**Week5\_Ensemble**

20기 김은서

앙상블은 여러 개별 모델을 조합하여 하나의 강력한 모델을 만드는 ML기법이다. 여러 개의 모델을 조합하는 방식은 각 모델이 서로 다른 관점에서 데이터를 해석하고 예측하는 방식을 활용하여 개별 모델의 약점을 보완하는데 중점을 둔다. 이러한 접근으로 단일 모델보다 더 높은 성능을 달성할 수 있고 일반화 능력을 제고할 수 있다.

앙상블 모델의 성능을 이해하기 위해선 분산(Variance)과 편차(Bias)의 개념을 이해하는 것이 중요하다. 이 두 요소가 모델의 복잡도와 예측 성능 간의 관계를 설명하는데 도움이 되기 때문이다. 우선 모델의 복잡도란 모델이 데이터의 패턴을 표현하거나 학습하는데 사용하는 파라미터 수 혹은 모델의 구조와 관련된 개념이다. 모델의 복잡도가 높다는 의미는 더 많은 파라미터나 구조를 가지며 데이터에 대한 학습을 더 정확하게 수행할 수 있지만, 동시에 과적합의 위험도 증가한다. 이러한 모델의 복잡도를 분산과 편차의 trade-off관계로 설명할 수 있다. 분산은 모델이 다양한 데이터셋에서 학습될 때 예측값의 변동성을 나타낸다. 그러나 높은 분산을 가진 모델은 학습 데이터에 과적합될 가능성이 크다. 과적합될 경우 새로운 데이터에서 예측 성능이 떨어질 수 있다. 따라서 분산이 높은 모델은 모델의 복잡도를 낮추어야 한다. 편차는 모델의 예측값과 실체값의 차이를 의미한다. 모델이 얼마나 실제 데이터의 패턴을 잘 잡아내는지를 측정한다. 편차가 클 경우 실제 데이터의 복잡한 패턴을 제대로 학습하지 못하고 예측이 부정확할 수 있다. 이를 과소적합이라고 한다. 따라서 편차가 높은 모델은 모델의 복잡도를 높여야 한다. 이처럼 편차와 분산이 trade-off관계를 가짐에 따라 앙상블 모델은 이를 극복하기 위해 여러 개별 모델을 조합한다. 다양한 모델을 결합함으로써 편차를 줄이고, 각 모델의 분산을 평균화하여 전체 앙상블의 분산을 낮춘다. 이로써 예측 성능과 일반화 능력을 향상시키는 것이다.

배깅과 부스팅 두 모델 모두 여러 개의 기본 모델을 조합하여 강력한 모델을 만드는 앙상블 기법이다. 그러나 모델을 조합하는 방식에 차이가 존재한다. 우선 배깅은 여러 개의 기본 모델을 학습시킬 때, 각 모델마다 다른 부분 데이터셋을 사용하여 학습하는 기법이다. 주어진 데이터셋으로부터 무작위로 복원추출하여 각 모델에 사용하는 데이터를 생상한다. 생성된 다양한 데이터셋을 사용하여 동일한 모델을 여러 번 학습시키고 각 모델의 예측을 평균하거나 투표를 통해 최종 예측을 결정한다. 가장 대표적인 예시로 Random Forest 기법이 있다. 부스팅은 약한 모델들을 순차적으로 학습시키는 기법으로, 이전 모델의 오차에 집중하여 새로운 모델을 학습시킨다. 이전 모델이 잘못 예측한 데이터에 가중치를 부여하여 다음 모델이 더 잘 예측하도록 돕는 방식이다. 대표적인 부스팅 알고리즘은 AdaBoost, Gradient Boosting, XGBoost, LightGBM 등이 있다.

두 기법의 차이점은 크게 4가지로 분류된다. 첫번째, 데이터 사용방식이다. 배깅은 당야한 부분 데이터셋으로 각 모델을 학습시키기 위해 복원추출을 사용한다. 부스팅은 이전 모델의 오차에 초점을 맞추기 위해 가중치를 조절하여 데이터를 사용한다. 두번째로 모델의 학습 순서다. 배깅은 병렬로 여러 모델을 학습시키고 각 모델은 독립적으로 작동한다. 부스팅은 순차적으로 모델을 학습시켜 이전 모델의 예측 오차를 보완한다. 세번째로 예측 결합 방식이다. 배깅은 다수결 투표나 평균 등을 통해 모델 예측을 결합한다. 부스팅은 각 모델의 예측을 가중합하여 최종 예측을 만든다. 마지막으로 모델 복잡도이다. 배깅은 일반적으로 복잡한 모델을 사용할 수 있다. 그러나 부스팅은 약한 모델, 즉 단순한 결정 트리 등을 사용하며 이를 순차적으로 강화시킨다. 두 기법 모두 예측 성능 향상을 목표로 하지만 배깅은 분산을 줄이고 모델 안정성을 높이는데 중점을 두며, 부스팅은 편차를 줄이고 예측 성능을 높이는 데 중점을 둔다.