데이터 처리와 관련 연구 & 데이터의 주요 문제 유형

202304708 한서현

목차

데이터 처리와 관련 연구

전처리

> "기계가 데이터를 빠르게 해석할 수 있도록 하는 머신러닝의 첫 단계 "

50-80%

분석 과정 비중

전체 머신러닝 프로젝트에서 데이터 전처리가 차지하는 시간

데이터 전처리는 감독학습 기계학습 알고리즘에 일반화 성능에 가장 중요하고 영향력이 큼

데이터 처리와 관련 연구

데이터 변환 공식

$$B_{ij} = T(A_{ik})$$

데이터 처리의 목적

1. 데이터와 관련된 문제를 해결하지 않으면 불일치 발생, 데이터 분석 수행에도 방해
2. 데이터의 특징과 본질 이해
3. 주어진 데이터 세트에서 더 필요하고 의미 있는 정보 추출

데이터의 주요 문제 유형

과잉 데이터

빅데이터 환경에서 데이터 규모, 속도 증가로 처리 비용과 시간 급증

해결책: 데이터 차원 줄이기

부족한 데이터

누락된 데이터가 많으면(20% 이상) 신뢰도 저하와 모델 오류 유발

권장: 20% 이상 누락 시 데이터 제거

분열된 데이터

다양한 출처의 데이터는 형식과 품질 차이로 인한 불일치 문제 발생

필요: 통합·표준화 프로세스