데이터 분석론

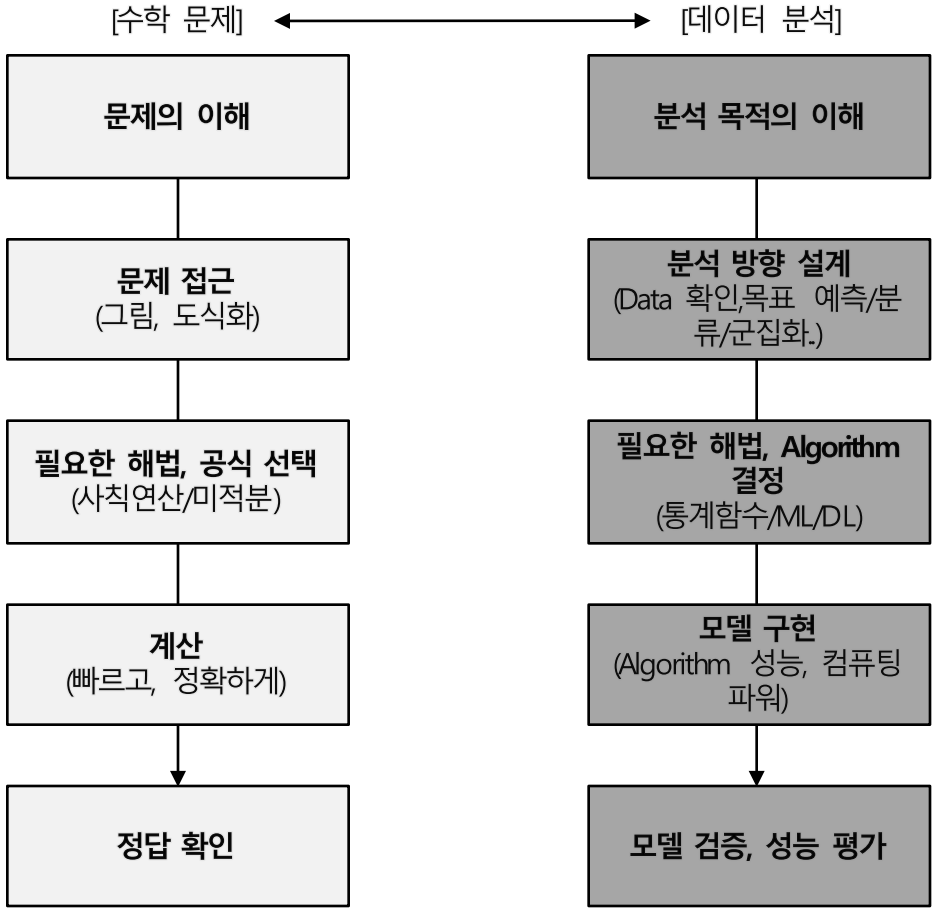
2019.07

황소희

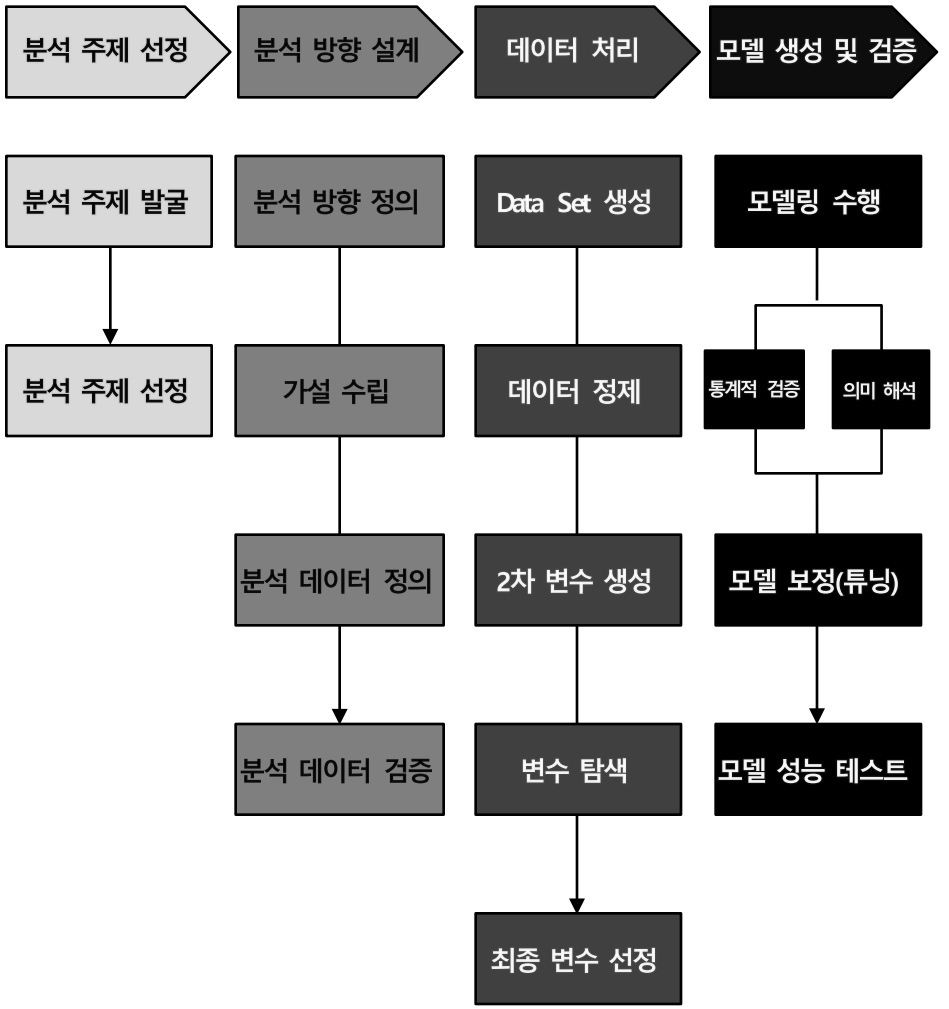
데이터 분석이란?

수학 문제처럼 Data&Data Analytics 기술을 활용하여 Question을 해결해나가는 과정

Data Analysis



Process



활용 목적에 따른 예시

활용 목적	
예측/추정	• 따릉이 요일별 수요 예측 • SK Telecom 통신 이상 징후 예측
분류	• SK Telecom 데이터 과다/정상 이용자 분류
군집	• 따릉이 이용자의 용도 별 군집 (출퇴근용/취미용) • 웹 로그를 이용하여 방문자의 행동 패턴 별 군집
패턴/구조 발견	• 웹 로그를 이용하여 방문자의 행동 패턴 발견
차원 축소	• 사진 파일 용량 축소 • 일반적으로 3차원 이상의 데이터를 표현할 때 사용

활용 목적에 따른 모델링 기법

사용 기법	
예측/추정	• Linear Regression (선형 회귀분석) • LASSO (Least Absolute Shrinkage and Selection Operator) • 시계열 분석
분류	• Decision Tree • Logistic Regression • K-NN (k-Nearest Neighbor)
군집	• K-Means Clustering (비계층적 군집분석) • Hierarchical Clustering (계층적 군집분석)
패턴/구조 발견	• Association Rule Analysis • Sequence Analysis
차원 축소	• PCA (Principal Component Analysis) • Factor Analysis • SVD (Singular Value Decomposition)

M/L, D/L 알고리즘의 활용으로 많은 시간과 노력이 소요되던 '데이터 처리' 단계가 손쉬워 짐

Basic

"컴퓨터가 수행해야 할 가이드라인(논리)을 사람이 제시해주고, 컴퓨터는 그 가이드라인을 따라가는 것"

Example, 자율주행차

다음과 같은 지시사항들을 사람이 직접 프로그램 안에 다 넣어주어야 함. 추가/변경 지시사항의 경우, 다 직접 넣어주어야 함.

① 비가 올 때는 속도를 늦춰야 한다.
② 신호등이 빨간 불이면 멈추고 초록 불이면 지나간다.
③ 차선 변경시에는 반드시 깜빡이를 켜야 한다.

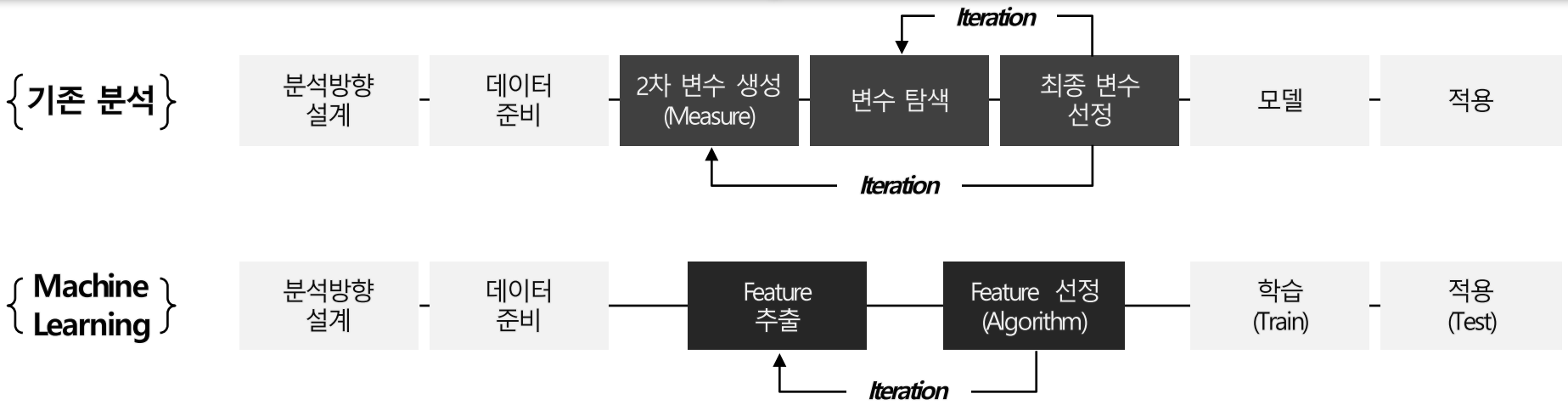
Machine Learning

"가이드라인(논리)을 컴퓨터가 스스로 찾아 내는 것"

Example, 자율주행차

머신 러닝의 경우, **기계가 수 천, 수 만 가지의 지시사항들을 스스로 학습**해 나감. (데이터만 충분하게 있다면!)
머신러닝에서 필요로 하는 것은 학습에 필요한 데이터(입력값)와 라벨(출력값) 뿐임.

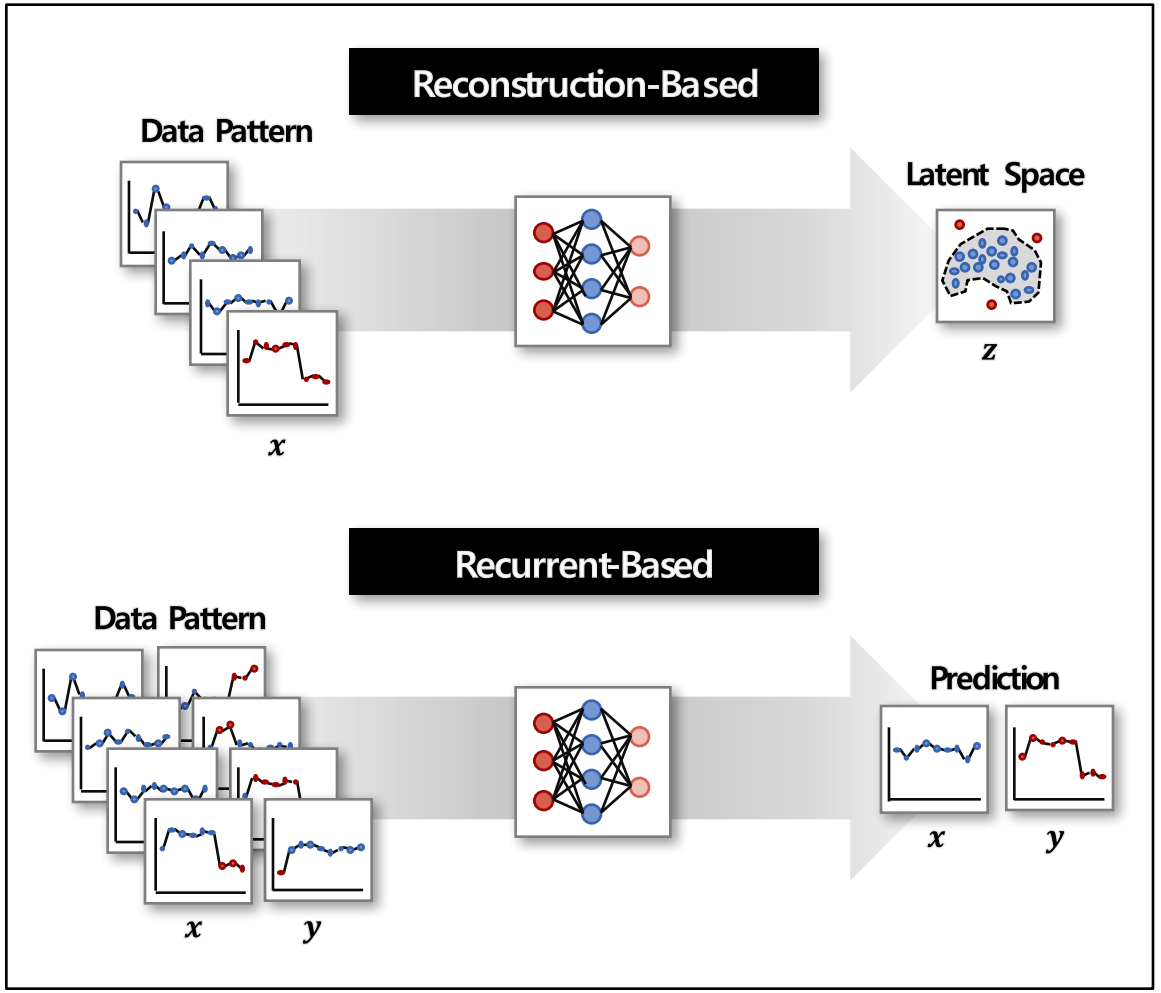
Machine Learning in Data Analysis



D/L 사례 (SK telecom anomaly detection)

통신의 이상이 발생했을 경우, 통신 이상을 감지하여 예측 및 감지를 할 수 있도록 하고 싶었고, Data 고유의 Pattern을 학습하여 예상 수치 및 정상/이상 여부를 판단

알고리즘



접근방법

1 자기복제를 통한 Pattern 학습

- Reconstruction 기법으로 Pattern이 분포하는 Latent Space를 탐색
  → **Data의 분포 학습**
- 이상 탐지: 정상/이상 빈도의 불균형을 이용
  - 정상 Case에 편향된 Training
  - **학습된 분포에서 벗어난 Data → 이상 Pattern**

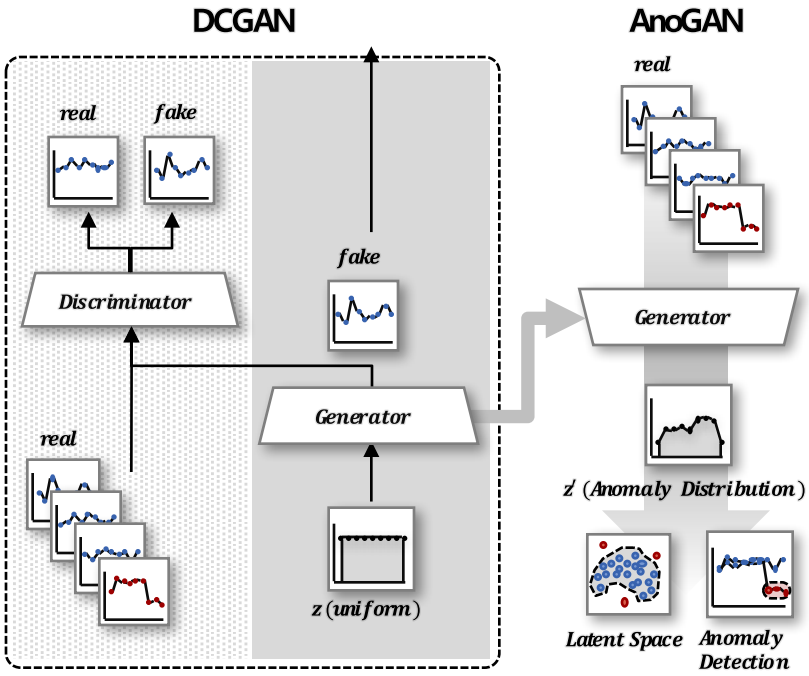
2 Many-to-Many 방식 Pattern 학습

- Time Sequence를 기준으로 전/후 Pattern 학습
  - **현재 Pattern에 대응하는 이후 Pattern 예측**
- 이상 탐지 : 예측값과 실제값 비교
  - 예측값 대비 실제값 편차 Thresholding
  - **기준치를 초과하는 차이 → 이상 Pattern**

Deep Learning Algorithm 자세한 설명 (pass해도 돼!)

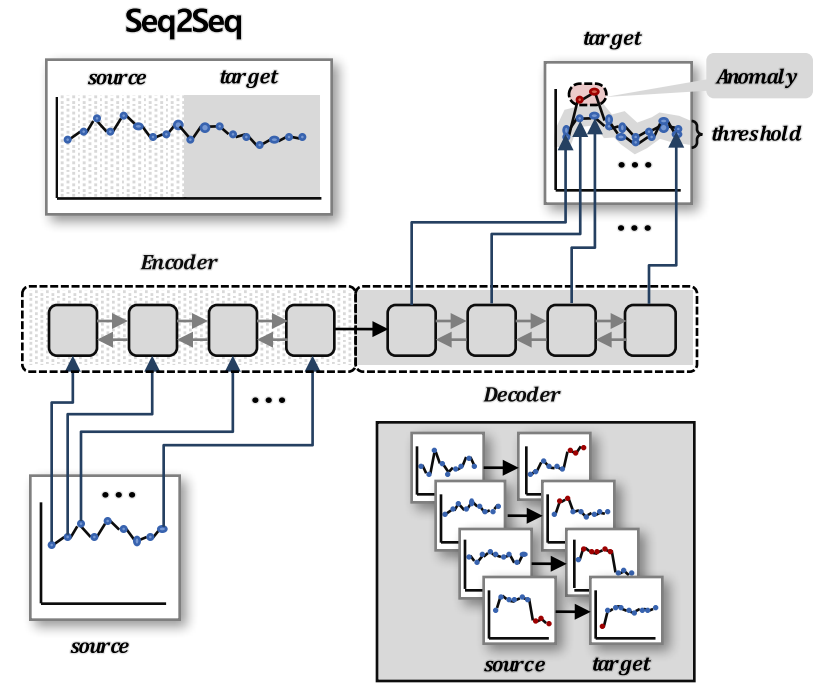
Reconstruction-Based

- **AnoGAN**
  - DCGAN 기반 **Unsupervised Anomaly Detection Algorithm**
  - ① DCGAN의 Generator로 Data Pattern을 학습
  - ② 기학습된 Generator를 이용해 *Anomaly z* 분포를 학습
  - ③ *Anomaly z* 분포를 활용하여 Anomaly Detection



Recurrent-Based

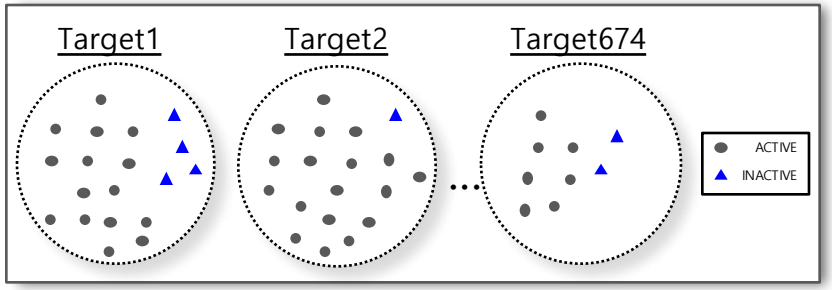
- **Bi-LSTM Seq2Seq**
  - LSTM Cell 기반 **Supervised Sequential Prediction Algorithm**
  - ① *Encoder* 를 통해 *source sequence* 입력
  - ② *Decoder* 는 *target sequence* 출력
  - ③ *target* 예측값과 실제 측정값의 편차를 이용해 Anomaly Detection



D/L 사례 (SKBP Target Deconvolution)

Target Deconvolution은 신약 개발 시, 미지의 화합물이 어떤 약효를 갖는지 즉, 어떤 Target Protein에 반응하는지를 예측하는 프로젝트로, Target Deconvolution의 issue인 데이터의 불균형을 해소할 수 있는 one-shot learning 방법을 모델로 채택함.

Data Imbalance



- Active/Inactive 데이터 불균형

→ Active 데이터 활용

- Target 별 데이터 불균형

→ One shot learning 활용

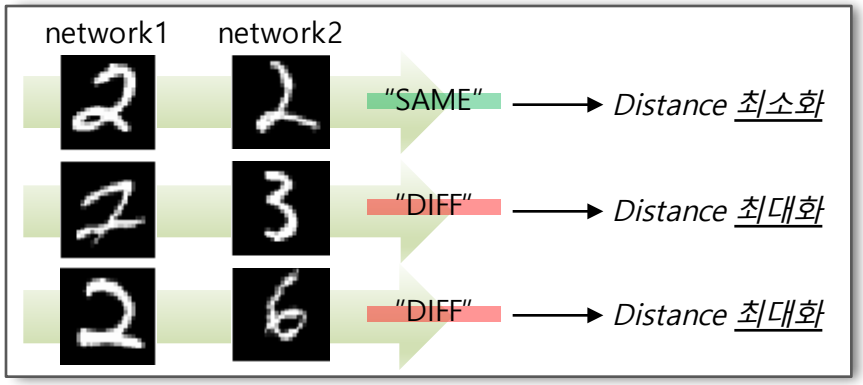
Modeling

A ***One Shot Learning***

- Classification 시, training data가 적은 경우 적용 가능한 방법론
- 각 클래스에 속하는 데이터와 분류하고자 하는 데이터를 비교하여 분류
- Distance 측정이 가능하여, Unknown Data가 어느 target에 active한지 ranking 가능

B ***Siameses Network***

- class간의 구별되는 feature를 생성하도록 학습
- 2개의 input data의 Class 일치 여부로 distance space 생성
- 2 input class 일치: 추출된 feature의 distance 최소화
  2 input class 불일치: 추출된 feature의 distance 최대화



End of Document