📌

EURON_2주차 STUDY

chapter 4 분류

01 분류의 개요

분류 : 학습 데이터로 주어진 데이터의 피처와 레이블값을 머신러닝 알고리즘으로 학습해 모델을 생성하고, 이렇게 생성된 모델에 새로운 데이터 값이 주어졌을 때 미지의 레이블 값을 예측하는 것

- 베이즈 통계와 생성 모델에 기반한 나이브 베이즈
- 독립변수와 종속변수의 선형 관계성에 기반한 로지스틱 회귀
- 데이터 균일도에 따른 규칙 기반의 결정 트리
- 개별 클래스 간의 최대 분류 마진을 효과적으로 찾아주는 서포트 벡터 머신
- 근접 거리를 기준으로 하는 최소 근접 알고리즘
- 심층 연결 기반의 신경망
- 서로 다른 머신러닝 알고리즘을 결합한 앙상블

02 결정트리

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

리프 노드 : 결정된 클래스 값

그러나 많은 규칙이 있으면 과적합으로 이어질수있음

정보의 균일도를 측정하는 대표적인 방법은 엔트로피를 이용한 정보 이득 지수와 지니 계수가 있다.

결정 트리 모델의 특징

장점 : 정보의 균일도 라는 룰을 기반으로 하고 있어서 알고리즘이 쉽고 직관적이다. 시각화 표현 가능. 각 피처의 스케일링과 정규화 같은 전처리 작업이 필요 없다.

단점 : 과적합으로 정확도가 떨어짐

트리의 크기를 사전에 제한하는 것이 오히려 성능 튜닝에 도움이 됨

결정 트리 파라미터

min_samples_split

min_samples_leaf

max_features

max_depth

max_leaf_nodes

결정 트리 모델의 시각화

Graphviz 패키지 사용

결정 트리 과적합

학습 데이터에만 지나치게 최적화된 분류 기준은 오히려 테스트 데이터 세트에서 정확도를 떨어뜨릴 수 있다

03 앙상블 학습

앙상블 학습 개요

앙상블 학습을 통한 분류는 여러 개의 분류기를 생성하고 그 예측을 결합함으로써 보다 정확한 최종 예측을 도출하는 기법

앙상블 알고리즘의 대표적인 랜덤포레스트와 그래디언트 부스팅 알고리즘은 뛰어난 성능과 쉬운 사용으로 인기가 많다.

전통적으로 보팅, 배깅, 부스팅의 세 가지로 나눌 수 있다.

보팅 유형 - 하드 보팅과 소프트 보팅

하드 보팅 : 예측한 결과값들 중 다수의 분류기가 결정한 예측값을 최종 보팅 결괏값으로 선정하는 것.

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

보팅 분류기

VotingClassifier

04 랜덤 포레스트

랜덤 포레스트이 개요 및 실습

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

여러 개의 데이터 세트를 중첩되게 분리하는 것을 부트스트래핑 분할 방식이라고 함

랜덤 포레스트 하이퍼 파라미터 및 튜닝

n_estimators

max_features

max depth

min_samples_leaf

GridSearchCV를 이용하면 랜덤 포레스트이 하이퍼 파라미터 튜닝을 할 수 있다.