머신러닝 7주차 과제

박민서

1. 결정트리의 단점: 높은 분산-훈련 데이터의 작은 변화에도 매우 민감함. random_state를 지정하지 않으면서 동일한 모델을 훈련시키면 다른 결과가 나온다.
2. 앙상블 학습: 여러 개의 모델을 훈련시킨 결과를 이용한 기법
   (1) 배깅 기법: 여러 개의 예측기를 독립적으로 학습시킨 후 모든 예측기들의 예측값들의 평균값을 최종 모델의 예측값으로 사용한다. 적은 분산을 갖는 모델을 구현한다-분산 줄이기, 예측기 병렬 적용
   (2) 부스팅 기법: 여러 개의 예측기를 순차적으로 훈련시킨 결과를 예측값으로 사용하여 보다 적은 편향을 갖는 모델을 구현한다-편향 줄이기, 예측기 순차 적용
3. 가장 좋은 성능을 내는 앙상블 학습 모델-XGBoost, 랜덤 포레스트, 그래디언트 부스팅=>앙상블 학습 모델은 특히 표 형식으로 저장될 수 있는 정형 데이터의 분석에 유용하다.
4. 앙상블 학습의 핵심-편향과 분산 줄이기
   (1) 편향: 예측값과 정답이 떨어져 있는 정도, 정답에 대한 잘못된 가정으로 발생하며, 편향이 크면 과소적합 발생
   (2) 분산: 샘플의 작은 변동에 반응하는 정도, 정답에 대한 너무 복잡한 모델을 설정하는 경우 발생하며, 분산이 크면 과대적합 발생
5. 편향과 분산의 트레이드오프: 편향과 분산을 동시에 좋아지게 할 수는 없음.
6. 회귀모델의 평균제곱오차: 편향의 제곱과 분산의 합으로 근사됨
7. 투표식 분류기: 동일한 훈련셋에 대해 여러 종류의 분류기를 이용하여 앙상블 학습을 적용한 후 직접 또는 간접 투표를 통해 예측값을 결정
   (1) 직접 투표: 앙상블에 포함된 예측기들의 예측값들의 다수로 결정
   (2) 간접 투표: 앙상블에 포함된 예측기들의 예측한 확률값들의 평균값으로 예측값 결정, 모든 예측기가 predict_probal() 메서드와 같은 확률 예측 기능을 지원해야 함, 높은 확률에 보다 비중을 두기 때문에 직접투표 방식보다 성능이 좋음.
   (3) 투표식 분류기의 확률적 근거: 이항 분포의 누적분포함수를 이용하여 앙상블 학습의 성능이 향상되는 이유를 설명할 수 있음, 다수결을 따를 때 성공할 확률, 즉 다수결 의견이 보다 정확할 확률인 반환값을 사용
8. 배깅과 페이스팅: 여러 개의 동일 모델을 하나의 훈련셋의 다양한 부분집합을 대상

으로 학습시키는 방식, 부분집합을 임의로 선택할 때 중복 허용 여부에 따라 앙상블 학습 방식이 달라짐

(1) 배깅: 중복 허용 샘플링, 부트스트래핑: 통계에서 중복허용 리샘플링을 의미함./ 페이스팅: 중복 미허용 샘플링

(2) 예측값

-분류 모델: 직접 투표 방식 사용. 즉, 수집된 예측값들 중에서 최빈값을 선택

-회귀 모델: 수집된 예측값들의 평균값 선택

(3) 배깅/페이스팅 방식으로 훈련된 모델의 편향과 분산

-개별 예측기의 경우에 비해 편향은 조금 커지거나 비슷하지만, 분산은 줄어든다, 배깅이 표본 샘플링의 다양성을 보다 많이 추가하기 때문이다, 배깅이 과대적합의 위험성을 줄여주어, 배깅 방식이 기본으로 사용된다.

-개별 예측기: 배깅/페이스팅 방식으로 학습하면 전체 훈련셋을 대상으로 학습한 경우에 비해 편향이 커짐, 과소적합 위험성이 커짐

9. OOB 평가

(1) OOB 샘플: 배깅 모델에 포함된 예측기로부터 선택되지 않은 훈련 샘플, 평균적으로 훈련셋의 약 37%

(2) OOB 평가: 각각의 샘플에 대해 해당 샘플을 훈련에 사용하지 않은 모델들의 예측값을 이용하여 앙상블 학습 모뎀을 검증하는 기법

10. 랜덤 패치와 랜덤 서브스페이스: BaggingClassifier는 특성에 대한 샘플링 기능 지원, 이미지 등 매우 높은 차원의 데이터셋을 다룰 때 유용, 더 다양한 예측기를 만들며 편향이 커지지만 분산은 낮아짐.

(1) max_features: 학습에 사용할 특성 수 지정, 특성 선택은 무작위(정수의 경우 지정된 수만큼 특성 선택/부동소수점인 경우 지정된 비율만큼 특성 선택), max_samples와 유사 기능 수행

(2) bootstrap_features: 학습에 사용할 특성을 선택할 때 중복 허용 여부 지정, 기본값은 False로 중복을 허용하지 않음, Bootstrap과 유사 기능 수행

(3) 랜덤 패치 기법: 훈련 샘플과 훈련 특성 모두를 대상으로 중복을 허용하며 임의의 샘플 수와 임의의 특성 수만큼 샘플링해서 학습하는 기법

(4) 랜덤 서브스페이스 기법: 전체 훈련 세트를 학습 대상으로 삼지만, 훈련 특성은 임의의 특성 수만큼 샘플링해서 학습하는 기법

11. 랜덤 포레스트: 배깅/페이스팅 방법을 적용한 결정트리의 앙상블을 최적화한 모델

(1) 랜덤 포레스트 하이퍼파라미터: 결정트리에 비해 편향은 크고, 분산은 낮게.

(2) 엑스트라 트리: 무작위로 선택된 일부 특성에 대해 특성 임계값도 무작위로 몇 개 선택한 후 그 중에서 최적 선택, 일반적인 랜덤 포레스트보다 속도가 훨씬 빠름, 이 방식을 사용하면 편향은 늘고, 분산은 줄어듦.

(3) 특성 중요도: 해당 특성을 사용한 마디가 평균적으로 불순도를 얼마나 감소시키는지를 측정. 즉, 불순도를 많이 줄이는 특성은 그만큼 중요도가 커짐.

12. 부스팅: 성능이 약한 모델을 순차적으로 보다 강한 성능의 모델로 만들어 가는 기법, 순차적으로 이전 학습기의 결과를 바탕으로 예측값의 정확도를 조금씩 높혀감 (편향을 줄여나감)
    (1) 그레디언트 부스팅: 이전 모델에 의해 생성된 잔차를 보정하도록 새로운 예측기 훈련
    (2) 잔차: 예측값과 실제값 사이의 오차
    (3) 모델은 결정트리 사용
    (4) 학습률: 훈련된 결정 트리 모델 각각이 최종 예측값을 계산할 때의 기여도 결정, 경사하강법의 학습률과 다르지만 최종 모델에 수렴하는 속도를 조절한다는 차원에서 동일한 기능 수행
    (5) 수축 규제: 훈련에 사용되는 각 모델의 기여도를 줄이는 방식으로 훈련 규제, 학습률을 낮게 정하면 많은 수의 결정트리가 필요하지만 성능은 일반적으로 좋아짐.
    (6) 조기 종료: 원래 500번 연속 결정트리를 훈련시켜야 하지만 검증셋에 대해 연속적으로 10번 제대로 개선되지 못하는 경우 훈련 자동 종료
    (7) 확률적 그레디언트 부스팅: 각 결정트리가 훈련에 사용할 훈련 샘플의 비율을 지정하여 학습, 훈련 속도 빨라짐, 편향이 높아지지만 분산이 낮아짐.
    (8) 히스토그램 그레디언트 부스팅: 대용량 데이터셋을 이용하여 훈련해야 하는 경우 사용, 훈련 샘플의 특성값을 max_bins 개의 구간으로 분류, 모델의 정확도는 떨어지며 경우에 따라 과대적합을 방지하는 규제역할 수행, 과소적합 발생 가능
    (9) XGBoost: 결정트리 학습에 사용되는 노드 분할을 통해 낮춰야 하는 비용함수가 다름, 불순도 대신 mse, logloss 등 모델 훈련의 목적에 맞는 손실함수 사용, 생성되는 결정트리의 복잡도가 비용함수에 포함되어 최종적으로 생성되는 모델에 사용되는 결정트리의 복잡도를 가능한 낮추도록 유도/빠른 속도, 확장성, 결측치 포함 데이터 처리 가능