The Apprentice Project

Lec04: How to manage Dataset

충북대학교

문성태 (지능로봇공학과)

stmoon@cbnu.ac.kr

01

Introduction

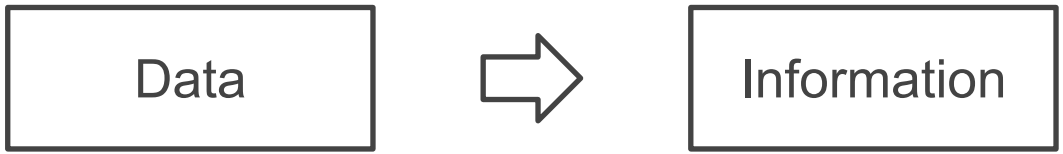
What is Data?

”data are a set of values of qualitative or quantitative variables about one or more persons or objects.”

데이터란 사람이나 물건 등에 대한 사실을 나타내는(표현하는) 값

Data vs Information

- Data is a collection of facts, while information puts those facts into context.
- While data is raw and unorganized, information is organized.
- Data, on its own, is meaningless. When it's analyzed and interpreted, it becomes meaningful information.



https://bloomfire.com/blog/data-vs-information/

Dataset site

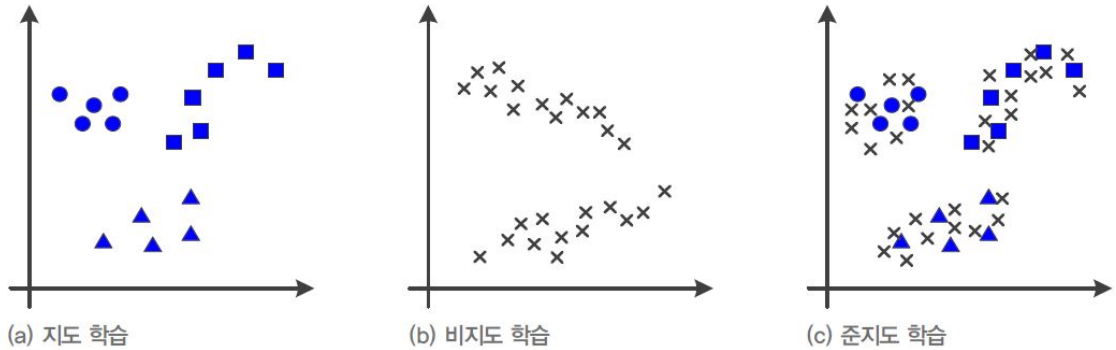
- (해외) Kaggle
  - https://www.kaggle.com/
- (국내) 공공데이터
  - https://www.data.go.kr
- (국내) AI 허브
  - https://aihub.or.kr/

02

Training set vs Test set

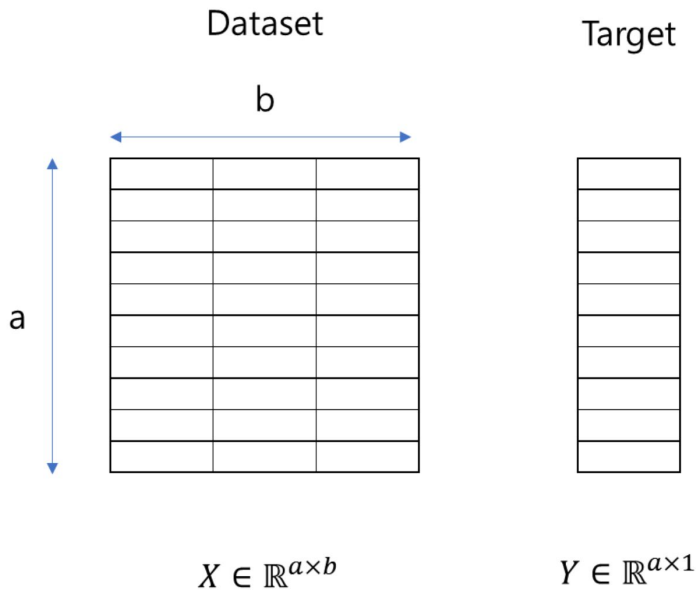
학습 방법

- 지도 학습 (Supervised Learning)
  - 모든 훈련 샘플이 레이블 정보를 가짐
- 비지도 학습 (Unsupervised Learning)
  - 모든 훈련 샘플이 레이블 정보를 가지지 않음
- 준지도 학습: 레이블을 가진 샘플과 가지지 않은 샘플이 섞여 있음



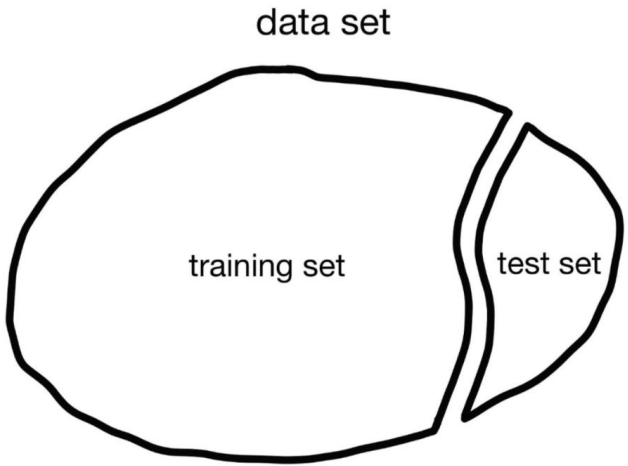
Dataset

- Dataset 구성
  - Input
  - Target (지도 학습의 경우 존재)
- Sample
  - input과 target으로 이루어진 하나의 데이터



Training set vs Test set

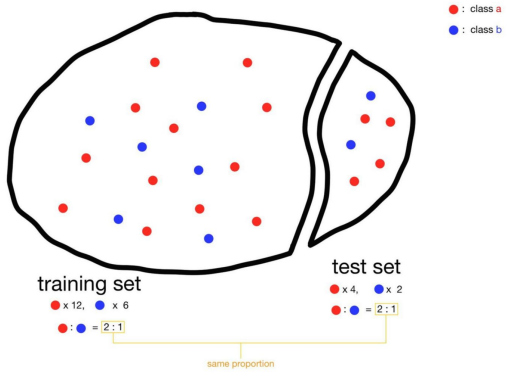
- Training set
  - 훈련을 위해 사용되는 dataset
  - 모델이 학습할 데이터
- Test set
  - 평가를 위해 사용되는 dataset
  - 모델의 성능을 테스트하기 위해 사용할 데이터



Dataset 생성 규칙

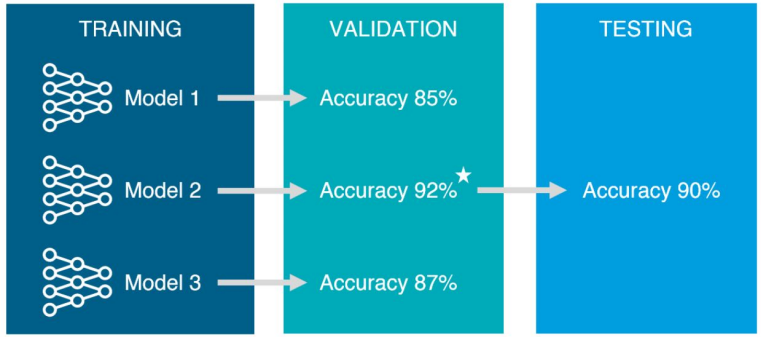
- Training set의 sample 갯수>> Test set의 sample 갯수
- 훈련 세트와 테스트 세트가 동일한 비율의 데이터 분포
  - 분포에 문제가 발생한 경우 샘플링 편향 (sampling bias) 발생
  - Shuffling을 통한 sampling bias 해소
- 훈련 세트와 테스트 세트의 중복 데이터 최대한 제거

Training set : Test set = 7 : 3 (or 8:2)



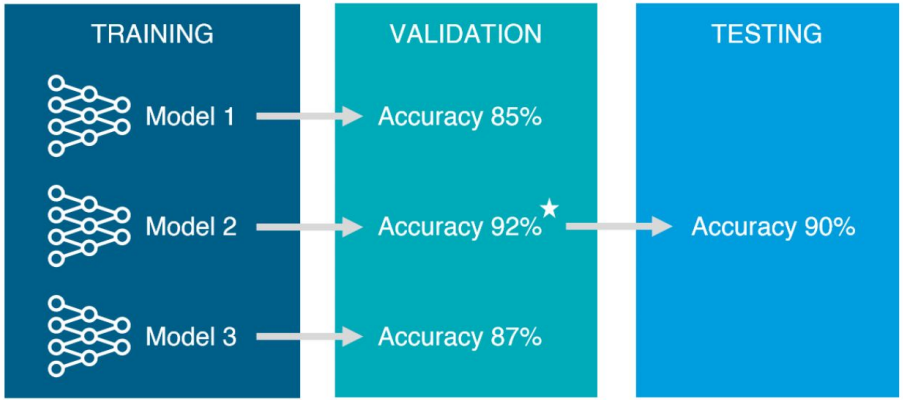
How can I validate the training results before testing?

Validation Set



Validation Set

- Test set과 같이 모델의 학습에 직접적으로 관여하지 않음
- 학습이 끝난 모델에 적용시켜 test set을 이용한 모델의 평가로 넘어가기 이전에 최종적으로 모델을 fine tuning 하는데에 사용

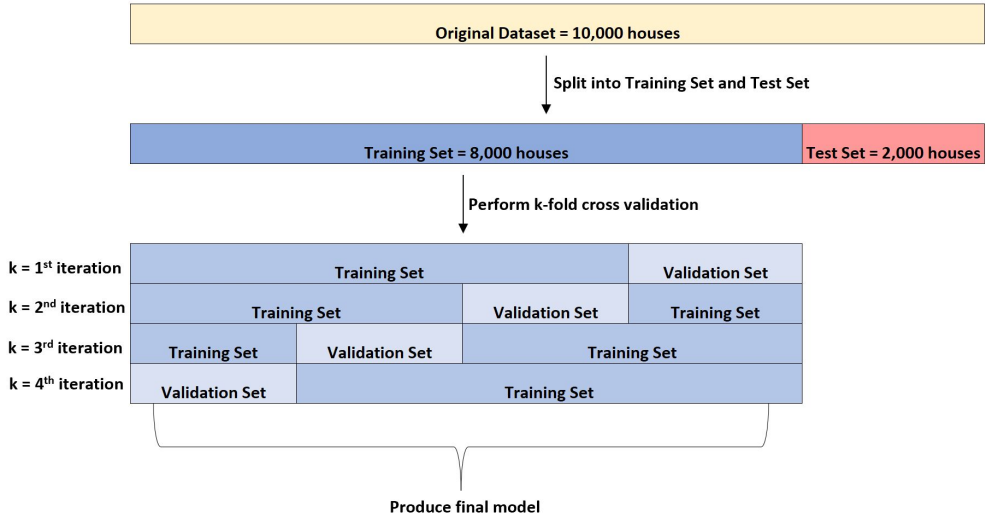


How to split a dataset

- 6:2:2 ?
- 6:1:3 ?

K-fold cross validation (교자 검증)

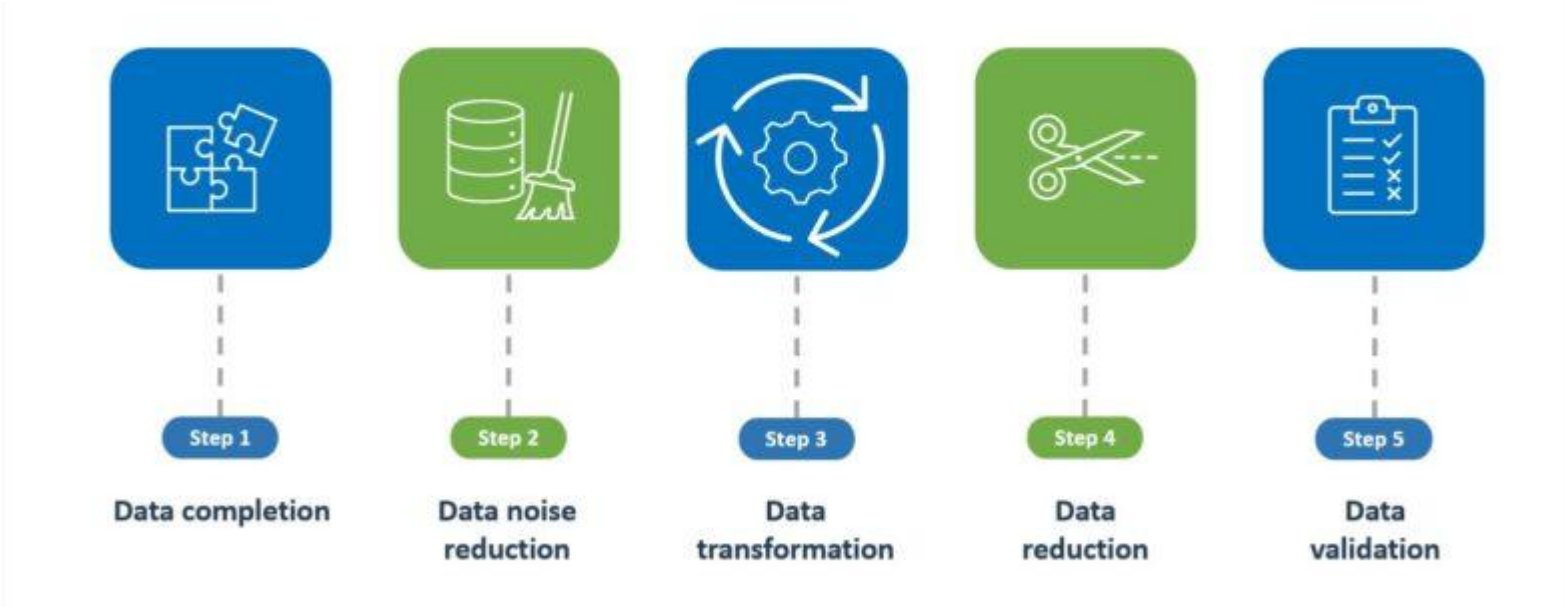
- 집합을 체계적으로 바꿔가면서 모든 데이터에 대해 모형의 성과를 측정하는 검증 방식
  - 데이터 셋이 적은 경우 정확도를 향상 시키는 방법으로 효과적
- 각 데이터를 학습하고 validation으로 평가를 한 다음 5개의 결과에 대해 평균을 내어 최종 성능 획득



03

Data Preprocessing

Step for Data Preprocessing



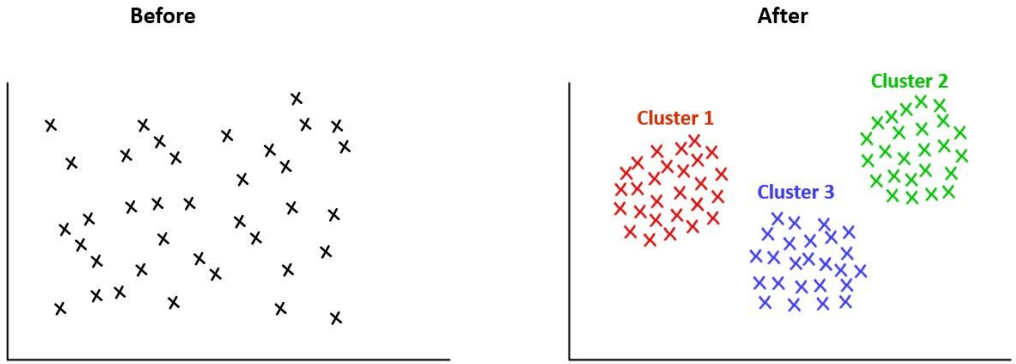
Ref: Data Preprocessing 5 Steps

Data Completion

- 손실된 데이터 추가
  - 손실된 데이터 중요도를 파악하여, 채워넣기
- Method
  - Interpolation
  - average (mean)

Data noise reduction

- Method
  - Binning method:
  - Regression method
  - Clustering: method



Data Noise Cleansing

- Binning method
  - This method involves arranging the data in different segments and binning it. Then the data in the bins can be replaced by their average, medium, or minimum and maximum values.
- Regression method
  - A linear or multiple variable regression function is used to smooth the data.
- Clustering
  - method Clustering helps smoothen the data by identifying similar data groups in a dataset and adding them to separate clusters (groups).

Data Transformation

- Normalization
  - 데이터의 범위 통일
- Attribute Selection
- Aggregation
- Concept hierarchy generation

Normalization (평균과 분산)

- 평균 (Mean)
  - 산술평균 (average, arithmetic mean), 기하평균, 조화 평균

$$\text{average} = \frac{\text{sum of values}}{\text{number of values}}$$

- 분산 (Variance)
  - 데이터가 평균을 기준으로 얼마나 퍼져 있는가를 나타내는 척도
  - 편차 (Deviation)
    - 평균과 데이터 값들의 차이
  - 분산은 편차 제곱의 평균

Normalization vs Standardization

- 정규화 (Normalization)
  - 공통 간격으로 데이터 자체를 늘이거나 줄이는 방법
  - Ex) Min-Max 정규화 (0 ~ 1 사이의 공통 간격으로 재배치)

$$\text{MinMax} = \frac{\text{data} - \text{data.min}}{\text{data.max} - \text{data.min}}$$

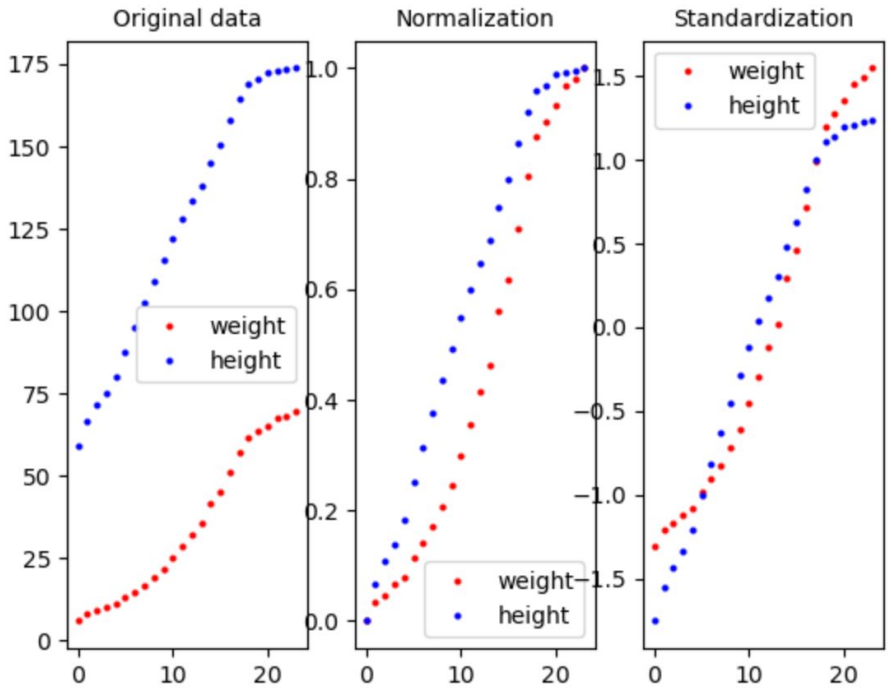
- 표준화 (Standardization)
  - 공통 척도로 데이터 자체를 표준화하는 방법
  - Ex) Z score: 평균 0, 표준편차 1인 공통 척도

$$\text{Zscore} = \frac{\text{data} - \text{data.mean}}{\text{data.std}}$$

weight	height
5.9	59.1
8	66.7
8.9	71.4
10.1	75
10.9	80.1
13.2	87.8
14.8	95.2
16.7	102.3
19.1	109
21.5	115.5
24.9	122
28.5	127.8
32.3	133.3
35.4	138
41.5	144.9
45.2	150.7
51	158.2
57.1	164.7
61.7	169.2
63.4	170.3
65.2	172.5
67.4	172.9
68.3	173.4
69.5	173.8

Normalization vs Standardization

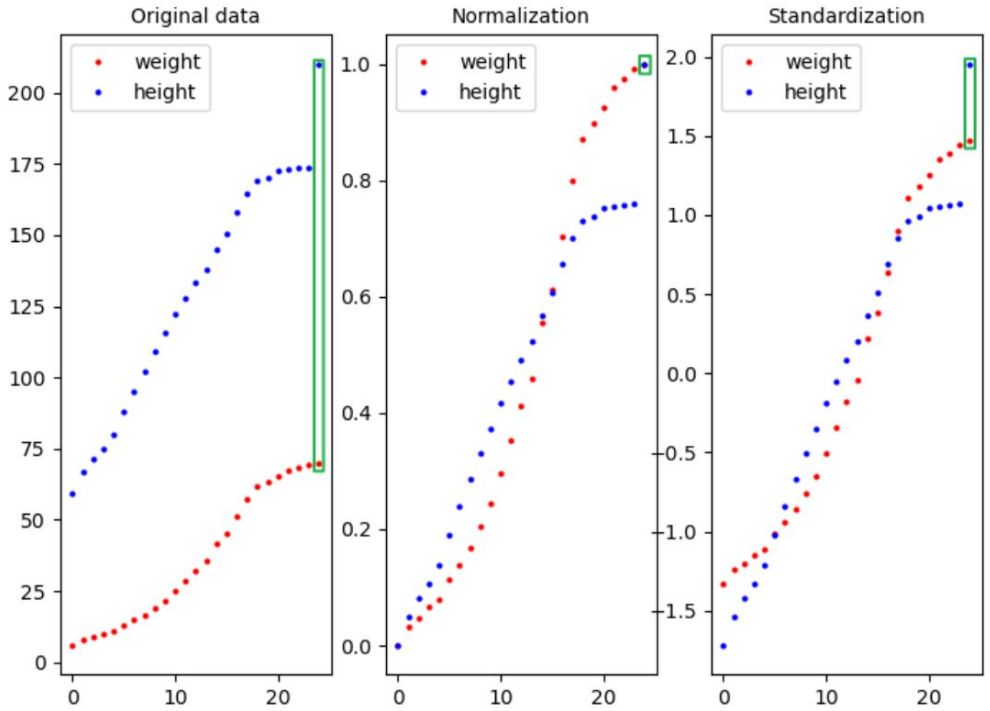
- Standardization의 경우 데이터의 특징이 그대로 살아 있음



weight	height
5.9	59.1
8	66.7
8.9	71.4
10.1	75
10.9	80.1
13.2	87.8
14.8	95.2
16.7	102.3
19.1	109
21.5	115.5
24.9	122
28.5	127.8
32.3	133.3
35.4	138
41.5	144.9
45.2	150.7
51	158.2
57.1	164.7
61.7	169.2
63.4	170.3
65.2	172.5
67.4	172.9
68.3	173.4
69.5	173.8

Normalization vs Standardization

- Outlier가 있는 경우 Min-Max Normalization은 Outlier 표현이 어려움



Data Reduction

- Dimensionality reduction
- Data Compression
  - 유사한 데이터를 압축
    - Non-lossy way
    - Lossy way
- 대표예
  - PCA (Principal Component Analysis)

Data Validation

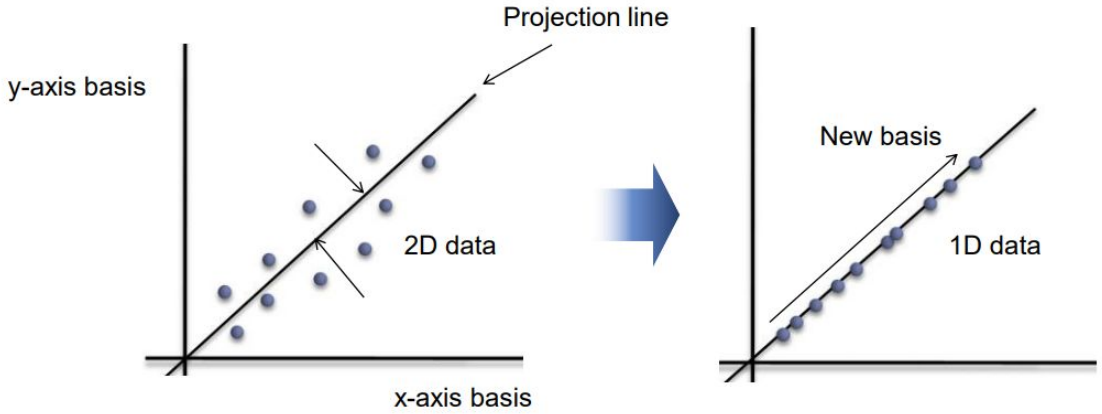
- 데이터의 품질 검증
  - 만족되지 않은 경우 다시 재 작업 필요

03

PCA

PCA (Principle Component Analysis)

- 분산이 최대화되는 방향으로 데이터를 줄임으로써, 데이터의 주요 패턴을 캡처하며 차원을 줄이는 기법
- 차원 축소(Dimensionality Reduction)
  - 데이터의 차원(특성의 수)을 줄이며 데이터의 중요한 정보를 최대한 보존



PCA (Principle Component Analysis)

❖ 장점

- 계산 효율성 증가
  - 데이터의 차원이 감소하면, 모델 학습과 예측에 필요한 계산량 감소
- 데이터 시각화
  - 차원을 축소함으로써, 데이터를 시각적 이해도 증가
- 노이즈 제거
  - 노이즈와 불필요한 정보를 제거
- 과적합 방지
  - 차원의 저주 문제를 해결하여 모델의 과적합을 줄임

차원의 저주(Curse of dimensionality)란 데이터 과학과 머신 러닝에서 데이터의 차원의 증가할수록 해당 공간의 크기가 기하급수적으로 증가하며, 데이터 분석이나 모델 학습에 어려움을 초래하는 현상

PCA (Principal Component Analysis)

❖ 단점

- 데이터의 일부 정보 손실 발생
- 어떤 차원을 유지하고 제거할지 결정하는 것이 복잡
- 정보 손실로 인해 모델은 데이터의 복잡성을 충분히 표현하지 못해, 이는 under-fitting의 원인

Target

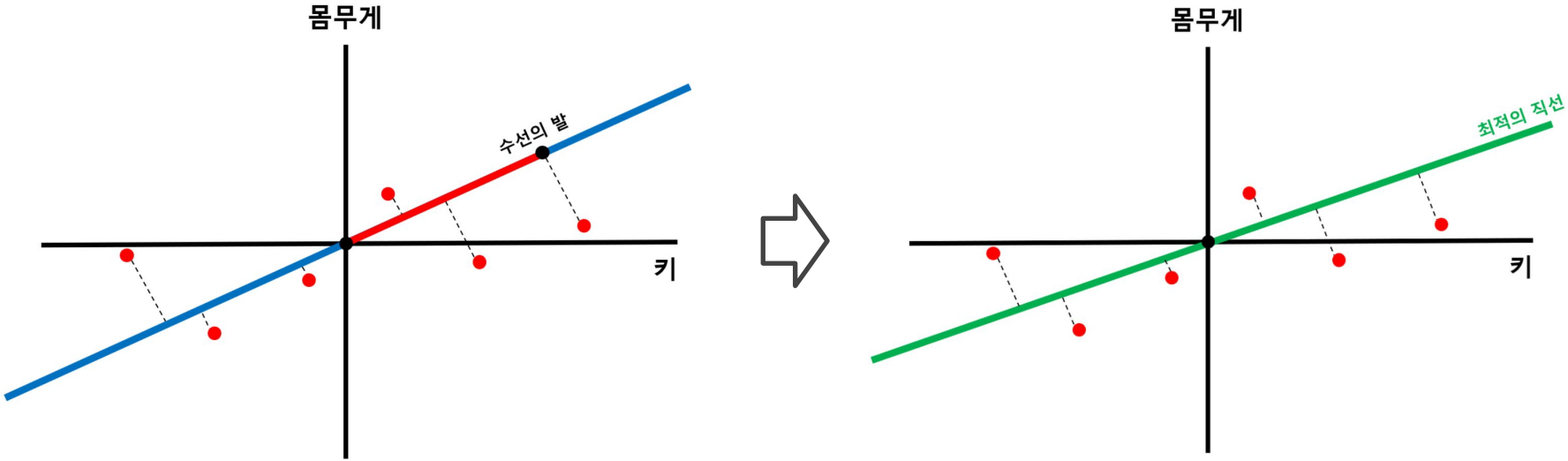
❖ Idea

- 데이터의 분산을 최대화하는 주성분을 찾자
- 분산은 데이터가 얼마나 특정 방향으로 퍼져 있는지를 나타내며, 큰 분산은 해당 방향에 데이터의 주요 정보나 패턴이 포함되어 있음을 의미
- 분산이 큰 주성분 방향으로 데이터를 투영함으로써 차원을 줄이면, 정보의 손실을 최소화할 수 있고 데이터들 사이의 차이점이 명확해진다!

Target

- Sum of Square의 최소화

$$\min \sum_{i=0}^{n} \left(X_i - \bar{X}\right)^2$$



PCA

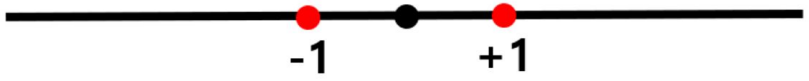

Normalization
Or
Standardization
Covariance
Matrix
Eigen Vector
Eigen Value
Principal
Component
Reconstructing
the original data

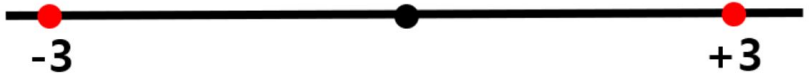
Variance & Covariance

$$\Sigma = \begin{pmatrix} \mathrm{Var(X)} & \mathrm{Cov(X,Y)} \\ \mathrm{Cov(X,Y)} & \mathrm{Var(Y)} \end{pmatrix}$$

$$\mathrm{Cov}(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}$$

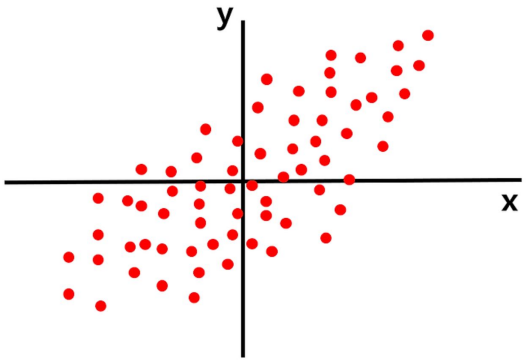
$$\mathrm{Var}(x) = \frac{\sum(x-\bar{x})^2}{n}$$



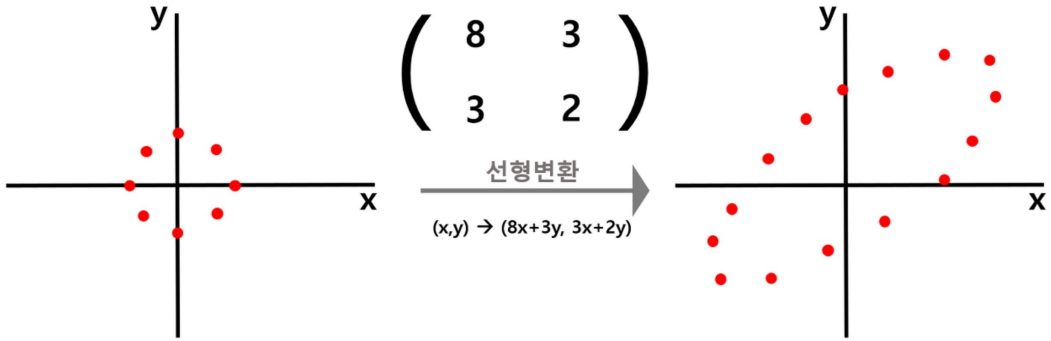


Covariance Matrix

$$\Sigma = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{pmatrix}$$



X축 방향으로 퍼진 정도



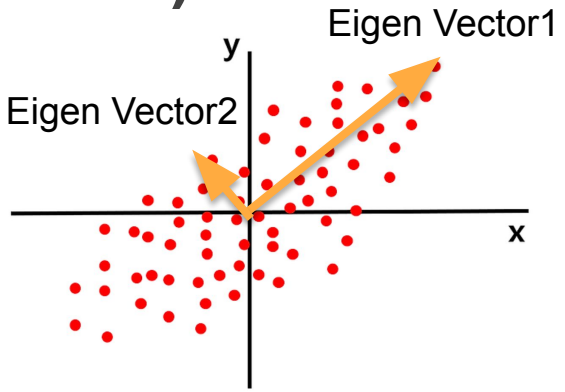
$$\Sigma = \begin{pmatrix} 8 & 3 \\ 3 & 2 \end{pmatrix}$$

XY축 방향으로 퍼진 정도

Y축 방향으로 퍼진 정도

Eigenstuff (Eigen Vector & Eigen Value)

- Eigen Vector
  - 선형 변환을 한 이전과 이후의 방향이 같은 벡터
  - direction of maximum variance
- Eigen Value
  - Eigen Vector의 변환 전과 후의 길이 변화 비율
  - the magnitude of this variance

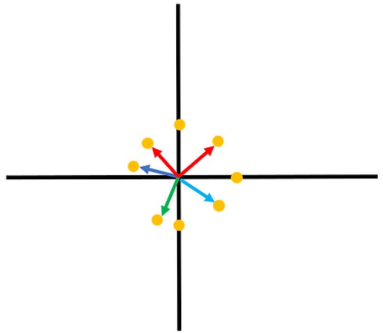


$$\begin{pmatrix} 8 & 3 \\ 3 & 2 \end{pmatrix}$$

Eigenvector

$$v = \lambda v$$

Eigenvalue

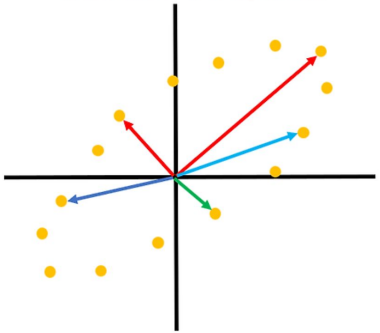


$$\begin{pmatrix} 8 & 3 \\ 3 & 2 \end{pmatrix}$$

선형변환

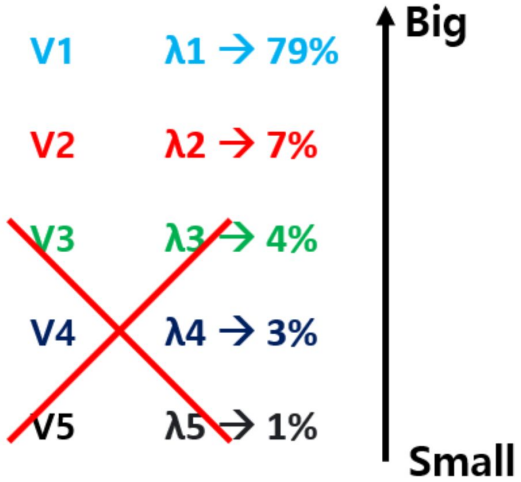
(x,y) → (8x+3y, 3x+2y)

빨간 벡터 2개는 선형변환 전후에도 방향이 동일하다. → Eigenvector



Principal Component

- Principal component 추출
  - 원하는 만큼 Eigenvector를 쳐냄으로써 차원 축소를 해준다



Example

	키	몸무게
사람 1	170	68
사람 2	174	72
사람 3	172	84
사람 4	176	76
사람 5	168	60
사람 6	166	74