기초 이론부터 실무 실습까지

머신 러닝 익히기

programmers

Part 06. 비지도학습 알아보기

정 정 민

programmers

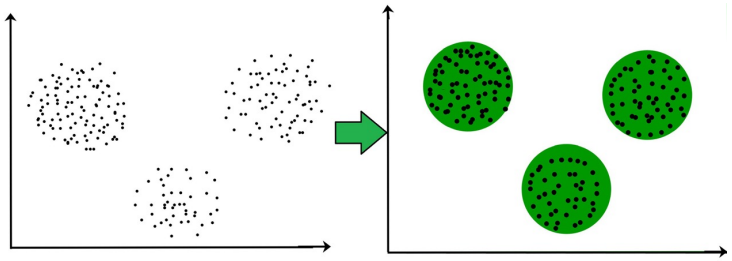
Chapter 15. 비지도학습의 개념

1. 비지도학습
2. 대표 알고리즘

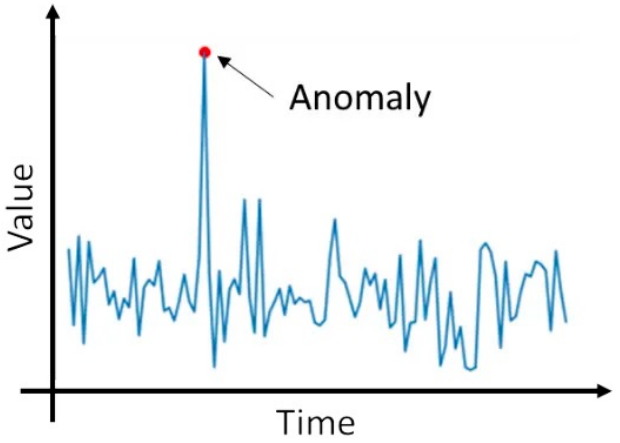
비지도학습

비지도 학습, Un-Supervised Learning

- 정답 레이블이 지정되지 않은 데이터로부터 패턴을 찾아내는 학습 방법론
- 이 방법으로 학습되는 알고리즘은 입력으로 주어진 데이터 내부에서 데이터의 구조나 패턴을 자동으로 탐색하는 목적을 갖고 있음
- 고객 세분화, 이상 탐지, 대규모 데이터셋의 구조 파악과 같은 다양한 활용이
- 특징 및 장점
  - 수동으로 데이터의 정답을 생성할 필요가 없어 비용과 시간이 절약
  - 데이터 내부의 구조를 탐색 → 다양한 통찰
  - 다양한 데이터 유형과 복잡한 구조에도 적용 가능
- 단점
  - 결과를 해석하기 어려운 상황이 있을 수 있음
  - 명확한 정답이 없으므로 모델의 성능 객관화와 평가가 어려움
  - 노이즈에 매우 민감함



클러스터링, Clustering



이상 탐지, Anomaly Detection

https://www.geeksforgeeks.org/clustering-in-machine-learning/
https://www.analyticsvidhya.com/blog/2021/06/univariate-anomaly-detection-a-walkthrough-in-python/

대표 문제

- 비지도 학습에서 다루는 대표 문제
  - 군집화 (Clustering)
    - 데이터를 유사한 특성을 가진 하위 그룹(Sub-group) 또는 클러스터(Cluster)로 분할
  - 차원 축소 (Dimensionality Reduction)
    - 고차원 데이터의 특성을 줄여 더 낮은 차원의 표현으로 만드는 과정
  - 이상 탐지 (Anomaly Detection)
    - 데이터에서 비정상적인 패턴, 이상치, 또는 예외적인 사례를 탐지

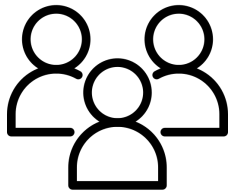
군집화, Clustering

- 데이터를 유사한 특성을 공유하는 하위 그룹(Sub-group) 또는 클러스터 (Cluster)로 분할하는 과정
  - 그래서 이름이 Clustering 이라고 합니다!

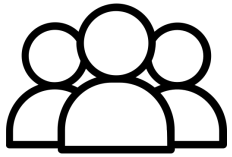
- 레이블이 없는 데이터에서 내재된 구조를 발견
- 유사한 데이터 포인트를 그룹화하여 패턴을 이해하는 데 사용

- 군집화 한 데이터의 의미는 사람이 부여해야 하는 경우가 많음

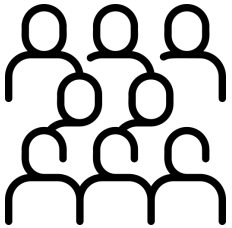
- 예를 들어,
  - 고객들의 구매 데이터를 활용해 유사 고객 그룹을 만들어보기!
  - 글을 유형 별로 나눠 주제별로 나누기
  - 유사한 특성을 갖은 이미지를 그룹핑 하기



충성 고객



잠재 고객



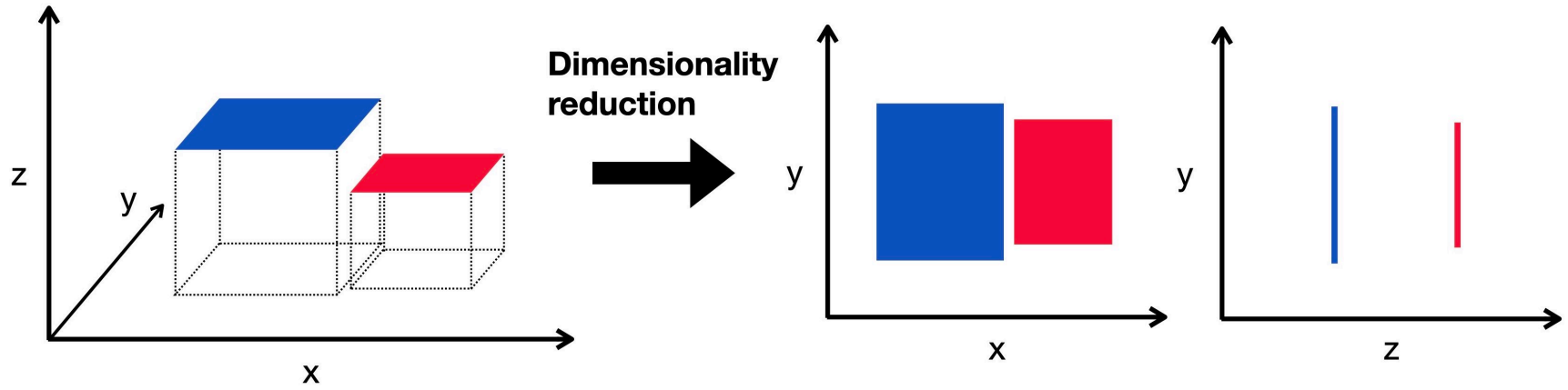
일반 고객

차원 축소, Dimensionality Reduction

- 고차원 데이터를 보다 낮은 차원으로 표현하여 데이터의 핵심적인 특성을 유지하는 기법

- 데이터 시각화, 노이즈 감소, 계산 효율성 증가, 그리고 더 나은 데이터 해석 등의 목적
- 차원 축소를 통해 복잡한 데이터 구조를 간소화하고, 중요한 정보를 강조하는 데 사용

- 예를 들어,
  - 분석할 데이터의 차원을 낮춰 2차원 혹은 3차원의 그래프로 확인하거나
  - 데이터의 주된 특성을 유지하면서 세밀하게 표현된 노이즈를 제거



https://www.sc-best-practices.org/preprocessing_visualization/dimensionality_reduction.html

이상 탐지, Anomaly Detection

- 데이터에서 비정상적인 패턴, 이상치, 또는 예외적인 사례를 탐지하는 과정

- 데이터에서 일반적으로 볼 수 있는 특성에서 많이 벗어난 데이터를 식별하는 과정에서 사용
- 보안, 금융, 의료 등의 분야에서 중요한 역할

- 예를 들어,
  - 나의 계좌가 갑자기 외국 어딘가에서 로그인 하려는 시도가 포착되었거나
  - 충격파 그래프를 이용해 물체 혹은 건물 내부의 균열을 찾아낸다거나

대표 알고리즘

군집화, Clustering

- **K-평균 (K-means)**
  - 데이터를 K개의 클러스터로 그룹화
  - 각 클러스터의 중심을 계산하고, 각 데이터 포인트를 가장 가까운 클러스터 중심에 할당하는 방식으로 작동
  - 반복적인 과정을 통해 클러스터 중심을 업데이트하며 최적화

- **계층적 군집화 (Hierarchical Clustering)**
  - 데이터 포인트를 개별 클러스터로 가정하여 시작
  - 점차 유사한 클러스터를 병합하거나 큰 클러스터를 세분화하는 방식으로 진행

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**
  - 데이터가 모여있는 밀도를 기반으로 클러스터를 형성
  - 고밀도 지역과 저밀도 지역을 이용해 군집화를 진행

차원 축소, Dimensionality Reduction

- **주성분 분석 (PCA)**
  - 데이터의 분산을 최대한 보존하는 방향의 축을 찾고
  - 해당 축을 기준으로 고차원 데이터를 저차원으로 변환
  - 주로 데이터의 주요 특성을 추출하고 시각화 하는 과정에서 사용

- **t-SNE**
  - 고차원 데이터의 구조를 보존하면서 저차원으로 매핑하는 기법
  - 시각화 과정에서 매우 유용하게 사용됨

- **오토인코더 (Autoencoder)**
  - 신경망(딥러닝)을 이용한 차원 축소 기법으로
  - 입력 데이터를 저차원으로 압축 후, 다시 원래 차원으로 복원하는 방식으로 핵심 특징을 만들어냄

이상 탐지, Anomaly Detection

- **Isolation Forest**
  - Tree를 기반으로하며 특정 데이터 포인트를 격리시키는 데 필요한 분할 수를 기준으로 이상치를 탐지
  - 이상치 데이터의 경우 더 적은 분할로 격리되는 경향이 있음을 활용

- **One-Class SVM**
  - 정상 데이터만을 활용해 “정상” 이라는 class로 SVM을 학습하고
  - SVM이 정상 패턴에서 벗어나는 데이터를 보고 출력하는 결과를 보고 이상치를 판단

- **LOF (Local Outlier Factor)**
  - 주어진 데이터 주변의 데이터 밀도를 계산
  - 정상 데이터는 주변에 높은 데이터 밀도를 갖고 있고,
  - 이상치 데이터는 주변에 데이터가 적어 낮은 밀도를 갖고 있음을 활용

E.O.D