1. 머신러닝 알고리즘

-KNN Imputer

-정의

지도학습의 한 종류로 거리기반 분석모델이다. (Y값이 존재하므로 비지도 학습과는 차이를 보임)

유사속성에 따라 분류하여 라벨링하는 알고리즘이다.

-장점

알고리즘이 간단하여 구현하기가 쉬움

훈련단계가 빠름

-단점

다른 머신러닝 알고리즘과 달리 모델을 생성하지 않아 특징와 클래스간 관계를 이해하는데 제한적이며 데이터가 많아질수록 분류속도가 느려지고 계산량이 많아진다.

또한 적절한 k의 수 선택이 필요하다.

-주의 점

함수의 표준화 정규화 작업을 진행해야함

-K-means

-정의

비지도 학습에 속한다.

k개의 군집으로 묶는 알고리즘 이다. means는 클러스터의 중심과 데이터들의 평균거리를 의미한다. 가깝게 위치하는 데이터를 비슷한 특성을 지닌 데이터로 여기고 군집화 한다.

k-nn알고리즘과는 비지도, 지도 학습에서 차이를 보인다. k-nn은 분류 알고리즘, k-means는 군집화 알고리즘 이다.

알고리즘 원리 순서로는

군집의 개수 설정-초기 중심점 설정-데이터를 군집에 할당-중심섬 재설정-데이터를 군집에 재할당 이다.

-단점

k의 값에 따라 클러스터링 결과가 극명히 바뀐다.

-Random Forest

-정의

모 데이터에서 n개 샘플 데이터를 중복을 허용해 무작위로 추출해서 여러개의 의사 결정나무 학습기에서 동시에 학습이 이루어진다. 전체 특성의 제곱근 수 만큼 특성을 무작위로 골라 계산한다.

-장점

상관관계가 없는 트리의 평균을 내어 전체 분산과 예측 오류를 낮추므로 과대 적합의 리스크

가 감소한다.

-단점

대규모의 데이터 세트를 처리하므로 시간이 많이 소요 된다.

대규모 데이터 세트를 처리하므로 데이터를 저장하기 위해 더 많은 리소스가 필요하다.

-XGBoost

-정의

지도학습 알고리즘

이전모형의 약점을 보완하는 방식으로 학습한다.

-장점

과적합 방지가 잘 되어있다.

예측성능이 좋다.

-단점

작은 데이터에 대하여 과적합 가능성이 있다.

해석이 어렵다.

-Grid Search CV

-정의

머신러닝 모델의 최적의 파라미터를 검색할 수 있는 클래스이다.

머신러닝에서 모델의 성능향상을 위해 쓰이는 기법중 하나이다.

사용자가 직접 모델의 하이퍼 파라미터 값을 리스트로 입력하면 경우의 수마다 예측성능을 측정, 평가한다.

-단점

시간이 오래걸린다.

+ 프로젝트에 어떻게 사용되는지