TCT-기술인증테스트

AI

서술형 해답지

[2020년 #차]

LG CNS

[서술형 1번] 딥러닝을 활용한 이진 분류 문제 해결 과정 (5점)

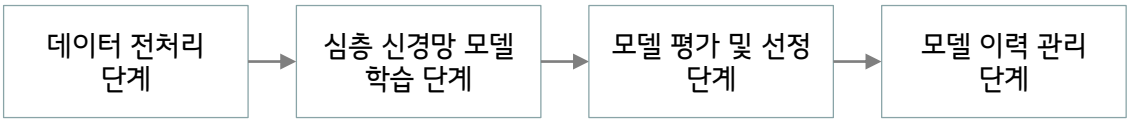
- 데이터 준비하기 – 데이터셋의 양품과 불량품의 비율을 생산 공정과 동일하게 유지할 경우 비대칭 데이터 문제가 발생하여 단순히 우세한 클래스를 선택하는 모델이 만들어질 가능성이 높다. 비대칭 데이터는 다수 클래스 데이터에서 일부만 사용하는 언더샘플링이나 소수 클래스 데이터를 증가시키는 오버 샘플링을 사용하여 데이터 비율을 유사하도록 만든다. (keyword: 데이터셋의 비대칭 언급 및 그에 대한 해결책 제시). 학습 및 검증 데이터셋은 양불 비슷한 비율로, 테스트 데이터셋은 실제 비율로 해야 할 것을 명시해야 함
- 성능 지표 정의하기 – 비대칭 데이터셋으로 테스트를 진행하였으므로 높은 accuracy가 하자 판정의 정확성을 보장하지 않는다. Confusion matrix를 통해 불량품 판정에 대한 recall과 양품 판정에 대한 precision으로 모델 성능을 평가해야 한다. (keyword: confusion matrix, recall, precision 관련 언급)
- 학습할 심층 신경망 모델 선정하기 – deep neural network 모델을 선정함에 있어 적절한 capacity의 모델을 찾기 위한 과정이 없었다. 깊은 구조의 모델은 높은 capacity를 가지기 때문에 과대적합(Overfitting)의 가능성이 높고, 더 많은 데이터, 더 많은 학습 시간 및 자원을 요구하기 때문에 최소 비용으로 시스템을 구축하려는 의도에도 맞지 않는다. 작고 단순한 모델에서 시작하여 순차적으로 capacity가 높은 모델을 시도하는 방향이 바람직하다.

[서술형 2번] 딥러닝의 실용적 방법론 점검 (4점)

- 1번 : 옳음(O) – 학습 데이터가 실세계의 분포를 잘 반영할수록 학습된 모델은 처음 보는 데이터에 대해서도 더 나은 추론이 가능하게 되는데, 실제 데이터를 더 추가하는 방법은 학습 데이터의 분포를 최대한 실제와 유사하게 만드는 가장 직접적이고 효과적인 방법이다.
- 2번 : 틀림(X) – 서로 다른 도메인의 데이터는 그 특징이 다르기 때문에 동일한 학습률 감쇠로 동일한 효과를 얻을 수 없다.
- 3번 : 옳음(O) – 무작위 탐색은 그리드 탐색에 비해 불필요한 반복 수행 횟수를 대폭 줄이면서 동시에 정해진 간격 사이에 위치한 값들에 대해서도 확률적으로 탐색이 가능하므로 최적 하이퍼파라미터 값을 더 빨리 찾을 수 있다.
- 4번 : 옳음(O) – early stopping 은 regularization의 한 방법으로, validation error 가 감소하다가 증가하는 시점에 모델 학습을 중단함으로써 과대적합(Overfitting)을 방지할 수 있다.

[서술형 3번] 딥러닝 응용 시스템 구축 (5점)

- 1) 프로세스 다이어그램 그리기:



- 2) 각 단계별 주요 기능 설명:

① 데이터 전처리 단계: 데이터 검증, 클린징, 증폭 등 학습을 위한 데이터 처리

② 모델 학습 단계: 학습 데이터로 심층 신경망 모델 학습

③ 모델 평가 및 선정 단계: 검증 데이터로 학습된 모델 평가, 최고 성능 모델 선정

④ 모델 관리 단계: 모델 이력 및 생명주기 관리, 모델 배포

[서술형 4번] 딥러닝 적용 여부 판단 (6점)

- 1) 데이터 차원(dimension) 또는 특징(feature)의 조건:
  분석하고자 하는 데이터 또는 학습 데이터가 고차원 데이터(high dimensional data)인 경우
  또는 고차원 데이터를 잘 표현하기 위해서 고차원 특징이 필요한 경우

- 2) 위에서 설명한 조건에서 딥러닝이 높은 성능을 보이는 이유

딥러닝의 심층 신경망 모델은 단계별 은닉층을 거칠 수록 고차원 데이터의 잠재적 특징(feature)을 포착하여 고도로 추상화된 표현(representation)을 할 수 있기 때문임.