2022 Deep Learning Study

Deep Learning Ch3-4

D조 Ch3-4 김도희 류하영 설현기 이지수 최동석 하정원

CH4 머신 러닝의 기본 요소

Universal -Workflow

머신 러닝 문제를 해결하기 위해 사용 가능한 **보편적 청사진 제시**

=> 이 청사진은 해당 장에서 배운 문제 정의, 평가, 특성 공학, 과대적합의 개념과 연결

4.5.1

문제 정의와 데이터셋 수집

주어진 문제 정의해야함
가용한 훈련 데이터가 있어야 어떤 것을 예측하도록 학습 가능, 입력/출력이 무엇인지 어떤 데이터 사용할 것인지 알기 전까지는 다음 단계로 넘어갈 수 X

* 필요한 가설
① 주어진 입력으로 출력을 **예측**할 수 있다
② 가용한 데이터에 입력/출력 사이의 관계를 학습하는 데 **충분한 정보**가 있다
=> 모델 작동 전까진 가설에 불과, 검증될지 아닐지 기다려야함

* 풀기 어려운 종류의 문제 : 시간에 따라 변하는 문제
시간 따라 모델 바꾸기 위해 **최근 데이터를 이용해 주기적으로 모델을 재훈련**하거나,
시간 분포에 맞게 데이터 수집해 시간 따라 변하지 않는 문제로 바꿈
=> 1년 중 언제인지, 기록시간도 모델에 입력해야함

머신러닝은 훈련 데이터에 있는 패턴을 기억하기 위해서만 사용
=> 이미 보았던 것만 인식 가능

미래 예측 위해 과거 데이터에서 훈련한 머신 러닝 사용하는 것
=> 미래가 과거처럼 움직인다고 가정한 것

4.5.2

성공 지표 선택

제어하기 위해선 관측 가능해야함

클래스 분포 균일한 분류 문제에서 일반적 지표
: 정확도, ROC AUC
클래스 분포 불균일한 분류 문제에서 일반적 지표
: 정밀도, 재현율

랭킹 문제/다중 레이블 문제 : 평균 정밀도

* 용어 설명

① ROC(Receiver Operating Characteristic curve)
거짓양성비율에 대한 재현율의 곡선

② ROC AUC
ROC 곡선의 아랫부분 면적

③ **정밀도-재현율 곡선** : 정밀도에 대한 재현율의 곡선

④ **평균 정밀도** : 위 곡선의 아랫부분 면적

=> 사이킷런은 ROC AUC를 위한 roc_auc_score() 함수와
평균 정밀도를 위한 average_precision_score() 함수 제공

4.5.3

평가 방법 선택

목표를 정했다면 **현재의 진척 상황을 평가할 방법** 정해야 함

* 가장 잘 알려진 3가지의 평가 방식

① **홀드아웃 검증 세트 분리** : 데이터가 풍부할 때 사용

② **K-겹 교차 검증** : 홀드아웃 검증을 사용하기에 샘플 수가 너무 적을 때 사용

③ **반복 K-겹 교차 검증** : 데이터가 적고 매우 정확한 모델 평가가 필요할 때 사용

=> 위 3가지 방법 중 택1하면 됨, 대부분 ①번 방법으로 충분

4.5.4

데이터 준비

무엇을 훈련, 최적화, 평가할지 정했다면 거의 모델 훈련 준비 완료

=> but, 머신 러닝 모델에 주입할 데이터 구성해야 하는데 여기서 머신 러닝 모델=심층 신경망 가정

* 머신 러닝 모델에 주입할 데이터 구성

① 데이터는 **텐서**로 구성

② 이 텐서에 있는 값은 일반적으로 **작은 값으로 스케일 조정** 되어 있음 ex). [-1. 1] or [0, 1]

③ 특성마다 범위가 다르면(=여러 종류의 값으로 이루어진 데이터라면) 정규화

④ **특히 데이터가 적을 때**, **특성 공학** 수행 가능

=> 입력/타깃 데이터의 텐서가 준비되면 모델 훈련 가능

4.5.5

기본보다 나은 모델 훈련하기

해당 단계의 목표 : 통계적 *검정력을 달성하는 것
=> 아주 단순 모델보다 나은 수준의 작은 모델 개발
ex).
MNIST 숫자 이미지 분류 : 0.1보다 높은 정확도를 내는 모델
IMDB : 0.5보다 높은 정확도를 갖는 모델
=> 통계적 검정력 가짐

통계적 검정력 달성이 늘 가능하진 X

***검정력**=가설이 참일 때 이를 채택할 확률*
여기선 적어도 데이터셋의 클래스별 분포<모델의 정확도
여야 가설이 옳다

* 통계적 검정력 달성 항상 가능 X

=> 여러 개 타당성 있는 네트워크 구조 시도해보고 무작위 예측하는 모델보다 낫지 않다면
입력 데이터에 존재하지 않는 것을 얻으려고 한다는 신호

* 첫 번째 모델 만들기 위해 중요한 3가지 선택 필요

① 마지막 층의 활성화 함수 : 네트워크 출력에 필요한 제한을 가함.
ex). IMDB 분류는 마지막 층에 시그모이드 함수 사용, 회귀는 마지막 층에 활성화 함수 사용 X

② 손실 함수** : 풀려고 하는 문제의 종류에 적합해야함
=> but, 주어진 문제의 성공 지표를 직접 최적화하는 것이 항상 가능하지 X
ex). IMDB는 binary_crossentropy 사용, 회귀는 mse 사용

③ 최적화 설정 : 주로 rmsprop 옵티마이저, 기본 학습률을 사용하는 게 일반적
=> 자주 사용하는 옵티마이저 'rmsprop', 'adam' & 기본 학습률 0.001 / 'sgd', 'adagrad' & 기본 학습률 0.01

***손실 함수는 주어진 미니 배치 데이터에서 **계산/미분 가능**해야함*
*ex). ROC AUC는 직접 최적화 X, 0-1 손실이라고도 부르는 정확도 비롯해 ROC AUC 계산에 사용되는 거짓양성비율과 재현율 모두 예측 맞은/틀린 개수 헤아리는 방식 => **미분 불가능***

4.5.5

기본보다 나은 모델 훈련하기

해당 단계의 목표 : 통계적 *검정력을 달성하는 것
=> 아주 단순 모델보다 나은 수준의 작은 모델 개발

ex).
MNIST 숫자 이미지 분류 : 0.1보다 높은 정확도를 내는 모델
IMDB : 0.5보다 높은 정확도를 갖는 모델
=> 통계적 검정력 가짐

통계적 검정력 달성이 늘 가능하진 X

***검정력**=가설이 참일 때 이를 채택할 확률*
여기선 적어도 데이터셋의 클래스별 분포<모델의 정확도
여야 가설이 옳다

* 자주 등장하는 문제 유형에 따라 선택가능한 마지막 층의 활성화 함수

표 4-1 모델에 맞는 마지막 층의 활성화 함수와 손실 함수 선택

문제 유형	마지막 층의 활성화 함수	손실 함수
이진 분류	시그모이드	binary_crossentropy
단일 레이블 다중 분류	소프트맥스	categorical_crossentropy
다중 레이블 다중 분류	시그모이드	binary_crossentropy
임의 값에 대한 회귀	없음	mse
0과 1 사이 값에 대한 회귀	시그모이드	mse 또는 binary_crossentropy

4.5.6

몸집 키우기
: 과대적합 모델 구축

통계적 검정력을 가진 모델 얻었다면 **모델이 충분히 성능을 내는지 질문**해봐야함

머신러닝은 최적화와 일반화 사이의 줄다리기
=> 과소/과대적합 사이 = 과소/과대용량의 경계에 적절히 위치한 모델이 이상적

* 과대적합된 모델 만들기

① 층 추가
② 층 크기 키우기
③ 더 많은 에포크 동안 훈련

=> 항상 관심 대상인 훈련/검증 지표 + 훈련/검증 손실을 모니터링,
검증 데이터에서 모델 성능이 감소하기 시작했을 때 과대적합에 도달한 것

4.5.7

모델 규제와 하이퍼파라미터 튜닝

해당 단계가 대부분의 시간 차지

① 반복적으로 모델 수정+훈련
② 검증 데이터 평가 (이때, 테스트 데이터 사용 X)
③ 재수정거쳐 가능한 좋은 모델 얻을 때까지 반복

* 모델 수정할 때 적용해볼 것

- 드롭아웃 추가
- 층 추가/제거하다가 다른 구조 시도
- L1/L2 또는 두 가지 모두 추가
- 최적 설정 찾기 위해 하이퍼파라미터 바꾸어 시도*(층 유닛 수나 옵티마이저의 학습률 etc.)*
- 선택적으로 특성 공학 시도, 새로운 특성 추가/유용하지 않을 것 같은 특성 제거

검증 과정에서 얻은 피드백 사용하여 **모델 튜닝할 때마다** 모델에 검증 과정에 대한 **정보를 누설**
=> 多 반복 시, 모델이 검증 과정에 과대적합되어 검증 과정의 신뢰도 감소

만족스러운 모델 설정 얻었다면 가용한 모든 데이터(훈련, 검증) 사용해 제품 투입할 최종 모델 훈련
+ **마지막 딱 한 번** 테스트 세트에서 평가

* 테스트 세트 성능이 검증 데이터 측정 결과보다 많이 나쁠 경우

① 검증 과정 신뢰성 X
② 모델의 하이퍼파라미터를 튜닝하는 동안 검증 데이터에 과대적합 된 것
=> 이럴 경우 좀 더 신뢰할 만한 평가 방법으로 교체 *ex).반복 K-겹 교차 검증*

Summary

① **주어진 문제와 훈련할 데이터 정의** - 이 데이터를 수집하고 필요하면 레이블을 태깅

② **성공을 어떻게 측정할지 선택** - 검증 데이터에서 모니터링할 지표는 무엇?

③ **평가 방법 결정** - 홀드아웃 검증? K-겹 교차 검증? 검증에 사용할 데이터 양?

④ 단순 랜덤 선택 모델보다 **나은 통계적 검정력이 있는 첫 번째 모델 제작**

⑤ **과대적합된 모델** 제작

⑥ **검증 데이터 성능에 기초**하여 모델에 규제를 적용하고 하이퍼파라미터 튜닝

(머신 러닝 연구 대부분 이 단계에 집중)

THANK YOU

D조 Ch3-4 김도희 류하영 설현기 이지수 최동석 하정원