<Chapter 4>

1. 분류의 개요

- 지도학습이란?

  명시적인 정답이 있는 데이터가 주어진 상태에서 학습하는 머신러닝 방식

  분류는 지도학습의 대표적인 유형으로, 나이브 베이즈/로지스틱 회귀/결정 트리/앙상블 등 다양한 머신러닝 알고리즘으로 구현할 수 있다.

- 앙상블

  서로 다른/또는 같은 알고리즘을 단순히 결합한 형태

  - 배깅(Bagging) 방식

    랜덤 포래스트: 뛰어난 예측 성능, 상대적으로 빠른 수행시간, 유연성 등으로 많은 분석가가 애용하는 알고리즘

  - 부스팅(Boosting) 방식

    그래디언트 부스팅: 뛰어난 예측 성능을 가졌지만 수행시간이 오래 걸리는 단점 존재

    XgBoost & LightBGM: 기존 그래디언트 부스팅의 예측 성능을 한단계 발전시키면서도 수행시간을 단축시키는 알고리즘

- 결정 트리

  앙상블의 기본 알고리즘으로 일반적으로 사용하는 것

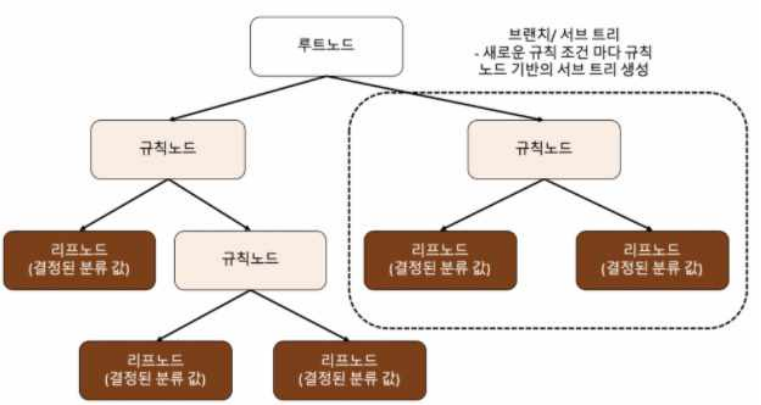
  데이터 스케일링이나 정규화 등 사전 가공의 영향이 매우 적어 쉽고 유연하게 적용될 수 있는 알고리즘

  예측 성능을 향상시키기 위해 복잡한 구조를 가져야 하며 이로 인해 과적합이 발생할 수 있음

  ➔ 이러한 단점은 앙상블에서 장점으로 작용: 앙상블은 매우 많은 여러 개의 약한 학습기를 결합해 확률적 보완과 오류가 발생한 부분에 대한 가중치를 계속 업데이트하면서 예측 성능을 향상시키는데, 결정 트리가 좋은 약한 학습기가 됨

2. 결정 트리(Decision Tree)

- 결정 트리 구조



데이터에 있는 규칙을 학습을 통해 자동으로 찾아내 트리기반의 분류 규칙을 만드는 것

규칙이 많을 수록, 트리의 깊이가 깊어질수록 과적합으로 이어질 가능성이 높음

가능한 적은 결정 노드로 높은 예측 정확도를 가지려면 데이터를 분류할 때 최대한 많은 데이터 세트가 해당 분류에 속할 수 있도록 결정 노드의 규칙이 정해져야 함

이를 위해서는 어떻게 트리를 분할할 것인가가 중요한데 최대한 균일한 데이터 세트를 구성할 수 있도록 분할하는 것이 필요함

- 균일도

데이터 세트 안에서 비슷한 데이터로 구성될수록 데이터 균일도가 높음

데이터 세트의 균일도는 데이터를 구분하는 데 필요한 정보의 양에 영향을 미침

결정 노드는 정보 균일도가 높은 데이터 세트를 먼저 선택할 수 있도록 규칙 조건을 만듦

정보의 균일도를 측정하는 대표적인 방법

- 정보 이득 계수: 주어진 데이터 집합의 혼잡도를 의미하는 엔트로피를 이용하 개념. 서로 다른 값이 섞여 있으면 엔트로피가 높고, 같은 값이 섞여 있으면 엔트로피가 낮음. 1- 엔트로피 지수 = 정보 이득 계수
- 지니 계수: 원래 경제학에서 불평등 지수를 나타낼 때 사용하는 개념. 머신러닝에 적용될 때는 지니 계수가 낮을수록 데이터 균일도가 높음

결정 트리 알고리즘을 사이킷런에서 구현한 DecisionTreeClassifier는 기본으로 지니 계수를 이용해 데이터 세트를 분할 함

- 결정 트리 특징

  장점: 쉽다. 직관적이다. 피처의 스케일링이나 정규화 등 사전 가공 영향도가 크지 않다.

  단점: 과적합으로 알고리즘 성능이 떨어진다. 이를 극복하기 위해 트리의 크기를 사전에 제한하는 튜닝이 필요하다.

- 결정 트리 파라미터

  - Min_samples_split

    노드를 분할하기 위한 최소한의 샘플 데이터 수로 과적합을 제어하는 데 사용됨

    디폴트는 2이고 작게 설정할수록 분할되는 노드가 많아져서 과적합 가능성 증가

  - Min_samples_leaf

    분할이 될 경우 왼쪽과 오른쪽의 브랜치 노드에서 가져야 할 최소한의 샘플 데이터 수

    큰 값으로 설정될수록, 분할될 경우 왼쪽과 오른쪽의 브랜치 노드에서 가져야 할 최소한의 샘플 데이터 수 조건을 만족시키기가 어려우므로 노드 분할을 상대적으로 덜 수행함

    Min_samples_split와 유사하게 과적합 제어 용도, 그러나 비대칭적 데이터의 경우 특정 클래스의 데이터가 극도로 작을 수 있으므로 이 경우는 작게 설정 필요

  - Max_features

    최적의 분할을 위해 고려할 최대 피처 개수

  - Max_depth

    트리의 최대 깊이를 규정

    깊이가 깊어지면 min_samples_split 설정대로 최대 분할하여 과적합 할 수 있으므로 적절한 값으로 제어 필요

- 결정 트리 과적합

  결정 트리 생성에 별다른 제약이 없도록 하이퍼 파라미터가 디폴트인 Classifier를 학습하면, 일부 이상치 데이터까지 분류하기 위해 분할이 자주 일어나서 결정 기준 경계가 매우 많아진다. 이렇게 복잡한 모델은 학습 데이터 세트의 특성과 약간만 다른 형태의 데이터

세트를 예측하면 예측 정확도가 떨어지게 된다. (과적합)

Min_samples_leaf=6을 설정해 6개 이하의 데이터는 리프 노드를 생성할 수 있도록 리프 노드 생성 규칙을 완화하면, 이상치에 크게 반응하지 않으면서 일반화된 분류 규칙에 따라 분류된다.

3. 앙상블 학습

- 앙상블 학습 개요

앙상블 학습의 목표는 다양한 분류기의 예측 결과를 결합함으로써 단일 분류기보다 신뢰성이 높은 예측 값을 얻는 것

- 앙상블 학습의 유형

  - 보팅(Voting)
  - 배깅(bagging)
  - 부스팅(boosting)

보팅과 배깅은 여러 개의 분류기가 투표를 통해 최종 예측 결과를 결정하는 방식

차이점: 보팅=서로 다른 알고리즘을 가진 분류기가 결합하는 것

배깅= 각각의 분류기가 모두 같은 유형의 알고리즘 기반이지만, 데이터 샘플링은 서로 다르게 가져가면서 학습을 수행해 보팅을 수행하는 것

- 보팅 유형

  - 하드 보팅

예측한 결괏값들 중 다수의 분류기가 결정한 예측값을 최종 보팅 결괏값으로 선정

  - 소프팅 보팅

분류기들의 레이블 값 결정 확률을 모두 더하고 이를 평균해서 이들 중 확률이 가장 높은 레이블 값을 최종 보팅 결괏값으로 선정

4. 랜덤 포레스트

- 랜덤 포레스트의 개요 및 실습

배깅은 보팅과 다르게 같은 알고리즘으로 여러 개의 분류기를 만들어서 보팅으로 최종 결정하는 알고리즘

배깅의 대표적인 알고리즘이 랜덤 포레스트

앙상블 알고리즘 중 비교적 빠른 수행 속도를 가지고 있으며, 다양한 영역에서 높은 예측 성능을 보임

여러 개의 결정 트리 분류기가 전체 데이터에서 배깅 방식으로 각자의 데이터를 샘플링해 개별적으로 학습을 수행한 뒤 최종적으로 모든 분류기가 보팅을 통해 예측 결정을 하게 됨

- 부트스트래핑(Bootstrapping)

랜덤 포레스트에서 개별 트리가 학습하는 데이터 세트는 전체 데이터에서 일부가 중첩되게 샘플링된 데이터 세트임

여러 개의 데이터 세트를 중첩되게 분리하는 것을 '부트스트래핑' 분할 방식이라고 함

- 랜덤 포레스트 하이퍼 파라미터 및 튜닝

트리 기반의 앙상블 알고리즘의 단점은 하이퍼 파라미터가 너무 많고, 그로 인해 튜닝을 위한 시간이 많이 소모된다는 것

그나마 랜덤 포레스트는 적은 편에 속함

- n_estimators

  랜덤 포레스트에서 결정 트리의 개수를 지정. 많이 설정할수록 좋은 성능을 기대할 수 있지만 계속 증가시킨다고 성능이 무조건 향상되는 것은 아니며 늘릴수록 학습 수행 시간이 오래 걸림

- max_features

  결정 트리에서 사용된 파라미터와 같음

- max_depth, min_samples_leat, min_samples_split

  결정 트리에서 과적합을 개선하기 위해 사용되는 파라미터 랜덤 포레스트에서도 똑같이 적용 가능