(2). Random forest(RF)

RF는 다수의 독립적인 decision tree를 훈련해 이들의 다수결을 바탕으로 예측치를 도출한다. 먼저, decision tree는 여러 개의 분류 규칙을 통해 데이터를 여러 개의 부분집합으로 나눠 예측치를 도출한다. Figure 4는 일반적인 decision tree의 구조를 나타낸다.

먼저 데이터는 초기 분리지점인 root node를 시작으로 나눠지기 시작하며 분기가 거듭될수록, 즉 중간 마디(intermediate node)들을 계속해서 지날수록 계속해서 나눠지며 각 집합에 속한 데이터의 개수는 줄어든다. 분류 문제에 적용되는 경우 끝마디(terminal node)에 모인 데이터들의 범주에 따라 데이터를 각각의 집단으로 분류한다.

decision tree는 각각의 노드에서 데이터를 분류할 때 분류 뒤 각 영역의 불순도(impurity)를 최소화하는 방향으로 학습을 하며, 불순도는 엔트로피(entropy), 지니계수(Gini index), 또는 오 분류 오차(misclassification error)를 이용해 측정한다.

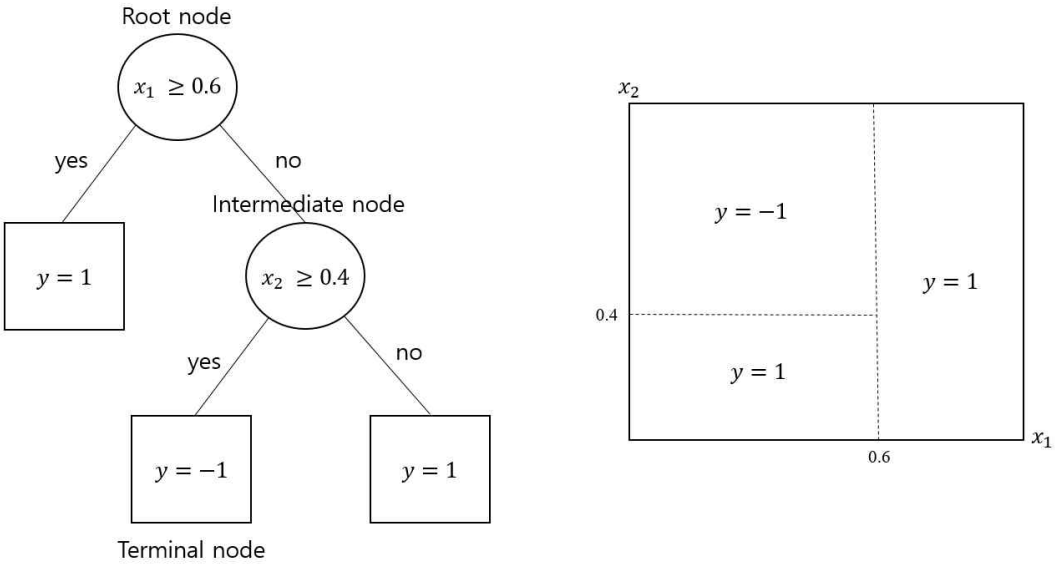

Figure 4. Structure of decision tree.

Decision tree는 모형의 학습이 간단하고 모형의 구조를 시각화할 수 있다는 장점이 있지만, 학습에 사용된 데이터가 조금만 달라지더라도 모형의 구조가 달라질 수 있다는 불안정성 및 표본 외 예측에서 높은 예측 오차를 보이는 과적합(overfitting)[1] 문제가 발생할 확률이 높다는 단점 또한 지닌다.

---

1) 과적합은 머신러닝 모형이 훈련 데이터를 과하게 학습해, 잡음(noise)에 해당하는 데이터까지 모형화하는 현상이다. 과적합이 발생하면 훈련 데이터에 대해서는 높은 예측 정확도를 보이지만, 모형화에 이용하지 않은 새로운 데이터가 주어졌을 때 예측 오차가 크게 나타나게 된다.

RF는 개별적인 decision tree가 지닌 불안정성 및 과적합 문제를 보완하기 위해 다수의 훈련된 decision tree를 사용하는 앙상블(ensemble)[2] 학습 방법이다. RF는 배깅(bagging)[3] 알고리즘을 통해 각 decision tree를 훈련할 때 무작위성을 도입한다. 즉, 각각의 decision tree들은 같은 데이터를 바탕으로 훈련되는 것이 아니라 전체 데이터에서 무작위로 추출된 표본들을 바탕으로 훈련된다. 이렇게 서로 다른 데이터들을 통해 훈련된 decision tree들을 사용해 예측치를 도출함으로써 데이터에 따라 모형에 달라지는 불안정성을 해소한다.

더 나아가, 각각의 decision tree를 훈련할 때 사용되는 데이터와 마찬가지로 예측변수들 또한 무작위로 선택된 일부분만 사용하게 된다. 예를 들어 자산의 가격 변동성을 예측하는데 거래량이 가장 높은 예측력을 가지는 예측변수라고 하자. 이런 경우에 모든 예측변수를 이용해 decision tree들을 훈련하는 경우 대부분 거래량을 중심으로 모형이 구성되어 decision tree들이 유사한 구조를 가지게 된다. RF는 각 마디에서 데이터를 나눌 때 무작위적으로 선택된 예측변수들을 이용함으로써 나무들 사이의 유사도를 낮춘다.

요약하면, decision tree는 일반적으로 훈련 데이터(training data)에서의 성능은 탁월하다. 그러나 실험 데이터(test data)에서의 예측 성능은 떨어지는 과적합 문제와 예측분산이 높아지는 경향을 보인다. RF는 배깅 알고리즘을 통해 decision tree를 다수 생성함으로써 표본 외 예측에서의 예측 오차를 줄이는 것으로 이해될 수 있다.

---

2) 앙상블 학습은 다수의 머신러닝 모형을 학습해 그 모델들의 예측 결과를 기반으로 하나의 예측치를 도출하는 방법론이다.

3) 배깅 알고리즘은 부트스트래핑(bootstrapping)을 통해 훈련데이터에서 샘플을 여러 번 추출해 각 모형을 학습시켜 결과를 집계하는 알고리즘이다.