분류

분류의 개요

- 분류 : 지도학습의 대표적 유형
  - 지도학습 : 명시적인 정답이 있는 데이터(레이블)가 주어진 상태에서 학습하는 머신러닝 방식
  - 분류 :
    - 학습 데이터로 주어진 데이터의 피처와 레이블값을 머신러닝 알고리즘으로 학습해 모델을 생성
    - 생성된 모델에 새로운 데이터 값이 주어졌을 때 미지의 레이블 값을 예측
- 다양한 종류가 존재
  - 나이브 베이즈
  - 로지스틱 회귀
  - 결정 트리
  - 서포드 벡터 머신
  - 최소 근접 알고리즘
  - 신경망
  - **앙상블**
- 앙상블 : **배깅**과 **부스팅** 방식으로 나뉨
  - 배깅 : 랜덤 포레스트
  - **부스팅** : 그래디언트 부스팅
- 앙상블의 기본 알고리즘으로는 **결정 트리**를 사용
  - 결정 트리
    - 장점
      - 매우 쉽고 유연하게 적용 가능, 사전 가공의 영향이 매우 적음
    - 단점
      - 예측 성능 향상 위해 복잡한 규칙 구조를 가져야 함 -> 과적합 발생 가능

결정 트리

- 결정트리란?
  - 데이터에 있는 규칙을 학습을 통해 자동으로 찾아내 트리 기반의 분류 규칙을 만드는 것
  - 어떤 기준을 바탕으로 규칙을 만드느냐가 알고리즘의 성능을 좌우
- 결정 트리의 구조
  - 규칙 노드 : 규칙 조건이 되는 것
  - 리프 노드 : 결정된 클래스 값
  - 서브 트리 : 새로운 규칙 조건이 생성될 때마다 생성
  - 규칙이 많다 == 트리가 깊어진다 == 분류를 결정하는 방식이 복잡해진다 == 과적합으로 이어지기 쉽다 == **예측 성능이 저하될 가능성이 높다**
  - 최대한 많은 데이터 세트 노드가 분류에 속할 수 있도록 규칙을 정하는 것이 필요 -> **트리 분할이 중요**
- 균일도 : 해당 데이터 세트에서 무작위로 무언가를 뽑았을 때 예측하기 쉬울수록 균일도가 높음
- 결정 노드 : 균일도가 높은 데이터 세트를 먼저 선택할 수 있도록 규칙 조건을 만듬
- 균일도 측정법
  - 정보 이득 : 엔트로피를 이용

    - 지니 계수 : 낮을 수록 균일도가 높음
- DecisionTreeClassifier : 지니계수 이용

결정 트리 모델의 특징

장점

- 알고리즘이 쉽고 직관적
- 전처리 작업이 필요 없음. (균일도만 신경쓰면 되므로)

단점

- 과적합으로 성능이 떨어질 수 있음 -> 트리의 크기를 제한하는 튜닝이 필요

결정 트리 파라미터

- min_samples_split : 노드를 분할하기 위한 최소한의 샘플 데이터 수, 과적합 제어
- min_samples_leaf : 말단 노드가 되기 위한 최소한의 샘플 데이터 수, 과적합 제어
- max_features : 최적의 분할을 위하 고려할 최대 피처 개수
- max_depth : 트리의 최대 깊이
- max_leaf_nodes : 말단 노드의 최대 개수

결정 트리 모델 직관적으로 이해하기

- 시각화 : Graphviz 패키지 사용
- feature_importances_ : 어떤 피처가 규칙을 정하는 데 중요한 영향을 끼쳤는지 ndarray 형태로 중요도 반환

앙상블 학습

앙상블 학습 개요

- 앙상블 학습 : 여러 개의 분류기를 생성, 그 예측을 결합 -> 정확한 최종 예측을 도출
- 정형 데이터 분류 시 뛰어난 성능 나타냄
    - XGBoost, LightGBM, 스태킹
- 보팅, 배깅, 부스팅으로 나눌 수 있음

보팅

- 서로 다른 알고리즘을 가진 분류기를 결합

배깅

- 같은 알고리즘 기반의 분류기를 사용하지만 데이터 샘플링을 서로 다르게 가져감
- ex) 랜덤 포레스트

부스팅

- 여러 개의 분류기가 순차적으로 학습을 수행하되, 앞선 분류기가 예측이 틀린 데이터에 대해서는 올바르게 예측할 수 있도록 가중치를 부여

- ex) 그래디언트 부스트, XGBoost, LightGBM

하드 보팅과 소프트 보팅

- **하드 보팅** : 다수의 분류기가 결정한 예측값을 최종 보팅 결괏값으로 선정
- **소프트 부팅** : 레이블 값 결정 확률을 모두 더하고 이를 평균해서 확률이 가장 높은 레이블 값을 최종 보팅 결과값으로 선정
- 주로 **소프트 보팅** 사용

랜덤 포레스트

랜덤 포레스트의 개요 및 실습

- **배깅** : 같은 알고리즘으로 여러 개의 분류기를 만들어서 보팅으로 최종 결정하는 알고리즘
- **랜덤 포레스트**
    - 비교적 빠른 수행속도를 가짐
    - 다양한 영역에서 높은 예측 성능을 보임
    - 기반 알고리즘 : 결정 트리
    - 트리가 학습하는 데이터 세트 : 전체에서 일부가 중첩되게 샘플링된 데이터 세트 -> 부트스트래핑

랜덤 포레스트 하이퍼 파라미터 및 튜닝

- 트리 기반 앙상블 알고리즘의 단점 :
    - 하이퍼 파라미터가 너무 많음 -> 시간을 많이 소모
- 랜덤 포레스트 : 적은 편