[7주차 과제1] DT의 하이퍼파라미터 변경에 따른 성능결과 분석

태양광 발전량 예측을 통한 분석

18011480 이진형 I 컴퓨터공학과

문제 분석

문제에서 주어진 데이터 셋(train.csv, test.csv)는 날짜, 시간, 날씨 정보로 구성되어 있는 기상 데이터셋이다. 데이터 셋을 기반으로 태양광 발전량을 판단하는 것이 목적이다. 당 데이터를 이용, 결정 트리 학습법과 랜덤 포레스트를 통해 모델 학습을 진행했다.

실험 조건

랜덤 포레스트는 각 트리는 데이터의 일부에 오버피팅하는 경향을 가진다는 데 기초한다. 랜덤 포레스트 모델에서의 주요 파라미터는 생성할 트리의 개수인 n_estimators와 선택할 최대 특성의 수인 max_features이다. 따라서 이 두가지 파라미터의 변화를 통해 실험을 진행했다.

실험

진행한 실험의 조건과 결과는 다음과 같다.

n_estimators	max_features	Score
1000	auto	0.99846
1000	sqrt	0.99244
1000	log2	0.99588
100	auto	0.99832
100	sqrt	0.99188
100	log2	0.99572
10	auto	0.99782
10	sqrt	0.98761
10	log2	0.99430
DecisionTree(with base parameters)		0.99611
baseline(randomforest)		0.99833

실험 결과

위와 같은 조건으로 실험을 진행한 결과, 최고 점수인 0.99846을 기록한 파라미터는 다음과 같았다.
1000/auto
이번 실험에서는 실험 진행 전에 예상했던 바와 같이 트리를 더욱 많이 생성할 수록, 선택하는 특성이 많은수록 높은 점수가 나온 것을 알 수 있었다.