



의료 인공지능 머신러닝 - 앙상블 모델

고려대학교 의료빅데이터연구소
채민수(minsuchae@korea.ac.kr)

1. 앙상블 모델이란

- 앙상블 모델(Ensemble)
 - 종전의 머신러닝 알고리즘의 성능을 개선하기 위한 방안
 - 여러 개의 머신러닝 알고리즘을 통해 얻어진 결과를 바탕으로 추론

1. 앙상블 모델이란

◦ 투표 기반 앙상블

- 다양한 알고리즘들을 결합하여 투표를 통해 그 결과를 반영
 - 다수 투표(majority voting) : 더 많이 예측한 것을 정답으로 반영. 각 알고리즘마다 동일한 가중치.
 - 가중 다수 투표(weighted majority voting) : 투표 수를 보는 것이 아닌 확률을 고려함. 확률이 높은 것에 대한 가중치 부여.
- Scikit-learn 에 서 는 VotingClassifier 와 VotingRegressor 를 지 원 하 며 , voting 파라미터에 hard와 soft에 따라 투표 수 혹은 확률을 고려함

2. 투표 기반 앙상블 모델 실습

- 물고기 데이터셋을 활용하여, Decision tree, Voting classifier 실습
 - 데이터셋 로드
 - 데이터 전처리
 - 입력 특징과 타겟 구분
 - 훈련 데이터와 테스트 데이터 나누기
 - Decision tree, Voting classifier를 활용한 실습

3. 데이터 리샘플링 기반 앙상블

- 데이터 리샘플링

- 성능 개선을 위해 훈련 데이터의 변화를 줌
 - Bootstrap : 데이터셋에서 중복을 허용하며 임의의 샘플을 선택하여 학습
 - Bootstrap aggregating(bagging) : Bootstrap을 확장한 투표 기반 앙상블
 - Pasting : 데이터셋에서 중복을 허용하지 않고 임의의 샘플을 선택하여 학습
- Scikit-learn에서는 BaggingClassifier와 BaggingRegressor를 지원하며, bootstrap에 True을 넣으면 Bootstrap, False를 넣으면 Pasting으로 수행

4. 데이터 리샘플링 기반 앙상블 모델 실습

- 물고기 데이터셋을 활용하여, Decision tree, Voting classifier 실습
 - 데이터셋 로드
 - 데이터 전처리
 - 입력 특징과 타겟 구분
 - 훈련 데이터와 테스트 데이터 나누기
 - Decision tree, Voting classifier를 활용한 실습

5. 랜덤 포레스트와 엑스트라 트리

- Random forest
 - Bagging을 이용한 Decision tree 기반의 앙상블 모델
 - 각 Decision tree에 사용하는 특징을 변화를 줌.
 - 특징 중 최적의 특성을 찾으려 함
 - Scikit-learn에서는 RandomForestClassifier와 RandomForestRegressor를 지원
- Extra tree(extremely randomized tree)
 - Random forest와 유사함. 특징을 임의로 분할 후 최적의 특징을 선택
 - Scikit-learn에서는 ExtraClassifier와 RandomForestRegressor를 지원

6. 랜덤 포레스트 실습

- 피마 인디언 당뇨병 데이터셋을 이용한 랜덤 포레스트 실습
 - 데이터셋 로드
 - 데이터 전처리
 - 입력 특징과 타겟 구분
 - 훈련 데이터와 테스트 데이터 나누기
 - RandomForestClassifier를 활용한 실습

7. 오버 샘플링과 언더 샘플링

- 오버 샘플링과 언더 샘플링
 - 데이터셋의 불균형을 해결하는 방안
 - 오버 샘플링의 경우 데이터셋을 복제, 언더 샘플링의 경우 제거를 수행
 - 오버샘플링은 과적합의 문제점이 발생할 수 있으며, 언더 샘플링은 정보 손실이 발생함
 - 오버 샘플링과 언더샘플링은 훈련 데이터셋에서만 적용해야 함
 - <https://imbalanced-learn.org/stable/>

7. 오버 샘플링과 언더 샘플링

◦ 오버 샘플링과 언더 샘플링

Over-sampling methods

RandomOverSampler

SMOTE

SMOTENC

SMOTEN

ADASYN

BorderlineSMOTE

KMeansSMOTE

SVM SMOTE

Under-sampling methods

ClusterCentroids

CondensedNearestNeighbour

EditedNearestNeighbours

RepeatedEditedNearestNeighbours

AllKNN

InstanceHardnessThreshold

NearMiss

NeighbourhoodCleaningRule

OneSidedSelection

RandomUnderSampler

TomekLinks

<https://imbalanced-learn.org/stable/references/index.html>

7. 오버 샘플링과 언더 샘플링 실습

- 피마 인디언 당뇨병 데이터셋을 이용한 오버 샘플링과 언더 샘플링 실습
 - 데이터셋 로드
 - 데이터 전처리
 - 입력 특징과 타겟 구분
 - 훈련 데이터와 테스트 데이터 나누기
 - 훈련 데이터셋에 오버 샘플링과 언더 샘플링 수행
 - RandomForestClassifier를 활용한 실습

7. 오버 샘플링과 언더 샘플링 실습

- breast-cancer-wisconsin.data를 이용한 오버 샘플링과 언더 샘플링 실습
 - 데이터셋 로드
 - 데이터 전처리
 - 입력 특징과 타겟 구분
 - 훈련 데이터와 테스트 데이터 나누기
 - 훈련 데이터셋에 오버 샘플링과 언더 샘플링 수행
 - RandomForestClassifier를 활용한 실습

8. 엑스트라 트리 실습

- wdbc.data를 이용한 위스콘신 유방암 진단 예측 실습
 - 데이터셋 로드
 - 데이터 전처리
 - 입력 특징과 타겟 구분
 - 훈련 데이터와 테스트 데이터 나누기
 - 훈련 데이터셋에 오버 샘플링과 언더 샘플링 수행
 - RandomForestClassifier와 ExtraTreesClassifier를 활용한 실습

8. 엑스트라 트리 실습

- wpbc.data를 이용한 위스콘신 유방암 예후 예측 실습
 - 데이터셋 로드
 - 데이터 전처리
 - 입력 특징과 타겟 구분
 - 훈련 데이터와 테스트 데이터 나누기
 - 훈련 데이터셋에 오버 샘플링과 언더 샘플링 수행
 - RandomForestClassifier와 ExtraTreesClassifier를 활용한 실습

9. 회귀 실습

- exams.csv 데이터셋을 이용한 오버 샘플링과 언더 샘플링 실습
 - 데이터셋 로드
 - 데이터 전처리
 - 입력 특징과 타겟 구분
 - 훈련 데이터와 테스트 데이터 나누기
 - RandomForestRegressor를 활용한 실습

10. Homework

- 스스로 해보기

- 물고기 데이터셋의 분류의 예측 정확도의 성능을 80%를 초과하도록 하여라.