Big Data Analytics Programming

Week-10. Anomaly Detection

Jungwon Seo, 2020-Fall

Anomaly Detection

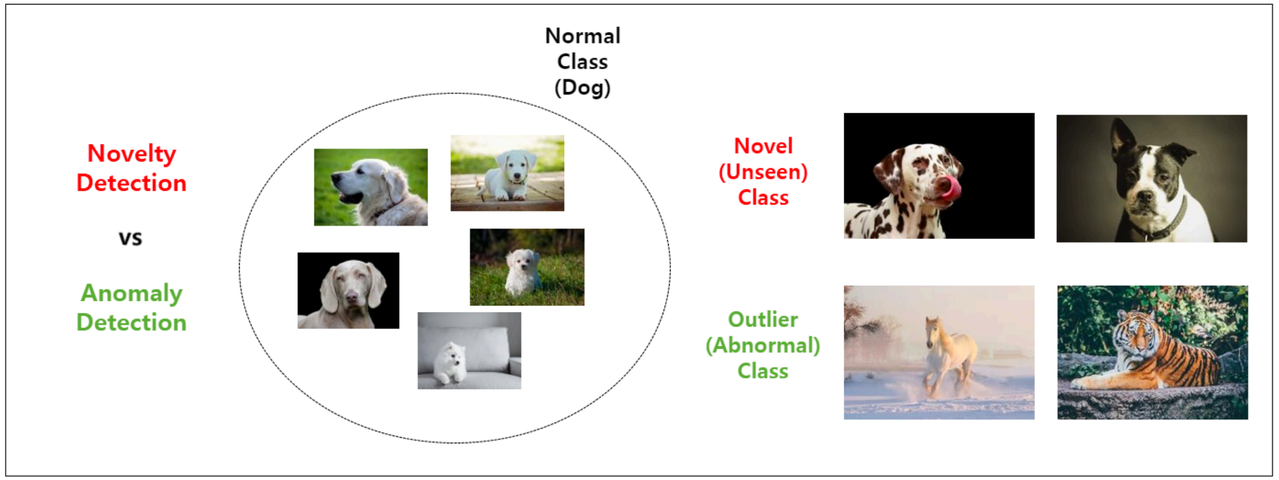
Overview

- Anomaly란?
- Conventional Approach
- Deep Neural Networks Approach

What is an anomaly?

Anomaly
이상치의 정의

- Anomaly
  - 대부분의 데이터와 **다른** 희귀 데이터
  - Novelty, Outlier로도 불리울수 있지만, 약간의 뉘앙스 차이가 존재
    - Novelty: 같은 부류지만 이전에 본적이 없는 (Unseen)
    - Outlier: 전혀 관련이 없는



https://hoya012.github.io/blog/anomaly-detection-overview-1/

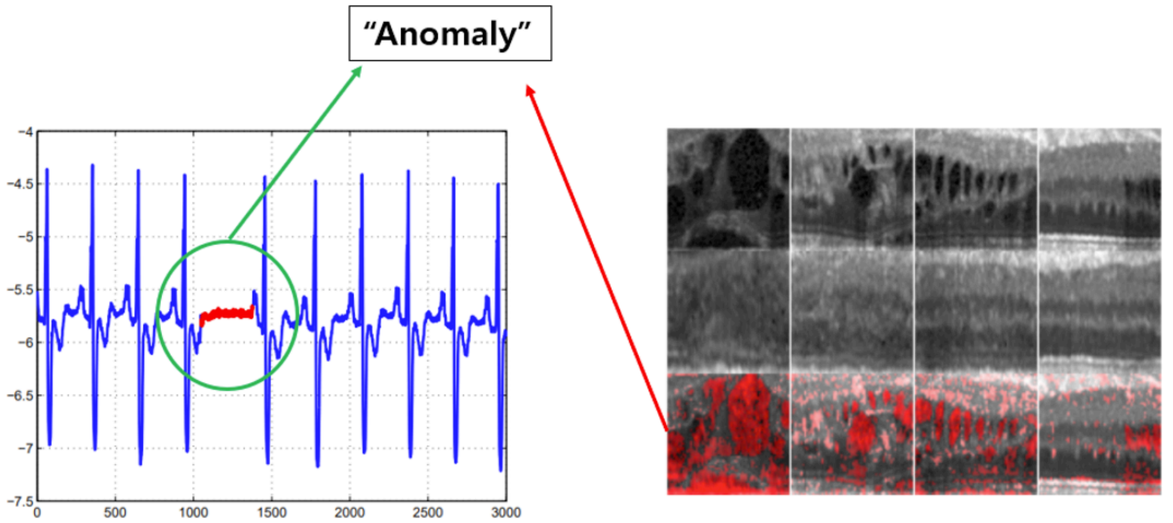
Anomaly
Data point of view

- Point anomaly
  - 데이터 셋 하나가 이상
  - 독립적으로 여러개
- Collective anomaly
  - 데이터 셋 내의 여러 **관련된/연결된** 데이터셋이 이상
- Contextual anomaly
  - 전체적인 데이터셋의 맥락을 고려했을 때 이상
  - 예: 시계열 데이터의 이상 Peak

Anomaly Detection

Anomaly Detection은 어디에 쓸 수 있을까?

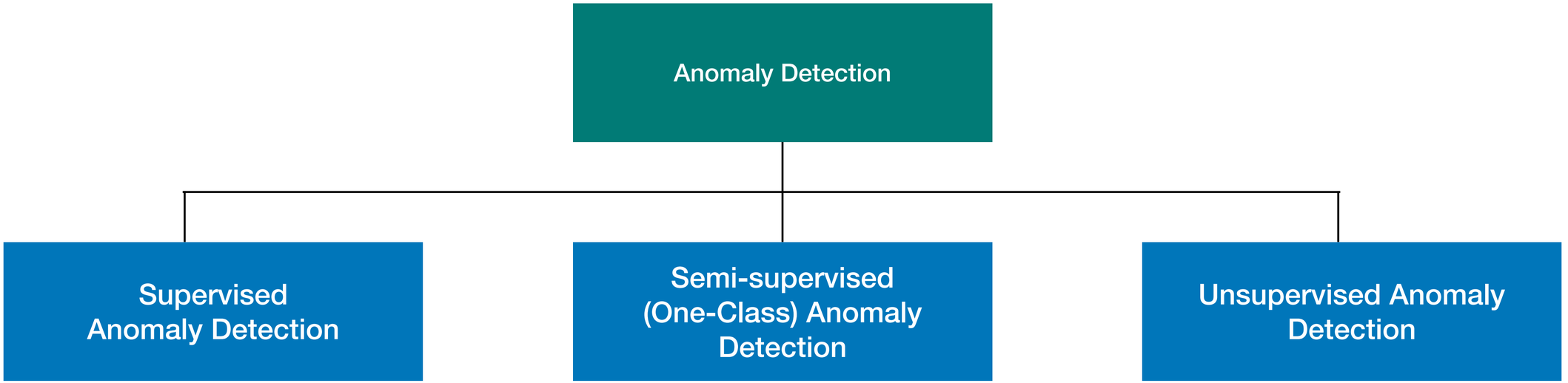
- 사례
  - 신용 카드 사기 탐지
  - 통신 사기 탐지
  - 네트워크 침입 탐지, 결함 탐지
  - Video Surveillance
  - 제조업 공정과정에서 이상탐지



Reference
[1] Anomaly Detection of Time Series, 2010
[2] Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery, 2017

Anomaly Detection

Anomaly Detection의 접근법



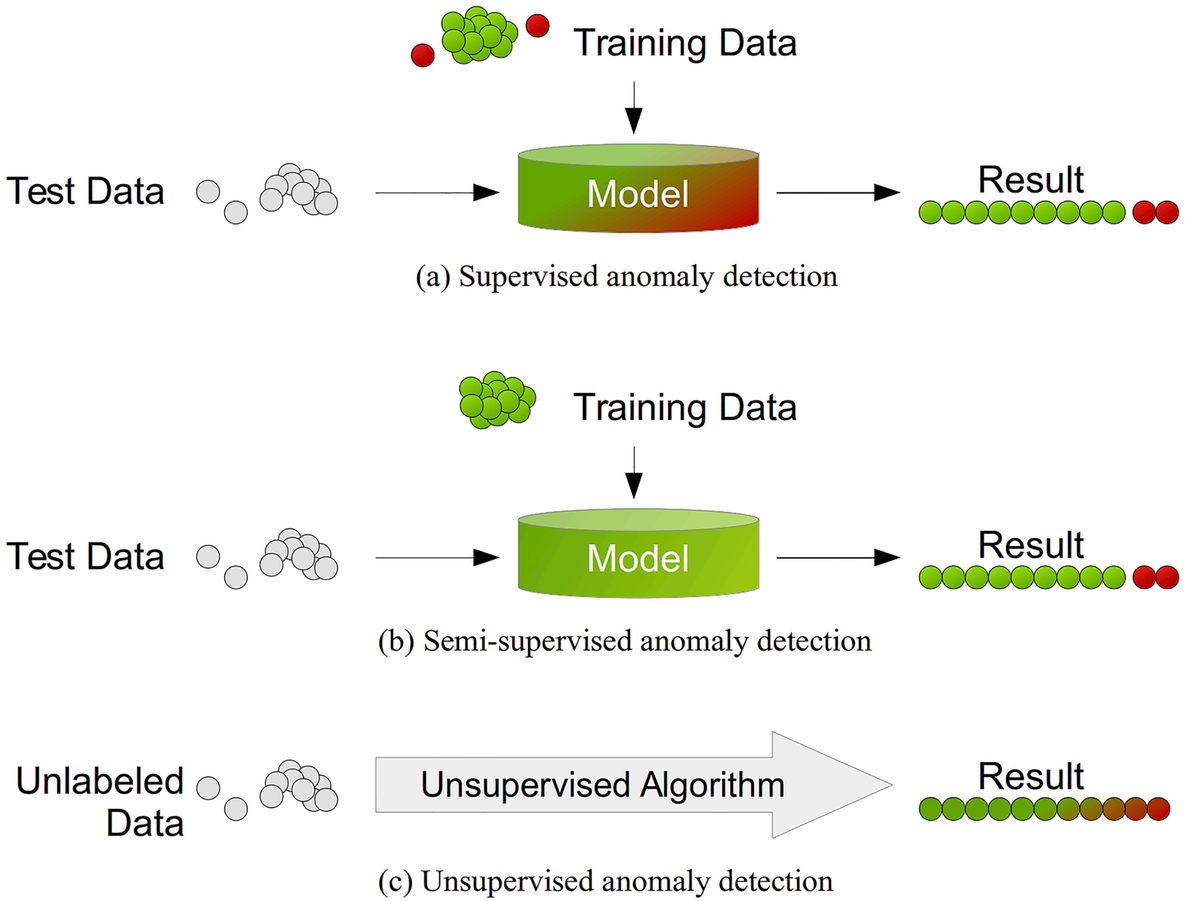
- 정상/비정상 Label이 주어진 경우
- 지도학습이므로 정확도가 높은편
- 일반적으로 비정상샘플이 정상샘플에 비해 적으므로 Class-Imbalance 문제에 직면함

- 정상/비정상 Label이 주어진 경우
- **정상 샘플만을** 갖고 학습하여, 정상의 범주(boundary) 를 결정
- 정상 샘플만을 활용하기 때문에, Class-Imbalance 문제에 직면하지 않음
- 지도 이상 탐지에 비해 **성능이 떨어짐**

- **Label 없이 모두 정상이라고 가정**
- 클러스터링 및 거리 기반의 비지도학습 알고리즘 사용
- PCA나 AutoEncoder를 이용하여, 원본과 복원본을 비교하여 차이를 기준으로 판별하는 방법도 많이 쓰임
- 정확도가 높지 않고, hyperparameter에 의해 영향을 많이 받음

Anomaly Detection

Anomaly Detection의 접근법



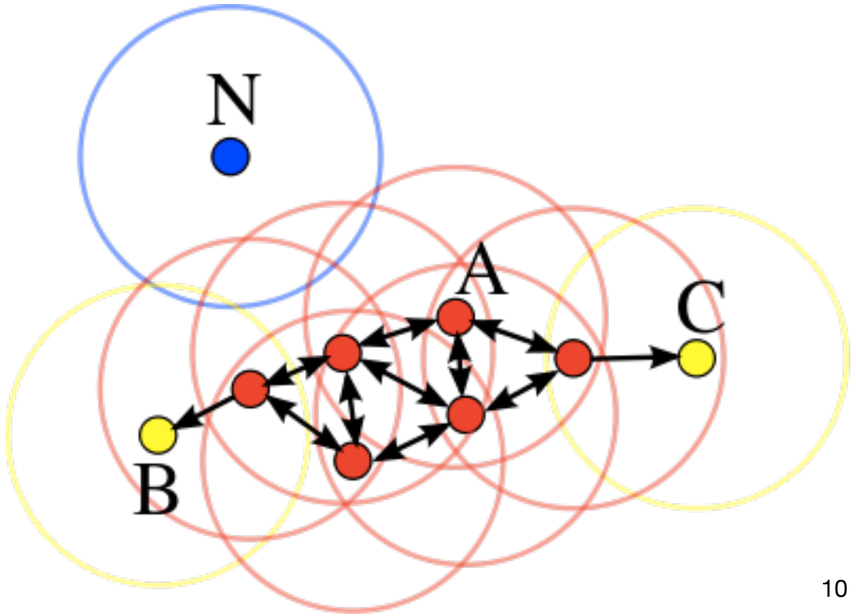
doi: https://doi.org/10.1371/journal.pone.0152173.g001

Conventional Techniques

DBSCAN

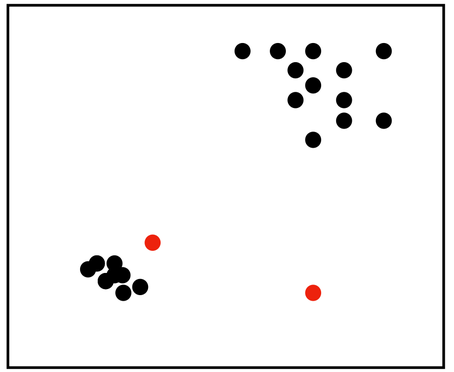
밀도기반 클러스터링 (Unsupervised)

- 데이터의 분포와 밀도를 고려하여 클러스터를 구성
- 클러스터링에 사용되었지만, 과정 중에 이상치를 탐색 가능
- 이상 데이터의 기준: 핵심 및 외각 데이터에 속하지 못한 데이터



Local Outlier Factor

지역의 밀도를 같이 고려하자



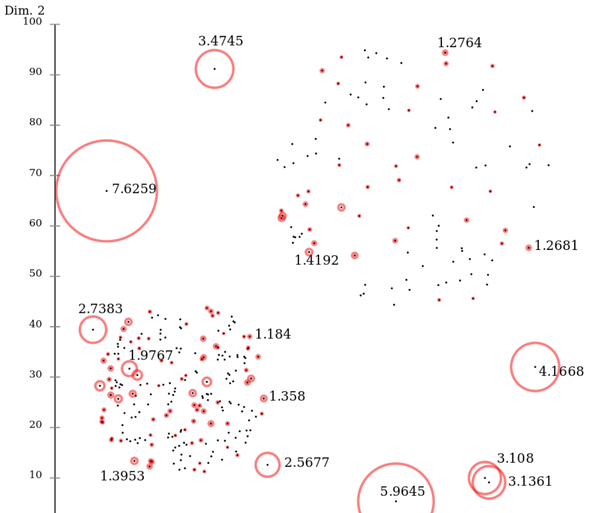
- 기존 밀도기반 알고리즘들의 한계
  - 밀도가 상이한 클러스터가 존재하는 경우 파라미터를 결정하기가 어렵다 (우상단 그림)
    - 반경, 반경 내 속해야하는 데이터 포인트의 수
- 구성요소
  - **k_distance(p)** : k 번째로 가까운 데이터와의 거리
  - **reachability distance(p,o)** : 주변 데이터 o를 고려한 거리
    - reachability-distance_k(A,B)=max{k-distance(B), d(A,B)}
  - **local reachability density(p)** : p주변의 k-neighbor들과의 reach dist의 역수

$$\mathrm{lrd}_k(A) := 1/\left(\frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A, B)}{|N_k(A)|}\right)$$

- **Local Outlier Factor(p)** :

$$\mathrm{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\mathrm{lrd}_k(B)}{\mathrm{lrd}_k(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \mathrm{lrd}_k(B)}{|N_k(A)| \cdot \mathrm{lrd}_k(A)}$$

- LOF(k) ~ 1 means Similar density as neighbors,
- LOF(k) < 1 means Higher density than neighbors (Inlier),
- LOF(k) > 1 means Lower density than neighbors (Outlier)

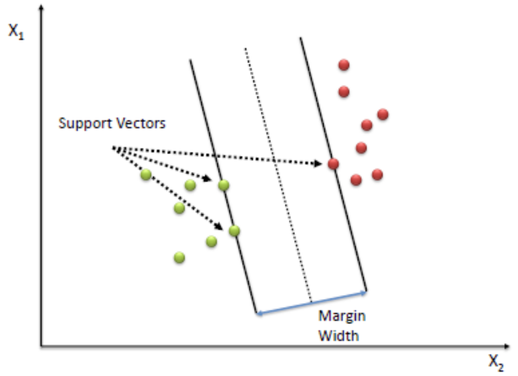


https://commons.wikimedia.org/wiki/File:LOF.svg#/media/File:LOF.svg

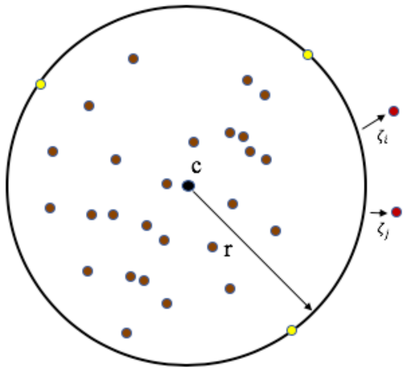
One-Class SVM

정상의 Boundary를 구할 수 있다면?

- Support Vector Machine
  - Margin이 최대가 되는 경계면(hyperplane)을 찾는 방식으로 분류 문제에 활용
  - DNN 이전의 최강자
- One-class SVM
  - Unsupervised Learning
    - But semi-supervised actually..
  - 정상 데이터로만 훈련을 진행하므로써, 정상의 **영역**을 계산
    - Finding the smallest hypersphere
  - 정상 Boundary의 밖에 위치하는 데이터 포인트들은 이상으로 간주



SVM Classifier



One-class SVM

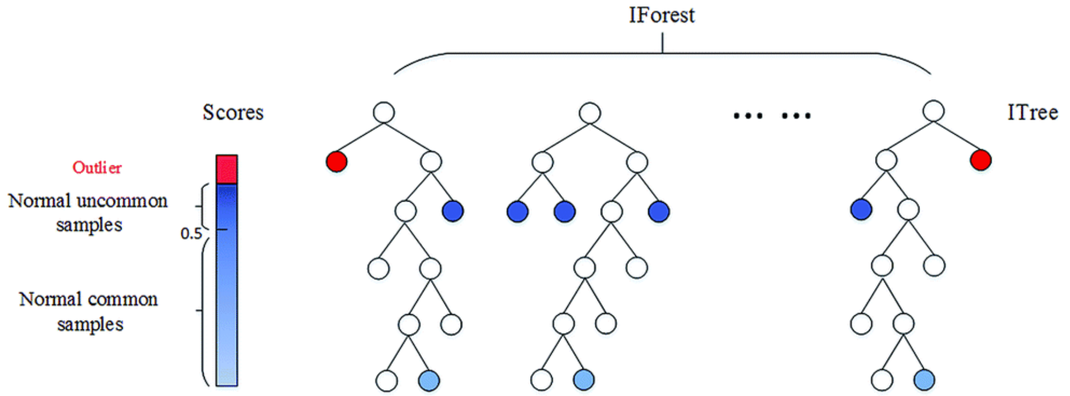
https://ratsgo.github.io/machine%20learning/2017/05/23/SVM/

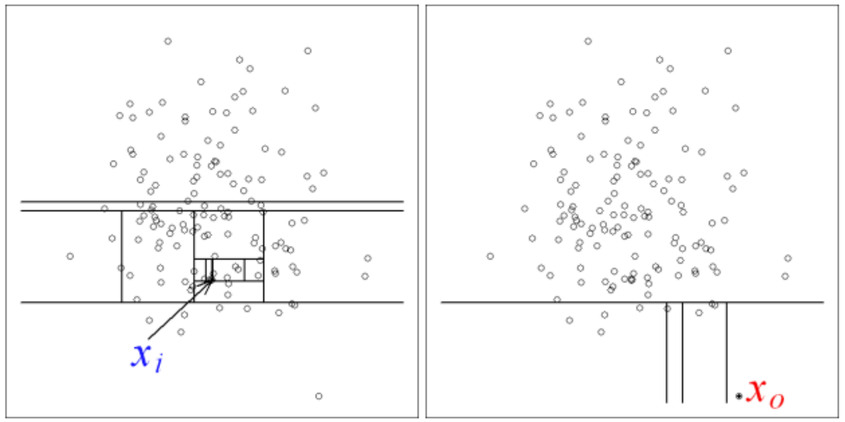
https://en.wikipedia.org/wiki/File:One-class_data_description_TAX.png#/media/File:One-class_data_description_TAX.png

Isolation Forest

정상에 집중하기 보다 비정상에 집중하는..

- Decision Tree
  - Feature 단위로 데이터셋을 분류하는 기법
    - If x1>0, if x2==“animal” ,..
- Random Forest
  - Decision Tree의 앙상블 버전
  - 앙상블 기법: **서로독립적인** 모델들의 집단 지성을 이용
- Isolation Forest
  - Isolation : ‘separating an instance from the rest of the instances’
  - 비정상적 데이터의 경우 Isolate하는데 더 적은 파티션이 필요
    - Tree기준으로 보면 Root에 더 가깝다.





(a) Isolating x_i (b) Isolating x_o

Chen, Wo-Ruo, et al. "Representative subset selection and outlier detection via isolation forest." *Analytical Methods* 8.39 (2016): 7225-7231.

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008.

Other techniques

- Distribution Based
  - Gaussian Mixture Model
  - Elliptic Envelop
- Dimensionality Reduction Based
  - Linear Dimensionality Reduction : PCA
  - Non-linear Dimensionality Reduction : Manifold Learning

Deep Learning for Anomaly Detection

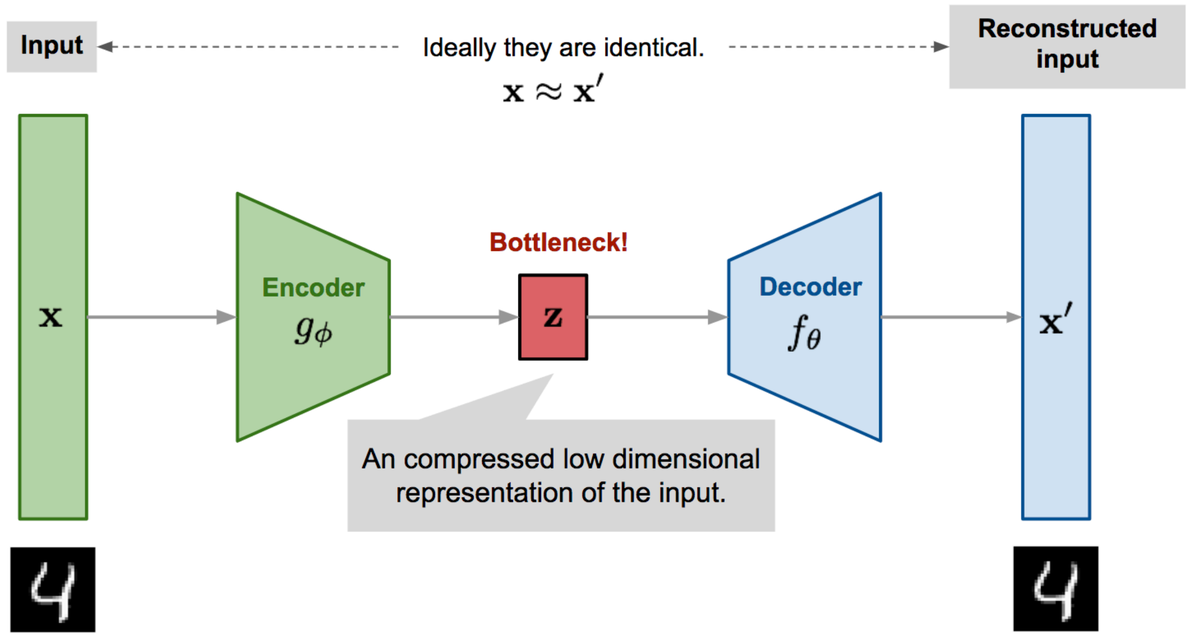
Deep Learning for Anomaly Detection
딥러닝을 활용한 이상탐지

- Why Deep Learning?
  - 딥러닝을 활용하는 다른 분야와 같은 이유
  - 대량의 Feature와 Non-linearty가 존재하는 데이터셋에 대해서 이상탐지를 하기 위해
- How?
  - Auto-encoder Approach
  - One-class Neural networks (OC-NN)

Auto Encoder

인코딩한 것과 디코딩한 것의 차이가 크다면?

- Motivation
  - DNN을 어떻게 비지도학습에 활용 할 수 있을까?
    - 차원축소
    - Representation
- Encoding
  - Input to representation
- Decoding
  - Representation to output
- 만약에 X와 X'이 차이가 크다면,
  이상치라고 볼 수 있지 않을까?

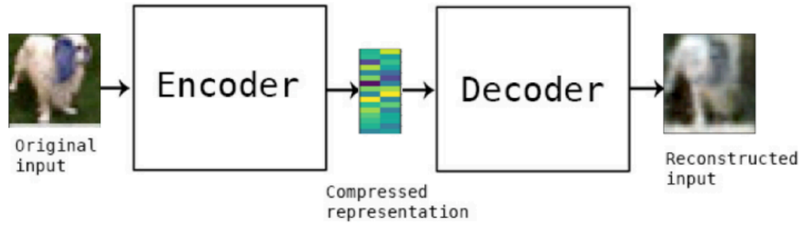


https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

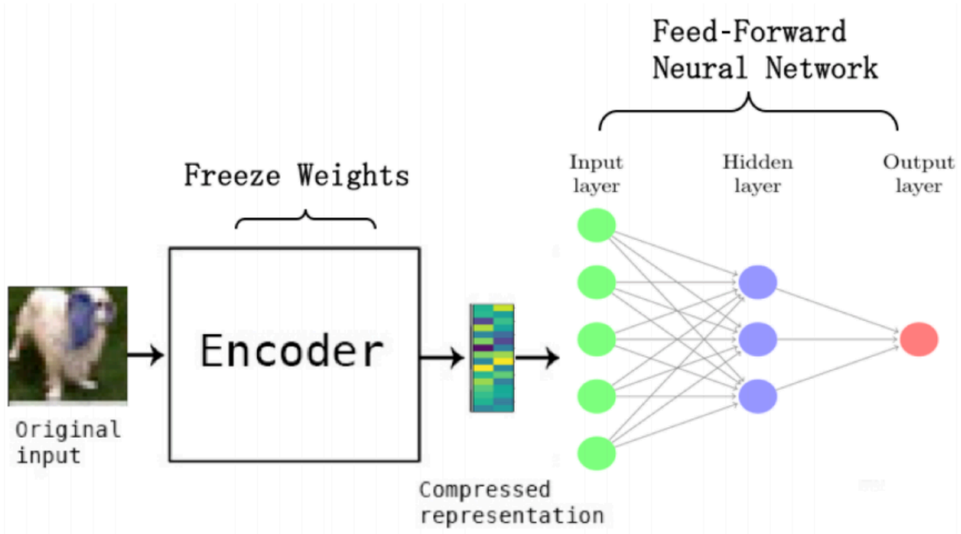
One-class Neural networks (OC-NN)

SVM 대신 Neural Networks로

- Motivation
  - DNN이 비선형성을 띄는 데이터셋에 잘 동작한다면, SVM 말고, NN을 쓰는게 더 좋지 않을까?
- 동작방식
  - 오토인코더 이용한 인코더 확보
  - 인코더를 이용해, Input의 차원축소
  - 축소된 인풋을 이용해 One-Class 학습



(a) Autoencoder.



(b) One-class neural networks.

Ruff, Lukas, et al. "Deep one-class classification." *International conference on machine learning*. 2018.

E.O.D