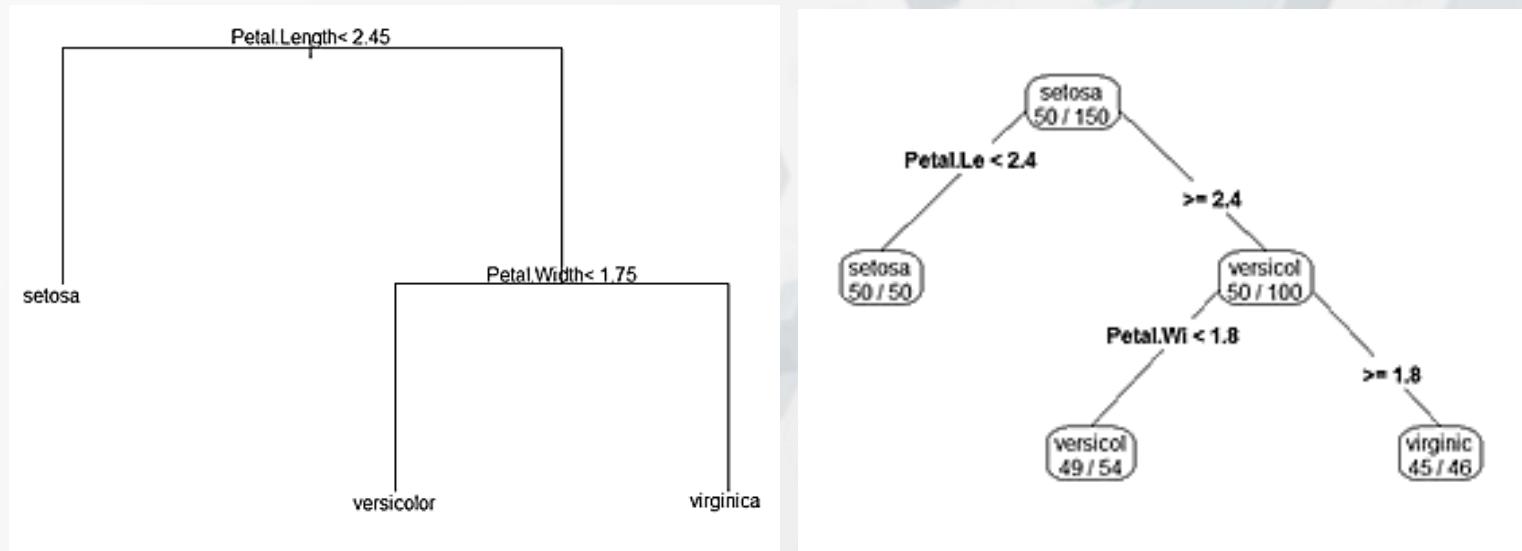
2-2. 의사결정나무모형

■ 알고리즘과 분류 기준변수의 선택법

	이산형 목표변수	연속형 목표변수
CHAID(다지분할)	카이제곱 통계량	ANOVA F-통계량
CART(이진분할)	지니지수	분산감소량
C4.5	엔트로피지수	

2-2. 의사결정나무모형

- 1) root 150 100 setosa (0.33333333333 0.33333333333 0.33333333333)
- 2) Petal.Length < 2.45 50 0 setosa (1.00000000000 0.00000000000 0.00000000000) *
- 3) Petal.Length >= 2.45 100 50 versicolor (0.00000000000 0.50000000000 0.50000000000)
- 6) Petal.Width < 1.75 54 5 versicolor (0.00000000000 0.90740740741 0.09259259259) *
- 7) Petal.Width >= 1.75 46 1 virginica (0.00000000000 0.02173913043 0.97826086957) *



2-3. 신경망모형

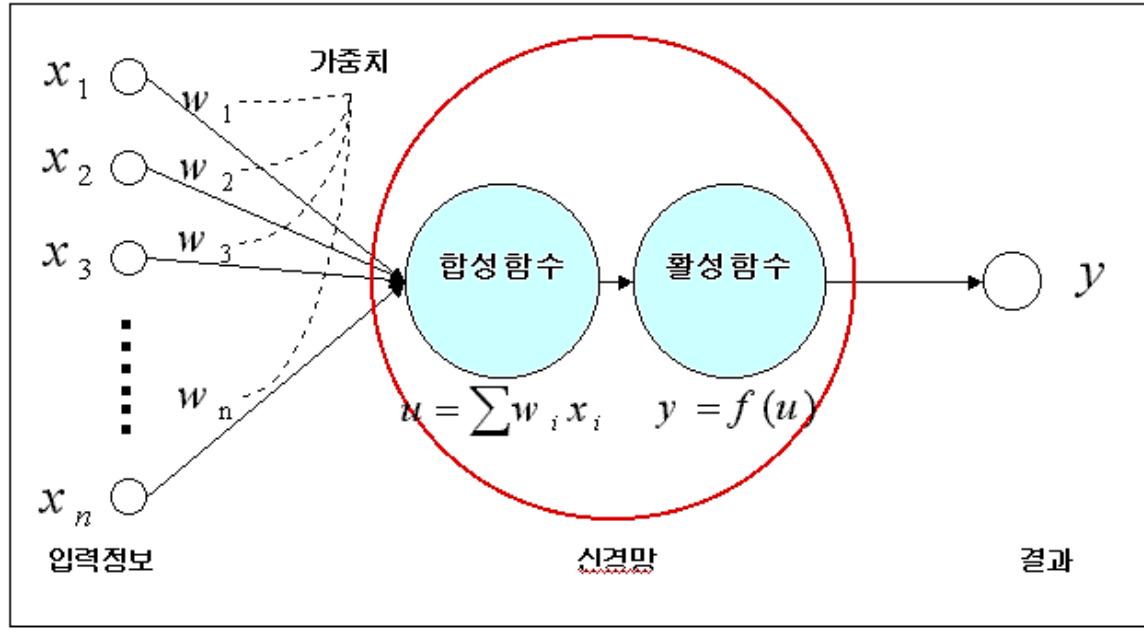
■ 신경망모형

- ✓ 뇌신경계를 모방하여 분류(또는 예측)을 위해 만들어진 모형
- ✓ 빅데이터에 대해 학습(learning 또는 training)을 거쳐 원하는 결과가 나오도록 가중치 조정



2-3. 신경망모형

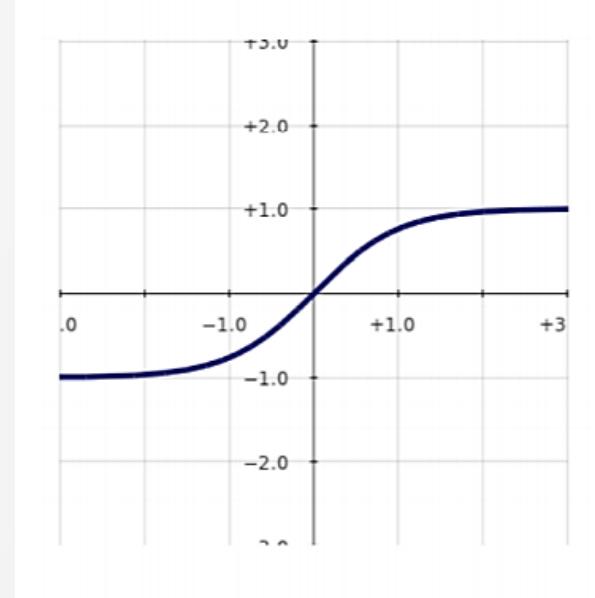
■ 신경망모형



2-3. 신경망모형

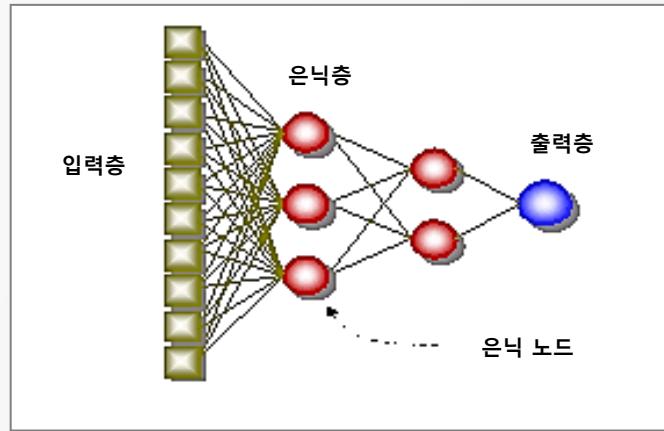
■ 활성함수

- ✓ 부호 함수
- ✓ 계단 함수
- ✓ Sigmoid 함수
- ✓ Softmax 함수
- ✓ Tanh 함수
- ✓ 가우스 함수



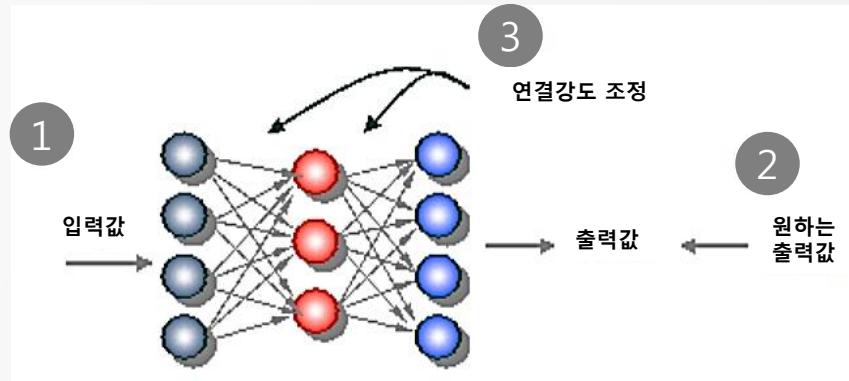
2-3. 신경망모형

■ 신경망모형 : 단층신경망, 다층신경망



2-3. 신경망모형

■ 신경망모형 훈련 : 역전파 알고리즘



2-3. 신경망모형

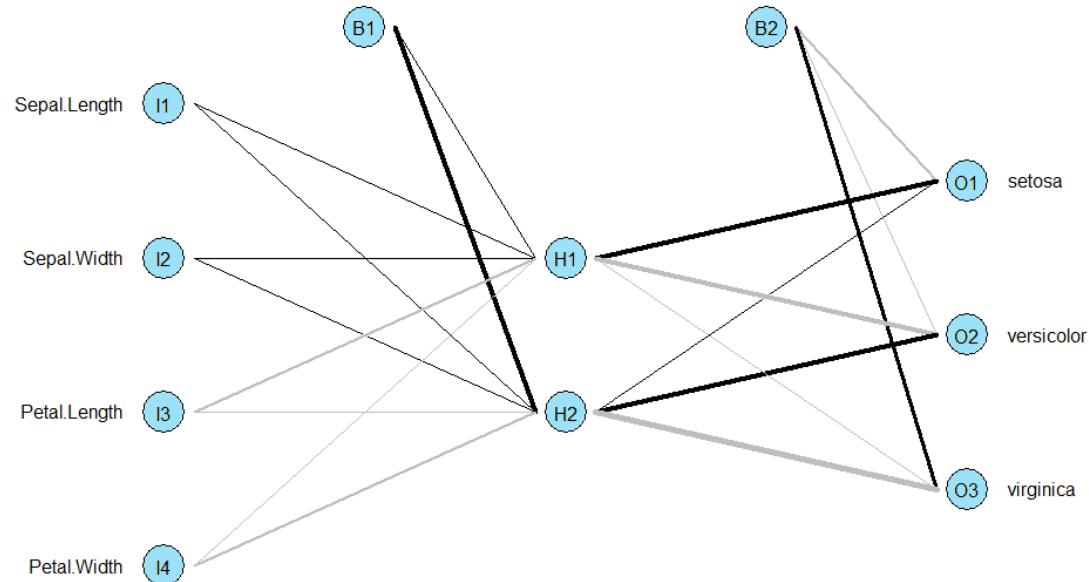
■ 신경망모형 : iris데이터

```
> library(nnet)
> nn.iris <- nnet(Species~., data=iris, size=2, rang=0.1, decay=5e-4, maxit=200)
> summary(nn.iris)

a 4-2-3 network with 19 weights
options were - softmax modelling decay=5e-04
b->h1 i1->h1 i2->h1 i3->h1 i4->h1
  0.48   0.68   1.82  -3.17  -1.52
b->h2 i1->h2 i2->h2 i3->h2 i4->h2
  8.85   0.52   1.40  -1.97  -3.81
b->o1 h1->o1 h2->o1
 -4.80   9.94   1.34
b->o2 h1->o2 h2->o2
 -2.53  -8.91   9.11
b->o3 h1->o3 h2->o3
  7.33  -1.03  -10.45
```

2-3. 신경망모형

■ 신경망모형 : iris데이터



2-3. 신경망모형

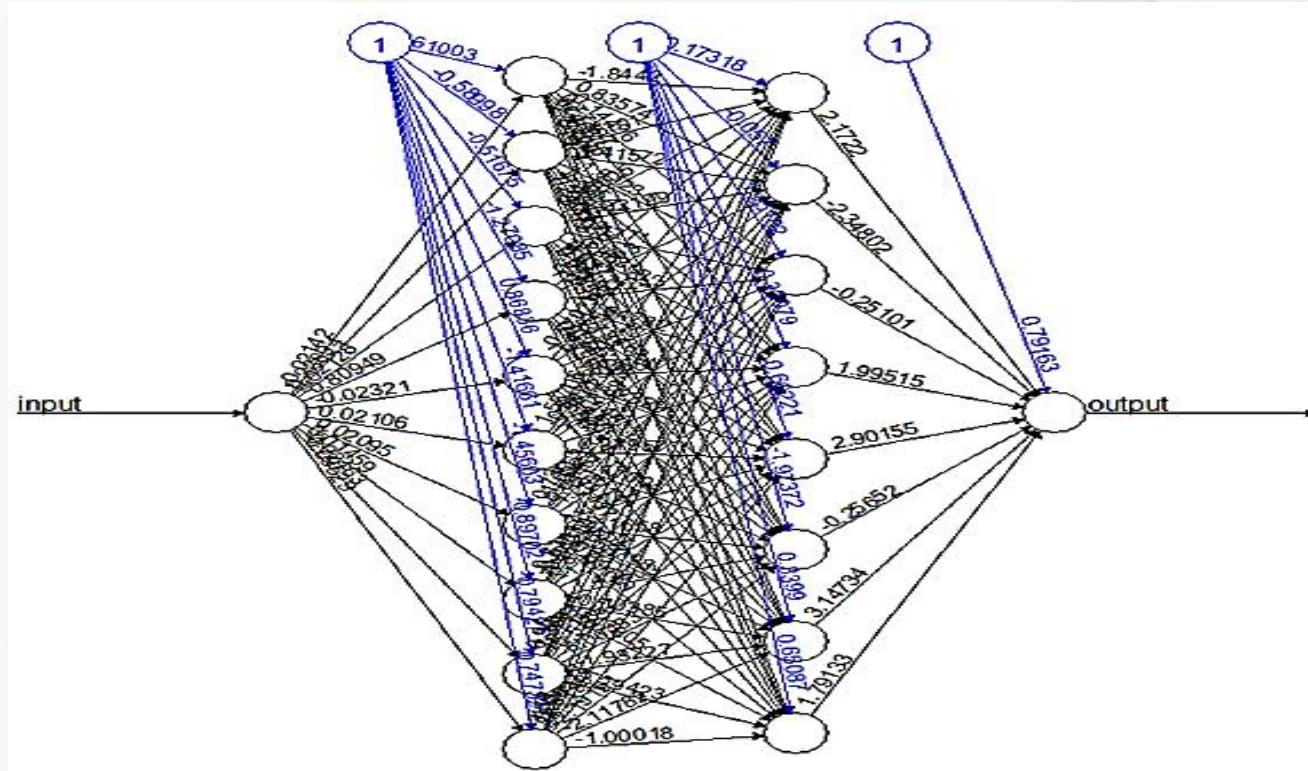
■ 신경망모형

```
> table(iris$Species, predict(nn.iris, iris, type="class"))
```

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	49	1
virginica	0	1	49

2-3. 신경망모형

■ 신경망모형

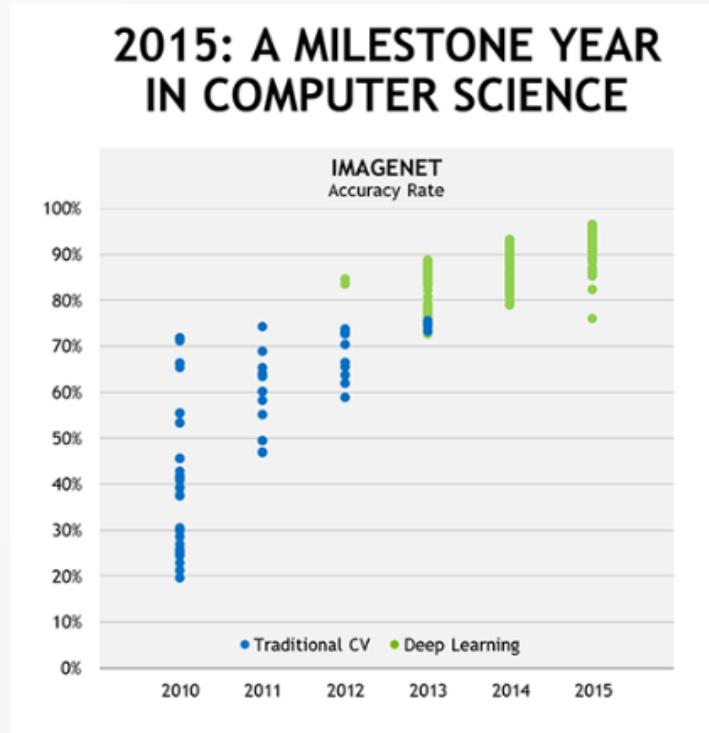


2-3. 신경망모형

- 신경망의 문제 : 추정의 문제(과적합), 수렴의 문제
컴퓨터 성능의 문제, 데이터의 문제

2-3. 신경망모형

■ IMGENET 2012년 대회 : 획기적인 인식률 제고(15.3%)



출처 : <https://blogs.nvidia.com/blog/2016/01/12/accelerating-ai-artificial-intelligence-gpus/>

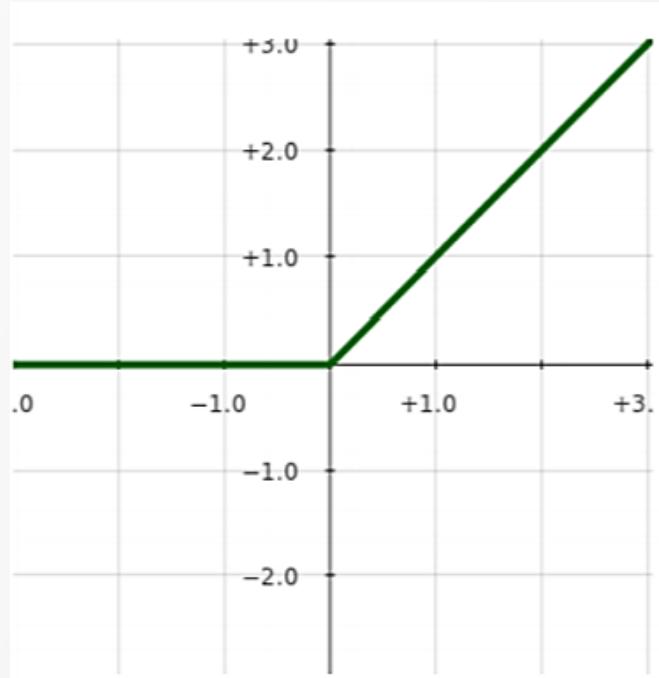
2-3. 신경망모형

■ 딥러닝(Deep Learning) : 신경망의 층수가 10개 이상인 모형

- ✓ 기존 신경망 모형의 알고리즘 문제를 해결 : 2006년 Hinton 교수 등
- ✓ 볼츠만 머신을 도입하여 경사감소소멸현상 해소
- ✓ ReLu 활성함수 도입
- ✓ Dropout 알고리즘 도입
- ✓ GPU 도입, 빅데이터

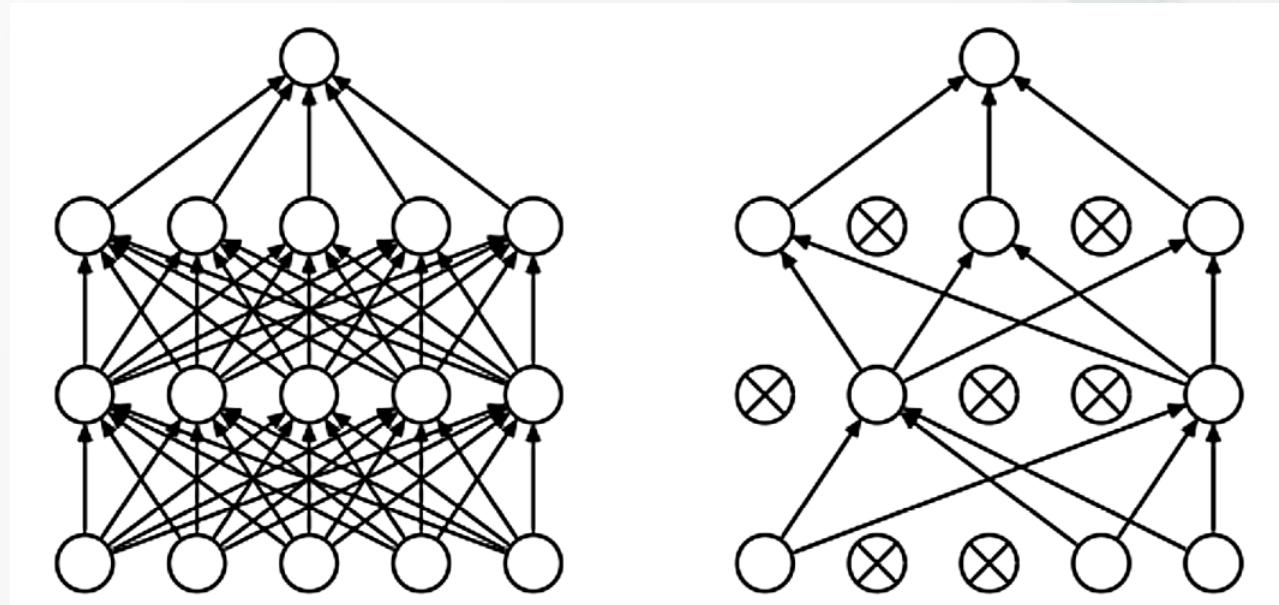
2-3. 신경망모형

■ ReLu 함수



2-3. 신경망모형

■ Dropout



2-4. 앙상블모형

■ 앙상블(ensemble) 모형

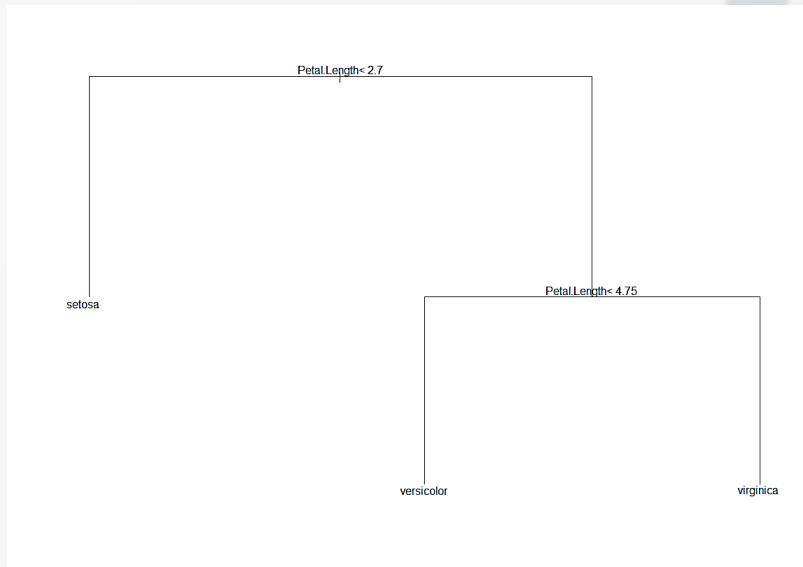
의사결정나무는 학습 데이터의 작은 변화에도 예측 및 분류모형이 크게 변화 → 여러 개의 모형에 의한 결과를 종합하여 분류의 정확도를 높일 필요

- ✓ 배깅(bagging)
- ✓ 부스팅(boosting)
- ✓ 랜덤포레스트 (random forest)

2-4. 앙상블모형

■ 배깅(bagging, bootstrap aggregation)

- ✓ 원 데이터로부터 크기가 같은 표본을 여러 번 단순 임의 복원 추출한 후, 각 표본에 대한 분류기를 생성한 후 그 결과를 결합하는 방법



2-4. 양상블모형

■ 배깅(bagging)

```
> pred <- predict(iris.bagging, newdata=iris)
> table(pred$class, iris[,5])
```

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	5
virginica	0	2	45

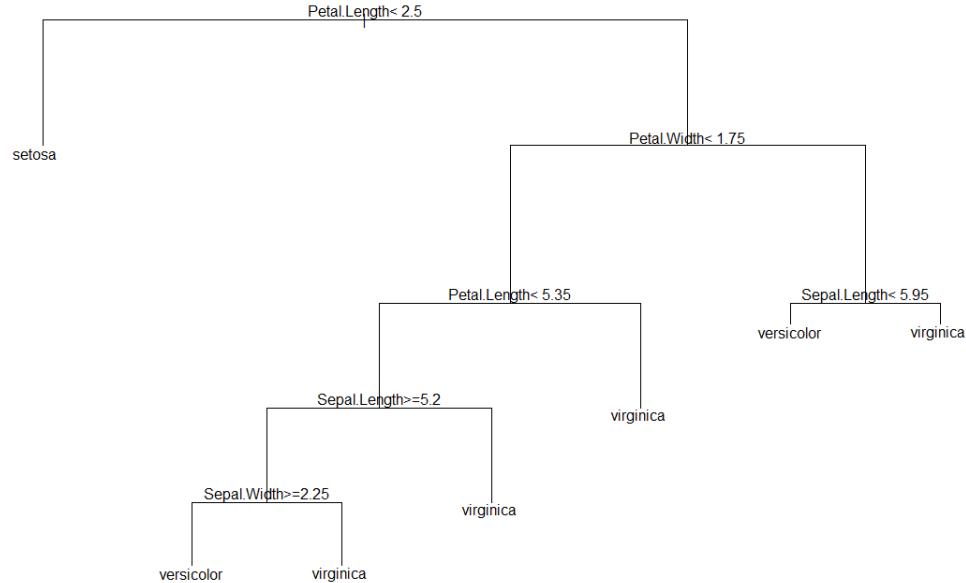
2-4. 양상블모형

■ 부스팅(boosting)

- ✓ 븁스트랩 표본을 구성하는 재표본(re-sampling) 과정에서 분류가 잘못된 데이터에 더 큰 가중을 주어 표본을 추출
- ✓ 아다부스팅(AdaBoosting : adaptive boosting)이 가장 많이 이용되는 알고리즘

2-4. 양상블모형

■ 부스팅(boosting)



2-4. 양상블모형

■ 부스팅(boosting)

```
> pred <- predict(boo.adabag, newdata=iris)  
> table(pred$class, iris[,5])
```

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	50	0
virginica	0	0	50

2-4. 양상블모형

■ 랜덤포레스트(random forest)

- ✓ 배깅에 랜덤 과정을 추가한 방법
- ✓ 원자료로부터 븋스트랩 표본에 대해 의사결정나무를 형성
- ✓ 예측변수들을 임의로 추출, 추출된 변수 내에서 최적의 분할을 만들어 나가는 방법

강의를 마쳤습니다

수고하셨습니다.

13차시 | 빅데이터 분석방법2

이긍희 교수



정형데이터분석(2)

1. 군집분석

1-1. 군집분석 개요

1-2. 계층적 군집화

1-3. k-평균 군집화

1-4. 혼합분포군집

2. 연관분석

1-1. 군집분석 개요

■ 군집과 군집화

군집 : 구성개체들이 유사한 집단

군집화 : 자료를 몇 개의 부분 집단으로 나누는 과정 → 자율학습

- ▶ 군집화의 효과 : 전체 자료를 간단히 요약
- ▶ 이용사례 : 타겟 마케팅

1-1. 군집분석 개요

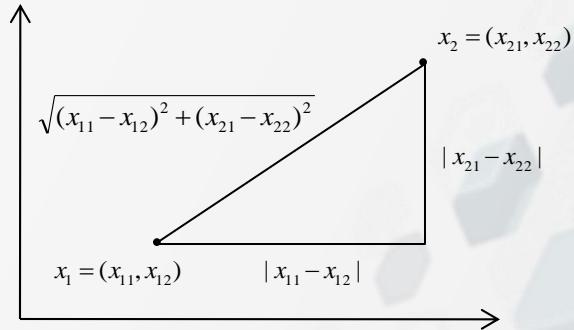
- 군집화를 위한 자료구조
 - ✓ 3개 개체 2개 변수

$$\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix}$$

1-1. 군집분석 개요

■ 두 개체간 유사성(거리)의 계산

✓ 두 개체 : $x_1 = (x_{11}, x_{12}), x_2 = (x_{21}, x_{22})$



✓ 두 개체 : $x_i = (x_{i1}, x_{i2}, \dots, x_{ip}),$
 $x_k = (x_{k1}, x_{k2}, \dots, x_{kp})$

✓ 유클리디안 거리 : $d = (x_i, x_k) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2}$

1-1. 군집분석 개요

■ 두 개체간 유사성(거리)의 계산 (계속)

- ✓ 맨해튼 거리: $d = (x_i, x_k) = \sum_{j=1}^p |x_{ij} - x_{kj}|$
- ✓ 민코브스키 거리: $d = (x_i, x_k) = \left(\sum_{j=1}^p |x_{ij} - x_{kj}|^m \right)^{1/m}$
- ✓ 표준화 거리, 마할라노비스 거리
- ✓ 비유사성 행렬 : 개체들간 거리를 행렬 형태로 표현

1-1. 군집분석 개요

■ 변수간 군집

✓ 상관계수

- › ±1에 가까우면 유사성 높음
- › 0에 가까우면 유사성 낮음

✓ 거리의 계산 : $d(x, y) = 1 - |corr(x, y)|$

1-2. 계층적 군집화

■ 군집화의 종류

계층적 군집화

최단 연결법, 최장연결법,
평균연결법, 와드 연결법

비계층적 군집화

k-평균 군집화, 혼합군집분포

1-2. 계층적 군집화

■ 알고리즘

- ① 개체의 개수 만큼 군집을 형성한다.
- ② 가장 가까운 두 개체를 하나의 군집으로 합친다.
- ③ ②의 과정을 반복하여 군집을 계속 합친다.

1-2. 계층적 군집화

■ 나무형 그림(dendrogram) :

계층적 군집화에서 군집화의 과정을 나무와 같은 형태로 나타내면 전체군집화 과정을 쉽게 이해할 수 있도록 하는 그림

1-2. 계층적 군집화

최단(단일)연결법

두 군집 거리 중에서 가장 작은 거리를 두 군집간의 거리로 정함

최장(완전)연결법

두 군집 거리 중에서 가장 긴 거리를 두 군집간의 거리로 정함

평균연결법

단일연결법과 완전연결법을 절충한 방법

와드연결법

군집내 오차 제곱합에 기초하여 군집 형성

1-2. 계층적 군집화

■ 계층적 군집화 예 : 최단 연결법, 최장 연결법

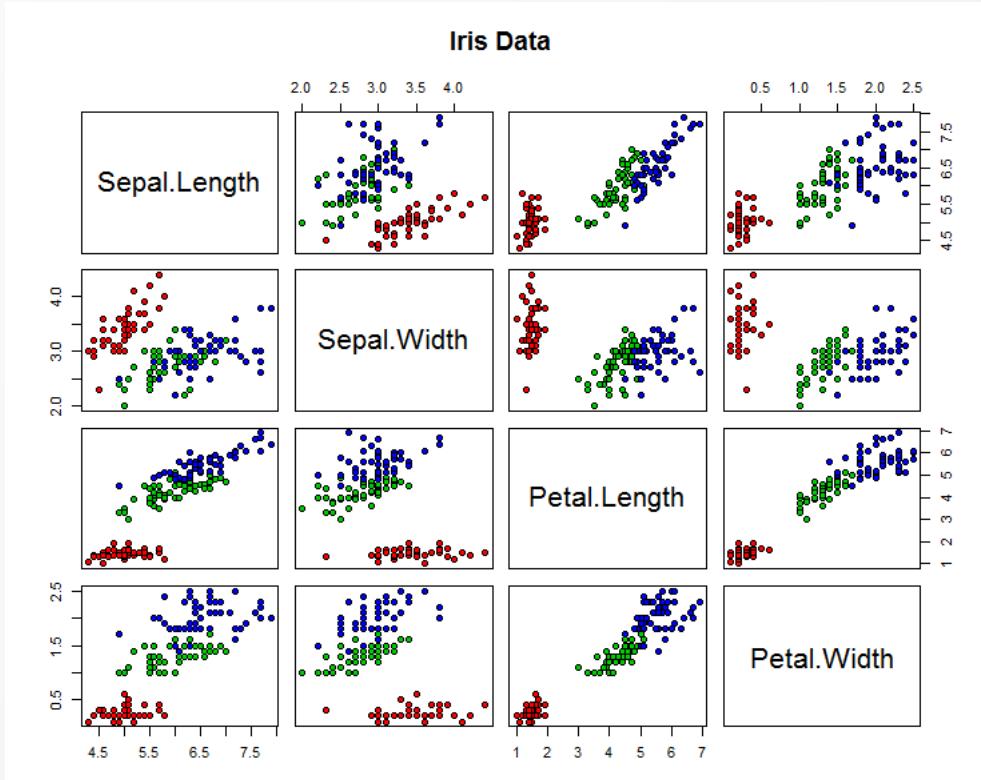
0				
7	0			
1	6	0		
9	3	8	0	
8	5	7	4	0

0				
6	0			
8	3	0		
7	5	4	0	

0				
7	0			
9	3	0		
8	5	4	0	

1-2. 계층적 군집화

■ iris 데이터



Setosa



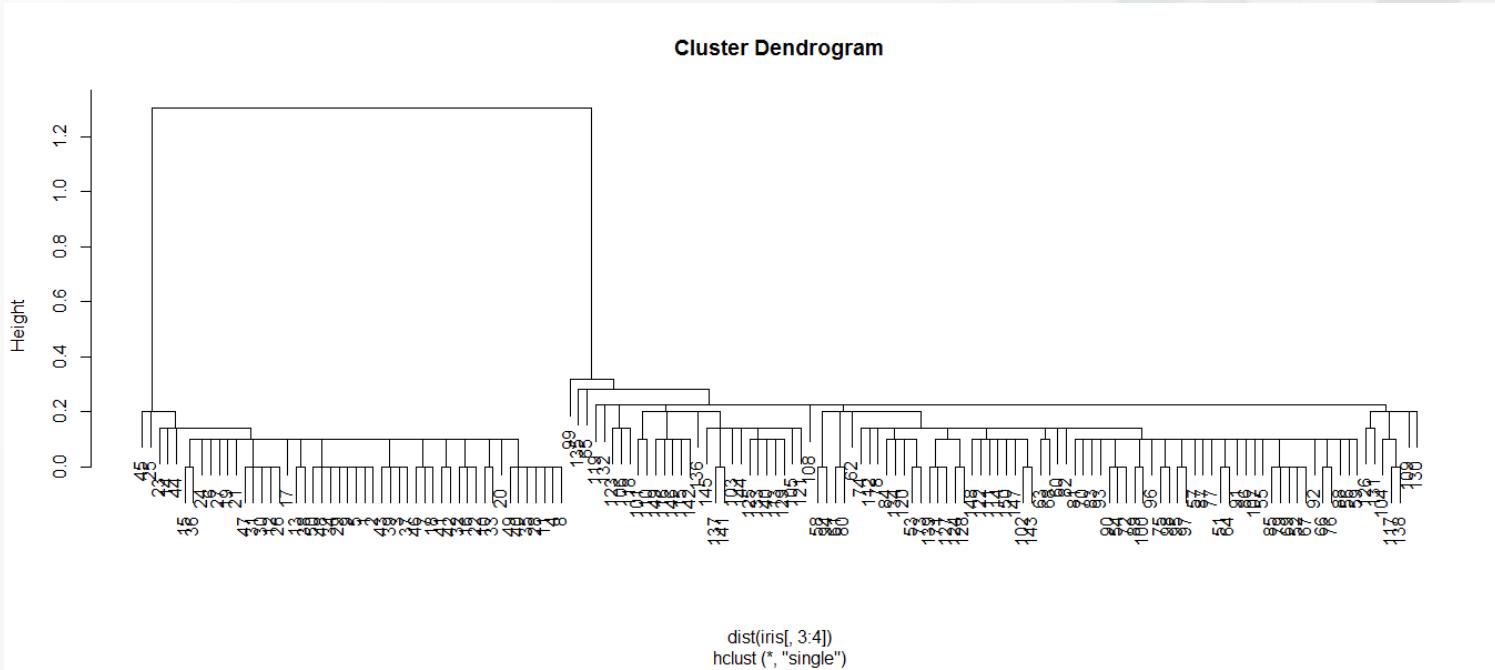
Versicolor



Virginica

1-2. 계층적 군집화

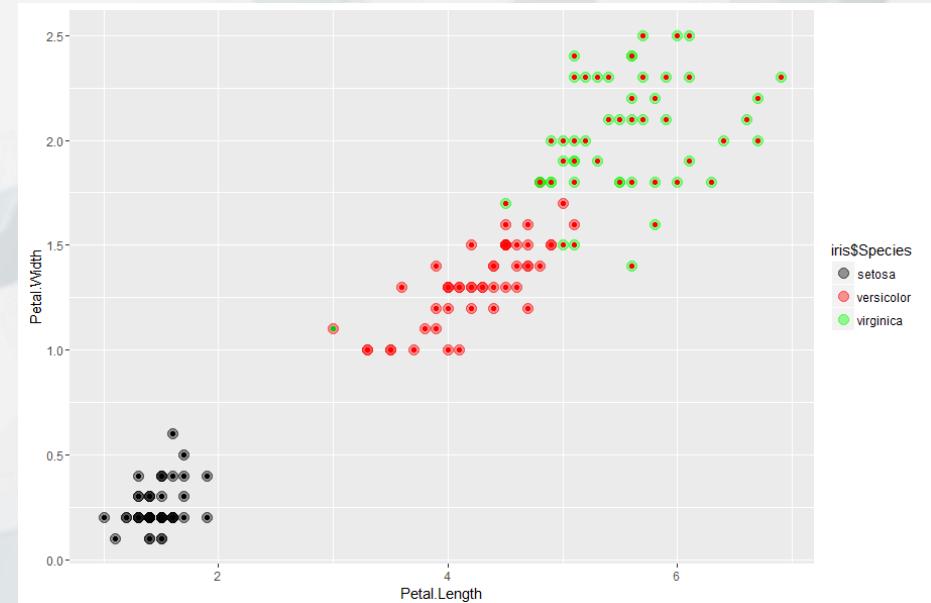
■ iris 데이터 (최단연결법)



1-2. 계층적 군집화

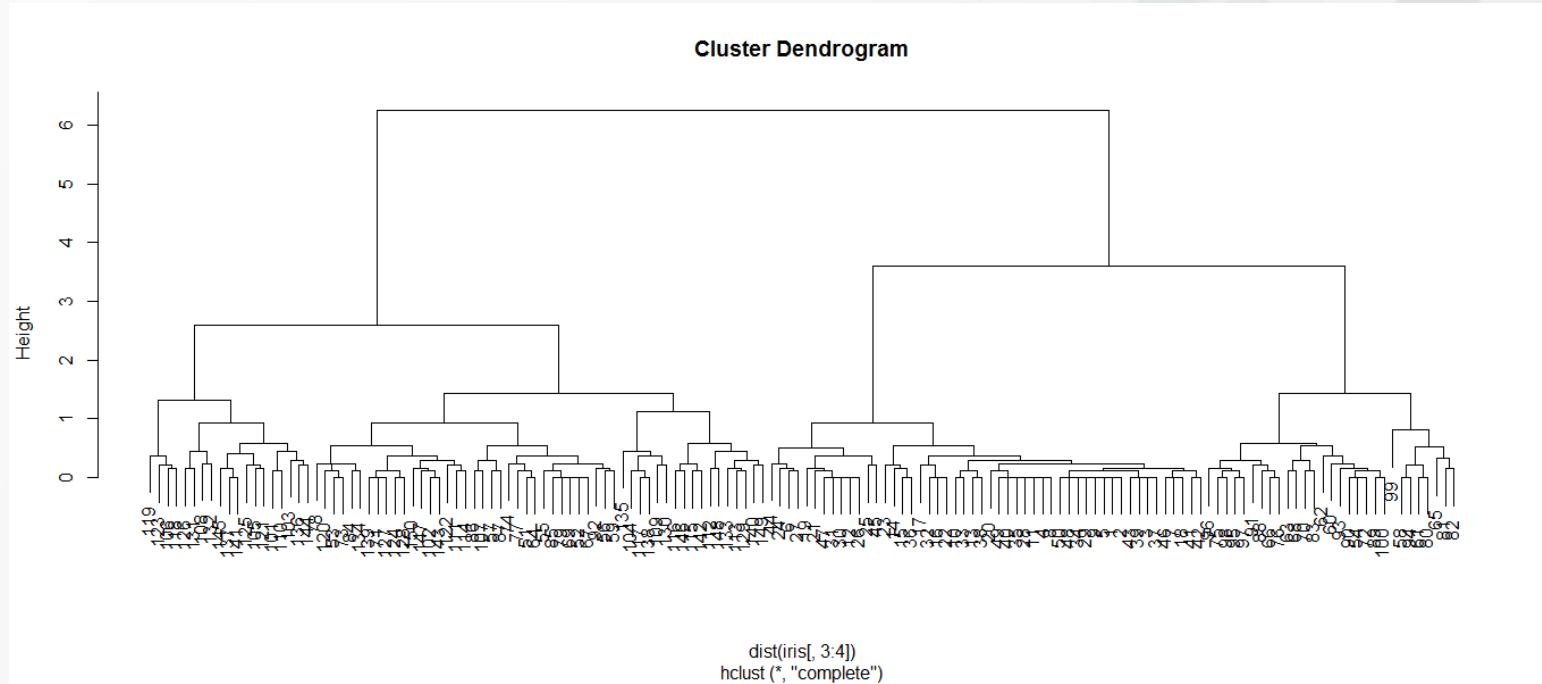
■ iris 데이터 (최단연결법)

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	49	50
Virgnica	0	1	0



1-2. 계층적 군집화

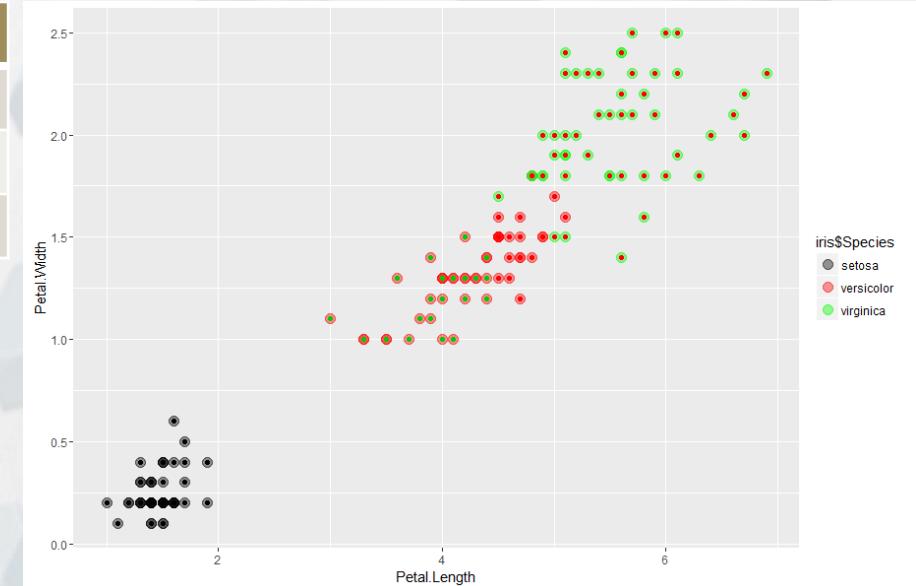
■ iris 데이터 (최장연결법)



1-2. 계층적 군집화

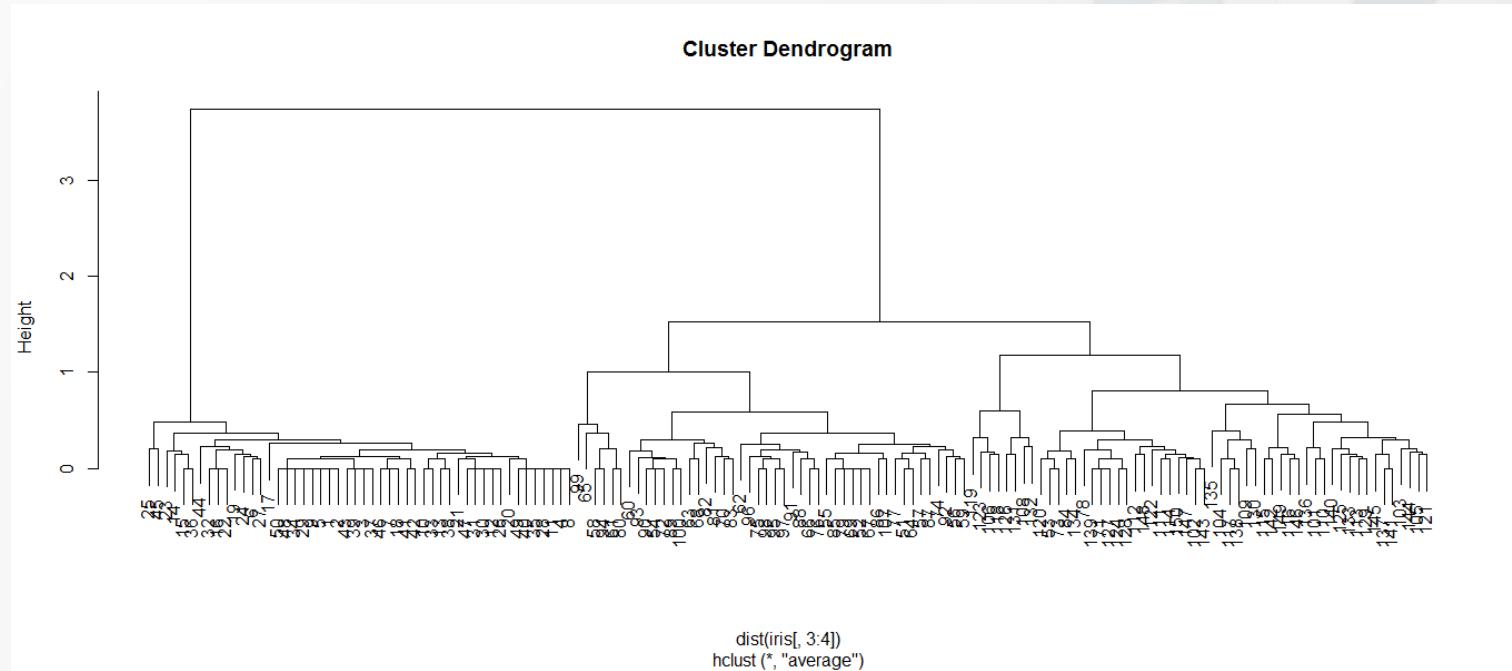
■ iris 데이터 (최장연결법)

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	21	50
Virginica	0	29	0



1-2. 계층적 군집화

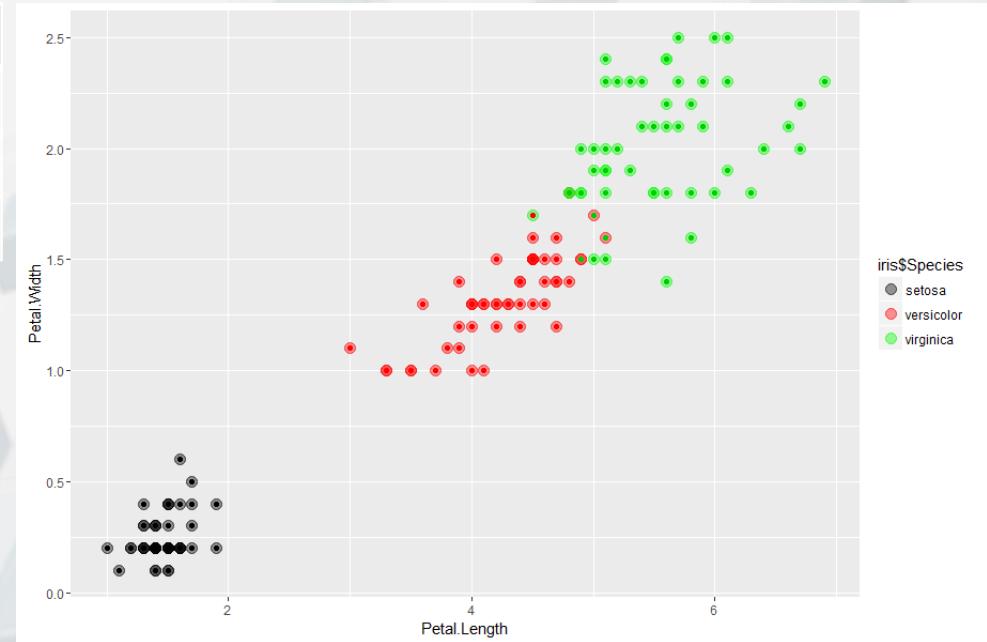
■ iris 데이터 (평균연결법)



1-2. 계층적 군집화

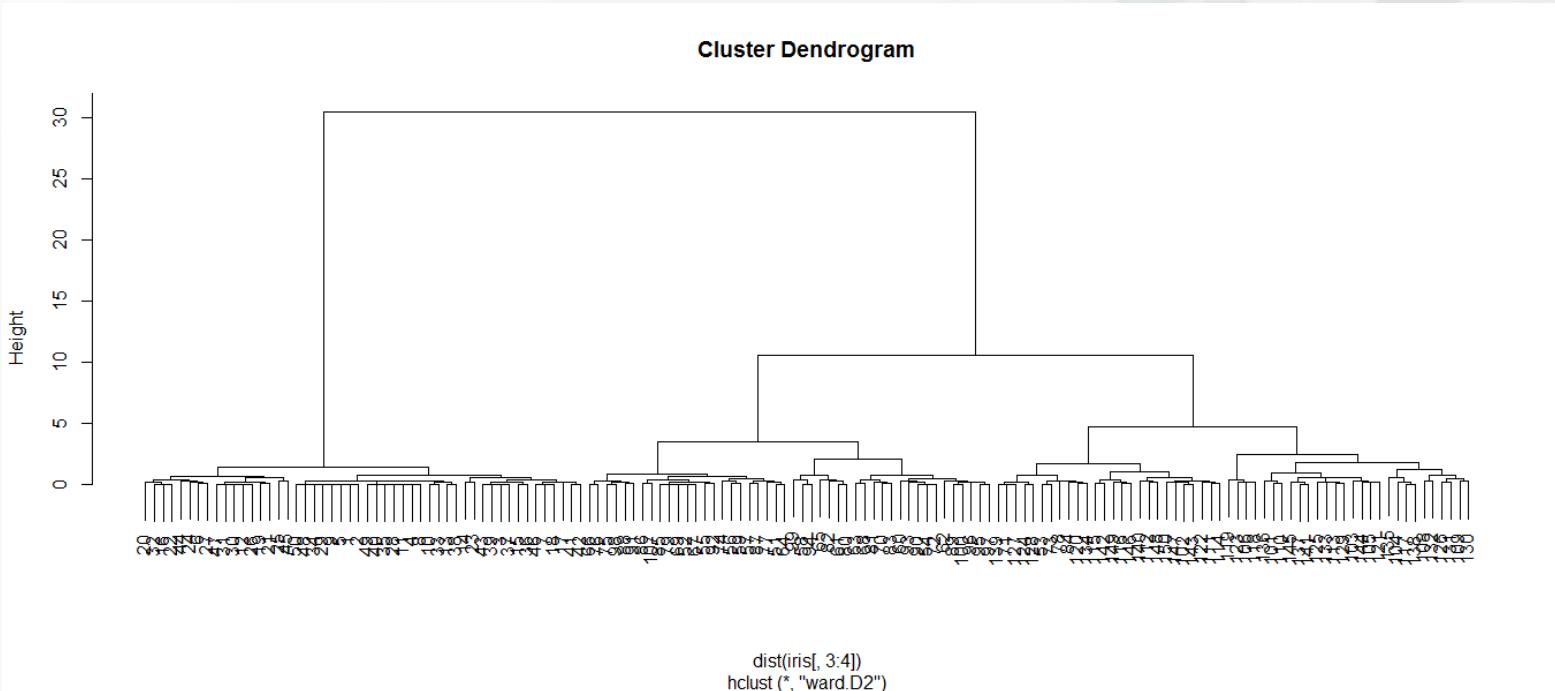
■ iris 데이터 (평균연결법)

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	45	1
Virginica	0	5	49



1-2. 계층적 군집화

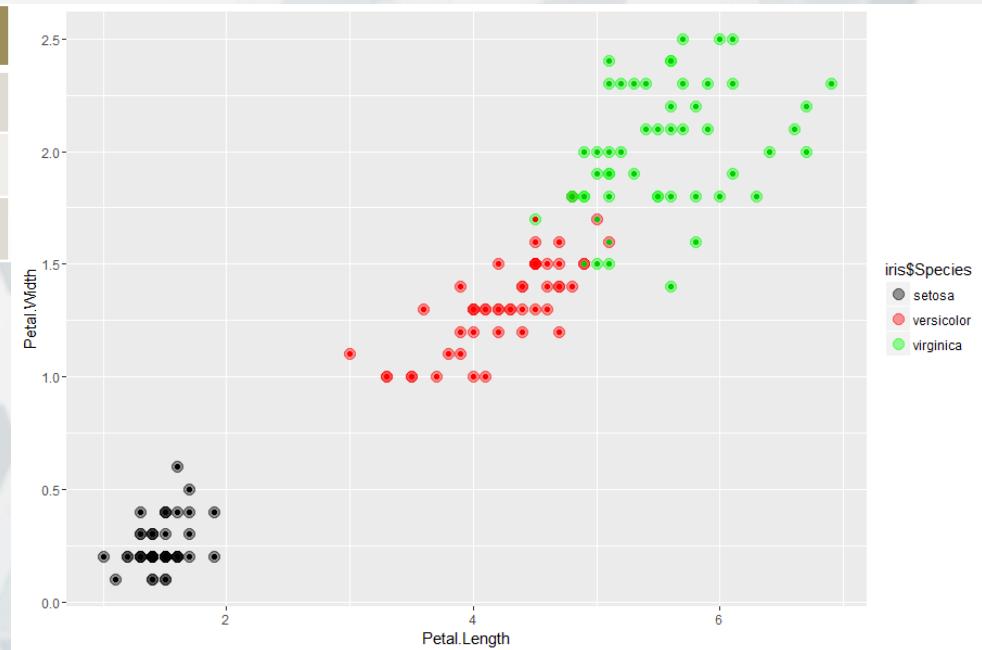
■ iris 데이터 (Ward 연결법)



1-2. 계층적 군집화

■ iris 데이터 (Ward 연결법)

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	45	1
Virginica	0	5	49



1-3. k-평균 군집화

- 비계층적 군집화는 군집수를 정하고 이에 대하여 군집화를 수행
- K-평균 방법, PAM (Partitioning Around Medoids)

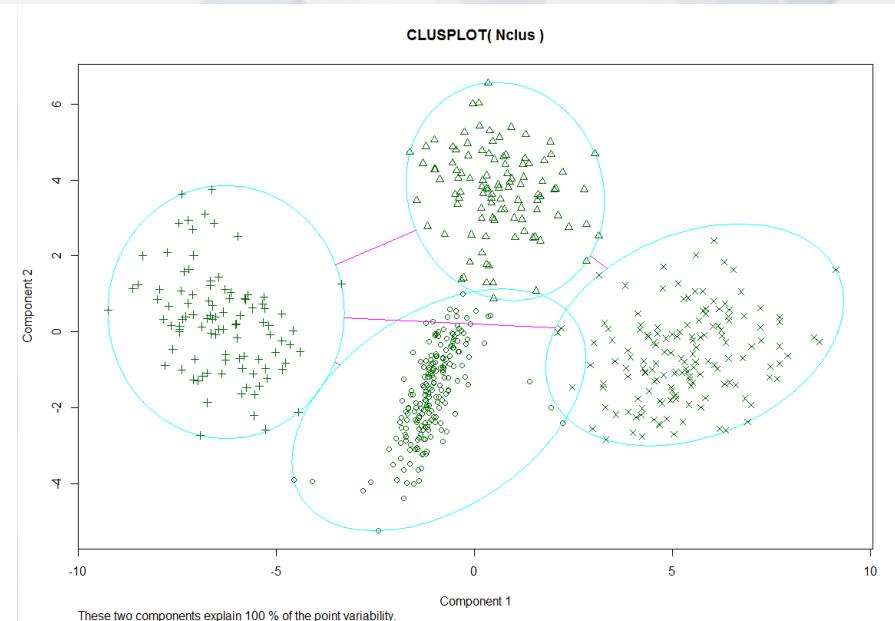
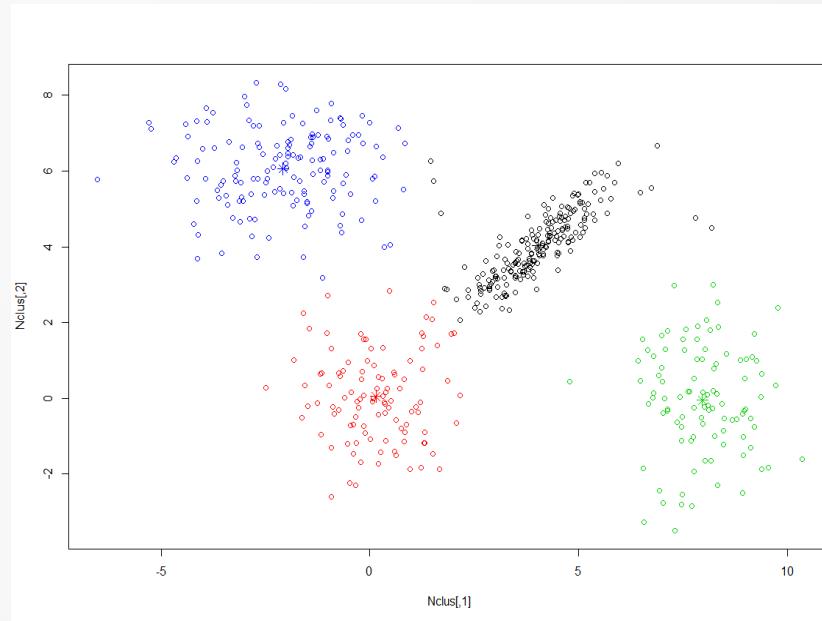
1-3. k-평균 군집화

■ k-평균 알고리즘

- ✓ 군집수 k 를 미리 정하고 중심점으로부터 거리를 계산하여 군집을 구하는 방법
- ✓ 알고리즘
 - ① 각 군집에 대하여 중심점의 초기값을 구함
 - ② 관측값에서 k 개의 중심점까지의 거리를 계산
→ 중심점이 가까운 군집으로 관측값을 재할당
 - ③ 군집의 변화가 없을 때까지 ②를 수행

1-3. k-평균 군집화

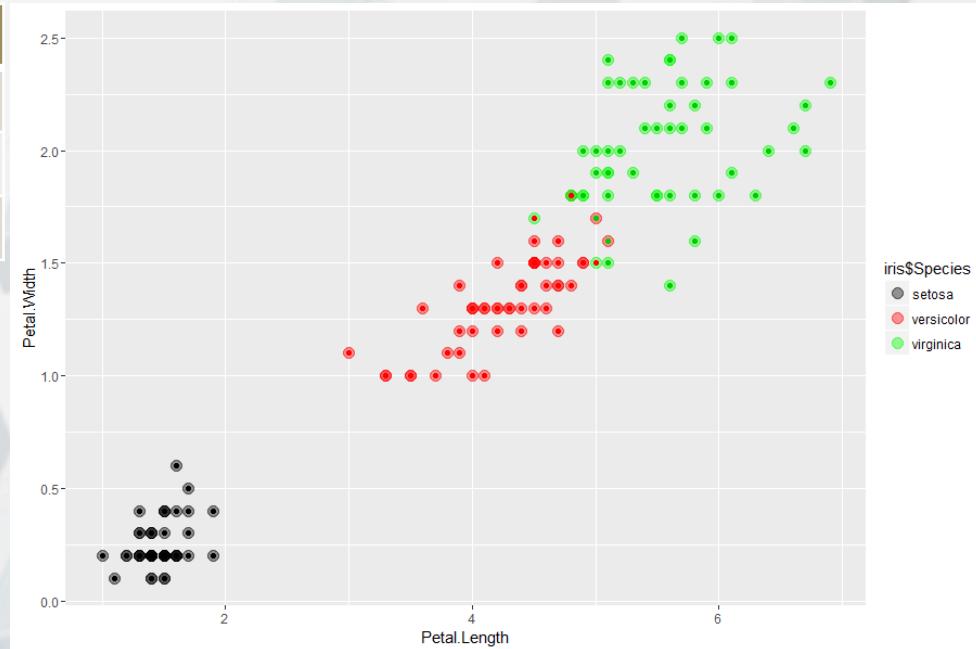
■ 4개의 2변량 정규분포로부터 발생된 난수



1-3. k-평균 군집화

■ iris 데이터 (k-평균 연결법)

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	4
Virginica	0	2	46



1-3. k-평균 군집화

■ 새 개체의 군집할당

✓ 계층적 군집화

- › 새 개체과 다른 군집 간 거리를 계산 → 가장 가까운 군집에 할당

✓ 비계층적 군집화

- › 중심점부터의 거리를 계산 → 가장 가까운 중심점의 군집으로 할당

1-3. k-평균 군집화

■ 군집화의 비교

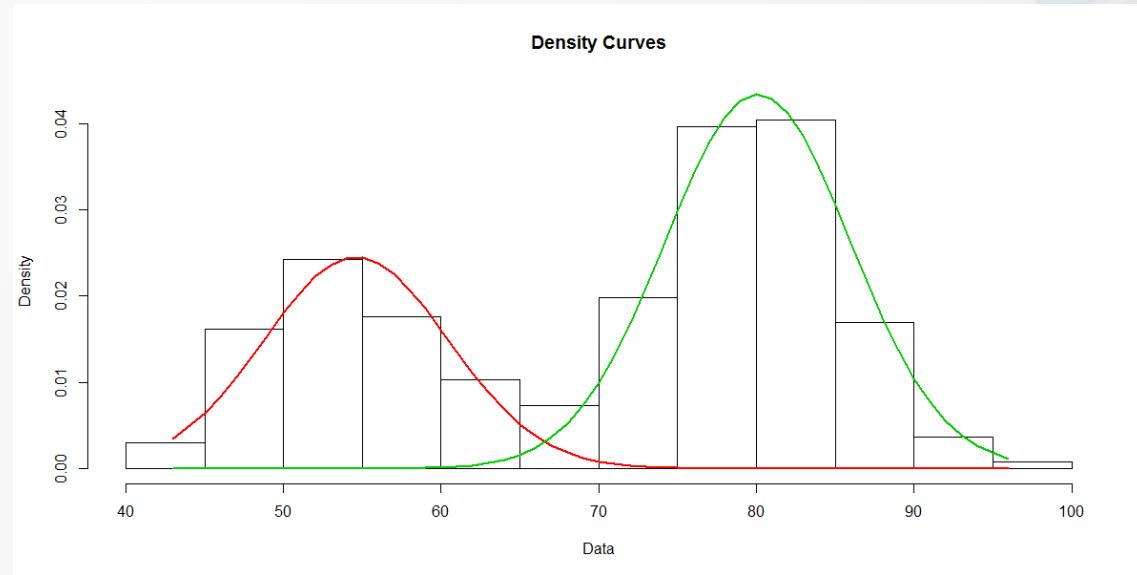
	장점	단점
계층적 군집화	간단, 이해하기 쉬움	군집을 수정할 수 없음
비계층적 군집화	계산이 빠르고 안정적	초기값 및 군집수 결정이 어려움

1-4. 혼합분포군집

- 모형기반의 군집방법으로 데이터가 k개의 모수적 모형에서 나왔다는 가정하에 모수와 가중치를 데이터로부터 추정하는 방법
- 추정방법으로 EM알고리즘이 이용됨
 - ① E-단계 : 잠재변수의 기댓값 계산
 - ② M-단계 : 잠재변수의 기댓값을 이용하여 모수 추정

1-4. 혼합분포군집

■ faithful 데이터의 분류



1-4. 혼합분포군집

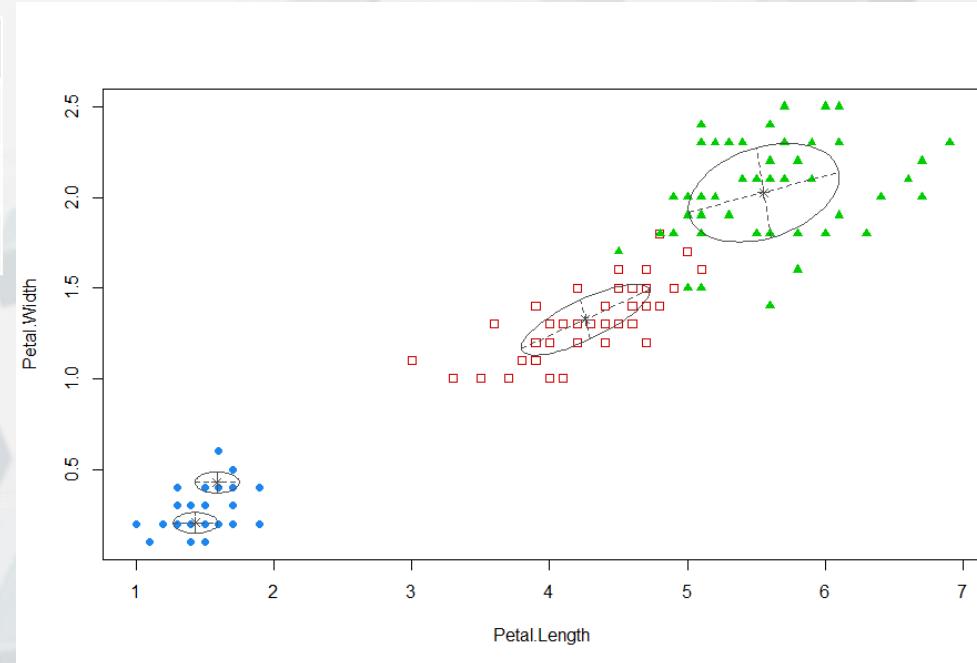
■ 혼합분포군집모형의 특징

- ✓ k-평균 군집과 유사하나 확률분포를 도입하여 군집을 수행하는 모형기반 군집방법

1-4. 혼합분포군집

■ iris 데이터의 분류

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	49	2
Virginica	0	1	48



2. 연관분석

- 고객이 상품을 구매한 기록이 계산대의 바코드(bar code)를 통해 해당 기업 데이터베이스에 쌓이게 되며 질문이 생김
 - ① 고객들은 어떤 상품을 동시에 구매하게 되는가?
 - ② A을 구매한 고객은 다른 어떤 상품을 주로 구매하는가?
→ 질문에 대한 답을 찾는 방법 : 연관규칙, 추천시스템

2. 연관분석

■ 연관규칙 : 특정 상품 구매가 얼마나 자주 동시에 발생하는가를 표현하는 규칙 또는 조건을 의미

- ✓ (예) 고객이 와인을 구매한다면 고객의 15%가 치즈를 구매한다.
컴퓨터를 구매하면 마우스와 키보드를 산다.

2. 연관분석

■ 편의점에서의 거래내역(transaction) 예

고객 번호	품목
1	오렌지쥬스, 사이다
2	우유, 오렌지쥬스, 식기세척제
3	오렌지쥬스, 세제
4	오렌지쥬스, 세제, 사이다
5	식기세척제, 사이다

2. 연관분석

■ 편의점 동시구매표 : “사이다를 구입하는 고객은 오렌지쥬스를 산다”

	오렌지쥬스	식기세척제	우유	사이다	세제
오렌지쥬스	4	1	1	2	2
식기세척제	1	2	1	1	0
우유	1	1	1	0	0
사이다	2	1	0	3	1
세제	2	0	0	1	2

2. 연관분석

■ 연관규칙은 자율학습(unsupervised learning) 분석방법

- ✓ 고객이 구매하는 품목 사이에 연관이 있는지를 파악하여 규칙을 도출하지만 목적변수(target variable) 없이 분석

2. 연관분석

■ 연관규칙 ("if X, then Y")의 지지도와 신뢰도

- ✓ 연관규칙 의 지지도 : 전체 거래들 중 품목 X와 품목 Y를 동시에 포함하는 거래의 비율

$$P(X \text{ I } Y) = \frac{X \text{와 } Y \text{가 동시에 포함된 거래수}}{\text{전체 거래수}}$$

- ✓ 연관규칙 의 신뢰도 : 품목 X를 포함하는 거래들 중 품목 Y를 포함하는 거래의 비율

$$\frac{P(X \text{ I } Y)}{P(X)} = \frac{\text{품목 } X \text{와 } Y \text{ 동시에 포함 거래수}}{\text{품목 } X \text{ 거래수}}$$

2. 연관분석

■ 편의점 거래내역 예

- ✓ 연관규칙 “오렌지쥬스를 구매하면 사이다를 구매한다”의
지지도와 신뢰도
- ✓ 연관규칙 “우유와 오렌지쥬스를 사면 식기세척제를 산다”의
지지도와 신뢰도

고객 번호	품목
1	오렌지쥬스, 사이다
2	우유, 오렌지쥬스, 식기세척제
3	오렌지쥬스, 세제
4	오렌지쥬스, 세제, 사이다
5	식기세척제, 사이다

2. 연관분석

■ 지지도와 신뢰도만으로는 유용한 규칙인지 판단하기 어려움

→ 향상도를 고려

✓ 연관규칙 $X \Rightarrow Y$ 의 향상도

$$\frac{P(X \text{ I } Y)}{P(X)P(Y)} = \frac{P(Y|X)}{\text{전체거래수}} = \frac{\text{신뢰도}}{P(Y)}$$

- ✓ 품목 X가 주어지지 않았을 때 품목 Y의 확률 대비
품목 X가 주어졌을 때의 품목 Y의 확률의 증가 비율
 - › 값이 클수록 X의 구매여부가 Y의 구매 여부에 큰 영향을 미침

2. 연관분석

■ 연관규칙 $X \Rightarrow Y$ 의 향상도

- ✓ 향상도 = 1 : 품목 X와 품목 Y의 구매가 상호 관련이 없는 경우
→ 두 품목이 서로 독립적인 관계
- ✓ 향상도 > 1 : 결과를 예측하는데 있어서 규칙은 우연적 기회보다 우수
(양의 상관관계)
- ✓ 향상도 < 1 : 규칙은 우연적 기회보다 나쁘다는 의미 (음의 상관관계)

2. 연관분석

- 연관분석 Apriori 알고리즘 : 최소 지지도보다 큰 집합을 대상으로 높은 지지도를 가지는 품목집합을 찾는 것
 - ① 최소 지지도를 설정
 - ② 개별 품목중 최소 지지도를 넘는 모든 품목을 찾음
 - ③ ②에서 찾은 품목에서 최소 지지도를 넘는 2가지 품목 집합을 찾음
 - ④ ②,③에서 찾은 품목 집합을 결합하여 최소 지지도를 넘는 3가지 품목집합을 찾음
 - ⑤ 이상을 반복하여 최소 지지도가 넘는 빈발 품목을 찾음

2. 연관분석

■ 연관규칙의 장점 : 다른 데이터 마이닝 도구에 비하여 사용하기 쉽고 결과가 명확하며 이해하기가 쉬움

- ✓ 결과를 “만약 ~ 이면, ~이다”라는 형태로 나오므로 다른 데이터 마이닝 방법의 결과보다 쉽게 이해할 수 있음
- ✓ 알고리즘이 신경망 등 다른 알고리즘에 비교하여 단순하고 간단

2. 연관분석

- 연관규칙의 단점 : 품목의 수가 증가함에 따라 계산량이 매우 증가
 - ✓ 연속형 변수 등에서 연관규칙을 구하기 어려움
 - ✓ 항목 수를 정하기 어렵고 거래가 드문 품목에 대한 규칙을 찾기 어려움

2. 연관분석

```
> summary(Groceries)
```

transactions as itemMatrix in sparse format with
9835 rows (elements/itemsets/transactions) and
169 columns (items) and a density of 0.02609146

most frequent items:

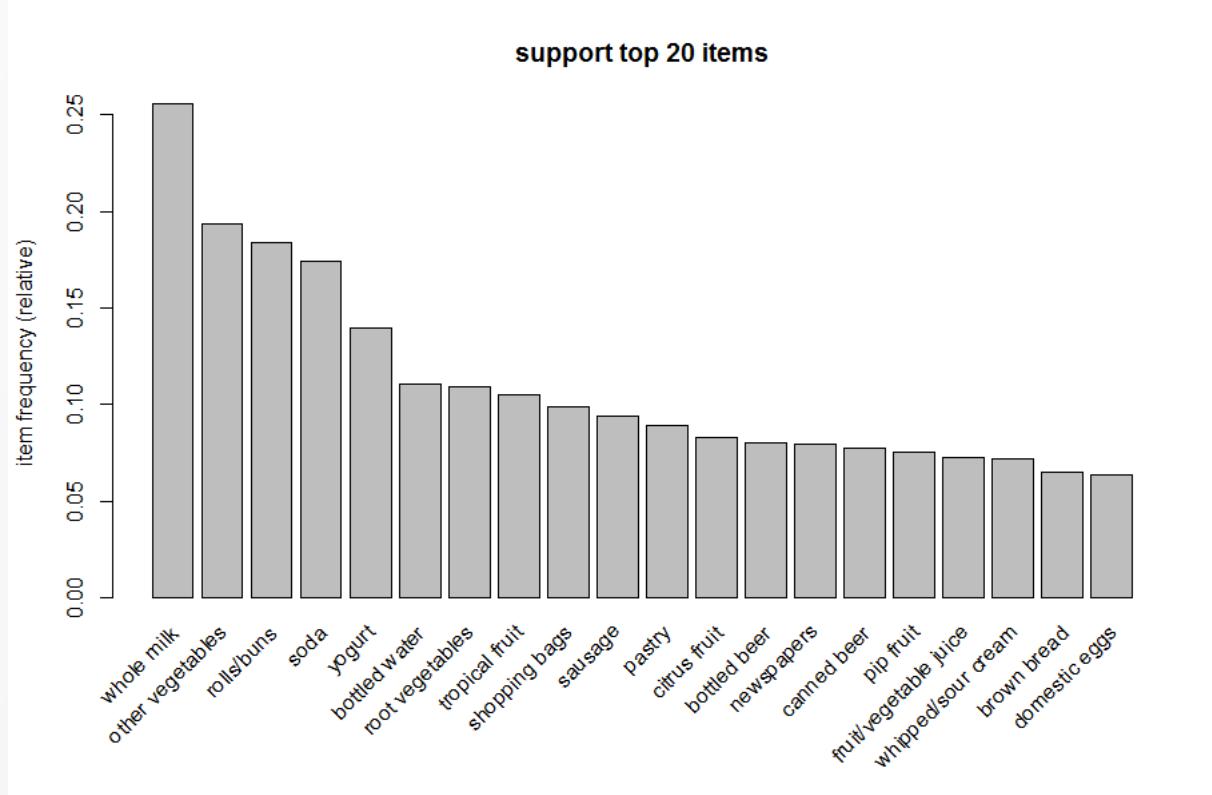
	whole milk	other vegetables
	2513	1903
rolls/buns		soda
	1809	1715
yogurt		(Other)
	1372	34055

```
> inspect(Groceries[1:5])
```

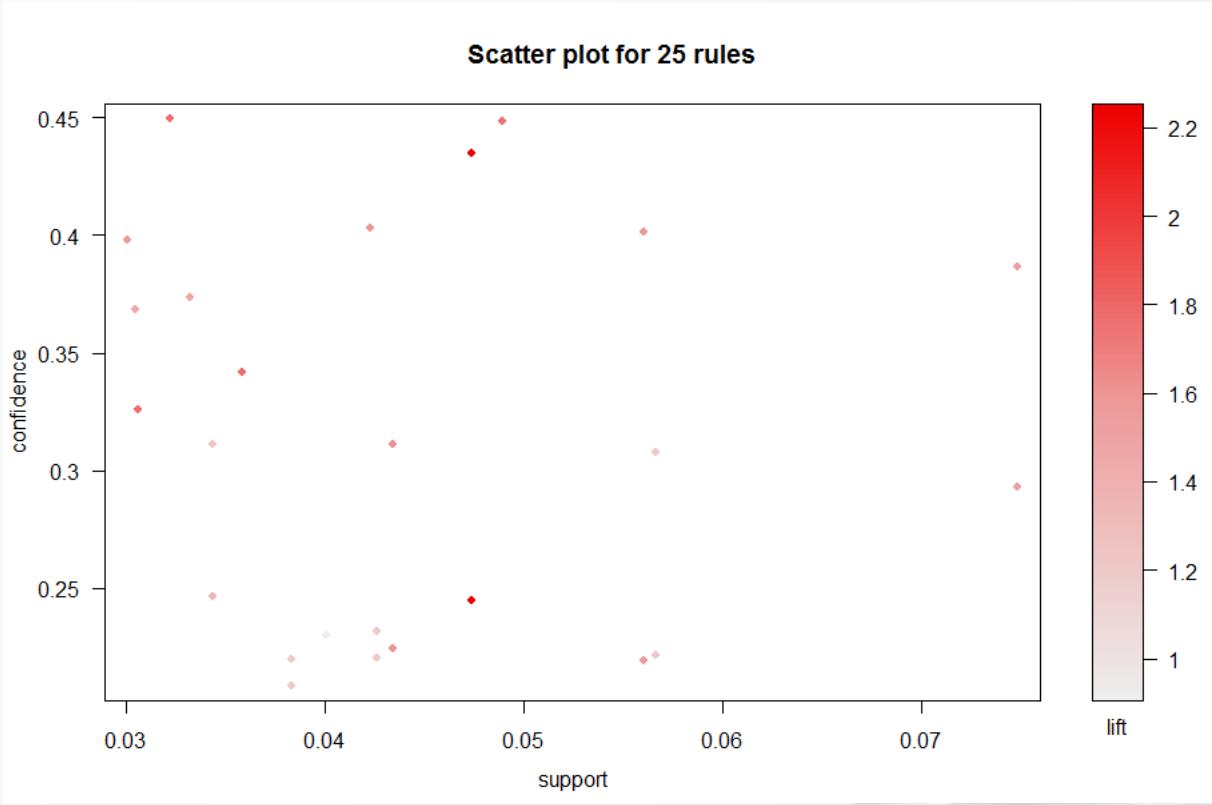
items

- [1] {citrus fruit,
semi-finished bread,
margarine,
ready soups}
- [2] {tropical fruit,
yogurt,
coffee}
- [3] {whole milk}
- [4] {pip fruit,
yogurt,
cream cheese ,
meat spreads}
- [5] {other vegetables,
whole milk,
condensed milk,
long life bakery product}

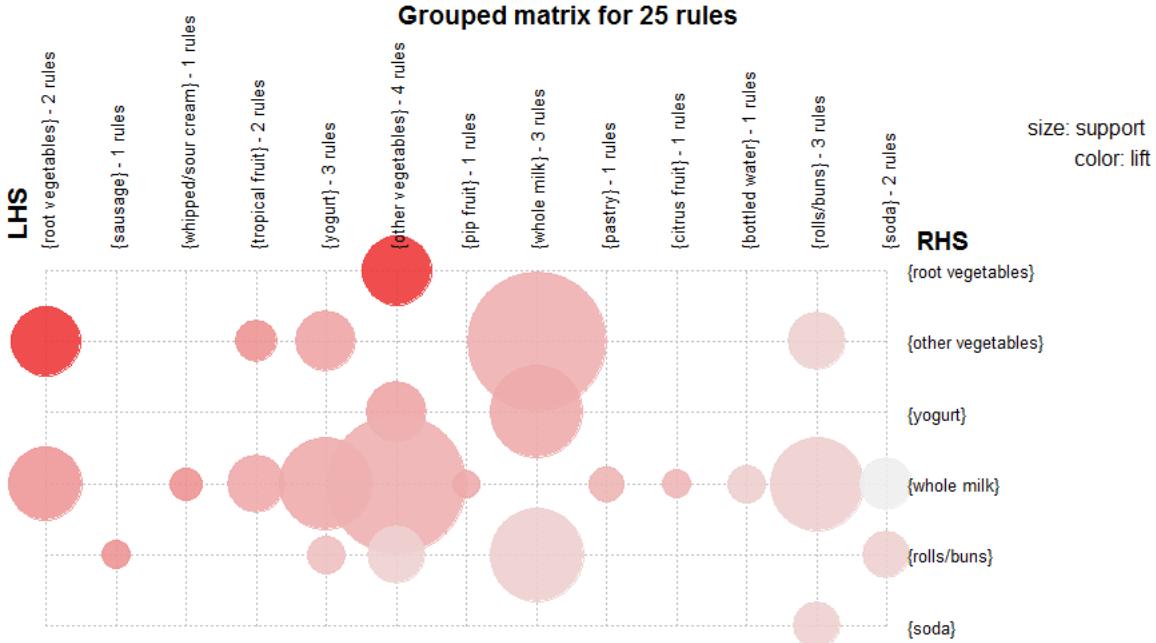
2. 연관분석



2. 연관분석



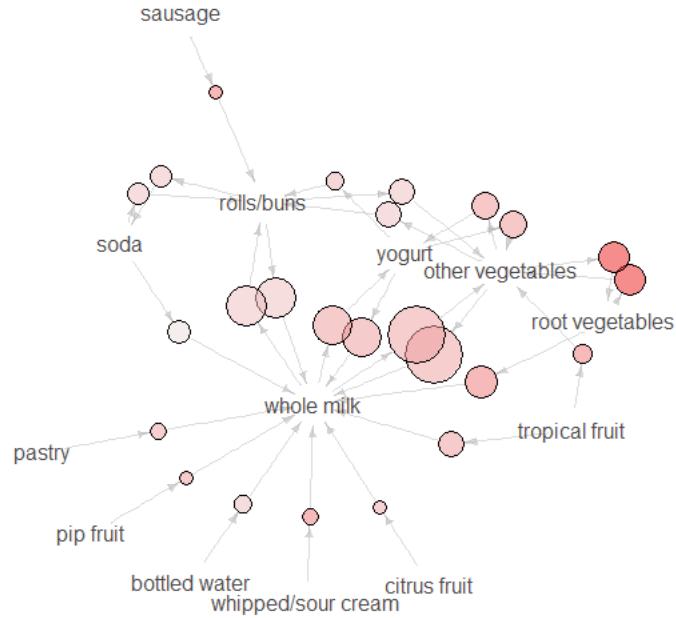
2. 연관분석



2. 연관분석

Graph for 25 rules

size: support (0.03 - 0.075)
color: lift (0.899 - 2.247)



강의를 마쳤습니다

수고하셨습니다.

14차시 | 빅데이터의 활용

이긍희 교수



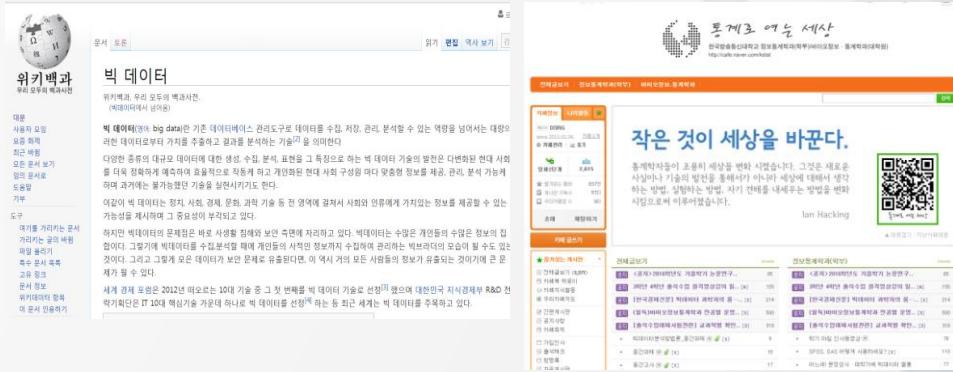
빅데이터의 활용

1. 텍스트 마이닝
2. 소셜네트워크분석
3. 모형의 평가

1. 텍스트 마이닝

■ 텍스트마이닝의 정의 : 다양한 형식의 문서로부터 자연어로 구성된 데이터를 획득한 후 이들을 데이터마이닝 방법을 적용해 데이터의 패턴 등을 추출하여 의미있는 정보를 얻는 방법

- ✓ 논문 등 문서는 물론 카카오톡, 트위터, 페이스북, 블로그, 카페, 온라인 뉴스, 웹페이지 등의 비정형데이터가 분석의 대상
- ✓ 분석대상 언어, 문화와 관습을 이해



The screenshot shows the 'Big Data' section of the Naver News website. It features a large image of a globe with the text '작은 것이 세상을 바꾼다.' (Small things change the world). Below the image is a QR code. The main content area displays a list of news articles from various sources, such as '국내외 뉴스' (Domestic and International News) and '인물·역사' (Personality·History). The articles are presented in a grid format with titles like '국내외 뉴스' (1), '인물·역사' (2), and '과학·기술' (3).

1. 텍스트 마이닝

■ 텍스트 마이닝의 기능

- ✓ 문서 요약
- ✓ 문서 분류
- ✓ 문서 군집
- ✓ 특성 추출

1. 텍스트 마이닝

■ 텍스트 마이닝 응용분야

- ✓ 스팸 필터링
- ✓ 감성 분석
- ✓ 기계 번역
- ✓ 추천시스템

1. 텍스트 마이닝

■ 텍스트 마이닝 과정

- ✓ 데이터 수집
- ✓ 데이터 전처리 및 가공
- ✓ 자연어처리
- ✓ 분석과 시각화

1. 텍스트 마이닝

■ 데이터의 수집

- ✓ 다양한 문서를 텍스트로 전환
- ✓ 웹페이지의 HTML을 가져와 파싱(parsing)하거나 API를 이용

1. 텍스트 마이닝

■ 데이터 전처리 및 가공 : 데이터 분석에 이용되지 않는 문장부호, 숫자, 단어 등을 제거

① Corpus 생성

- › Corpus : 전자적으로 저장되어 있는 텍스트의 묶음으로 데이터분석에 이용 될 수 있는 상태의 데이터
- › Corpus : VCorpus(메모리)와 PCorpus(DB)로 구분
- › 원천 : 파일, 폴더, Web 데이터
- › 형태 : 일반 텍스트, HTML, XML, Word, 한글, PDF

② 토큰화(tokenizing) : 텍스트 데이터를 작은 단위로 잘게 분리하는데 스페이스를 중심으로 잘라내며 영어의 관사 등 불필요한 단어 삭제

- › 특수문자, 구두점, 공백, 불용어 등을 제거
- › 대소문자 구분 해제

1. 텍스트 마이닝

■ 자연어처리 : 기본적으로 형태소 분석

① 스테밍(stemming) 기법 : 영어 단어를 처리할 때 단어의 기본형을 추출하는 과정

- ▶ 영어 단어의 goes, went, gone 등의 동사의 기본형은 go이기 때문에 미리 만들어진 사전을 이용해 단어의 기본형을 찾음

② 형태소 분석 : 하나의 어절에서 의미를 갖는 최소 단위인 각 형태소를 분석해 내는 기법

- ▶ 한국어 경우 어미의 변화가 심해 스테밍 알고리즘으로 기본형 처리 불가능
- ▶ 형태소 분석기는 문장을 명사 위주의 자립형태소, 조사인 의존 형태소, 동사를 포함한 실질형태소 등으로 분리해 냄 : 주로 동사와 명사를 추출
- ▶ 형태소 분석기는 단어의 등장 빈도수, 문서 내에서의 등장 위치 등을 분석해 주는데 이를 텍스트 시각화에 이용

1. 텍스트 마이닝

■ 한글의 텍스트 분석을 위해서는 형태소 분석기를 이용

- ✓ R에서는 KoNLP라는 형태소 분석 패키지를 통해 형태소를 분석
 - KoNLP는 카이스트의 “한나눔”이라는 형태소 분석기를 이용한 것으로 사용법이 간단하여 R을 이용하여 빅데이터 분석을 할 때 주로 사용

1. 텍스트 마이닝

■ TDM(document-term matrix)의 구축

	Term 1	Term 2	Term 3	Term 4
Ducoment 1				
Ducoment 2				
Ducoment 3				
Ducoment 4				

1. 텍스트 마이닝

■ TDM(document-term matrix)의 구축

```
> dtm <- DocumentTermMatrix(reuters)
> inspect(dtm[5:10, 740:743])
```

```
<<DocumentTermMatrix (documents: 6, terms: 4)>>
Non-/sparse entries: 8/16
Sparsity           : 67%
Maximal term length: 8
Weighting          : term frequency (tf)
```

Docs	Terms			
	one-week	opec	opec"s	opec's
211	0	0	0	0
236	0	6	2	0
237	0	1	0	0
242	1	2	0	0
246	0	1	0	1
248	0	6	0	0

1. 텍스트 마이닝

■ 분석 및 시각화

① 단어빈도수 분석 : TF, TF-IDF

> findFreqTerms(dtm,5)

```
[1] "15.8"      "abdur-aziz"   "ability"     "accord"
[5] "agency"     "agreement"    "ali"        "also"
[9] "analysts"   "arab"        "arabia"     "barrel."
[13] "barrels"    "billion"     "bpd"       "budget"
[17] "company"   "crude"       "daily"      "demand"
[21] "dlrs"       "economic"    "emergency"  "energy"
[25] "exchange"   "expected"    "exports"    "futures"
[29] "government" "group"       "gulf"       "help"
[33] "hold"       "industry"    "international" "january"
[37] "kuwait"     "last"        "market"    "may"
[41] "meeting"    "minister"   "mln"       "month"
[45] "nazer"      "new"         "now"       "nymex"
[49] "official"   "oil"         "one"       "opec"
[53] "output"     "pct"        "petroleum" "plans"
[57] "posted"     "present"    "price"     "prices"
[61] "prices,"    "prices."    "production" "quota"
[65] "quoted"     "recent"     "report"    "research"
[69] "reserve"    "reuter"     "said"      "said."
[73] "saudi"      "sell"       "sheikh"    "sources"
[77] "study"      "traders"   "u.s."     "united"
[81] "west"       "will"       "world"
```

1. 텍스트 마이닝

```
> findAsspcs(dtm, "opec", 0.8 )
```

```
$opec
  meeting emergency      oil    15.8 analysts   buyers     said   ability
  0.88       0.87      0.87    0.85      0.85      0.83    0.82      0.80
```

```
>inspect(DocumentTermMatrix(reuters, list(dictionary = c("prices", "crude", "oil"))))
```

```
<<DocumentTermMatrix (documents: 20, terms: 3)>>
Non-/sparse entries: 41/19
Sparsity           : 32%
Maximal term length: 6
Weighting          : term frequency (tf)
```

Docs	Terms		
	crude	oil	prices
127	2	5	3
144	0	11	3
191	2	2	0
194	3	1	0
211	0	1	0

1. 텍스트 마이닝

■ 분석 및 시각화

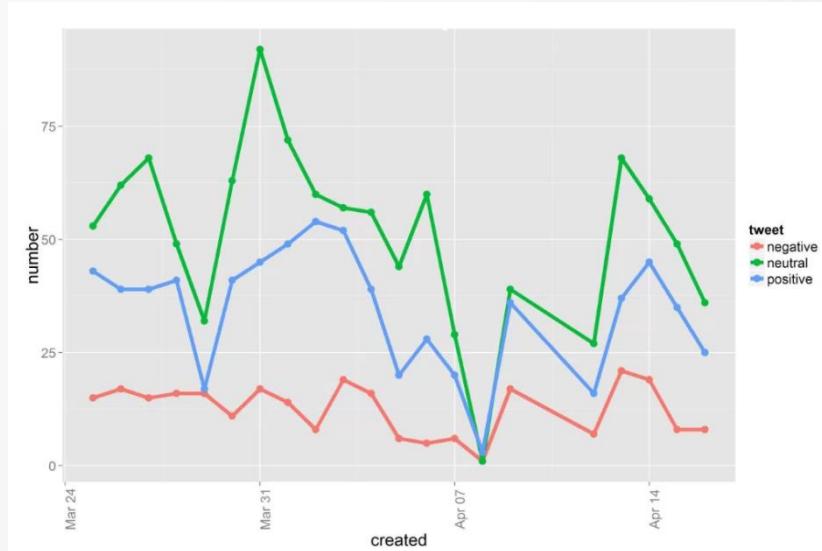
② 워드클라우드

prices
meeting production
agency revenue government reuter
output pumpin' dunnan per statesman
expenditure futures demand commitment
address west markets effective maller projected high
oil plan's produce cut position sources expo present
accord studi president test twain analysts change states
levels say sharp world tex
nazer arabian expected local intervention research 155 barrel can
state must intermediate sheikh mizan arabis announced
rise spa made ability all issue published bank week exploration
new posted pointed called problem xindex exchange finery estimate development
remain named four market total buyers' estimates currency com
named up report price galson protec strategic foreign measure main
month hold rep price galson over december average billion
yesterday says oil to cover three years
days sweet transaction gulf energy may but real one recent
year will country 5 reiterates abdulaziz self reserves
contract with arabia members march however
alkhalifa according owner companies group to appear
policy higher exports spokeswoman indonesia
compared says oil official trade emirates pvt add
saudi ymex quoted february daily
traders fair economy dls
minister openc economy
said oil around
indonesia imports brings ceiling
last petrolium help they
kuwaitis dollars
kuwait kingdom
mln agreement

1. 텍스트 마이닝

■ 분석 및 시각화

③ 감성분석

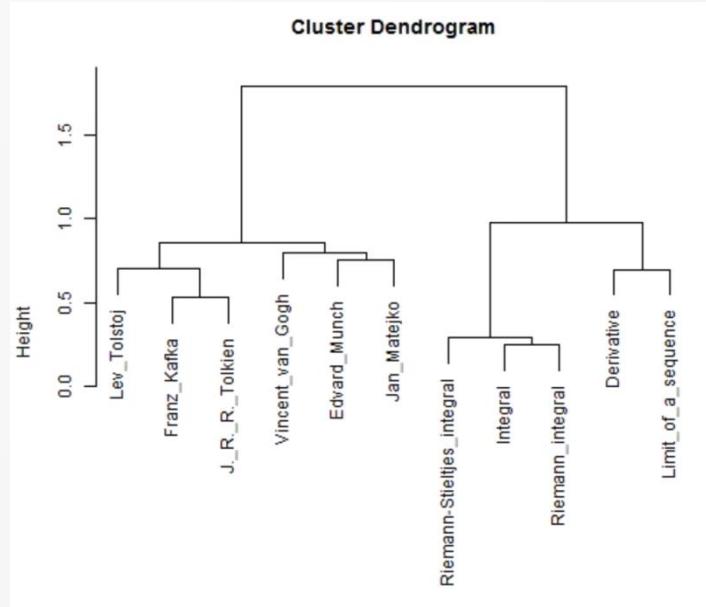


<출처> <http://analyzecore.com/2014/04/28/twitter-sentiment-analysis/>

1. 텍스트 마이닝

■ 분석 및 시각화

④ 군집화



<출처> <http://www.rexamine.com/2014/06/text-mining-in-r-automatic-categorization-of-wikipedia-articles/>

1. 텍스트 마이닝

■ Topic model

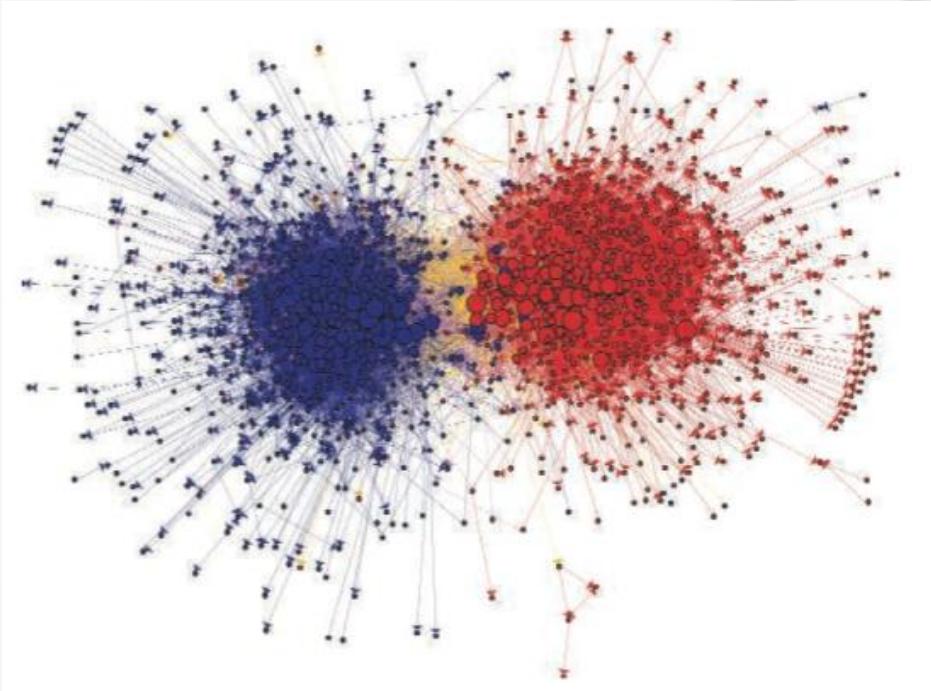
- ✓ 문서내의 단어의 분포로 표현되는 주제와, 해당 주제의 확률적 집합체로 표현되는 다수의 문서를 표현하는 기법

2. 소셜네트워크분석

- 소셜 네트워크(Social Network)에서 개인이 생산하는 데이터의 양이 현저하게 증가 → SNS상 이루어지는 사회현상을 탐색적으로 분석
 - ✓ 소셜 네트워크 시각화 → 중요한 통찰력을 제공

2. 소셜네트워크분석

- 미국 정치 블로거의 소셜 네트워크 연결망 구조를 시각화 :
파란색은 공화당, 빨간색은 민주당을 지지하는 블로거를 상징
 - ✓ 네트워크는 공평하게 연결되어 있지 않으며 서로 양극화되는 현상



2. 소셜네트워크분석

■ 소셜 네트워크의 시각화에서 행위자들은 버티스(vertice) 혹은 노드(node)로 표현

- ✓ 엣지(edge) 혹은 링크(link) : 노드와 노드 사이 연결하는 선
- ✓ 그래프(graph) : 노드와 엣지의 관계를 그림으로 표현한 것
- ✓ 각각의 노드 연결 구조를 인접 테이블로 만든 후 그래프 시각화

2. 소셜네트워크분석

- 노드의 연결 구조를 인접 테이블로 만든 후 그래프로 시각화
 - ✓ A는 B와 C와 연결구조를 가지고 있지만, B와 C와 서로 연결되어 있지 않음
→ 데이터를 그림으로 표현

인접테이블을 이용하여 소셜 네트워크 그래프 그리기

	A	B	C
A	0	1	1
B	1	0	0
C	1	0	0

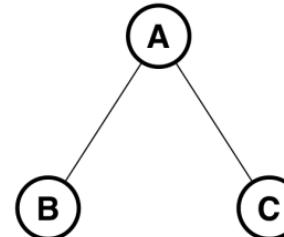
adjacency matrix

A : B, C

B : A

C : A

adjacency list



graph network

2. 소셜네트워크분석

- 그래프는 엣지의 연결 방향에 따라 디렉티드(directed)와 언디렉티드(undirected)로 구분
 - ✓ 디렉티드 그래프 : 각 노드의 연결 구조가 방향성을 가진 것을 의미
 - ✓ 언디렉티드 그래프 : 방향성은 중요하지 않음

2. 소셜네트워크분석

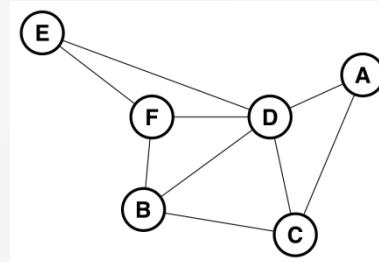
■ 노드의 디그리(degree)는 노드에 연결된 엣지의 숫자를 의미

- ✓ 노드 D는 다른 노드보다 많은 5개의 디그리

- › 소셜 네트워크 분석에 있어서 노드 D는 네트워크의 중심으로 이해

- ✓ 디렉티드 그래프 : 방향성에 따라 인디그리(in-degree)와 아웃디그리(out-degree)로 나눔

- › 인디그리는 노드를 향해 들어오는 엣지의 숫자, 아웃디그리는 노드에서 다른 노드로 나가는 엣지의 숫자

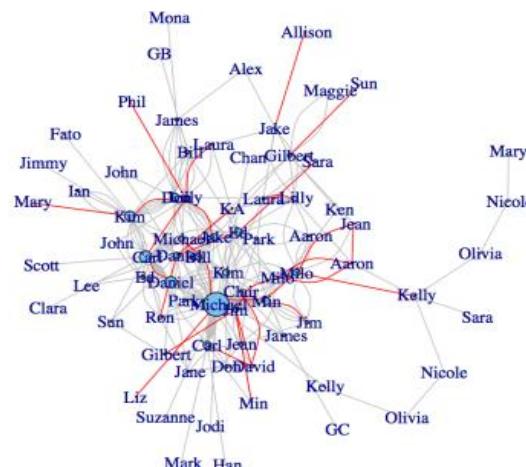


2. 소셜네트워크분석

■ 소셜 네트워크 시각화는 전체 네트워크를 이해하는데 매우 큰 도움

- ✓ 트위터에서의 소셜 네트워크 구조를 시각화 : 메시지의 전파구조, 전파속도, 메시지 확산의 중요한 역할을 하는 사람 등을 파악
→ 소셜 네트워크 시각화는 정치 캠페인 분야에서도 활발하게 이용
 - ✓ 스팸메일러 색출작업, 해킹 시도 방지 등 다양한 분야에서 활용
 - ✓ 페이스북 사용자 간 인터랙션을 담은 데이터를 소셜네트워크 그래프로 표현

Social network example



2. 소셜네트워크분석

■ 트위터 데이터 분석 :

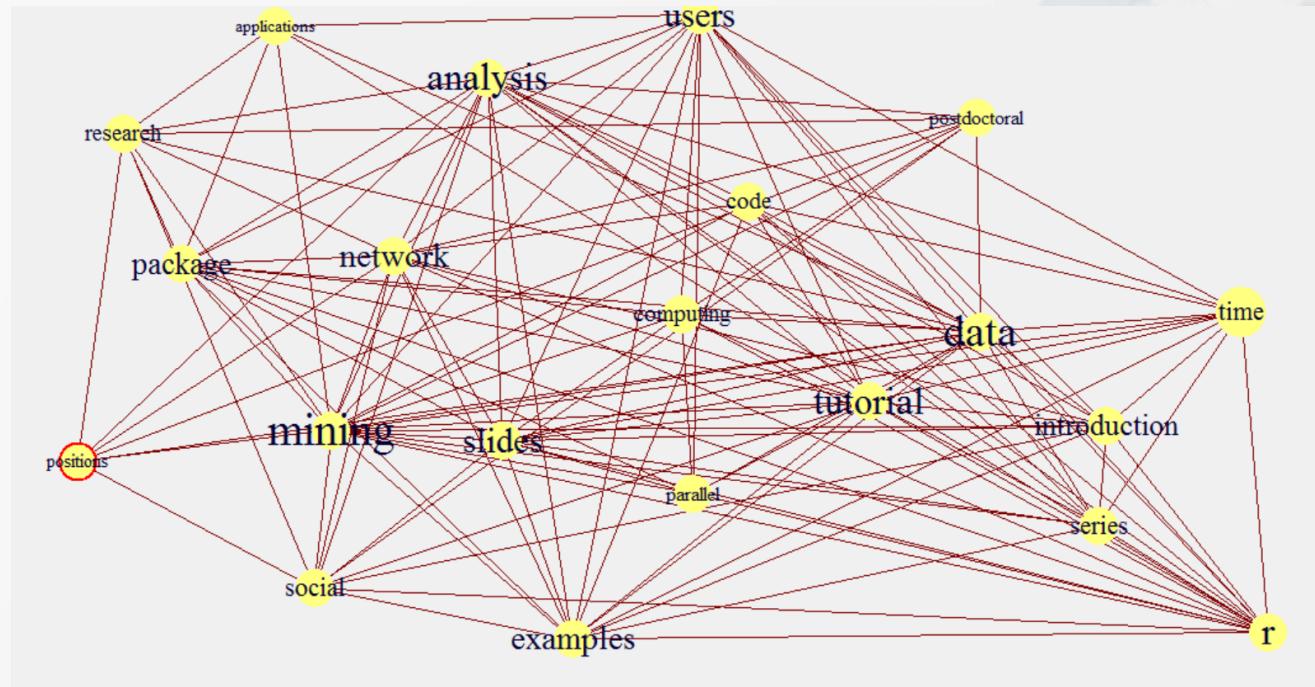
Terms	Docs	1	2	3	4	5	6	7	8	9	10
data		1	1	0	0	2	0	0	0	0	0
examples		0	0	0	0	0	0	0	0	0	0
introduction		0	0	0	0	0	0	0	0	0	0
mining		0	0	0	0	0	0	0	0	0	0
network		0	0	0	0	0	0	0	0	0	0
package		0	0	0	1	1	0	0	0	0	0

Terms	data	examples	introduction	mining
data	53	5	2	34
examples	5	17	2	5
introduction	2	2	10	2
mining	34	5	2	47
network	0	2	2	1
package	7	2	0	5

<출처> [https://rdatamining.wordpress.com/2012/05/17/
an-example-of-social-network-analysis-with-r-using-package-igraph/](https://rdatamining.wordpress.com/2012/05/17/an-example-of-social-network-analysis-with-r-using-package-igraph/)

2. 소셜네트워크분석

■ 트위터 데이터 분석



3. 모형의 평가

■ 모형의 평가

- ✓ 예측 및 분류를 위해 구축된 모형이 임의의 모형보다 더 우수한 성과를 보이는지와 고려된 모형들 중 예측 및 분류 성과를 가장 우수한지 비교 분석
- ✓ 분류 분석 모형의 평가를 위해서 검증용 자료 추출

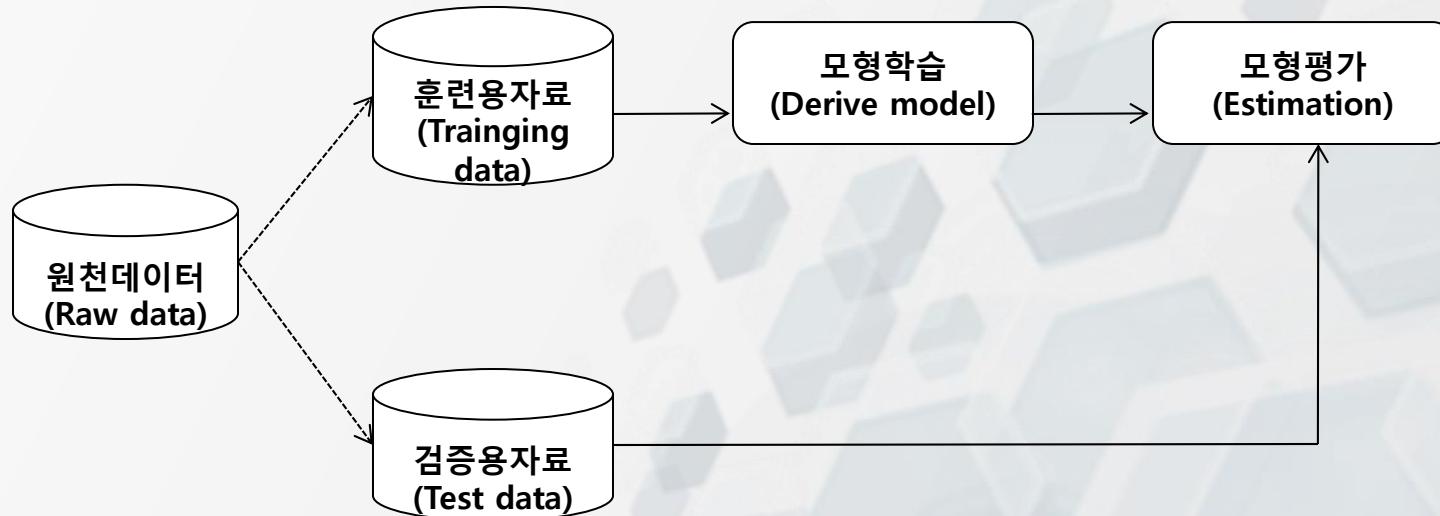
3. 모형의 평가

■ 추출방법

- ✓ 홀드아웃(hold-out)
- ✓ 교차검증 K-fold
- ✓ 봇스트랩

3. 모형의 평가

■ 홀드아웃(hold-out)



3. 모형의 평가

■ 교차검증 K-fold

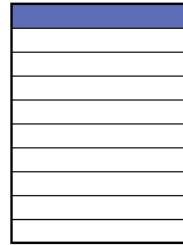


Training Data



Test Data

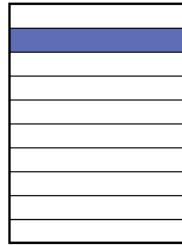
Round1



Accuracy:

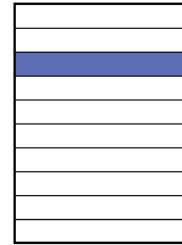
93%

Round2



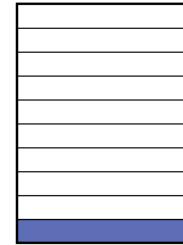
90%

Round3



91%

Round10



95%

• • •

3. 모형의 평가

- 봇스트랩 : 훈련용 자료를 복원추출

3. 모형의 평가

■ 오분류표(confusion matrix)

		예측치		합계
		True	False	
실제값	True	TP	FN	P
	False	FP	TN	N
합계		P'	N'	P+N

3. 모형의 평가

■ 오분류표(confusion matrix)

- ✓ 정분류율

$$accuracy = \frac{TP + TN}{P + N}$$

- ✓ 오분류율 : $1 - accuracy$

- ✓ 민감도와 특이도

$$sensitivity = \frac{TP}{P}$$

$$specificity = \frac{TN}{N}$$

3. 모형의 평가

■ 오분류표(confusion matrix)

- ✓ 모형의 완전성

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

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{TP}{P}$$

- ✓ F1지표와 F_β 지표

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F_\beta = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$

3. 모형의 평가

■ iris 데이터

```
150 samples  
 2 predictor  
 3 classes: 'setosa', 'versicolor', 'virginica'
```

Pre-processing: re-scaling to [0, 1] (2)

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 150, 150, 150, 150, 150, 150,

..

Resampling results across tuning parameters:

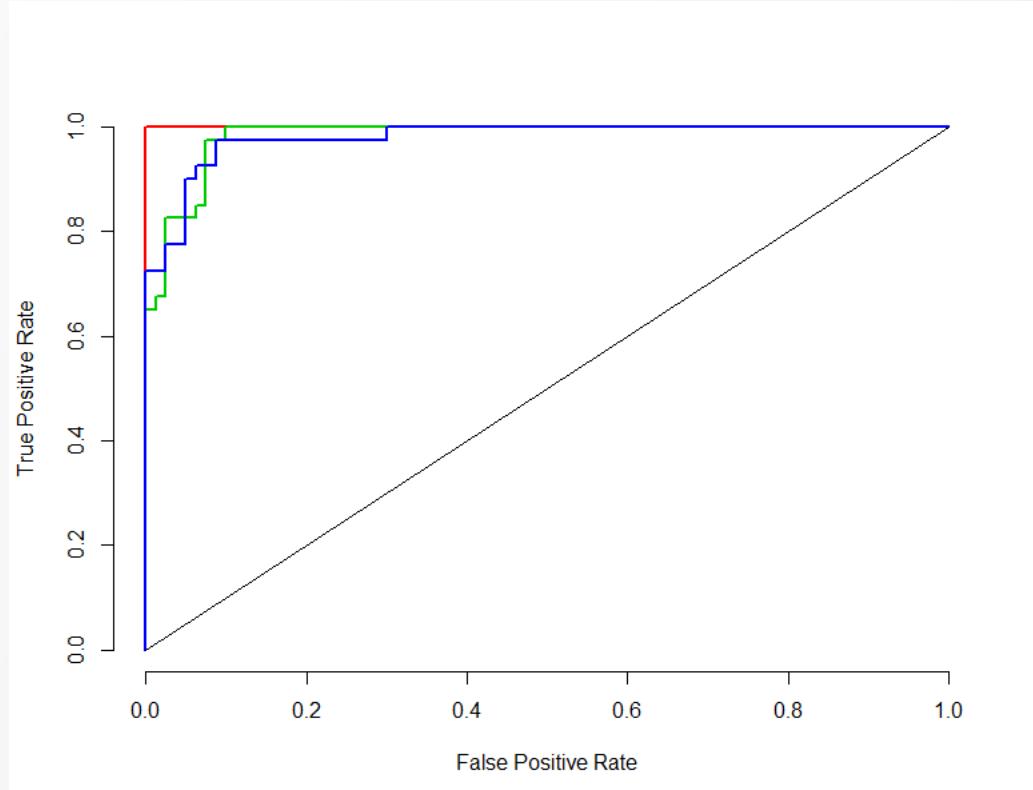
size	decay	Accuracy	Kappa
1	0.0	0.9073482	0.8624527
1	0.1	0.9546901	0.9315288
3	0.0	0.9595834	0.9389146
3	0.1	0.9630405	0.9441297

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were size = 3 and decay = 0.1.

3. 모형의 평가

■ ROC 커브



강의를 마쳤습니다

수고하셨습니다.

15차시 | 특강 : 빅데이터시대의 개인정보 보호

(기말고사)

이긍희 교수

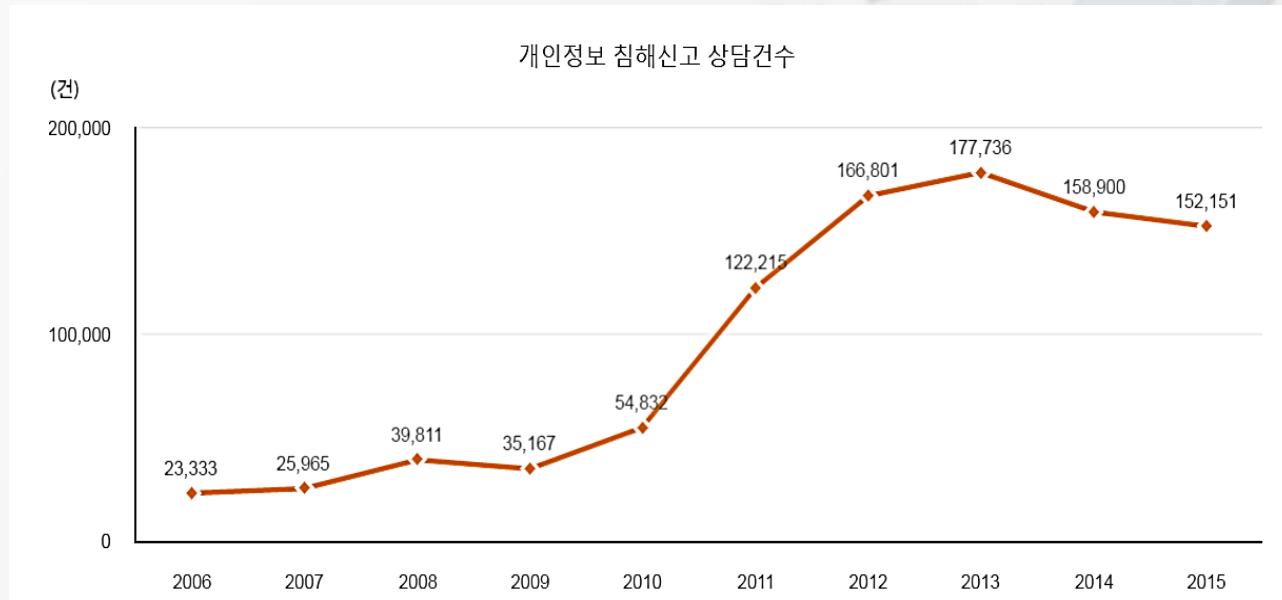


빅데이터 시대의 개인정보 보호

1. 빅데이터 시대와 개인데이터
2. 개인정보와 프라이버시
3. 개인 데이터의 수집
4. 개인정보의 침해사례
5. 개인정보의 보호제도
6. 개인정보의 기술적 보호

1. 빅데이터 시대와 개인데이터

- 빅데이터 시대 : 편리성은 증대되지만 개인정보 및 프라이버시 침해 등 사이버 범죄 증가
- 개인정보 침해 상담건수 추이



1. 빅데이터 시대와 개인데이터

- 카드사의 고객 개인정보 유출 : 1억4000만건
- 온라인 쇼핑몰의 개인정보 유출
- 사이버 망명 논란 : 카카오톡
- 습관의 파악 : Target, Amazon

1. 빅데이터 시대와 개인데이터

■ 빅데이터 시대의 개인데이터

- ✓ 개인정보가 스마트 기술 발전으로 인해
소셜미디어와 GPS, CCTV, NFC 등을 통해 축적

1. 빅데이터 시대와 개인데이터

■ 빅데이터 시대의 정부와 기업의 개인데이터 활용

정부

범죄자 식별, 세금 체납 방지, 테러 방지, 질병 확산 방지, 정책 효율화 등

기업

축적된 정보와 구매패턴을 조합하여 고객별 맞춤 마케팅을 실시

1. 빅데이터 시대와 개인데이터

■ 빅데이터 시대의 개인정보

- ✓ 빅데이터는 기업 이윤 창출, 정부 효율화 등의 동력
 - › 개인정보 유출과 프라이버시 침해 위험 상존
 - › 정부나 기업이 서비스 제공 등을 위해서 데이터를 과다 수집 이용 분석

1. 빅데이터 시대와 개인데이터

■ 빅데이터 시대의 개인데이터 활용과 문제점

- ✓ 데이터 분석 기술이 빠르게 발전

→ 분석 기술로 다양한 원천의 정보를 결합하여 식별된 개인의 다양한 정보를 축적, 이용

- › 비개인 정보가 전화번호부, 이메일 등 공개된 개인정보와 결합되어 개인을 식별
- › 이 때 알리기 원하지 않는 정보가 당사자 동의 없이 수집, 취합, 확산

1. 빅데이터 시대와 개인데이터

■ 빅데이터 시대의 개인데이터 활용과 문제점

- ✓ 정부는 산업육성을 위해 공공의 목적으로 수집한 개인 데이터 공개
- ✓ 기업은 보유 개인 데이터를 거래 시장을 통해 유통
 - › 개인 식별정보를 제거 → 프라이버시 문제
 - › 개인정보 및 프라이버시 보호 측면
 - › 사회적, 법적, 기술적 고려가 필요

2. 개인정보와 프라이버시

■ 개인정보의 정의

- ✓ 살아있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통해 개인을 알아볼 수 있는 정보(개인정보보호법 제2조 1호)
 - › 해당 정보만으로 특정 개인을 알아볼 수 없더라도
 - › 다른 정보와 쉽게 결합하여 알아볼 수 있는 것 포함
- ✓ OECD의 정의 : 개인정보를 식별되거나 식별될 수 있는 개인의 모든 정보

2. 개인정보와 프라이버시

■ 개인정보의 정의

- ✓ 법률적 개인정보 : 직접 식별할 수 있는 식별정보와 다른 정보와 연결되어 식별할 수 있는 속성 정보
- ✓ 식별정보 : 행정정보와 생체정보로 구분
 - › 행정정보 : 성명, 주민등록번호, 여권번호 등
 - › 2차 식별정보 : 은행 계좌번호, 휴대번호 등
 - › 공개 식별정보 : 명함, 홈페이지 등
 - › 생체정보 : 지문, 홍채, 유전자 정보 등

2. 개인정보와 프라이버시

■ 개인정보의 정의

- ✓ 속성정보 : 개인의 성장 및 경제활동에 따라 발생
 - › 성장 관련정보 : 학력정보와 건강정보
 - › 경제활동 관련 정보 : 채무정보, 신용정보 등
 - › 기타 정보 : 가입 단체, 인터넷 활동, 종교, 취미 등
 - › 속성정보 중 일부는 개인 식별성을 제거한 채
 - › 공개 : 속성정보는 식별성을 제거하고 공개
→ 비개인정보로 취급

2. 개인정보와 프라이버시

■ 프라이버시의 정의

- ✓ **프라이버시(privacy)** : 개인의 사생활이나 사적인 일, 또는 남에게 알려지지 않거나
간섭 받지 않을 권리
 - › 신체 프라이버시
 - › 정보 프라이버시
 - › 조직 프라이버시

2. 개인정보와 프라이버시

■ 프라이버시의 정의

- ✓ **프라이버시는 유엔 세계인권 선언은 물론 우리나라 헌법에서 보호되는 기본법**
 - ▶ 유엔 인권선언 제12조
어느 누구도 자신의 프라이버시, 가족, 가정 또는 서신을 임의적으로 간섭해서는 안 되며, 자신의 명예와 평판을 공격해서도 안 된다. 모든 사람은 위의 간섭과 공격에 대항할 수 있는 법의 보호를 받을 권리가 있다.

2. 개인정보와 프라이버시

■ 프라이버시의 정의

- ✓ 각 국가에서는 프라이버시 보호 법률을 제정
- ✓ 데이터의 국경이 없어지고 프라이버시 관련 데이터가 영구적으로 복사·저장·유통되면서 국가별로 적용되는 법 적용에 한계

2. 개인정보와 프라이버시

■ 프라이버시의 정의

- ✓ 개인정보는 프라이버시와 밀접한 관계를 가지고 있으나 개인정보 전체와 일치하지는 않음
 - › 사상, 종교, 신념, 질병정보, 재산상황, 범죄경력 등은 프라이버시, 공개되는 이름, 직장 등의 개인정보는 프라이버시에 속하지 않음
 - › 프라이버시는 시대, 지역별로 달라짐
 - › 독재국가에서는 이를 반대하는 글이나 영상이 프라이버시
 - › 병력, 사고전력은 구직자, 보험가입자에게 프라이버시

2. 개인정보와 프라이버시

■ 프라이버시의 정의

✓ 빅데이터 시대의 프라이버시 : 자신의 정보가 타인에게 동의 없이 공개되거나 침해되지 않는 권리

→ 자신 정보를 통제할 수 있는 권리(잊혀질 권리)로 확대

› 디지털 세탁소 : 과거 인터넷에 게시한 글을 관리해주는 사업

› 디지털 장례 : 세상을 떠난 이의 인터넷 계정, 접속기록, 콘텐츠 등을 삭제해주는 사업

3. 개인 데이터의 수집

■ 개인 데이터의 수집

- ✓ 개인데이터는 개인정보를 포함한 데이터로 다양한 방식으로 수집, 축적, 유통
 - › 당사자의 동의 하에 수집, 쿠키를 통해 자동 수집
 - › 웹 페이지로부터 웹 크롤러 등으로 수집

3. 개인 데이터의 수집

■ 정부와 기업의 개인정보 수집

- ✓ 성명, 주소, 이메일, 휴대폰번호, 개인식별번호 등 개인 식별정보 중심으로 수집
- ✓ 여론 파악, 사회 분위기 파악 등을 목적으로 비개인정보도 수집

3. 개인 데이터의 수집

■ 정부와 기업의 개인정보 수집

- ✓ 정부는 기초 개인정보를 생성관리하고, 동 정보를 바탕으로 행정통계를 작성
- ✓ 기업은 회원가입, 웹버그 등을 통해 수집
 - › 기업은 개인정보를 수집한 후 회원 가입 시 계열사와 제3자에게 유통되는 것을 동의
 - › 개인정보가 광범위하게 사용 됨

3. 개인 데이터의 수집

■ 개인정보 활용의 동의 방식

- ✓ **정보통신망 이용촉진 및 정보보호 등에 관한 법률**

정보통신서비스 제공자는 이용자의 개인정보를 이용하려고 수집하는 경우에는 개인정보의 수집, 이용 목적, 수집하는 개인정보의 항목, 개인정보의 보유, 이용 기간을 이용자에게 알리고 동의를 받아야 한다.

3. 개인 데이터의 수집

■ 개인정보 활용의 동의 방식

- ✓ 옵트인 방식 : 개인에게 개인정보 수집에 대해 사전에 동의를 받는 방식
- ✓ 옵트아웃 방식 : 거부의사를 표시하지 않는 한 동의한 것으로 간주

4. 개인정보의 침해 사례

■ 개인정보의 침해

- ✓ 개인정보의 침해 : 개인정보가 유출, 변경, 훼손, 도용되어 개인의 정보가 자기 통제권을 침해 받은 것을 의미
 - › 개인정보보호법에서의 개인정보의 침해 : 동의 없는 개인정보 수집, 수집 시 고지 의무 불이행, 과도한 수집, 제3자 제공, 취급자에 의한 훼손, 이용 후 개인정보 미파기 등

4. 개인정보의 침해 사례

■ 개인정보의 침해

✓ 개인정보는 단계에 따라 피해 유형이 달라짐

- › 수집 단계 : 동의 없는 수집, 과도한 수집
- › 저장 및 관리 단계 : 외부 해킹 및 내부자 유출
- › 이용 및 제공 단계 : 동의 없이 목적 외로 분석, 계열사 및 자회사와 개인정보를 공유
- › 파기 단계 : 파기하지 않거나 파기과정에서 유출

4. 개인정보의 침해 사례

■ 개인정보의 침해

- ✓ **빅데이터 시대 이전** : 침해정보가 개인 식별정보나 신용정보가 중요한 개인정보에 국한
- ✓ **빅데이터 시대** : 여러 가지 데이터를 결합하여 분석할 수 있으므로 개인의 생활패턴이나 성향까지 추가

4. 개인정보의 침해 사례

■ 개인정보 침해 유형과 통계

✓ 개인정보 침해 상담건수 증가

- › 새로운 스마트기기 및 서비스(클라우드, SNS등)가 빠르게 확산되고, 개인정보 유출사고, 이용자의 개인정보에 대한 관심증가로 인해 침해 상담건수 증가
- › 주민번호 등 타인정보도용이 가장 많았고 법적용 불가 침해사례가 많았음

4. 개인정보의 침해 사례

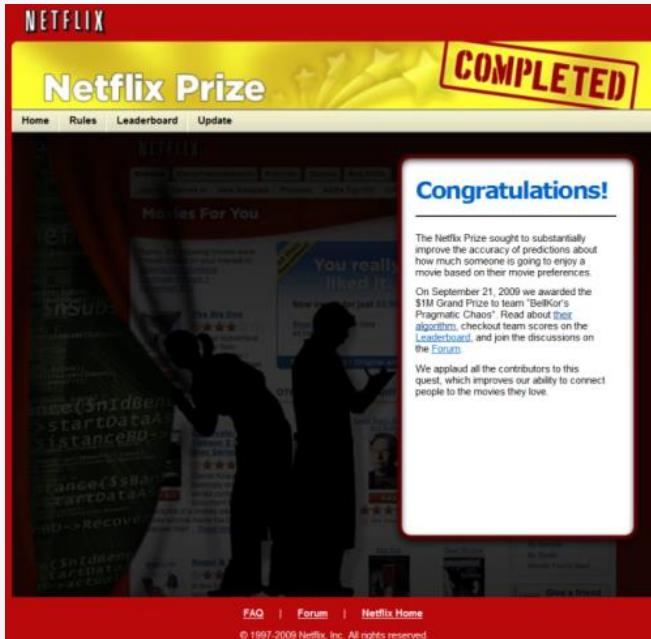
■ 고객정보의 유출

- ✓ 역대 개인 정보 유출 사건을 살펴보면 금융사 뿐만 아니라 인터넷 포탈 사이트, 게임 회사, 통신 회사, 방송국까지 다양
 - › 유출 방식은 내부자의 유출, 해킹에 의한 유출이었는데 내부자의 유출이 보다 광범위

4. 개인정보의 침해 사례

■ 넷플릭스의 고객정보 간접 유출

- ✓ 2006년 넷플릭스(Netflix)는 영화추천 서비스인 시네매치의 성능을 개선할 목적으로 100만 달러의 상금을 걸고 추천 알고리즘을 공모



4. 개인정보의 침해 사례

■ 넷플릭스의 고객정보 간접 유출

- ✓ 넷플릭스는 자체 방대한 데이터를 공모전에 참여하는 모든 사람들에게 알고리즘 작성을 위해 공개
 - › 1998년부터 6년간 480,189명이 17,770개 영화에 부여한 평점 데이터 100,480,507건
 - › 데이터는 <사용자 ID, 영화, 평점, 평점일>로 구성
 - › 참가자는 데이터를 바탕으로 알고리즘을 개발 후 추천데이터를 만들어서 넷플릭스에 보내면 그 회사에서는 실제 데이터와 비교하여 예측력을 평가

4. 개인정보의 침해 사례

■ 넷플릭스의 고객정보 간접 유출

- ✓ 2007년 텍사스 대학 연구팀이 넷플릭스의 정보를 Internet Movie Database의 영화 평점 데이터와 결합 관련된 고객 중 84%를 식별
 - › 넷플릭스는 2009년 집단소송에 말려들었음
 - › 알고리즘 공모와 데이터 공개가 중단

4. 개인정보의 침해 사례

■ 개인 게놈 프로젝트(PGP)에서의 개인정보 식별

- ✓ 개인 게놈 프로젝트(PGP)의 홈페이지

Personal Genome Project Log in ▶

Public data ▾ About ▾

Participant Profiles

The participants in the PGP have volunteered to share their DNA sequences, medical information, and other personal information with the research community and the general public.

Show	10 <input checked="" type="checkbox"/> entries	Search:					
PGP#	participant ID	Date enrolled	Received samples	Health records	Relatives enrolled	Whole genome datasets	Other genetic data
PGP1	hu43860C	2010-11-21	Whole Blood, Microbiome	Yes	1	1	1
PGP2	huC30901	2010-11-21		Yes		1	
PGP3	huBEDA0B	2010-11-21	Saliva, Whole Blood, Microbiome			3	
PGP4	huE80E3D	2007-04-02				1	
PGP5	hu9385BA	2010-11-21	Microbiome	Yes		4	
PGP6	hu04FD18	2010-11-21	Saliva	Yes		1	
PGP7	hu0D879F	2010-11-21	Saliva, Whole Blood			3	
PGP8	huAE6220	2010-11-21		Yes		1	
PGP9	hu034DB1	2010-11-21	Saliva, Whole Blood, Microbiome			3	2
PGP10	hu604D39	2010-11-21	Microbiome	Yes		4	

Showing 1 to 10 of 3,250 entries

4. 개인정보의 침해 사례

■ 개인 게놈 프로젝트(PGP)에서의 개인정보 식별

- ✓ 유전자 정보는 개인정보 중 가장 민감한 정보

- ▶ 본인이 공개에 동의해서 공개하더라도 그 유사성으로 그의 가족, 자녀, 친지의 개인정보를 침해할 가능성이 높음

4. 개인정보의 침해 사례

■ 개인 게놈 프로젝트(PGP)에서의 개인정보 식별

- ✓ 하버드 대학 교수인 스위니 등(2013) : 미국의 개인 게놈 프로젝트(PGP)에 익명으로 참가한 사람들의 신원을 찾음

5. 개인정보의 보호 제도

■ 개인정보의 보호 제도

- ✓ 2011년 9월 30일을 기준으로 “개인정보보호법”이 시행되면서
국내 개인정보 보호체계가 통합

5. 개인정보의 보호 제도

■ 개인정보 보호 원칙

- ✓ 1995년 유럽연합은 8가지 개인 데이터 보호지침

- ① 수집 제한의 원칙
- ③ 목적 명확화의 원칙
- ⑤ 안정성 확보의 원칙
- ⑦ 개인 참가의 원칙

- ② 내용의 정확성 원칙
- ④ 이용제한의 원칙
- ⑥ 개인정보 활용 정책 공개의 원칙
- ⑧ 책임의 원칙

5. 개인정보의 보호 제도

■ 개인정보 보호법

- ✓ 개인정보 보호법 : 개인정보의 수집, 유출, 오용, 남용으로부터 사생활의 비밀 등을 보호함으로써 국민의 권리와 이익을 증진, 개인의 존엄과 가치를 구현
 - › 개인정보처리에 관한 기본원칙, 절차 및 방법, 제한, 안전한 처리를 위한 관리, 감독, 정보주체의 권리, 개인정보 권리 침해에 대한 구제 등

5. 개인정보의 보호 제도

■ 개인정보보호와 익명성

- ✓ 빅데이터 시대에 있어 익명화는 개인정보 보호에 있어서 무엇보다 중요
 - › 익명화 : 자신의 본래 이름을 숨기거나 숨긴 이름을 쓰는 것
 - › 민감한 의료 데이터 등에서 식별자를 가공함으로써 누구 정보인지 알 수 없게 한 다음, 처리나 분석, 활용
 - 프라이버시 침해 우려 줄이고 활용 극대화
 - › 개인정보보호법에서는 익명화 처리를 요구

5. 개인정보의 보호 제도

■ 공공데이터의 개인정보보호 지침

- ✓ 정부 정책의 투명화 및 빅데이터 산업 육성 등을 목적으로 정부의 공공데이터 개방과 공유
 - 공공데이터의 제공 및 이용 활성화에 관한 법률 : 공공기관이 보유, 관리하는 데이터의 제공 및 그 이용 활성화에 관한 사항을 규정

5. 개인정보의 보호 제도

■ 공공데이터의 개인정보보호 지침

- ✓ 개인 데이터를 수집하거나 이용할 때 법령 근거 또는 정보주체 동의에 의해 수집, 이용
 - › 인터넷, 언론 등 공개된 개인정보는 사회 통념상 공개된 목적 범위 내에서만 수집 이용
- ✓ 정부 및 공공기관이 개인데이터 분석 시 개인 식별 가능 정보는 삭제, 비식별화 후 분석
 - › 분석 시 개인정보 활용이 불가피한 경우 당초 수집 목적 범위 내에서만 분석

5. 개인정보의 보호 제도

■ 공공데이터의 개인정보보호 지침

- ✓ 정부나 공공기관이 개인데이터를 공유 시 최소한으로 하고 목적 외로 이용할 때는 법률 근거 또는 별도 동의절차를 거쳐야 함
- ✓ 개인데이터를 공개할 때 개인정보를 배제, 비식별화 처리한 후 개방해야 하고 법률 근거 또는 정보주체 동의 하에 제한적으로 공개
- ✓ 정부와 공공기관이 개인데이터를 관리할 때 주민등록번호 등 중요 개인정보는 암호화해서 관리

6. 개인정보의 기술적 보호

■ 개인정보의 기술적 보호

- ✓ 개인정보와 프라이버시는 제도만으로 보호되기 어려우며 정보보호 기술이 필요
- ✓ 개인정보보호 기술은 수집, 저장, 관리, 이용, 제공, 분석, 파기 등 개인정보 수명주기에 따라 구분

6. 개인정보의 기술적 보호

■ 개인정보의 기술적 보호

- ✓ **수집단계 기술** : 익명화 기술은 이용자가 익명으로 서비스를 이용할 수 있도록 개인정보를 숨기는 기술
 - › 데이터 마스킹(Data Masking)
 - › 가명처리(Pseudonymisation)
 - › 데이터 범주화 방식

6. 개인정보의 기술적 보호

■ 개인정보의 기술적 보호

- ✓ 저장, 관리 단계 기술 : 방화벽(침입 차단), 침입탐지 및 방지, 데이터베이스 보안 및 접근통제 등
- ✓ 이용, 제공 단계 기술 : 데이터 암호화 기술, 데이터 접근통제 기술,
데이터 필터링 및 등급 분류 기술 등
- ✓ 데이터 처리 및 분석단계 기술 : 익명화된 데이터 처리기술, 암호화된 데이터 처리 기술 등

6. 개인정보의 기술적 보호

■ 개인정보의 기술적 보호

- ✓ 데이터 분석결과 가시화 및 이용단계 : 이용자 동의와 관련된 기술, 분석정보의 이용 모니터링 기술 등
- ✓ 파기단계 기술 : 삭제한 개인정보파일의 복원을 원천적으로 불가능하게 만드는 기술

6. 개인정보의 기술적 보호

■ 빅데이터 개인정보보호 가이드라인(방송통신위)

- ✓ 수집 시부터 개인식별 정보에 철저한 비식별화 조치
- ✓ 빅데이터 처리 사실·목적 등 공개를 통한 투명성 확보
- ✓ 개인정보 재식별시, 즉시 파기 및 비식별화 조치
- ✓ 민감 정보 및 통신비밀의 수집·이용·분석 등 처리 금지
- ✓ 수집 정보 저장·관리 시 '기술적·관리적 보호조치' 시행

강의를 마쳤습니다

수고하셨습니다.