데이터전처리(Preprocessiong)

: 학습 전에 학습에 용이하게 데이터를 처리하는 과정(데이터 품질 상승)

1. 데이터 실수화(Data Vectorization)

자료 → float

자료의 유형

연속형 자료(Continuous data)

범주형 자료(Categorical data)

텍스트 자료(Text data)

2차원 자료(행렬, 2차원 텐서) = [n_sample, n_features]

범주형 자료의 실수화: One-hot encoding

id	City
1	Seoul
2	Dubai
3	LA

⇒

id	City1	City2	City3
1	1	0	0
2	0	1	0
3	0	0	1

sklearn.feature_extraction.DictVectorizer

Input argument: 디폴트 옵션 Sparse = True

희소행렬(Sparse Matrix): 행렬의 값이 대부분이 0인 경우

프로그램 시 불필요한 0 값으로 메모리 낭비가 심함

행렬 크기가 커서 연산 시간이 많이 소요

해결 방법: COO 표현식, CSR 표현식

텍스트 자료의 실수화: 단어의 출현 횟수

id	
1	떴다 떴다 비행기 날아라 날아라
2	높이 높이 날아라 우리 비행기
3	내가 만든 비행기 날아라 날아라
4	멀리 멀리 날아라 우리 비행기

⇒

id	날아라	내가	높이	떴다	만든	멀리	비행기	우리
1	2	0	0	2	0	0	1	0
2	1	0	2	0	0	0	1	1
3	2	1	0	0	1	0	1	0
4	1	0	0	0	0	2	1	1

출현 횟수가 정보 양과 비례× → TF-IDF기법 이용

TF-IDF(Term Frequency Inverse Document Frequency)

: 자주 등장하고 분석에 의미 없는 단어 중요도 낮추기(관사, The, a)

가중치 재계산(높은 빈도는 낮은 가중치, 낮은 빈도에 높은 가중치)

sklearn.feature_extraction.text.CountVectorizer.fit_transform()

fit_transform(text).toarray(): CSR 표현의 압축을 풀기 위해 사용

2. 데이터 정제(Data Cleaning)

없으면 채우고: 불완전한 데이터 제거(NULL, NA, NAN)

잡음 지우고: 잡음 섞인 데이터 제거(과도하게 큰 나이, (-)값인 가격)

모순은 교정하기: 모순된 데이터 제거(남자 주민번호가 2)

결측 데이터 채우기(Empty Values)

결측 데이터: np, nan, npNaN, None

평균(mean), 중위수(median), 최빈수(most frequent value)로 대체

sklearn.impute.SimpleImputer(): 입력인자로 평균, 중위수, 최빈수 선택

pandas.DataFrame.dropna(): 결측 데이터 0으로 채우기

pandas.DataFrame.fillna(): 결측 데이터 0으로 채우기

3. 데이터 통합(Data Integration)

多 → 하나

pandas.DataFrame.merge(): 여러개의 데이터 파일을 하나로 합치기

pandas.DataFrame.dtypes: 변수의 자료 타입 확인

4. 데이터 축소(Data Reduction)

too big → 수↓(sampling), 차원↓

5. 데이터 변환(Data Transformation)

이유: 머신러닝은 데이터의 특성(feature)를 비교하여 패턴 찾는다.

특성 간 스케일 차이가 심하면 패턴 찾는데 문제 발생한다.

방법: 표준화, 정규화, 로그, 평균값, 구간화

표준화(Standardization): $x_{std} = \frac{x - mean(x)}{sd(x)}$

정규화(Normalization): $x_{nor} = \frac{x - \min(x)}{\max(x) - \min(x)}$

일반적으로 정규화가 표준화보다 유용

BUT 데이터 특성이 bell-shape이거나 이상치가 있으면 표준화가 유용

6. 데이터 균형(Data Balancing)

샘플링으로 클래스 비율 맞추기 = 불균형 데이터 해결

데이터 불균형(Data Imbalance)

불균형자료: 머신러닝의 목적이 분류일 때, 다른 클래스에 비해 관측지가 매우 낮게 나타난 클래스의 자료

해소 방법: 과소표집(Undersampling), 과대표집(Oversampling)

일반적으로 과대표집이 과소표집 보다 유용하다.

의사결정나무(Decision Tree)와 앙상블(Ensemble)은 상대적으로 불균형자료에 강인한 특성을 보인다.

과소표집: 다수클래스의 표본을 임의로 학습데이터에서 제거

NearMiss

imlearn.under_sampling.NearMiss()

과대표집: 소수클래스의 표본을 복제하고 학습데이터에 추가

SMOTE(Synthetic Minority Oversampling Technique)

imlearn.over_sampling.SMOTE()

ADASYN(Adaptive Synthetic Sampling Method)

imlearn.over_sampling.ADASYN()