Ch.6 Ensemble and Random Forest

1.1 Decision Tree의 기초

일반적으로 트리 기반의 방법들은 해석하기 쉽고, 용이하다.
하지만 성능이 비교적 낮기 때문에 Bagging, Random Forest, Boosting 등의 방법과 함께 사용된다.
이 기법들은 각각 다중트리를 생성하고, 최종적으로 하나의 합의 예측(Consensus Prediction)을 제공한다.
트리 모델은 너무 복잡하면, 데이터를 과적합 할 가능성이 높아진다.
이런 경우에는 트리 모델을 Pruning(가지치기) 해서 복잡도를 낮춰주도록 한다.
따라서 트리 모델은 어떻게 Partition 해서 Split 하는 것이 좋을까를 고민하는 모델이라고 정리할 수 있다.

1.2 Decision Tree - ID3 Algorithm

ID3 알고리즘을 통해서 '분류' 모델을 만들어보도록 한다.
이는 변별력이 좋은 질문을 위에서 하나하나 셋팅하는데, 이때는 Information Gain과 Entropy를 사용한다.
이 지표들은 특정 사건에 대한 혼잡도로 분류의 과정에서 T와 F가 많이 섞여 있다면, 이 지표의 값이 높다.

- T와 F가 많이 섞여 있지 않아서 확신도가 큰 경우: IG와 Entropy가 낮다. (확실한 경우)
- T와 F가 많이 섞여서 혼잡한 경우에 확신도가 작은 경우: IG와 Entropy가 높다. (불확실한 경우)

ID3 알고리즘은 각 Partition에서 IG와 Entropy가 낮은 것을 찾아서 나눠주도록 한다.

1.3 Decision Tree - CART Algorithm

CART 알고리즘에서는 IG나 Entropy 대신에 Gini-index를 사용한다.
이 지표 역시 많이 섞여있다면 gini-index가 높고, 섞여 있지 않다면 gini-index가 낮다.

2. Bias and Variance

Bias(편향)와 Variance(분산)는 단순함과 복잡함 사이에서의 균형을 이루는 것이라고 할 수 있다.
모델의 과소 및 과대적합 등을 확인할 수 있는 지표로 사용된다.

- 예측값과 정답이 멀리 떨어져 있는 경우에는 Bias가 높다고 이야기한다. = Underfitting
- 예측값이 대체로 서로 멀리 흩어져 있는 경우에는 Variance가 높다고 이야기한다. = Overfitting

일반적으로 이 둘 지표는 trade-off 관계에 있어서, 하나의 값이 올라가면 다른 값은 떨어지는 관계이다.
이를 위해 트리 모델에서는 pruning 등의 기법으로 regularization, boosting, bagging이 사용된다.

3. Bagging (Bootstrap Aggregating)

일반적인 트리 모델은 Variance가 높아서 Overfitting의 우려가 있는 한계가 있다.
따라서 주어진 데이터에 대해 여러 개의 Bootstrap 자료를 생성하고, 각각에 대해서 모델링하여 결합한다.
이를 통해 최종적인 예측 모형을 산출하는 방법을 Bagging이라고 한다.
여기서 Bootstrap 자료는 복원 임의추출로 원자료로부터 크기가 동일한 여러개 자료이다.
즉, 여러 번의 복원 Sampling을 통해 모델의 분산을 줄여서 예측력을 높이는 모델이라고 할 수 있다.
따라서 Overfitting 된 모델이나, 분산이 큰 모델에 사용하는 방식을 Bagging 이라고 할 수 있다.

- 다양한 종류의 데이터 `Sampling`을 통해 다양한 모델로 설계할 수 있다.
- 모델의 복잡함을 의미하는 `Variance`를 줄일 수 있다.
- 다양한 데이터를 통해 만든 모델을 결합하여 `Bias`를 줄일 수 있다.

일반적으로 Bagging 모델은 test data 없이, 트리를 만드는데 사용되지 않은 데이터로 성능을 평가한다.
이를 Out-Of-Bag (OOB) 관측치라고 하고, Random Forest가 Bagging의 한 방법이다.

4. Boosting

앞선 Bagging 방법은 Sampling을 동시에 여러 개를 만들어서 모델링 하는 방법이었다.
하지만 Boosting 방법은 트리들이 순차적으로 만들어지는 모델이라고 할 수 있다.
Low variance와 High bias 모델에 적합한 방식으로, 궁극적으로는 Bias를 감소시킨다.
Gradient Boosting이나 AdaBoost 등이 Boosting의 한 방법이라고 할 수 있다.

5. Random Forest

Random Forest는 Bagging 방법을 통해서 트리 기반의 앙상블 모델이라고 할 수 있다.
이는 각 트리들의 상관성을 제거하는 방법을 통해 더 나은 성능을 갖는다.
즉, 트리들의 상관성 제거를 위해 매번 분할을 수행할 때마다 설명변수들의 일부분만 고려한다.
이를 통해 최종적인 트리들의 평균은 변동성이 적어지고, 안정적이 된다.
이 방법은 과적합이 되지 않는 범위에서 트리의 크기를 키우는 방법이라고도 할 수 있다.
또한 최종 예측에 있어서 모든 트리들은 동등하게 중요도를 갖는 것이 특징이다.
그리고 Random Forest는 모든 트리가 독립적으로 만들어지는 특징이라고 할 수 있다.
Bagging과 Random Forest의 차이는 설명 변수의 부분집합 크기가 다르다는 것이다.

6. Estimated Error Rate, Variable Importance

모델을 생성할 때, Bootstrap 데이터에 포함되지 않은 데이터들을 OOB 라고 한다.
이는 Test data처럼 성능을 평가하는 등으로 사용된다.
또한 모델에 사용된 특정 변수를 permutation(섞어서) 기존 성능과 비교한다.
그래서 accuracy 등이 얼마나 차이가 나는지 확인하여, 모델에서 그 변수의 중요도를 확인한다.

7. Adaboost

Adaboost는 Random Forest와 다르게, 트리의 Stump(끝 부분)만 사용하는 방식으로 알려져 있다.
즉, 트리의 모양을 크게 키우는 것이 아니라, Simple한 단일 모델이라고 할 수 있다.
또한 트리들이 동등하게 중요도를 갖는 Random Forest와 다르게, 몇몇 Stump가 더 중요도를 갖는다.
마지막으로는 Adaboost는 트리들이 순차적으로 (Sequentially) 만들어지는 특징을 갖는다.

8. XGBoost

XGBoost는 Residual (잔차)에 대해서 tree를 만드는 모델이라고 할 수 있다.