

생성형AI

Day 2

데이터의 활용 및 구현 II



목차

1. 오늘의 목표 및 공지!
2. 데이터 전처리 개요
3. 데이터 수집 후 전처리 과정
4. 데이터 타입별 전처리 방법
5. 데이터 정제
6. 데이터 변환
7. 데이터 통합
8. Advanced 기법
9. 데이터 전처리 최적화
10. 실습과제



오늘의 목표 및 공지

오늘의 강의 목표

- 데이터 전처리의 개념 이해
- 데이터 전처리의 주요 단계 이해

공지

- 11시부터 2시까지 사전 기술 테스트
- 2시부터 1일차 공통 미션 피드백 – 미션 기획 문서 / 교육 전반적인 내용에 대한 질의응답

데이터 전처리 개요

데이터 전처리

- 원시 데이터를 분석 및 모델링에 적합하게 변환하는 과정.
- 데이터의 정제, 변환, 통합 등을 포함.
- 데이터 품질을 높여 신뢰성 있는 분석을 가능하게 함.

데이터 전처리의 중요성 및 필요성

- 데이터 품질 향상
- 모델 성능 최적화
- 분석의 신뢰성 향상
- 데이터 활용의 효율성 향상

데이터 전처리의 주요 단계

- 데이터 정제
 - 결측값 처리
 - 이상값 탐지 및 수정
 - 중복 데이터 제거
- 데이터 변환
 - 데이터 스케일링
 - 데이터 인코딩
- 데이터 통합 및 변형
 - 데이터 병합
 - 데이터 집계 및 변형

데이터 수집 후 전처리 과정

데이터 수집에서 전처리로의 연결

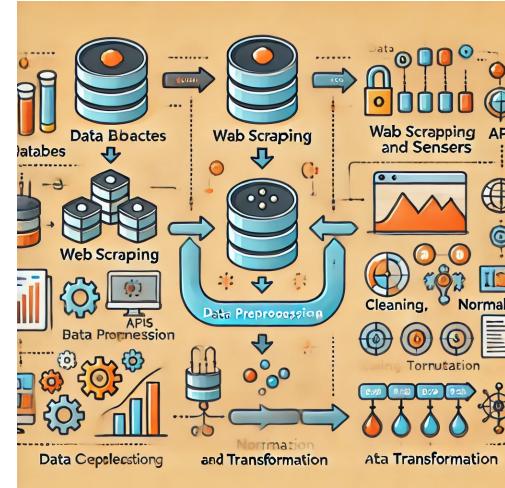
- 데이터수집
- 다양한 소스에서 데이터 수집(OpenAPI, 웹크롤링 등)
- 수집된 데이터는 대부분 정제가 필요

왜?

- 데이터는 활용을 목적으로 수집계획을 짠 경우는 거의 없기 때문
- 데이터 수집이 먼저이고, 수집된 데이터를 어떻게 하면 잘 활용할 수 있을까?는 나중에 고려

수집된 데이터의 초기 확인 및 분석

- 데이터 불러오기
- 기본정보확인(데이터 구조, 크기, 타입)
- 결측치 및 이상치 처리



데이터 타입별 전처리 과정

숫자형 데이터 전처리

- 결측값 및 이상값 처리
 - 평균값으로 대체
 - 중앙값으로 대체
 - 이웃값으로 대체
 - 제거
- 스케일링 및 정규화

범주형 데이터 전처리

- 결측값 및 이상값 처리
 - 최빈값으로 대체
- 레이블 인코딩

날짜 및 시간 데이터 전처리

- 형식 변환 및 추출
- 텍스트 데이터 전처리
- 텍스트 정제 및 벡터화

데이터 정제 - 결측치

결측치의 종류

MCAR (Missing Completely at Random):

- 데이터가 완전히 무작위로 누락된 경우
- 예: 설문 조사에서 무작위로 질문을 건너뛴 경우

MAR (Missing at Random):

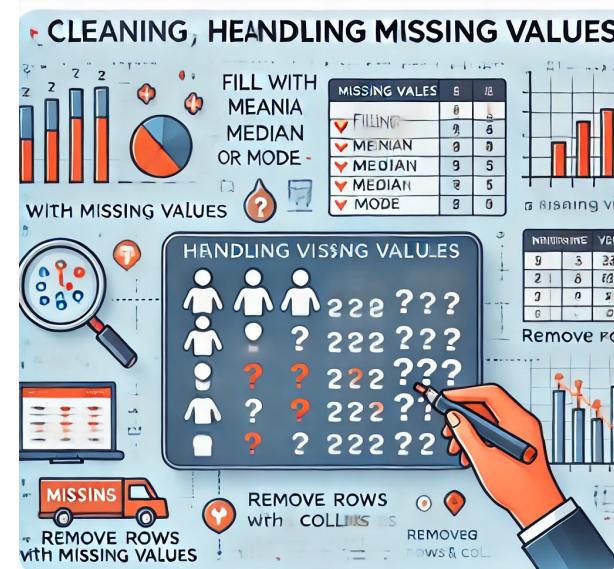
- 데이터 누락이 다른 관측 가능한 변수와 관련된 경우
- 예: 나이가 많은 응답자들이 소득 정보를 제공하지 않은 경우

MNAR (Missing Not at Random):

- 데이터 누락이 자체 변수와 관련된 경우
- 예: 높은 소득을 가진 사람들이 소득 정보를 제공하지 않은 경우

결측값의 원인

- 데이터 입력 오류: 수동 입력 과정에서 발생하는 실수
- 응답자의 미응답: 설문 조사나 인터뷰에서 특정 질문에 대한 응답 누락
- 데이터 수집 과정의 문제: 센서 오작동, 통신 오류 등



데이터 정제 – 이상치(Outlier)

이상치(Outlier)란?

- 데이터 분포에서 벗어난 극단적인 값
- 통계적 분석과 모델 성능에 영향을 줄 수 있음

이상치의 원인

- 데이터 입력 오류: 수동 및 자동화된 시스템에서 발생할 수 있는 오류
- 데이터 수집 문제: 잘못된 수집 방법이나 전송 과정에서의 오류
- 이질적인 데이터: 서로 다른 특성을 가진 그룹의 혼합, 드문 이벤트
- 자연적 변동성: 데이터의 자연스러운 변동
- 특이한 상황 또는 조건: 특정 상황에서만 발생하는 예외적인 값



데이터 변환

스케일링

- 데이터의 범위를 임의로 조정해주는 과정

표준화(Standardization)

- 데이터의 평균을 0, 표준편차를 1로 변환하여 데이터의 분포를 표준정규 분포로 만드는 과정.

정규화(Normalization)

- 데이터를 특정범위로 변환하는 과정(주로 0~1), regularization 이랑 다름

$$z = \frac{(x - \mu)}{\sigma}$$

x: 원본 데이터 값, μ : 평균, σ : 표준편차

$$x' = \frac{(x - x_{\min})}{(x_{\max} - x_{\min})}$$

x: 원본 데이터 값, x_{\min} : 최소값, x_{\max} : 최대값

데이터변환의 이유

- 다양한 특성의 단위 차이를 없애고, 동일한 스케일로 조정하여 모델 학습에 도움
- 데이터를 일정한 범위로 조정하여 계산의 안정성과 속도 향상

데이터 변환 – 데이터 인코딩

레이블 인코딩

- 범주형 데이터를 숫자로 변환하는 기법
- 각 범주를 고유한 정수로 맵핑.
- 수치형 데이터로 변환하여 모델 학습에 사용
- 범주 간 순서가 없는 경우 잘못된 관계를 학습할 수 있음

A diagram illustrating the concept of label encoding. On the left, there is a table with a header 'Colour' and three rows: 'red', 'green', and 'blue'. To the right of this table is a large black arrow pointing to the right. To the right of the arrow is another table with a header 'Colour' and three rows containing the integers '0', '1', and '2' respectively. This visualizes how categorical labels are mapped to a unique integer value.

Colour
red
green
blue

→

Colour
0
1
2

원-핫 인코딩 (One-Hot Encoding):

- 범주형 데이터를 이진 벡터로 변환하는 기법.
- 각 범주를 이진 벡터의 고유한 위치에 1로 표시.
- 범주 간 순서가 없는 데이터를 처리할 때 유용
- 모든 범주를 독립된 변수로 변환하여 모델 학습에 사용
- 많은 범주가 있는 경우, 고차원 데이터로 변환되어 메모리 사용량이 증가

A diagram illustrating the concept of one-hot encoding. On the left, there is a table with a header 'Colour' and three rows: 'red', 'green', and 'blue'. To the right of this table is a large black arrow pointing to the right. To the right of the arrow is another table with two columns: 'red' and 'green'. The first row has a '1' under 'red' and a '0' under 'green'. The second row has a '0' under 'red' and a '1' under 'green'. The third row has a '0' under both 'red' and 'green'. This visualizes how each category is represented as a binary vector where only one position is '1' and the others are '0'.

Colour
red
green
blue

→

red	green
1	0
0	1
0	0

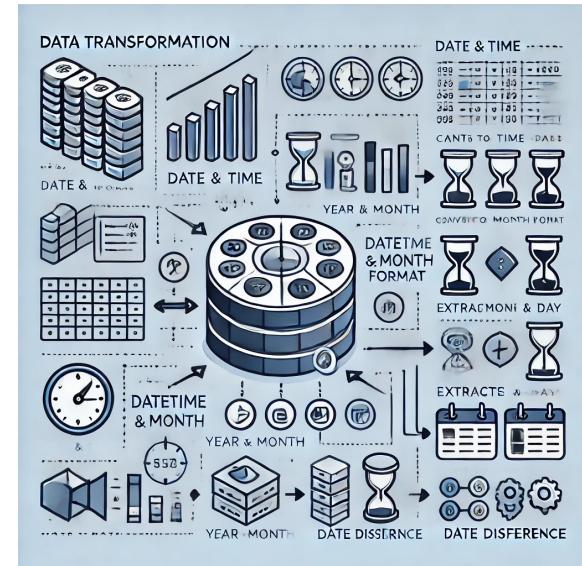
데이터 변환 – 날짜 및 시간 데이터

형식 변환

- 문자열 형식의 날짜 데이터를 datetime 형식으로 변환
- 날짜 연산 및 비교를 쉽게 수행할 수 있도록 하기 위해 진행

추출

- 날짜 객체에서 특정 정보(연도, 월, 일 등)를 추출
- 날짜 데이터에서 특정 정보를 추출하여 분석에 활용하기 위해 진행



데이터 통합 및 변형

데이터 합치기

- 여러 데이터를 하나로 결합하는 과정
- 공통된 키 또는 인덱스를 사용하여 데이터 간의 데이터를 결합
- 더 풍부한 인사이트를 도출하기 위해 사용
 - 고객 정보 데이터와 거래 내역 데이터를 결합하여 고객 행동 분석

데이터 집계

- 데이터의 특정 열을 기준으로 합계, 평균 등의 요약 통계를 계산하는 과정
- 그룹별로 데이터를 요약하여 더 높은 수준의 인사이트를 제공
- 데이터의 전반적인 경향을 파악하고, 주요 지표를 요약하여 분석의 효율성 향상
 - 월별 매출 데이터에서 각 월의 총 매출을 계산

데이터 그룹화

- 특정 열을 기준으로 데이터를 그룹화하여 집계하는 과정
- 그룹별로 데이터 집계를 수행하여 상세한 분석 가능
- 특정 그룹(예: 지역, 제품군 등)별로 데이터를 분석하여 세부적인 인사이트를 도출
 - 각 제품군별 평균 판매량을 계산



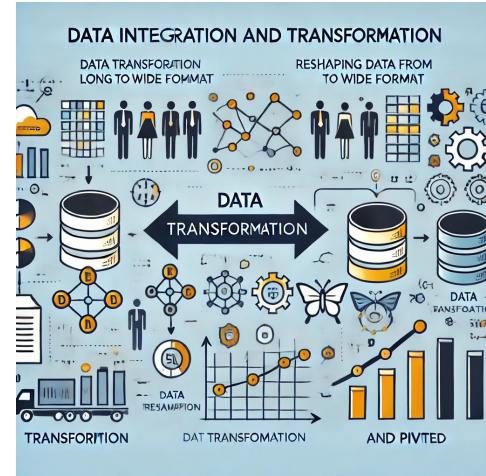
데이터 통합 및 변형

데이터 변형(재구조화):

- 데이터프레임의 형태를 변경하는 다양한 기법
- 데이터를 요약하거나 분석 목적에 맞게 재구성
- 분석의 편의성을 높이고, 특정 분석 목표에 맞는 데이터 구조를 만들기 위해
 - 긴 형식의 데이터를 넓은 형식으로 변환

데이터 피벗:

- 데이터프레임의 행과 열을 변환하여 데이터를 요약하는 과정
- 특정 열을 인덱스로 사용하고, 다른 열을 새로운 열로 변환
- 데이터의 특정 측면을 강조하고, 요약 통계를 쉽게 확인하기 위해
 - 월별 제품 판매 데이터를 피벗하여 각 제품의 월별 판매량을 열로 변환



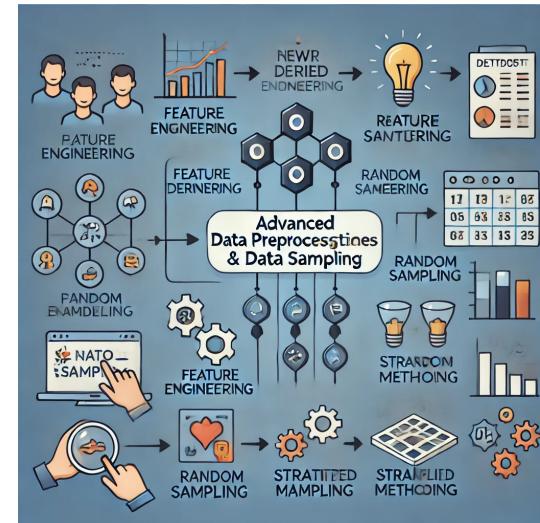
데이터 전처리 Advanced

파생 변수 생성

- 기존 데이터에서 새로운 변수를 생성하는 과정
- 기존 변수 간의 상호작용, 계산 또는 변환을 통해 새로운 정보를 추출
- 데이터의 정보량을 증가시켜 분석 및 모델링 성능을 향상.
 - 날짜 데이터에서 연도, 월, 일, 요일 등을 추출하여 분석의 깊이를 더함.
 - 텍스트 데이터에서 텍스트 길이, 단어 수 등을 파생변수로 생성
 - 키, 몸무게를 이용한 BMI데이터 ($\text{몸무게}/(\text{키}/100)^2$)

데이터 샘플링

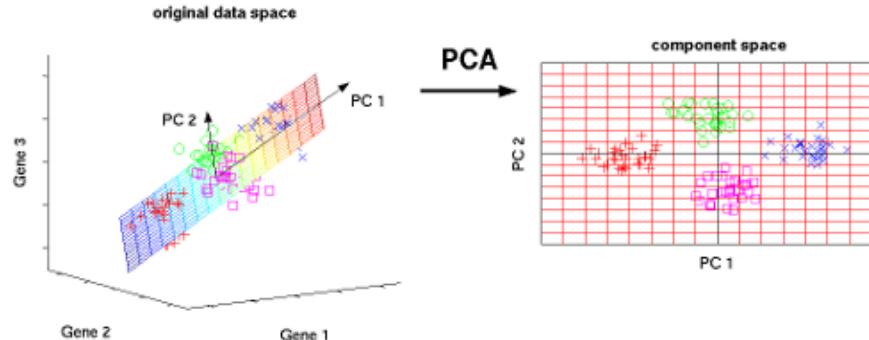
- 전체 데이터셋에서 일부 데이터를 선택하여 분석이나 모델링에 사용하는 과정
- 모집단의 특성을 대표하는 샘플을 추출
- 분석의 신속성, 대용량 데이터 처리 시간과 비용을 절감
- 랜덤 샘플링, 층화 샘플링



데이터 전처리 Advanced

차원 축소 기법

- 고차원 데이터를 저차원으로 변환하여 데이터의 복잡성을 줄이는 과정
- 데이터의 변동성을 최대한 보존하면서 주요 특성을 추출
- 데이터의 시각화와 이해를 용이하게 하고, 모델의 과적합을 방지하며 계산 효율성을 향상
- PCA, t-SNE



<http://phdthesis-bioinformatics-maxplanckinstitute-molecularplantphys.matthias-scholz.de/>

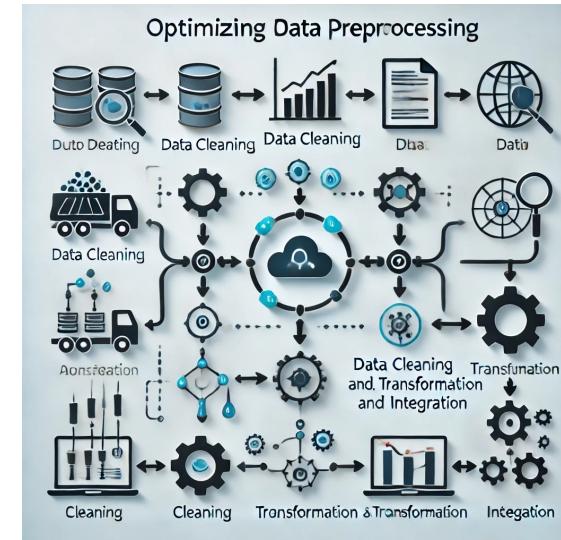
데이터 전처리 최적화

전처리 작업의 자동화

- 반복적이고 일관된 전처리 작업을 자동화하여 효율성을 높이는 과정
- 스크립트 및 파이프라인을 사용하여 전처리 작업을 자동으로 실행
- 수동 작업의 오류를 줄이고, 작업 시간을 단축하며, 일관성을 유지

파이프라인 구축

- 일련의 전처리 단계를 순차적으로 자동으로 실행하는 파이프라인을 구축
- 각 단계별로 전처리 작업을 정의하고, 이를 파이프라인으로 연결
- 전처리 작업의 일관성과 재현성을 보장
- 실제 코드상의 파이프라인, 의미/구조상의 파이프라인



데이터 전처리 최적화

대용량 데이터 처리

- 대용량 데이터를 효율적으로 처리하기 위한 전략과 기법
- 분산 처리, 병렬 처리, 메모리 관리 등의 기법을 사용하여 데이터 처리
- 데이터의 양이 클수록 처리 시간과 자원 소모가 증가하므로 효율적인 처리가 필요

분산 처리

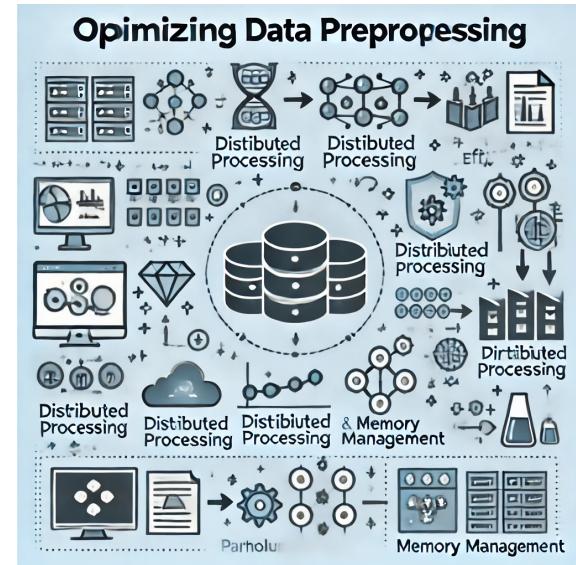
- 데이터를 여러 노드에 분산하여 동시에 처리
- Hadoop, Spark 등의 분산 처리 프레임워크를 사용

병렬 처리

- 데이터를 여러 프로세스에서 동시에 처리
- multiprocessing, Dask 등의 병렬 처리 라이브러리를 사용

메모리 관리

- 메모리 사용을 최적화하여 대용량 데이터를 효율적으로 처리
- 데이터 타입 최적화, 청크(chunk) 단위 처리



데이터 전처리 최적화

전처리 과정의 효율성 향상:

- 전처리 과정의 속도와 효율성을 높이기 위한 다양한 기법
- 최적화된 알고리즘, 코딩 스타일 개선
- 데이터 전처리 시간을 단축하고, 리소스 사용을 최적화

적절한 데이터 구조 사용:

- 작업에 적합한 데이터 구조를 사용하여 효율성을 높임
- pandas, NumPy, Dask 등의 라이브러리를 적절히 활용
- 데이터 구조의 특성을 활용하여 처리 속도 향상



실습 과제

- 이론 강의 후 제시하는 실습과제
- 강제성 없음!
- 이론 강의 복습 및 개인 실력 향상/유지
- github 정리 권장 (공유, 피드백 요청)
- 3가지 코스 공통으로 진행하는 실습과제
- 공통으로 진행하는 과제이기 때문에 꼭 참여 부탁
- 디스코드 공지 확인!

실습 과제

kaggle의 데이터를 선정해 데이터 전처리 파이프라인 구축해보기

- 전처리가 쉬운 데이터를 선택하기 보다 본인이 흥미있는 분야 데이터를 선정 권장
- 전처리 시나리오
- 전처리 코드
- 전처리 전후 데이터
- 전처리방법 적용 이유
- 기타

A large, bold, blue "kaggle" logo centered on the page.

1일차 미션 피드백

1일차 미션 피드백을 진행합니다.

- 피드백 주제: 미션 기획 문서 / 교육 전반적인 내용에 대한 질의응답
 - 장소: ZEP
 - 시간: 오후 2시 ~ 오후 6시
-
- ◆ 참여가 필수는 아니며, 피드백을 진행하는 팀(생성형AI 멤버)은 2시 이후 디스코드 공지 글의 스레드로 요청
 - ◆ 모든 과정에서 진행하는 세션이며, 각 과정의 내용 위주로 진행
 - ◆ 다른 실습 피드백은 요청 주시면 진행(github으로 코드까지 주세요!)

사전 기술 테스트 진행

오늘 실습 시간에 공지드린 사전 기술 테스트 진행합니다.

- 오전 11시 ~ 오후 12시 / 오후 1시 ~ 오후 2시 (점심시간 오후 12시 ~ 오후 1시)
- 앞으로 진행할 강의 수준을 조절하기 위한 목적
- 카카오테크 부트캠프 강의실에서 온라인으로 진행
 - ▶ <https://kakao-tech-bootcamp.goorm.io/>
 - ▶ 카카오테크 부트캠프 – 생성형 인공지능(AI) 과정 1기 → 시험/과제
- 데이터와 AI, 파이썬에 관한 지식과 개념, 약간의 코드 문제 출제
- 여러분의 평가에 활용되지 않습니다.

실습 진행