Data_preprocessing

원문: microsoft azure document

https://docs.microsoft.com/ko-kr/azure/architecture/data-science-process/prepare-data

데이터 전처리 및 정리가 필요한 이유

데이터 전처리 및 정리는 모델 학습에 데이터 세트를 사용할 수 있기 전에 수행해야 하는 중요한 작업입니다. 원시 데이터는 노이즈가 많고, 불안정하고, 값이 누락된 경우가 종종 있습니다. 이러한 데이터를 모델링에 사용하면 결과가 잘못될 수 있습니다.

실제 데이터는 다양한 소스 및 프로세스에서 수집되며 데이터 세트의 품질을 떨어트리는 이상값 또는 손상된 값이 포함될 수 있습니다. 다음과 같은 일반적인 데이터 품질 문제가 자주 발생합니다.

- 불완전 : 데이터에 특성이 없거나 값이 누락되었습니다.
- 노이즈가 많은 : 데이터에 잘못된 레코드 또는 이상값이 있습니다.
- 불일치 : 데이터에 충돌하는 레코드 또는 일치하지 않는 값이 있습니다.

가장 일반적으로 사용되는 데이터 상태 검사 방법으로는 어떤 것이 있습니까?

다음을 검사하여 데이터의 전체적인 품질을 확인할 수 있습니다.

- 레코드 수
- 특성 수
- 특성 데이터 유형
- 누락된 값의 수
- Well-Formed data
    - 데이터가 TSV 또는 CSV로 되어 있으면 열 구분 기호 및 줄 구분 기호가 열과 줄을 항상 올바르게 구분하는지 확인합니다.

데이터 전처리의 주요 작업

- 데이터 정리: 누락된 값을 입력하고, 노이즈 데이터 및 이상값을 검색하고 제거합니다.
- 데이터 변환: 차원 및 노이즈를 줄이기 위해 데이터를 정규화합니다.
- 데이터 감소: 데이터 처리를 용이하게 하는 샘플 데이터 레코드 또는 특성입니다.
- 데이터 불연속화: 특정 기계 학습 방법에서 쉽게 사용할 수 있도록 연속 특성을 범주 특성으로 변환합니다.
- 텍스트 정리: 탭으로 구분된 데이터 파일에 포함된 탭, 레코드 줄 바꿈 문제를 일으킬 수 있는 포함된 새 줄 등 데이터 정렬 문제를 일으킬 수 있는 포함된 문자를 제거합니다.

누락된 값을 처리하는 방법

- 삭제
- 더미 대체: 예를 들어 범주 값은 *알 수 없음*, 숫자 값은 0으로 대체합니다.
- 평균 대체
- 빈도 대체
- 회귀 대체: 회귀 메서드를 사용하여 누락된 값을 회귀된 값으로 대체합니다.

데이터를 정규화하는 방법

데이터 정규화는 숫자 값을 지정된 범위로 다시 스케일링합니다.

- 최소-최대 정규화: 0과 1 사이에서 데이터를 선형적으로 범위로 변환합니다. 여기서 최소값은 0, 최대값은 1로 조정됩니다.
- Z 점수 정규화: 평균 및 표준 편차를 기반으로 데이터 조정: 데이터와 평균의 차이를 표준 편차로 나눕니다.
- 소수점 배열: 특성 값의 소수점을 이동하여 데이터 크기를 조정합니다.

데이터를 분할하는 방법

- 동일 너비 범주화
- 동일 높이 범주화

데이터를 줄이는 방법

데이터 크기 및 도메인에 따라 다음 방법을 적용할 수 있습니다.

- 레코드 샘플링: 데이터 레코드를 샘플링하고 데이터에서 대표적인 하위 집합만 선택합니다.
- 특성 샘플링: 데이터에서 가장 중요한 특성의 하위 집합만 선택합니다.
- 집계: 데이터를 여러 그룹으로 나누고 각 그룹에 대한 숫자를 저장 합니다.
