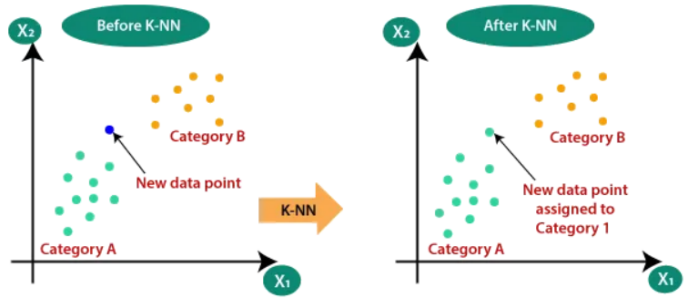
머신러닝 알고리즘

-KNN Imputer



출처: MEDIUM.COM

거리기반 분류분석 모델, 머신러닝에서 데이터를 가장 가까운 유사 속성에 따라 분류하여 데이터를 분류하는 기법

K값 정하는 법 : 작은 K값은 데이터 포인트와 가까운 이웃들의 특징 잘 반영 but, 노이즈 포함 쉬움. 큰 K값은 노이즈의 영향을 줄이지만 데이터 포인트 주변의 세부적인 특성 파악 어려움. → 교차 검증 등의 방법 사용, 데이터셋이 노이즈가 많은 경우 작은 K값 선택, 데이터셋이 복잡한 패턴 or 클래스 간 경계가 모호한 경우 큰 K값 선택

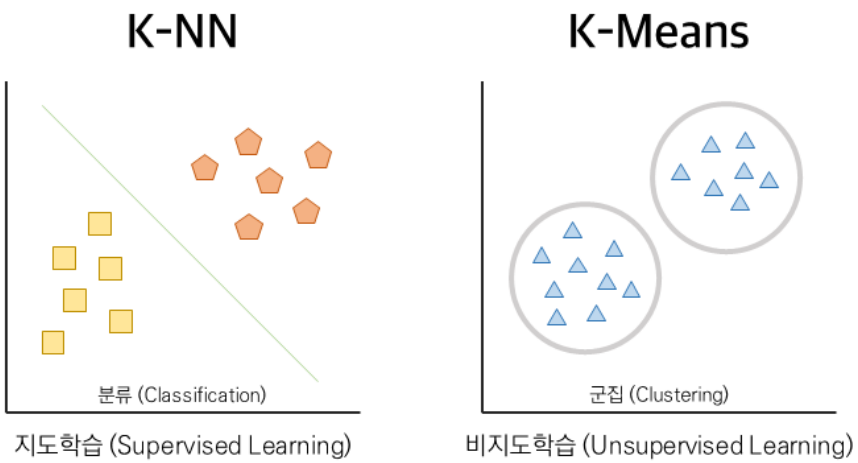
장점

1) 간단하고 이해하기 쉬움
2) 학습 데이터 분포 고려 X
3) 높은 분류 정확도
4) 적은 데이터셋에서도 잘 작동
5) 다목적 사용 가능

단점

1) 계산 복잡도 높음
2) 이상치에 민감
3) 클래스 불균형

-Kmeans



K-NN : 미리 레이브링 붙어 있는 데이터들을 학습 후 새로운 데이터에 대해 분류

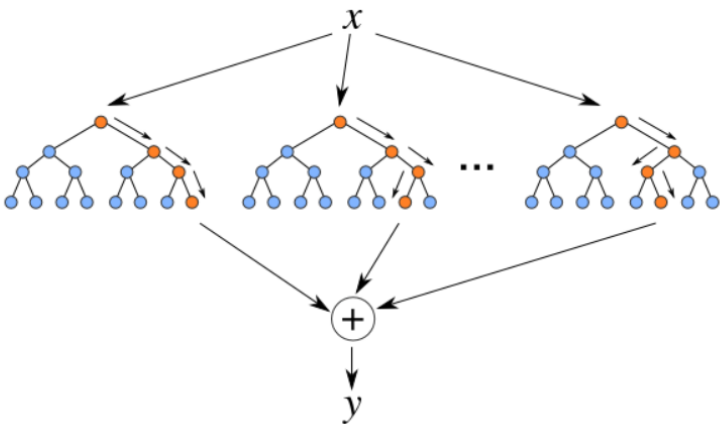
K-Means : 레이블을 모르더라도 비슷한 특징을 가진 데이터끼리 묶어주는 군집 수행

K-means 프로세스

1) 데이터셋에 K개의 중짐을 임의로 지정
2) 각 데이터들을 가장 가까운 중심이 속한 그룹에 할당
3) 분산을 계산하고 각 클러스트의 새 중심 배치
4) 중심이 더 이상 변하지 않을 때까지 2번 3번 반복
5)

-Random Forest

※무작위 숲의 이해를 돕기 위한 그림



Decision Tree(의사결정나무)의 Forest(숲)

하나의 결과에 도달하기 위해 여러 의사결정 트리의 출력을 결합

분류와 회귀 문제를 모두 다루며 사용 편의성과 유연성 뛰어남

배깅 방법의 확장, 배깅과 특성 무작위성을 모두 활용하여 상관관계가 없는 의사결정 트리의 포레스트를 만드는 것

-XGBoost

기본 학습기를 의사결정나무로 하며 Gradient Tree Boosting과 같이 Gradient(잔차)를 이용하여 이전 모형의 약점을 보완하는 방식으로 학습

Gradient Tree Boosting에 과적합 방지를 위한 파라미터(λ,γ)가 추가된 알고리즘

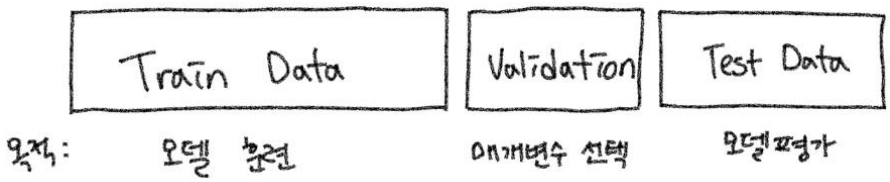
https://zephyrus1111.tistory.com/232

-Grid Search CV

관심있는 매개변수들을 대상으로 가능한 모든 조합을 시도해 보며 최적의 하이퍼 파라미터 튜닝을 하는 것

알고리즘에 사용되는 하이퍼 파라미터를 순차적으로 입력하면서 편리하게 최적의 파라미터를 도찰할 수 있는 방안을 제공해 주는 모듈

여러 가지 매개변수 값으로 많이 시도해 보고 정확도가 가장 높은 조합 선택해야 하는데 만약 데이터가 train과 test 데이터로만 나눴더라면 테스트 세트를 이미 사용했기 때문에 모델이 얼마나 좋은지 평가하는데 사용 X → rain, validation, test 셋으로 나눠, '모델 훈련', '매개변수 선택', '최종 모델 평가'를 각각 다른 데이터로 진행하면 더 정확한 모델 만들기 가능



교차 검증을 같이 하면서 그리드 서치를 한 번에 해주는 모델 = Grid Search CV