<4장 분류>

분류의 개요

나이브 베이즈, 로지스틱 회귀, 결정 트리, 서포트 벡터 머신, 최소 근접 알고리즘, 신경망, 앙상블

결정트리

결정트리: 데이터에 있는 규칙을 학습을 통해 자동으로 찾아내 트리 기반의 분류 규칙을 만드는 것
정보 균일도가 높은 데이터 세트를 먼저 선택할 수 있도록 규칙 조건을 만든다
정보의 균일도를 측정하는 대표적인 방법은 엔트로피를 이용한 정보 이득지수와 지니 계수가 있다
정보이득: 1 - 엔트로피 지수(데이터 집합의 혼잡도 의미)
지니계수: 불평등 지수를 나타낼 때 사용하는 계수, 1로 갈수록 불평등

- 결정 트리 모델의 특징

장점: 쉽다, 직관적이다, 사전 가공 영향도가 크지 않다.
단점: 과적합으로 알고리즘 성능이 떨어진다.

- 결정 트리 모델의 시각화

max_depth: 결정 트리의 최대 트리 깊이를 제어
min_samples_split: 자식 규칙 노드를 분할해 만들기 위한 최소한의 샘플 데이터 개수
min_samples_leaf: 리프 노드가 될 수 있는 샘플 데이터 건수의 최솟값을 지정
feature_importances: 피처의 중요도 추출

앙상블 학습

앙상블 학습: 여러 개의 분류기를 생성하고 그 예측을 결합함으로써 보다 정확한 최종 예측을 도출하는 기법
정형데이터 분류시에 뛰어난 성능을 보임

앙상블 학습의 유형: 보팅, 배깅, 부스팅
보팅과 배깅은 여러 개의 분류기가 투표를 통해 최종 예측 결과를 결정하는 방식으로 보팅은 서로 다른 알고리즘을 가진 분류기를 결합하는 것이고 배깅은 각각의 분류기가 모두 같은 유형의 알고리즘 기반이지만 데이터 샘플링을 다르게 가져가면서 학습을 수행한다
부스팅은 여러개의 분류기가 순차적으로 학습을 수행하되 앞에서 학습한 분류기가 예측이 틀린 데이터에 대해서는 올바르게 예측할 수 있도록 다음 분류기에게는 가중치를 부여하면서 학습과 예측을 진행한다

- 보팅 유형 - 하드 보팅과 소프트 보팅

하드보팅: 다수결 원칙과 비슷, 예측 결괏값들 중 다수의 분류기가 결정한 예측 값을 최종 보팅 결괏값으로 선정
소프트 보팅: 분류기들의 레이블 값 결정 확률을 모두 더하고 이를 평균해서 확률이 가장 높은 레이블 값을 최종 보팅 결괏값으로 선정(일반적인 보팅 방법)

랜덤 포레스트

- 랜덤 포레스트의 개요 및 실습

배깅의 대표적인 알고리즘은 랜덤 포레스트이다.
랜덤포레스트는 앙상블 알고리즘 중 비교적 빠른 수행 속도를 가지고 있으며 높은 예측 성능을 보인다.
랜덤포레스트의 기반 알고리즘은 결정 트리이다.
랜덤 포레스트는 여러개의 결정 트리 분류기가 전체 데이터에서 배깅 방식으로 각자의 데이터를 샘플링해 개별적으로 학습을 수행하고 최종적으로 모든 분류기가 보팅을 통해 예측 결정을 한다.

부트스트래핑: 여러개의 데이터 세트를 중첩되게 분리하는 것
랜덤 포레스트의 서브세트 데이터는 부트스트래핑으로 데이터가 만들어짐
데이터가 중첩된 개별 데이터 세트에 결정 트리 분류기를 각각 적용하는 것이 랜덤 포레스트이다

- 랜덤 포레스트 하이퍼 파라미터 및 튜닝

n_estimators: 결정 트리의 개수 지정(디폴트 10개)
max_features: 결정 트리에 사용된 max_features와 같다
max_dept, min_samples_leaf와 같이 결정 트리와 똑같이 적용