Chap 1
이상 탐지란?

20101659 이유경

기업 needs

- 제조

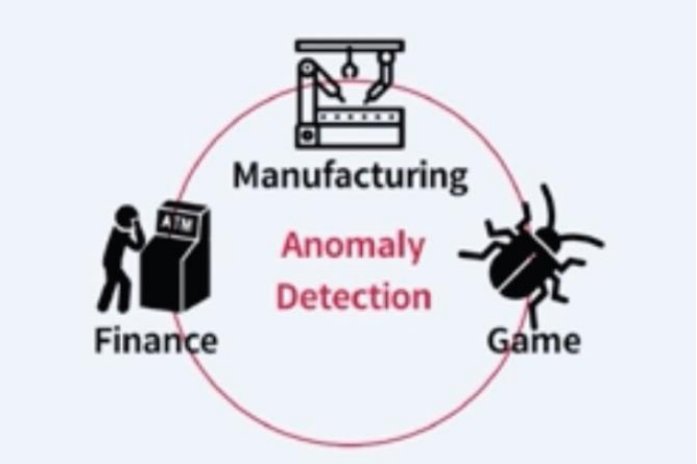
- 설비 예지 보전으로 고장 발생 시 loss 비용 절감

- 금융

- FDS을 통해 이상거래를 탐지, 사기 거래로 발생할 수 있는 리스크 절감

- 게임

- 어뷰징 해결



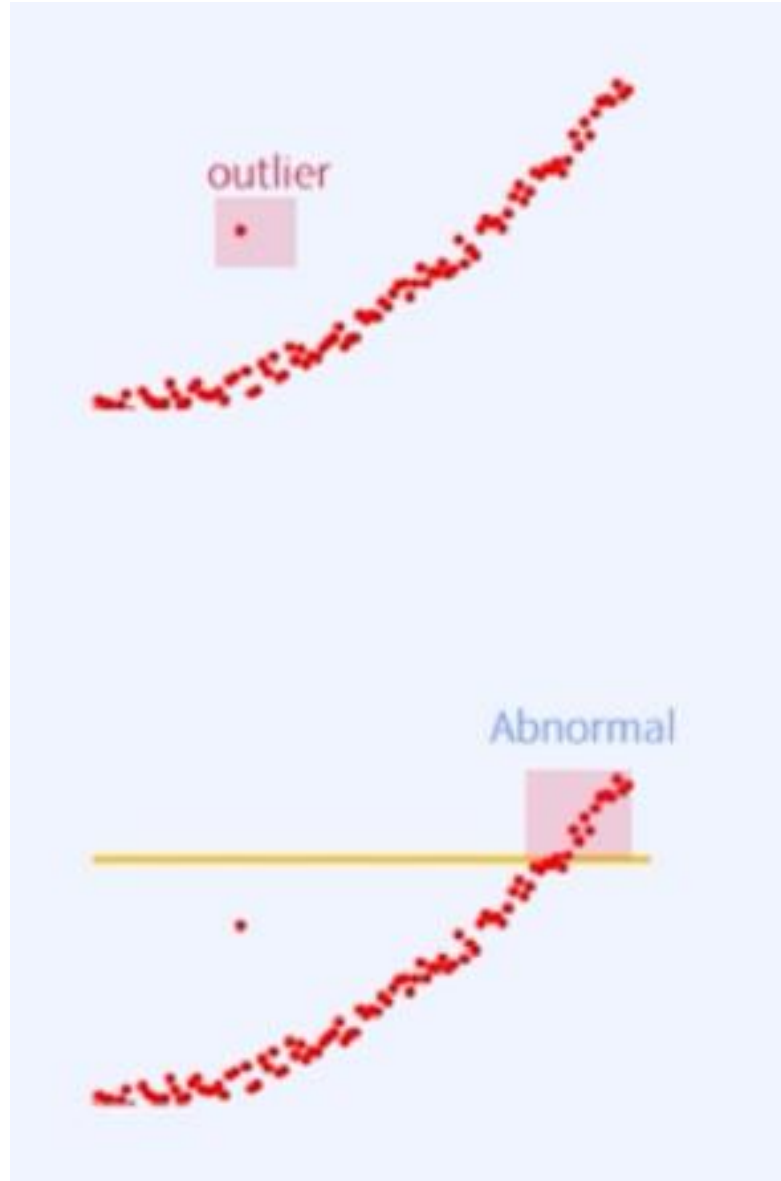
이상치(outlier) vs 이상(abnormal)

- 이상치(outlier)

- 관측된 데이터의 범위에서 많이 벗어난 아주 작은 값이나 큰 값
- 분석하고자 하는 데이터에서 적은 확률로 나타나는 데이터
- 분석 결과 해석 시 오해를 발생시킬 수 있기 때문에 사전 제거

- 이상(abnormal)

- 문제 발생 가능성이 높은 데이터
- 정상적인 범주의 데이터라도 이상으로 정의할 수 있음
- 일반적으로 자주 발생하지 않는 패턴이 이상일 확률이 높음


outlier
Abnormal

- 이상치(outlier)란 데이터 관점, 이상(abnormal)이란 현업의 문제해결 관점
- 이상 탐지(abnormal detection)은 이상이라고 정의한 사건 및 패턴을 탐지하는 활동
- 이상 탐지의 최종 목적

: 더 큰 risk가 발생하기 전에 피해를 최소화하기 위함

이상 데이터 발생 원인

1. 표본 추출 오류
2. 입력 오류
3. 실험 오류
4. 측정 오류
5. 데이터 처리 오류
6. 자연 오류

Data Type

1. Time series(sequential) vs static(정적인, point)
2. Unvariable(단변량) vs multivariable(다변량)
3. Binary/categorial/continuous/hybrid
4. Relational(상관관계가 있는) vs independent(독립적인)
5. Well-known or not(기존의 룰의 적용 가능한/알려져 있지 않은)

Data type에 따른 이상 탐지의 종류

1. Point anomaly detection - 정적인 점 분포
2. Contextual anomaly detection - sequential
3. Collective anomaly detection
4. Online anomaly detection – 실시간 데이터 수집 체계
5. Distributed anomaly detection

Label 유무에 따른 이상 탐지 방법론

현재 보유하고 있는 데이터의 성격에 따라 결정

1. Supervised anomaly detection (지도 이상 징후 탐지)
2. Semi-supervised anomaly detection (준지도 이상 징후 탐지)
3. Unsupervised anomaly detection (비지도 이상 징후 탐지)

Train data에 따른 이상 탐지의 종류

새로운 관측치가 기존 분포에 속하는지, 기존 분포를 벗어났는지 구분

학습할 데이터를 어떻게 정의하는지에 따라 문제의 성격과 해결 방법론이 달라짐

1. Outlier detection
2. Novelty detection