Chapter4. 분류

01.분류의 개요

지도학습: 명시적인 정답이 있는 데이터(레이블)이 주어진 상태에서 학습하는 머신러닝 방식

지도학습의 대표적인 유형: 분류

분류: 학습 데이터로 주어진 데이터의 피처와 레이블값을 머신러닝 알고리즘으로 학습해 모델을 생성,

생성한 모델에 새로운 데이터가 주어졌을때 미지의 레이블 값을 예측하는 것

—>결론:

분류: 기존 데이터가 어떤 레이블에 속하는지 패턴을 알고리즘으로 인지한 뒤, 새롭게 관측된 데이터에 대한 레이블 판별하는 것

다양한 알고리즘으로 구현 가능, 그 중에서 앙상블 방법 집중적으로 다룸

앙상블?

서로 같거나 다른 알고리즘을 단순히 결합한 형태

배깅

랜덤 포레스트

부스팅

그래디언트 부스팅

수행 시간이 너무 오래 걸림

기존 그래디언트 부스팅의 예측 성능을 발전시키면서 수행시간 단축시킨 알고리즘 등장

앙상블의 기본 알고리즘으로 일반적으로 사용하는건 결정 트리

02. 결정 트리

결정 트리: 데이터에 있는 규칙을 학습을 통해 자동으로 찾아내 트리 기반의 분류 규칙을 만드는 것

가장 쉽게 규칙을 표현: if else

어떤 기준을 바탕으로 규칙을 만들어야 가장 효율적일지가 성능에 크게 좌우

규칙노드: 규칙 조건

리프 노드: 결정된 클래스 값

새로운 규칙조건마다 서브트리가 생성됨

트리의 분할은 적고 정확도는 높이기 위해 데이터의 균일도가 중요

이 균일도를 측정하는 대표적인 방법으로 엔트로피를 이용한 정보 이득(information gain) 지수와 지니 계수가 있다.

사이킷런에서 구현돼 있는 DecisionTreeClassifier를 기본으로 지니 계수를 이용해 데이터 세트를 분할

결정 트리 모델의 특징

정보의 균일도만 신경 쓰면 되므로 특별한 경우를 제외하고는 각 피처의 스케일링과 정규화 같은 전처리 작업 필요 없

균일도를 기반으로 함으로 직관적이고 명확하다. 어떤 노드들이 만들어졌는지 알 수 있고, 시각화 표현 또한 가능

과적합으로 정확도가 떨어진다는 점. (서브 트리를 계속 만들다 보면, 피처가 많고, 균일도가 다양하게 존재할수록 깊이가 깊어져 복잡해지기 때문) 이 때문에 트리의 크기를 제한하는 등 튜닝이 필요

트리의 분할은 적고 정확도는 높이기 위해 데이터의 균일도가 중요.

이 균일도를 측정하는 대표적인 방법으로 엔트로피를 이용한 정보 이득(information gain) 지수와 지니 계수.

사이킷런에서 구현돼 있는 DecisionTreeClassifier를 기본으로 지니 계수를 이용해 데이터 세트를 분할.

결정 트리 모델의 특징

정보의 균일도만 신경 쓰면 되므로 특별한 경우를 제외하고는 각 피처의 스케일링과 정규화 같은 전처리 작업 필요 없.

균일도를 기반으로 함으로 직관적이고 명확하다. 어떤 노드들이 만들어졌는지 알 수 있고, 시각화 표현 또한 가능.

과적합으로 정확도가 떨어진다는 점. (서브 트리를 계속 만들다 보면, 피처가 많고, 균일도가 다양하게 존재할수록 깊이가 깊어져 복잡해지기 때문) 이 때문에 트리의 크기를 제한하는 등 튜닝이 필요.

03.앙상블 학습

앙상블 학습 : 여러 개의 분류기를 생성하고 그 예측을 결합함으로써 보다 정확한 최종 예측을 도출하는 기법

대표적인 앙상블 학습 유형

보팅(voting) : 여러개의 분류기를 결합해서 개별 예측을 해서 최종 예측 값을 내는 것이다.

배깅(bagging) : 보팅과 비슷하지만 배깅은 개별 데이터셋을 각각 분류기마다 부여해서 예측을 실행한다. (분할된 데이터셋은 중복이 허용된다.)

부스팅(boosting) : 분류기들이 순차적으로 학습을 수행하되, 분류기들의 학습이 끝나면 다음 학습으로 가중치를 넘겨주는 방식이다.

보팅 유형

하드 보팅 : 분류기 예측 값들을 종합해서 다수결로 레이블 예측 값을 결정한다.

소프트 보팅 : 분류기 예측 값들을 확률로 환산해서 평균을 구해 확률이 높은 예측 값으로 결정한다.

04.랜덤 포레스트

랜덤 포레스트 : 여러개의 결정 트리 분류기가 전체 데이터에서 각자의 데이터를 샘플링해 개별저긍로 학습을 수행, 모든 분류기가 보팅을 통해 예측 하는 방식이다.

부트스트래핑(Bootstraping) : 여러 개의 데이터 세트를 중첩되게 분리하는 기법, bagging 은 bootstrap aggragation의 줄임말이다.