데이터 전처리 = 데이터의 품질을 올리는 과정

전처리나 최적화 with scikit learn.

데이터의 품질은 데이터 분석의 90%를 좌우한다.

데이터 전처리 과정

- 데이터 실수화 : 컴퓨터가 이해할 수 있는 값으로의 변환
- 불완전한 데이터 제거
- 잡음 섞인 데이터 제거

  ex) 가격 데이터에 있는 -값 제거

  연령 데이터 중 과도하게 큰 값 제거
- 오류된 데이터 제거
- 불균형 데이터 해결

  과소표집 (undersampling), 과대표집 (oversampling)

주요 기법

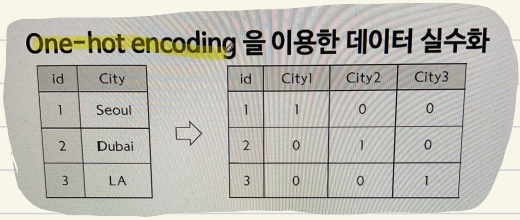
① 데이터 실수화 (Data Vectorization)
② 데이터 정제 (Data Cleaning)
③ 데이터 통합 (Data Integration)
④ 데이터 축소 (Data Reduction)
⑤ 데이터 변환 (Data Transformation)
⑥ 데이터 균형 (Data Balancing)

데이터 실수화 (Data Vectorization)

= 범주형 자료, 텍스트 자료, 이미지 자료 등을 실수로 구성된 형태로 전환하는 것.

자료의 유형

- 연속형 자료 Continuous data
- 범주형 자료 Categorical data
- 텍스트 자료 Text data



희소행렬 (Sparse Matrix) : 행렬의 값이 대부분 0인 행렬

- COO 표현식과 CSR 표현식을 통해 문제 해결

텍스트 자료의 실수화

- 단어의 출현 횟수를 이용한 데이터 실수화

  출현 횟수가 정보의 양과 비례하는 것은 아님. 때문에 TF-IDF 기법을 이용해야 함.

  자주 등장하며 분석에 의미를 갖지 못하는 단어의 중요도를 낮추는 개념

데이터 변환 (Data Transformation)

- 표준화 $x_{std} = \frac{x - mean(x)}{sd(x)}$
- 정규화 $x_{nor} = \frac{x - min(x)}{max(x) - min(x)}$

(정규화가 표준화보다 유용. 단, 데이터특성이 bell-shape 이거나 이상치가 있을 경우 표준화가 유용)

데이터 정제

- 결측 데이터 채우기 (Empty Values)
  - 결측 데이터 : np.nan, npNaN, None
  - 평균 (mean), 중위수 (median), 최빈수 (most frequent value)로 대체하는 기법

데이터 통합 (Data Integration)

여러 개의 데이터를 하나로 합치는 과정

데이터 불균형 (Data Imbalance)

= ML의 목적이 분류일때, 특정 클래스의 관측치가 다른 클래스에 비해 매우 낮게 나타나면 이러한 자료를 불균형 자료라 함.

해결 기법

- undersampling, oversampling
  일반적으로 oversampling이 통계적으로 유용
- decision tree와 ensemble은 상대적으로 불균형 자료에 강인한 특성을 보임