7주차. 앙상블 학습과 랜덤 포레스트

1. 앙상블학습

앙상블 학습

- 정의 : 여러개 모델 훈련 결과 이용 -> 좋은 성능

- 종류 : 배깅, 부스팅

- 표 형식으로 저장되는 정형 데이터 분석에 유용 (이미지, 오디오 등의 데이터는 딥러닝이 유리)

- 핵심 : 편향, 분산 줄이기

배깅

- 여러개 예측기 독립적으로 학습 후 평균값 사용 (병렬 적용)

- 분산 줄이기

- 랜덤포레스트, 엑스트라 트리

부스팅

- 에측기 순차적으로 훈련 후 적은 편향 모델 사용 (순차 적용)

- 편향 줄이기

- 그레이언트 부스팅, XGBoost

편향 : 에측값과 떨어진 정도, 과소적합 발생

분산 : 작은 변동에 반응하는 정도, 과대적합 발생

트레이드오프 : 편향과 분산 둘 다 좋아지게 할 수 없음

- 회귀모델 평균제곱오차 = 편향^2 + 분산

2. 투표식 분류기

투표식 분류기 : 여러 분류기로 앙상블 학습 후 투표를 통해 예측값 결정

직접투표 : 각각의 모델에 같은 권리 주어지는 다수결 투표 방식

간접투표 : 각 값들의 평균값으로 예측 (직접투표 방식보다 성능 좋음, 확률 예측 기능 지원할때만 사용가능)

튜표식 분류기의 확률적 근거 : 이항분표의 누적분포함수 이용

3. 배깅 vs 페이스팅

- **배깅**(bootstrap aggregation) : 중복 허용 샘플링
- **페이스팅** : 중복 미허용 샘플링
- **예측값**
    - 분류모델 : 직접 투표 방식, 최빈값 선택
    - 회귀모델 : 예측값 평균 선택
- **개별 예측기** : 배깅/페이스팅으로 학습시 편향커짐, 과소적합 위험성 커짐
- **OOB 평가**(out of bag) : 훈련에 사용하지 않은 모델로 학습 모델 검증

4. 랜덤패치 & 랜덤서브스페이스

- 특성에 대해서도 리샘플링함 (원래는 데이터에 대해서만 리샘플링)
- 높은 차원 데이터셋 다룰때 유용
- 더 다양한 예측기 만듬 (편향 커지고 분산 낮아짐)
- Max_features : 학습에 사용할 특성 수 지정, 특성 선택 무작위
- Bootstrap_features : 중복 허용 여부 지정
- 랜덤 패치 기법 : 중복 허용. 샘플수와 특성수 만큼 샘플링하는 학습 기법
- 랜덤 서브스페이스 기법 : 전체 훈련 세트 대상, 특성은 특성 수만큼 샘플링

5. 랜덤 포레스트

랜덤 포레스트

: 배깅/페이스트 방법 적용한 결정트리 앙상블 최적화 모델

- 결정트리에 비해 편향 크고 분산 낮음

- BaggingClassifier와 DecisionTreeClassifier의 옵션을 거의 모두 가짐

엑스트라 트리

- 익스트라 랜덤 트리 앙상블

- 무작위로 선택된 일부 특성에 대해 임계값도 무작위로 선택 후 최적 선택

- 일반 랜덤포레스트보다 속도 훨씬 빠름

- 평향 높이고 분산 줄어듬

특성 중요도

: 해당 특성 사용한 마디가 평균적으로 불순도 얼마나 감소 시키는지 측정

- 불순도 많이 줄이는 특성 -> 중요도 커짐

6. 부스팅

부스팅

- 성능이 약한 모델을 순차적으로 보다 강한 성능의 모델 만들어내는 기법
- 편향을 줄여나감
- 대표 모델 : 에이다부스트(AdaBoost), 그레디언트 부스팅, XGBoost

그레디언트 부스팅

- 이전 모델에 의해 생성된 잔차 보정하도록 훈련
- 잔차 : 예측값과 실제값 사이의 오차
- 결정트리 연속적으로 사용

학습률 : 훈련된 결정 트리 모델이 각각 최종 예측값 계산시의 기여도 결정

수축 규제 : 훈련에 사용되는 각 모델의 기여도 줄이는 방식으로 훈련 규제

조기종료 : 연속적으로 n번 제대로 개선되지 못하는 경우 훈련 자동 종료

확률적 그레디언트 부스팅

- 각 결정트리가 훈련에 사용할 훈련 샘플의 비율 지정하여 학습
- 훈련 속도 빨라짐
- 편향 높아짐 분산 낮아짐

히스토그램 그레디언트 부스팅

- 대용량 데이터셋 훈련시 사용
- 정확도 떨어짐
- 과대적합 방지 규제, 과소적합 발생 가능

XGBoost(Extreme gradient boosting)

- 그레디언트 부스팅과 차이점
  - 비용함수 다름
  - 불순도 대신 훈련 목적에 따른 손실 함수 사용(mse, logloss)
  - 결정트리 복잡도 낮추도록 유도
- 빠른속도, 확장성
- 결측치 포함 데이터 처리 가능