데이터 전처리 (Data Pre-processing)

practice_인공지능,머신러닝 · 2024. 10. 13. 14:09

데이터 전처리 (Data Pre-processing)

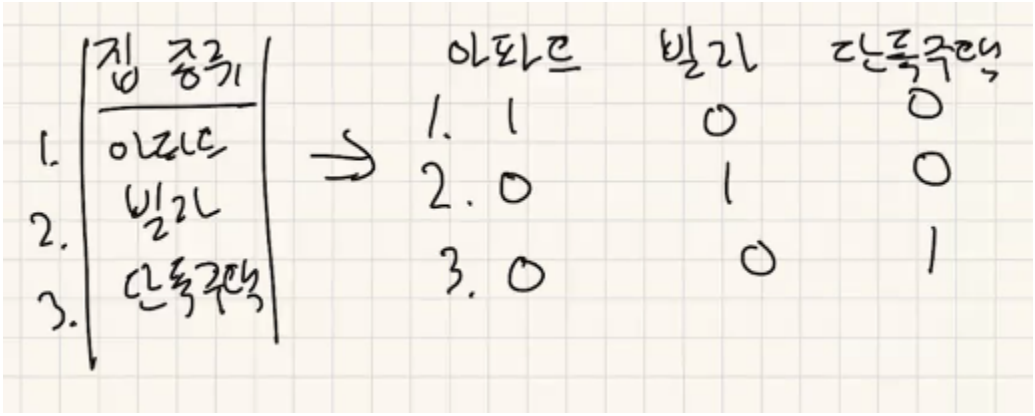
데이터 전처리는 모델 학습 전에 데이터를 정제(filter-이상한거 거름)하고 준비하는 과정입니다.

전처리가 잘 안되면 성능이 떨어질 가능성이 있다.

1. **데이터 정제**: 결측값, 이상치, 중복 데이터를 처리하는 과정입니다.
2. **데이터 변환**: 정규화(데이터 범위를 조정)나 표준화(평균 0, 분산 1로 조정)와 같은 변환을 수행하며, 범주형 데이터를 숫자로 변환(예: One-Hot Encoding)합니다.
3. **차원 축소**: 불필요한 변수를 줄이는 방법으로 주성분 분석(PCA) 등이 있습니다.
4. **데이터 분할**: 데이터를 학습용, 검증용, 테스트용으로 나누어 모델 성능을 평가합니다.
5. **특징 추출 및 생성**: 의미 있는 새로운 특징을 생성합니다.
6. **불균형 데이터 처리**: 클래스 불균형 문제를 해결하기 위한 방법으로 샘플링 기법을 사용합니다.

데이터 전처리는 데이터 품질을 높이고, 모델 성능을 향상시키기 위한 중요한 단계입니다.

예) one hot encoding



---

1. 데이터 정제(Data Cleaning)

데이터 정제는 결측값, 이상치, 중복 데이터 등을 처리하는 과정입니다.

- **결측값 처리(Missing Values)**: 결측값은 다양한 이유로 발생할 수 있으며, 이를 처리하는 방법은 여러 가지입니다. 대표적인 방법으로는 결측값을 제거하거나 평균, 중간값, 최빈값 등으로 대체하는 방식이 있습니다.
    - 예: df.fillna(mean_value) 또는 df.dropna()
- **이상치 처리(Outliers)**: 데이터에서 정상 범위에서 벗어난 값들을 감지하고 처리하는 단계입니다. 이상치를 제거하거나 수정해야 모델의 성능을 저해하지 않습니다.
- **중복 데이터 처리**: 동일한 레코드가 중복되었을 때 이를 제거하여 데이터의 일관성을 유지합니다.

2. 데이터 변환(Data Transformation)

데이터를 모델이 학습할 수 있는 형태로 변환하는 과정입니다.

- **정규화(Normalization)**: 데이터의 범위를 0과 1 사이로 조정하여 스케일의 차이를 줄이는 과정입니다. 특히 거리 기반 모델에서 중요합니다.
    - 방법: 최소-최대 정규화(Min-Max Normalization)

  Xnorm=X−XminXmax−XminX_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}Xnorm=Xmax−XminX−Xmin
- **표준화(Standardization)**: 데이터의 평균을 0, 분산을 1로 변환하여 분포의 차이를 줄이는 과정입니다. 특히, 선형 회귀나 신경망과 같은 모델에서 자주 사용됩니다.

- 방법: Z-스코어 정규화(Z-score Normalization)

  Xstd=X−μσX_{std} = \frac{X - \mu}{\sigma}Xstd=σX−μ여기서 μ\muμ는 평균, σ\sigmaσ는 표준 편차입니다.

- **범주형 데이터 인코딩(Categorical Data Encoding)**: 범주형 데이터를 수치 데이터로 변환하는 방법입니다.
  - **Label Encoding**: 범주형 값을 정수로 변환하는 방식입니다.
  - **One-Hot Encoding**: 범주형 데이터를 이진 벡터로 변환하는 방식입니다. 예: pd.get_dummies()

3. 차원 축소(Dimensionality Reduction)

많은 변수로 구성된 데이터셋의 경우, 불필요한 변수들을 제거하여 차원을 줄임으로써 모델의 성능을 높이고 계산 효율성을 향상시킵니다.

- **주성분 분석(PCA, Principal Component Analysis)**: 데이터를 주성분으로 변환하여 차원을 축소하는 기법입니다.
- **특징 선택(Feature Selection)**: 모델의 성능에 중요한 변수만을 선택하는 과정입니다.

4. 데이터 분할(Data Splitting)

모델을 학습하기 위해 데이터를 학습용과 평가용으로 분할하는 과정입니다.

- **훈련 데이터(Training Set)**: 모델을 학습시키는 데 사용됩니다.
- **검증 데이터(Validation Set)**: 모델을 평가하고 하이퍼파라미터 튜닝에 사용됩니다.
- **테스트 데이터(Test Set)**: 최종적으로 모델 성능을 평가하는 데 사용됩니다.
  - 일반적인 비율: 훈련 70%, 검증 15%, 테스트 15% 또는 훈련 80%, 테스트 20%.

5. 특징 추출 및 생성(Feature Extraction & Engineering)

특징 추출은 기존 데이터를 기반으로 새로운 특징을 생성하거나 필요한 정보를 추출하는 과정입니다. 도메인 지식에 기반하여 의미 있는 정보를 뽑아내는 것이 중요합니다.

6. 균형 조정(Handling Imbalanced Data)

클래스 간 데이터가 불균형할 때, 이를 해결하기 위해 언더샘플링, 오버샘플링, SMOTE 등의 기법을 사용하여 데이터를 균형 있게 조정합니다.

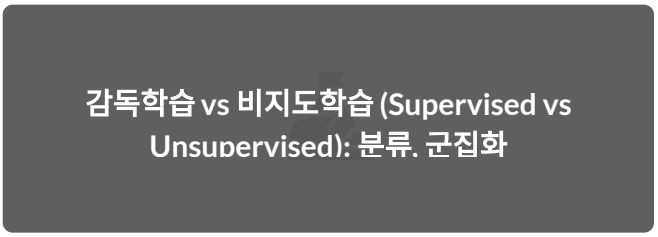
• • •

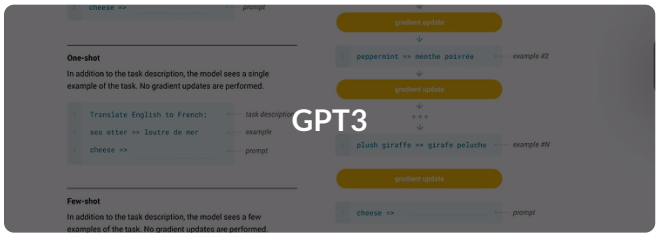
데이터 전처리는 모델 학습에 앞서 필수적으로 수행되어야 하며, 데이터 품질이 모델 성능에 직접적인 영향을 미치기 때문에 매우 중요합니다. 적절한 전처리 과정을 통해 모델이 더 정확하고 신뢰성 있는 결과를 낼 수 있도록 돕습니다.

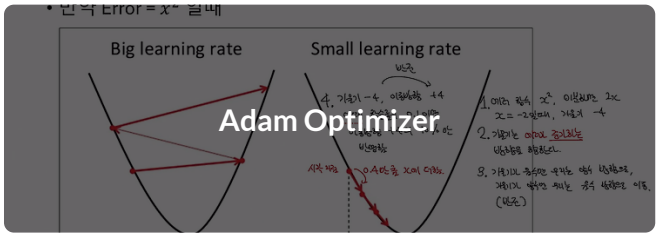
'practice_인공지능,머신러닝' 카테고리의 다른 글

감독학습 vs 비지도학습 (Supervised vs Unsupervised): 분류, 군집화 (0)	2024.10.13
GPT3 (0)	2024.10.12
Transformer, RNN 차이 (0)	2024.10.12
Adam Optimizer (1)	2024.10.12
Batch Normalization (3)	2024.10.12

관련글 관련글 더보기







자연어(NLP)

네이쳐2024 님의 블로그입니다.

구독하기 +

댓글 0

이름

비밀번호

내용을 입력하세요.

등록