2. Preprocessing

KUBIG 학술부

Contents

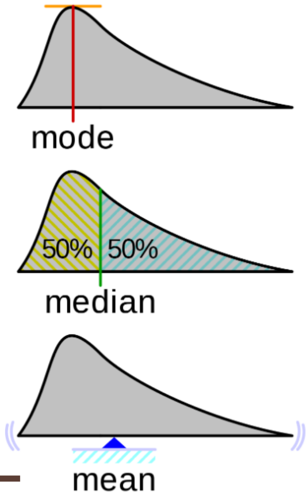
1. Descriptive Statistics

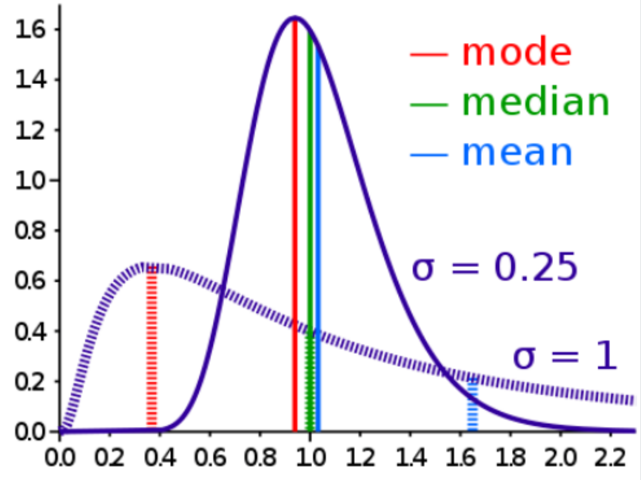
(1) 중심 경향 측정 : Mean, Median, Mode

Mean

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

Sensitive to Outliers

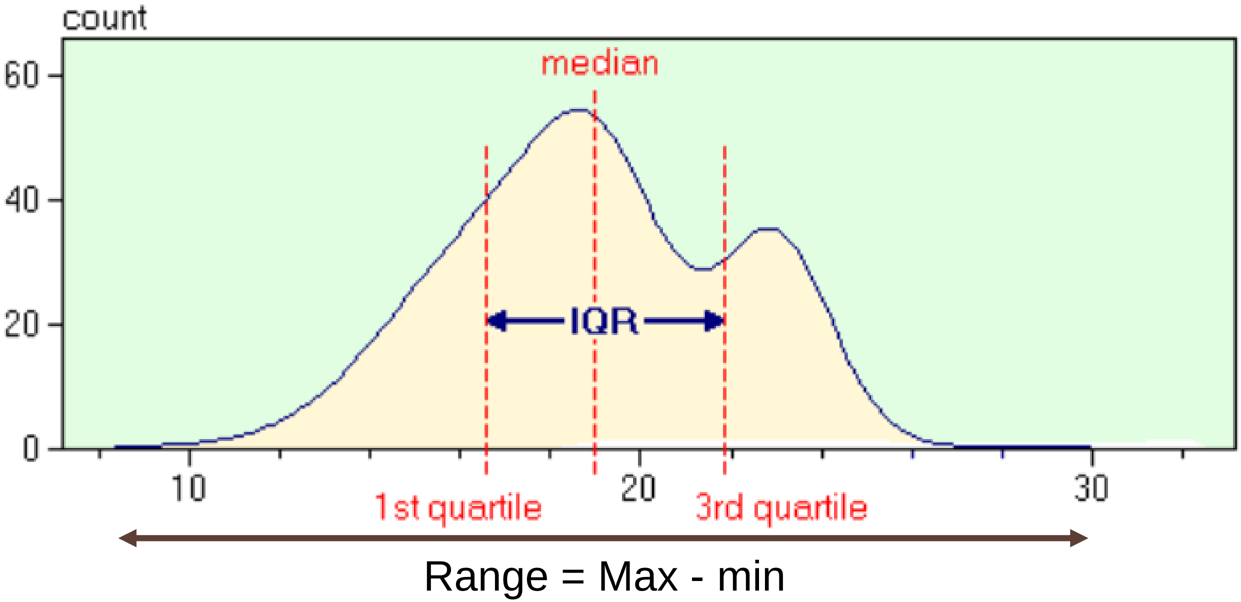




Comparison of mean, median and mode of two log-normal distributions with different skewness.

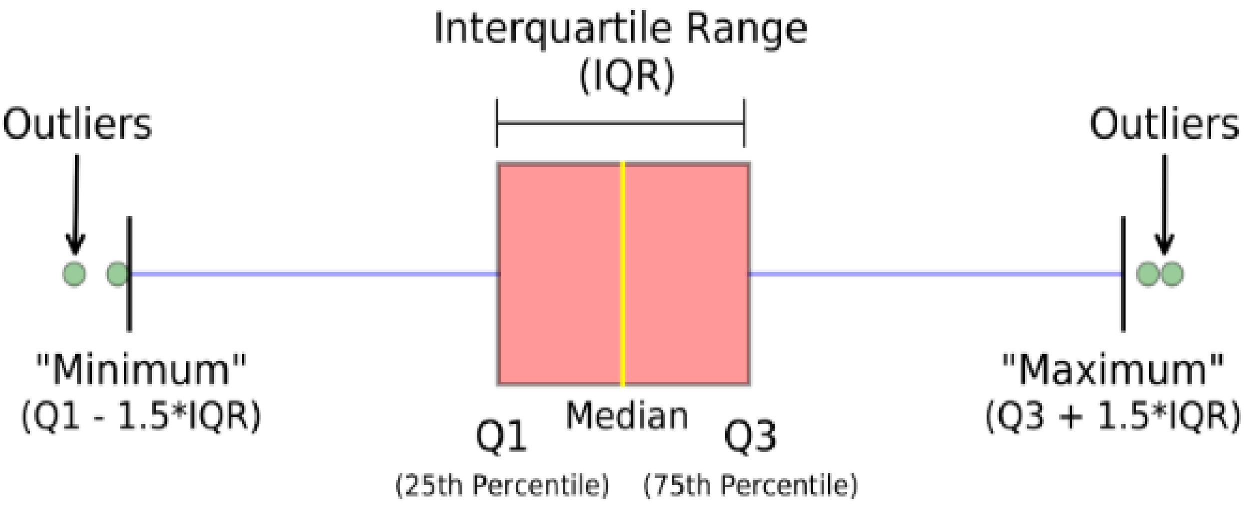
1. Descriptive Statistics

(2) 산포(Variation) 측정 : Range, Quartile, IQR



1. Descriptive Statistics

(3) Box Plot – Skewedness, and Symmetric

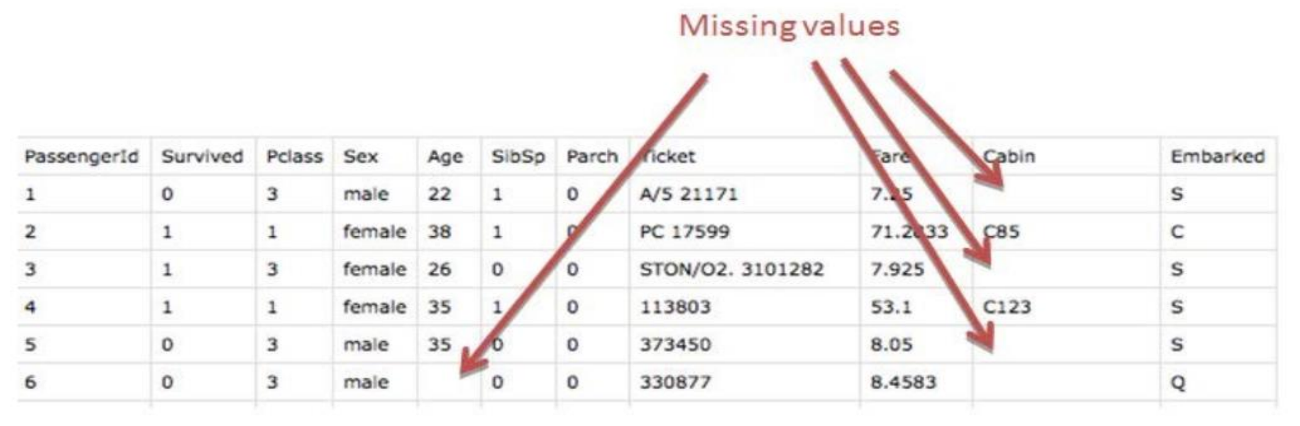


Contents

2. Data Cleansing

(1) Missing Imputation

- 행 제거 / 열 제거
- Global value로 결측치 대체



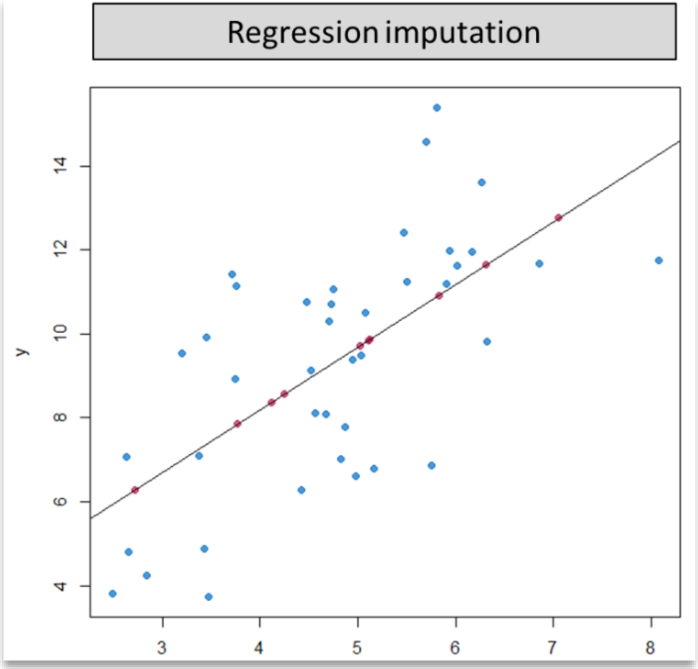
PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

2. Data Cleansing

(1) Missing Imputation

- 결측치 obs와 동일한 범주에 속하는 obs의 평균/중위수 활용
- **가장 가능성이 높은 값으로 결측치 채우기.**

  Regression, Bayesian, Decision Tree, KNN, or other models



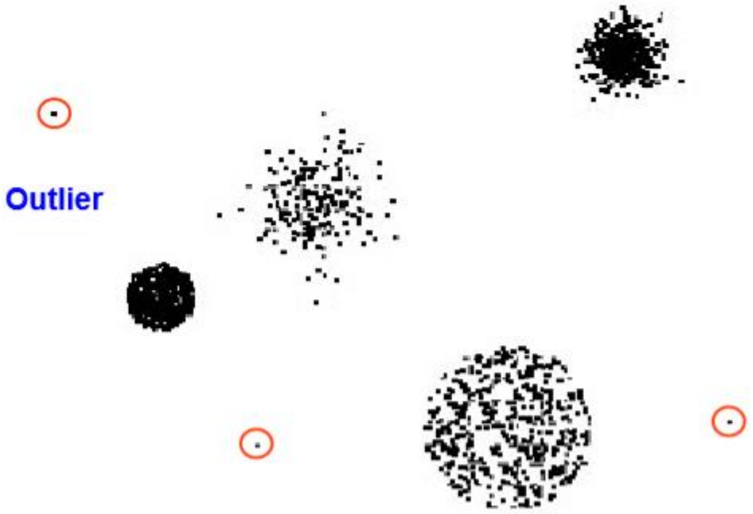
https://www.iriseekhout.com/missing-data/missing-data-methods/imputation-methods/

KU-BIG

2. Data Cleansing

(2) Outlier

- Outlier – data objects with characteristics that are considerably different than most of the other data objects in the data set
- Different from Noise



2. Data Cleansing

(2) Outlier Detection

- Internally/Externally studentized residual
- Leverage
- Cook's Distance
- 1.5*IQR …

- Anomaly/Novelty Detection

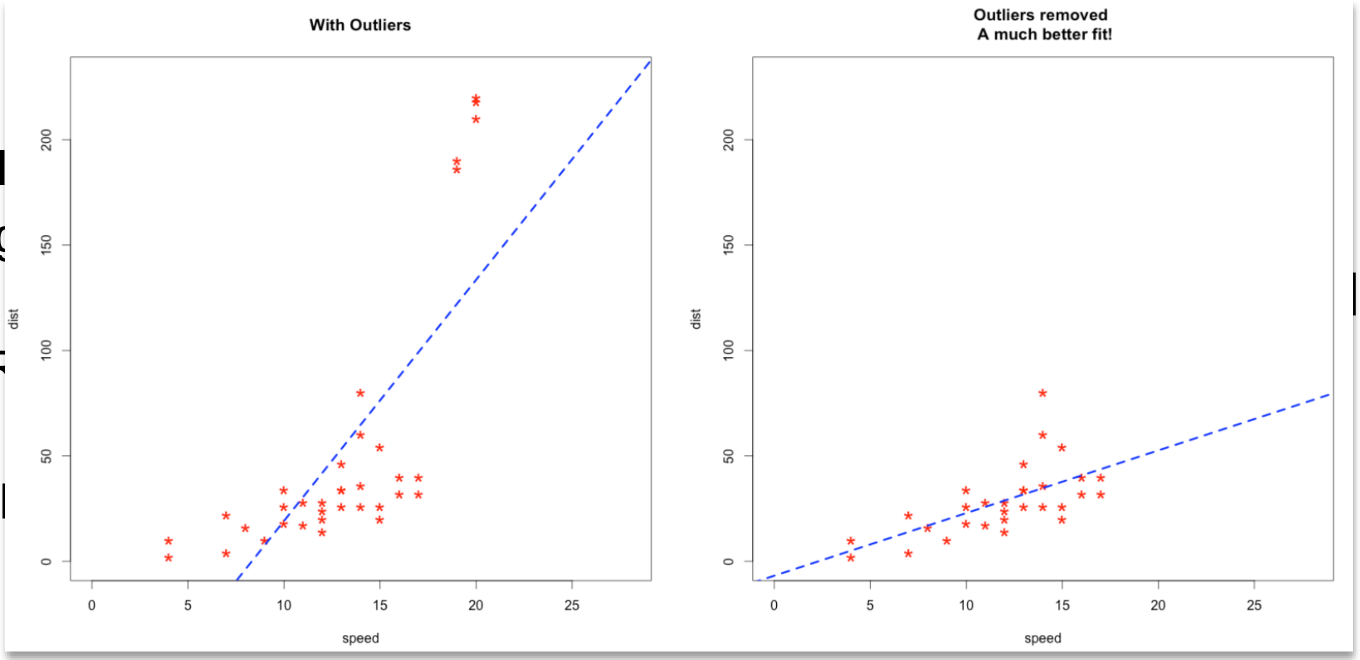
→

- 삭제
- 상, 하한선 제한
- 케이스 분리 분석

2. Data Cleansing

(2) Outlier

- Internal
- Leverag
- Cook's
- 1.5*IQR



ㅔ한

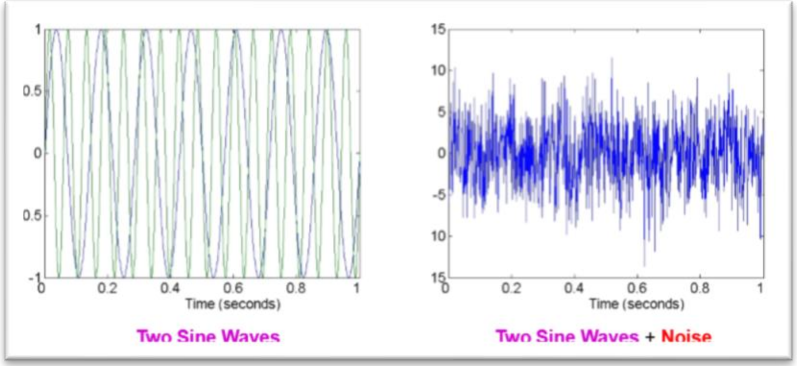
분석

- Anomal

2. Data Cleansing

(3) Noise

- Noise refers to modification of original values



(a) Three groups of points.

(b) With noise points (+) added.

2. Data Cleansing

(3) Noise

- Noise refers to modification of original values

↓ 평활화

- Binning : 근접한 다른 값을 참고하여 정렬한 데이터 값을 평활화
- Regression

가격으로 정렬 (달러 기준) : 4, 8, 15, 21, 21, 24, 25, 28, 34

동일빈도 빈으로 **분할** :
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

빈 평균으로 평활화 :
Bin 1: 9,9,9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

빈 경계로 평활화 :
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Contents

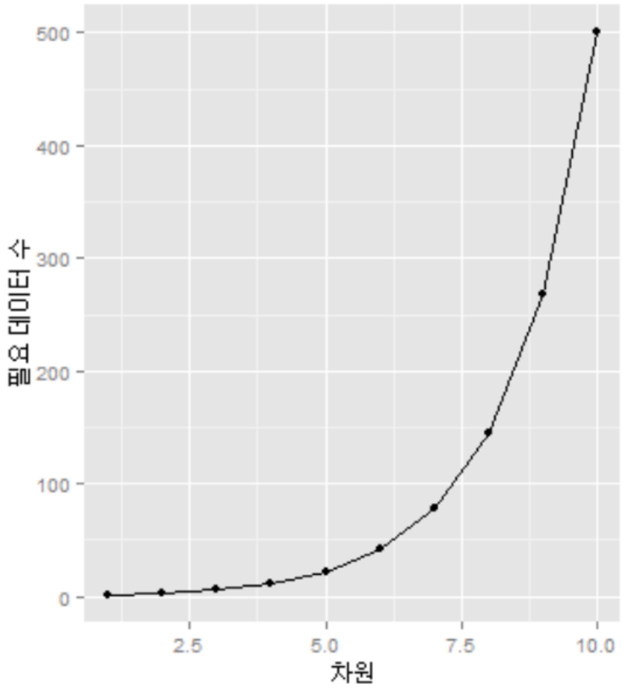
3. Data Reduction

(1) Dimension Reduction

- Curse of Dimensionality : dimensionality increases, data becomes sparse in data space

↓

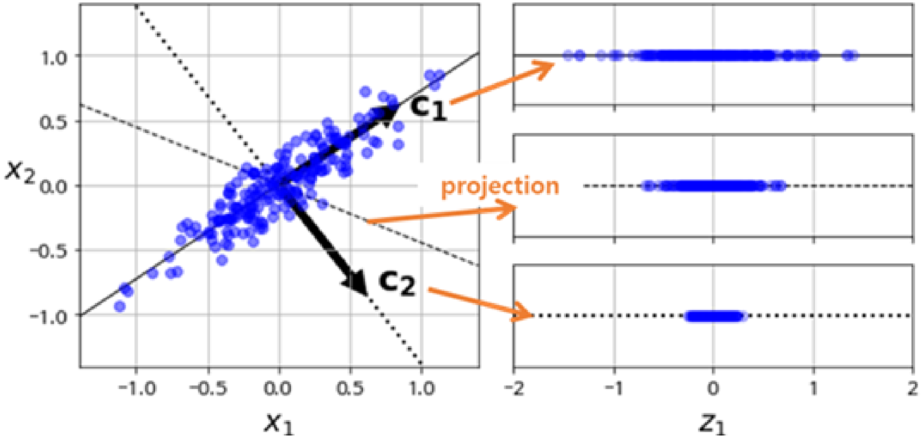
1. Density와 Distance의 정의에 따라, clustering과 outlier detection에 크게 영향을 준다.
2. Complexity -> infeasible algorithms



3. Data Reduction

(1) Dimension Reduction – PCA (Principal Component Analysis)

- PCA – find projection that captures the largest amount of variation in data



3. Data Reduction

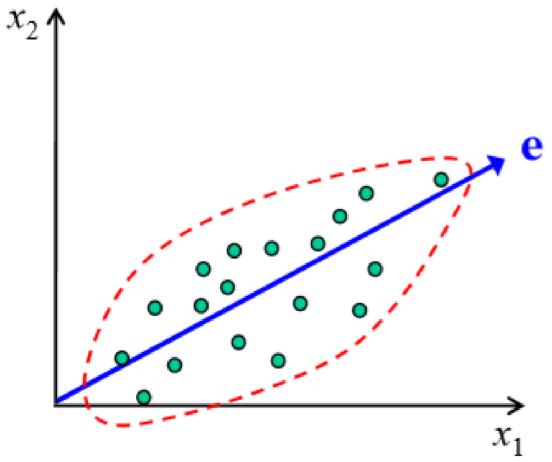
(1) Dimension Reduction – PCA (Principal Component Analysis)

- Find eigenvectors of the covariance matrix

$$Z = XU$$

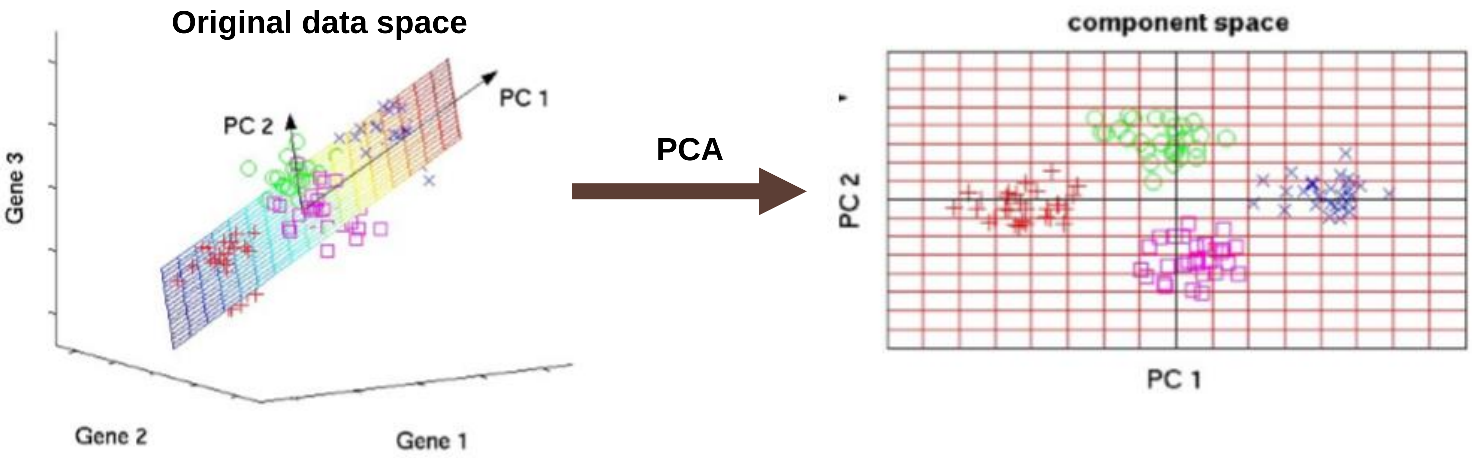
U : eigenvector matrix
Z : Principal Components

- Transform the original data based on the eigenvector (new space)



3. Data Reduction

(1) Dimension Reduction – PCA (Principal Component Analysis)



3. Data Reduction

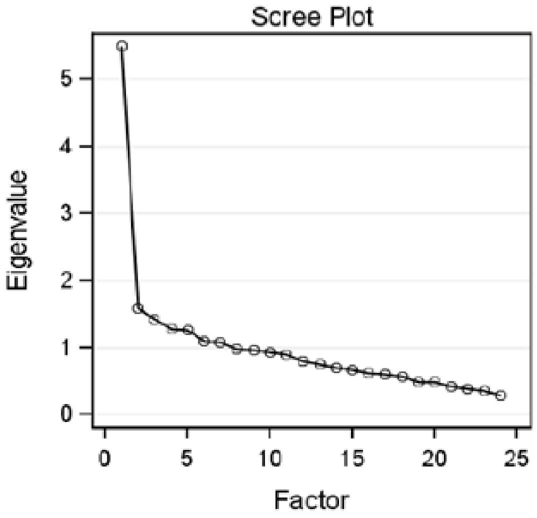
(2) PCA – Choosing number of Principal Components

- Scree plot

  Y축 : eigenvalue(Variance of PC’s)
  X축 : Principal Components

- Explains 90% of total variance

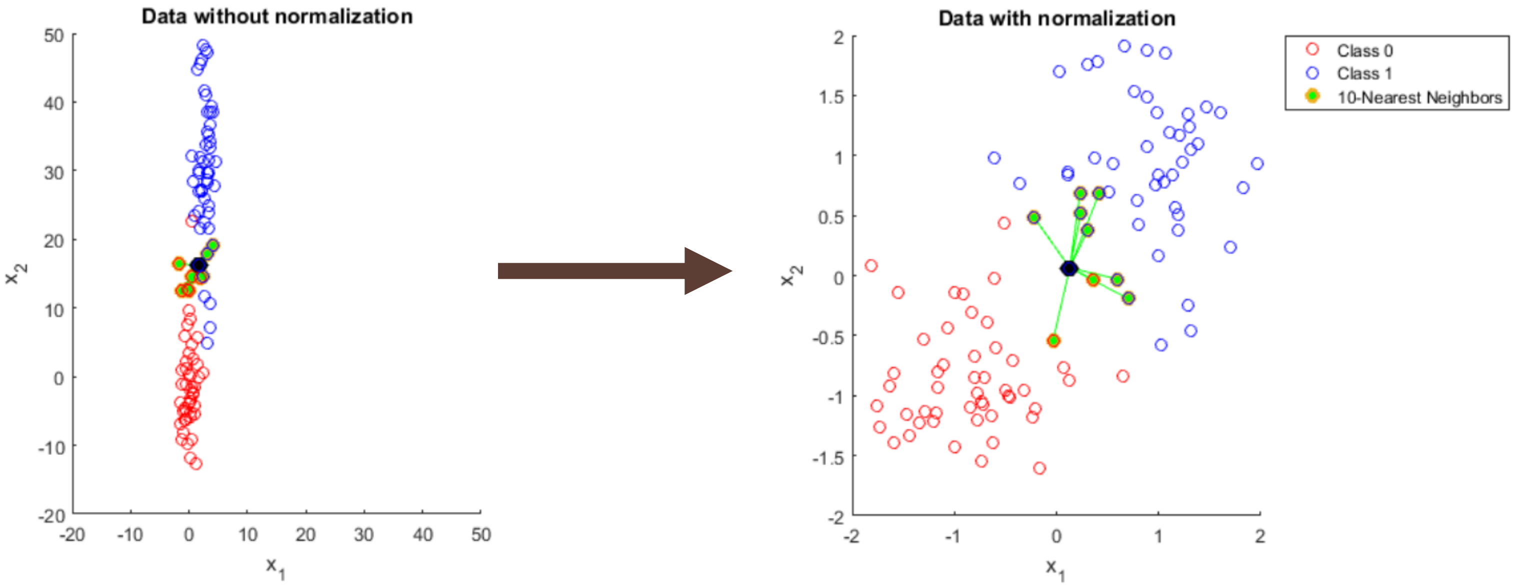
- 단점 : 각 Principal Components의
  해석이 어렵다.



Contents

4. Data Transformation

(1) Need of data transformation



KU-BIG

4. Data Transformation

(2) Scaling methods

Min-Max Normalization

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

원 데이터 값 간의 관계를 유지하면서 해당 attribute를 0~1 사이의 값으로 나타낸다.

Range : [0, 1]

Standardization

$$x_{new} = \frac{x - \mu}{\sigma}$$

값의 scale이 다른 두 변수가 있을 때, 이 변수들의 scale 차이를 제거해 주는 효과가 있다.

$N(0, 1)$

Decimal scaling

$$x_{new} = \frac{x_i}{10^j}$$

J는 $\max(|x_{new}|) < 1$ 를 만족하는 최소 정수.

Range : [-1, 1]

END