

3.1 이론

3.1.1 비지도학습이란?

비지도학습(Unsupervised Learning)은 데이터에 **정답(label)이 존재하지 않는 상황**에서 작동하는 학습 방법입니다. 즉, 데이터의 입력값만 주어지고, 모델은 **스스로 그 안의 패턴이나 구조를 찾아내는 방식**입니다.

지도학습과 비교

구분	지도학습 (Supervised Learning)	비지도학습 (Unsupervised Learning)
입력 데이터	입력값 + 정답(레이블) 존재	입력값만 존재
목적	정답 예측	패턴 발견, 구조 분석
예시	스팸 메일 분류, 가격 예측	고객 세분화, 데이터 시각화

3.1.2 비지도학습의 대표적인 목적

비지도학습은 다음과 같은 두 가지 대표 목적을 갖고 사용됩니다.

1. 군집화 (Clustering)

- 유사한 데이터를 서로 묶어 그룹화하는 작업
- 데이터에 숨겨진 잠재적인 클래스나 유형을 알아내는 데 효과적

활용 예시:

- 마케팅: 고객 세분화
- 생물학: 유전자 발현 패턴 그룹화
- 추천 시스템: 사용자 행동 기반 그룹 분류

2. 차원 축소 (Dimensionality Reduction)

- 고차원 데이터를 핵심 정보만 유지하면서 저차원으로 압축
- 분석 효율과 시각화 용이성 향상

활용 예시:

- 데이터 전처리: 노이즈 제거
- 시각화: 2D/3D로 투영하여 분포 파악
- 학습 효율: 모델 훈련 시간 단축

3.1.3 고차원 공간의 문제점과 차원 축소의 필요성

데이터의 **차원이 많아질수록** (즉, 특성의 개수가 많아질수록) 다음과 같은 문제가 발생합니다. 이를 "차원의 저주(Curse of Dimensionality)"라고 합니다.

차원의 저주(Curse of Dimensionality)

- 데이터가 희소(Sparse):** 고차원 공간에서는 데이터가 넓게 퍼져서 유사한 점을 찾기 어려움
- 거리 측정 불안정:** 유클리디안 거리 등이 신뢰성이 떨어짐
- 계산 비용 증가:** 차원이 높을수록 학습 시간이 급격히 늘어남
- 시각화 불가능:** 3차원 이상부터는 시각적으로 이해하기 매우 어려움

→ 따라서 차원 축소 기법을 통해 의미 있는 정보만 유지하면서 차원을 줄이는 것이 중요합니다.

3.1.4 군집화와 차원 축소의 관계

- 차원 축소 후 군집화:** 차원을 줄인 뒤 군집화를 하면 성능 향상
- 군집화 후 차원 축소:** 군집별로 어떤 특징이 있었는지 시각화 가능

예를 들어, 100차원의 고객 데이터를 먼저 PCA나 t-SNE로 2차원으로 줄이고, 그 위에 K-means 군집을 시각화하면 다음과 같은 이점이 있습니다:

- 사람이 직접 **클러스터 해석** 가능
- **고객 유형별 특성**을 쉽게 파악할 수 있음
- 시각적으로 **군집의 경계와 분포**를 파악 가능

3.1.5 주요 비지도 학습 알고리즘 분류

알고리즘	분류	특징
K-means	군집화	중심 기반 클러스터링
DBSCAN	군집화	밀도 기반, 이상치 탐지
PCA	차원 축소	선형 투영 기반
t-SNE	차원 축소	비선형 시각화 특화