데이터 전처리 중요성

2025. 05. 12(월)

학습 목표

01 데이터 전처리의 개념과 중요성을 이해합니다.

02 현실 데이터의 주요 문제점과 데이터 정제의 필요성을 파악합니다.

03 데이터 정제의 기본적인 작업들을 파악합니다.

목차

데이터 전처리란?

데이터 전처리란?

- 정의: 원시 데이터를 분석 및 모델링에 적합한 형태로 변환하는 과정
- 목적: 불완전하거나 부적절한 데이터를 모델 학습에 사용할 수 있도록 가공하여 성능 향상
- 머신러닝/데이터 분석 파이프라인의 가장 중요하고 시간이 많이 소요되는 단계 중 하나입니다.

왜 중요할까요?

- 모델 성능 향상: 불량 데이터는 모델 학습을 방해하고 잘못된 패턴을 학습하게 하여 예측 정확도 및 일반화 성능을 저하시킵니다.
- 데이터 품질 개선: 정확하고 일관성 있는 데이터는 분석 결과의 신뢰성을 높입니다.
- 알고리즘 효율성 및 요구사항 충족:
  - 거리 기반 모델: Feature 스케일 조정 필요 (K-Means, KNN)
  - 경사하강법 기반 모델: 스케일링 시 수렴 속도 향상 (신경망, 회귀)
  - 선형 모델: 데이터 분포(정규성), 이상치에 민감
  - 트리 기반 모델: 비교적 덜 민감하나, 인코딩 등 다른 전처리 필요
- 과적합 방지 및 계산 효율성 증대: 불필요한 Feature 제거, 차원 축소 등

현실 데이터의 문제점

실제 수집된 데이터는 이상적이지 않으며, 다음과 같은 문제점을 흔하게 포함하고 있습니다.

- 불완전성 (Incompleteness):
  - 결측치 (Missing Values): 일부 데이터 값이 누락되어 있음
- 불일치성 (Inconsistency):
  - 동일한 의미를 다른 형식으로 표현 (예: "남", "남자" / "서울", "Seoul")
  - 데이터 간 논리적인 모순 (예: 출생일보다 사망일이 빠른 경우)

- 노이즈 (Noise):
  - 이상치 (Outliers): 다른 대부분의 데이터와 동떨어진 극단적인 값
  - 측정 오류, 데이터 입력 오류 등
- 비표준화된 형식:
  - Feature 값 범위의 큰 차이 (예: 나이(0~100) vs 소득(0~1억))
  - 모델 입력에 적합하지 않은 데이터 타입 (예: 범주형 문자열, 텍스트)
- 중복 데이터 (Duplicates):
  - 동일한 내용의 데이터 레코드가 여러 개 존재

이러한 문제점들을 해결하지 않고 모델 학습에 사용하면 심각한 성능 저하를 초래합니다.

데이터 정제 개요

데이터 정제 (Data Cleaning)

- 정의: 데이터의 불완전성, 불일치성, 노이즈, 중복 등 데이터 품질 문제를 식별하고 수정하는 과정
- 전처리 과정의 가장 첫 단계로 간주되는 경우가 많습니다.
- 왜 필요한가?: 모델 학습 전 데이터의 신뢰성과 정확성을 확보하여 분석 결과의 왜곡을 방지합니다.

주요 데이터 정제 작업 (1)

1. 결측치 처리 (Handling Missing Values):
   - 누락된 데이터 값을 어떻게 처리할 것인가?
   - 방법:
     - 해당 데이터(행 또는 열) 제거
     - 평균(Mean), 중앙값(Median), 최빈값(Mode) 등 통계량으로 대체 (Imputation)
     - 다른 Feature나 모델을 사용하여 예측값으로 대체
     - 상수 값으로 대체
   - 어떤 방법을 사용할지는 데이터 특성, 결측 패턴, 결측 비율 등을 고려하여 결정

주요 데이터 정제 작업 (2)

2. 이상치 처리 (Handling Outliers):
   - 다른 값들과 극단적으로 다른 값을 어떻게 처리할 것인가?
   - 문제점: 모델 학습에 큰 영향을 미쳐 결과를 왜곡할 수 있습니다 (특히 평균 기반 통계, 선형 모델 등).
   - 탐지 방법: 시각화 (Box Plot, Scatter Plot), 통계적 방법 (Z-score, IQR), 모델 기반 방법 등
   - 처리 방법:
     - 이상치 데이터 제거
     - 다른 값으로 대체 (예: IQR 범위 경계 값으로 대체)
     - 변환 (Log Transform 등)을 통해 이상치의 영향 완화
     - 모델이 이상치에 강건한 알고리즘 사용 (예: 트리 모델, RobustScaler 사용)
   - 이상치가 단순 오류인지, 아니면 의미 있는 데이터인지 신중하게 판단해야 합니다.

주요 데이터 정제 작업 (3)

3. 중복 데이터 처리 (Handling Duplicates):
   - 동일한 관찰치(행)가 여러 번 기록된 경우
   - 문제점: 분석 결과 왜곡 (예: 빈도수 과대 계상), 모델 학습 방해
   - 처리 방법: 중복된 데이터 제거
   - 어떤 열을 기준으로 중복을 판단할지 정의하는 것이 중요합니다.

주요 전처리 기법 개요

주요 전처리 기법 개요

앞서 살펴본 데이터 정제 외에, 모델 학습 효율과 성능 향상을 위해 다음과 같은 전처리 기법들이 중요하게 사용됩니다.

- 특성 스케일링 (Feature Scaling):
  - 서로 다른 범위의 특성 값들을 유사한 Scale로 조정 (표준화, Min-Max 등)
  - 거리 기반 및 경사 하강법 기반 모델에 필수적
- 데이터 변환 (Data Transformation):
  - 데이터의 분포 또는 형태를 변경하여 모델 가정을 충족시키거나 이상치 영향 완화 (로그 변환, Box-Cox 등)
- 범주형 데이터 인코딩 (Categorical Data Encoding):
  - 성별, 지역 등 범주형 데이터를 모델이 이해할 수 있는 숫자형으로 변환 (One-Hot, Label Encoding 등)
- 특성 선택/추출 (Feature Selection/Extraction):
  - 모델 성능에 기여하는 중요한 특성만 선택하거나 새로운 특성을 만듦 (예: PCA)

요약 정리

- 데이터 전처리는 원시 데이터를 모델 학습에 적합하게 가공하는 필수 과정이며, 모델 성능에 큰 영향을 미칩니다.
- 현실 데이터는 결측치, 이상치, 중복 등 다양한 문제를 포함하므로 이를 해결하는 **데이터 정제**가 중요합니다.
- 데이터 정제의 주요 작업에는 결측치 처리, 이상치 처리, 중복 데이터 처리가 있으며, 데이터 특성에 맞는 방법을 선택해야 합니다.
- 데이터 전처리에는 데이터 정제 외에도 스케일링, 변환, 인코딩 등 다양한 기법이 있으며, 앞으로 이 기법들을 상세히 학습할 것입니다.