9주차 정리노트

202184026 서동현

201904025 김성욱

◉ 앙상블 학습

- 앙상블 : 여러 개의 예측기로 이루어진 그룹
- 앙상블 학습 : 예측기 여러 개의 결과를 종합하여 예측값을 지정하는 학습

◉ 투표 기반 분류기

: 동일한 훈련 세트에 대해 여러 종류의 분류기 이용한 앙상블 학습 적용 후 직접 또는 간접 투표를 통해 예측값 결정

1. 직접 투표 : 앙상블에 포함된 예측기들의 예측값들을 다수결 투표로 결정
2. 간접 투표 :
   - 앙상블에 포함된 예측기들의 예측한 확률값들의 평균값으로 예측값 결정
   - 높은 확률에 보다 비중을 두기 때문에 직접투표 방식보다 성능 좀 더 좋음

3. 투표식 분류기의 특징 :
   - 앙상블에 포함된 분류기들 사이의 독립성이 전재되는 경우 개별 분류기 보다 정확한 예측 가능
   - 독립성이 보장되지 못한다면 투표식 분류기의 성능이 더 낮아질 수 있음

4. 큰 수의 법칙
   - 반복 시행하는 횟수가 많거나 표본이 커질수록 일정한 수준으로 수렴되고 비교적 정확한 예측이 가능하다는 의미

◉ 배깅과 페이스팅 : 동일한 예측기를 훈련 세트의 다양한 부분집합을 대상으로 학습시키는 방식

1. 배깅(bootstrap aggregation의 줄임말) : 중복 허용 샘플링 방식
2. 페이스팅(pasting): 중복 미 허용 샘플링 방식

◉ 랜덤 패치와 랜덤 서브스페이스

1. 랜덤 패치 기법 : 훈련 샘플과 훈련 특성 모두를 대상으로 중복을 허용하며, 임의의 샘플 수와 임의의 특성 수만큼을 샘플링해서 학습하는 기법

2. 랜덤 서브스페이스 기법 : 전체 훈련 세트를 학습 대상으로 삼지만 훈련 특성은 임의의 특성 수만큼 샘플링해서 학습하는 기법

◉ 랜덤 포레스트

: 배깅/페이스팅 방법을 적용한 결정트리의 앙상블을 최적화한 모델

- 랜덤포레스트의 노드 분할 방식
  - 특성: 무작위 선택
  - 특성 임곗값: 무작위로 분할한 다음 최적값 선택
- 엑스트라 트리의 노트 분할 방식
  - 특성과 특성 임계값 모두 무작위 선택(근단적으로 무작위한 트리의 랜덤 포레스트)

◉ 부스팅 : 성능이 약한 학습기를 여러 개 연결하여 강한 성능의 학습기를 만드는 앙상블 기법

- 부스팅 방법
  - 에이다부스트 : 좀 더 나은 예측기를 생성하기 위해 잘못 적용된 가중치를 조정하여 새로운 예측기를 추가하는 앙상블 기법
  - 그레이디언트 부스팅 : 샘플의 가중치를 수정하는 대신 이전 예측기가 만든 잔여 오차(residual error)에 대해 새로운 예측기를 학습시킴

질문

김성욱: 투표 기반 분류기에서 직접 투표 방식을 사용하는 것보다 간접 투표 방식을 무조건 사용하는 게 좋은거야?

서동현: 상황에 따라 다를 수도 있겠지만 보통 확률값들의 평균값으로 예측값을 결정하는 간접 투표가 성능이 더 좋게 나오니 간접 투표 방식을 사용할 것 같아

서동현: 부스팅에서 성능이 약한 학습기를 여러 개 연결하여 강한 성능의 학습기를 만드는데 성능이 강한 학습기를 사용하면 연결하는 학습기의 개수도 줄어들어서 속도도 빨라지고 학습기의 성능이 더 좋아지지 않을까?

김성욱: 속도도 빨라지고 초기 학습기만으로도 높은 성능을 얻을 수 있겠지만 과적합으로 인해서 새로운 데이터에 대한 일반화 능력이 약해질 수 있어서 일반화 성능이 더 좋은 약한 학습기를 여려 개 연결하는 방식이 더 좋은 거 같아.