Model-based Learning
트레이닝 데이터가 주어졌을 때 분류 및 예측 모델을 구축을 한 이후 새로운 모델이 오면 분류 및 예측 모델을 이용하여 분류 및 예측 진행

Instance-based Learning
모델 존재 X
모델을 만들지 않고, 새로운 데이터가 왔을 때 트레이팅 데이터의 패턴을 보고 이 데이터가 패턴 상 어떤 데이터에 가장 가까운지 분류하고 예측함

1-nearest neighbor
새로운 데이터에 가장 가까운 하나의 이웃을 정의함
새로운 데이터와 다른 데이터들의 거리를 모두 구함
새로운 데이터의 class를 구함

KNN 알고리즘의 구분 및 특징
1. instance based learning에 속함
- 각각의 관측치(instance) 만을 이용하여 새로운 데이터에 대한 예측을 진행
2. Memory-based Learning
- 모든 학습 데이터를 메모리에 저장한 후, 이를 바탕으로 예측 시도
3. Lazy Learning
- 모델을 별도로 학습하지 않는 ㅌ스팅 데이터가 들어와야 비로서 작동하는 게으른 알고리즘

KNN은 비선형 모델
- 결정 경계선이 비션형으로 나타남

KNN 분류모델
k = 가장 가까운 이웃의 개수
k 의 개수를 확인 후 가까운 이웃에서 다수의 패턴으로 따라감

KNN 분류 알고리즘 요약
1. 분류할 관측치 x를 선택
2. x로부터 인접한 k개의 학습 데이터를 탐색
3. 탐색된 k개 학습 데ㅣ터의 majority class c를 정의
4. c를 x의 분류결과로 반환

KNN 예측모델
k = 가장 가까운 이웃의 수
k에 가장 가까운 이웃의 좌표값을 이용해서 새로운 관측지의 좌표를 예측함

KNN 예측 알고리즘 요약
1. 예측할 관측지 x를 선택

2. x로부터 인접한 k개의 학습데이터를 탐색
3. 탐색된 k개 학습 데이터의 평균을 x의 예측 값으로 반환

KNN 하이퍼파라미터
1. k
- 인접한 학습 데이터를 몇 개까지 탐색할 것인가?
2. Distance Measures
- 데이터간 거리는 어떻게 측정할 것인가?

k의 범위는 1개부터 전체 데이터 수

k가 매우 작을 경우 : 데이터의 지역적 특성을 지나치게 반영함(overfitting)
k가 매우 클 경우 : 다른 범주의 개체를 너무 많이 퐘하여 오분류할 위험(underfitting)

k 선택 방법
- 일정 범위 내로 k를 조정하여 (1~k^*), 가장 좋은 예측 결과를 보이는 k값을 선정함

분류 모델

$$MisclassDrror_k = \frac{1}{k}\sum_{i=1}^{k} I(c_i \neq \hat{c_i}) \text{ for } k = 1,2,\ldots,k^*$$

$I(\bullet)$: Indicator Function -> 0 : 거짓, 1 : 참
c_i : 실제 데이터, $\hat{c_i}$: 예측된 class
같으면 좋음, 다르면 다를수록 안좋음

예측모델

$$SSE_k = \sum_{i=1}^{k} (y_i - \hat{y_i})^2 \text{ for } k = 1,2,\ldots,k^*$$

트레이팅, 테스트 에러가 모두 가장 작아지는 k지점을 찾아야함

거리측도 (1-유사도)
데이터 내 변수들이 각기 다른 데이터 범위, 분산 등을 가질 수 있으므로, 데이터 정규화(혹은 표준화)를 통해 이를 맞추는 것이 중요함
ex) 거리를 계산할 때 단위가 큰 특정 변수(들)가 거리를 결정하는 것 방지

대표적인 거리측도
Euclidean Distance
Manhattan Distance
Mahalanobis Distance

Correlation Distance

KNN의 장점과 한계점

장점
- 데이터 내 노이즈에 영향을 크게 받지 않으며, 특히 MAhalanobis distance와 같이 데이터의 분산을 고려할 경우 강건함
- 학습 데이터의 수가 많을 경우 효과적임

한계점
- 파라미터 k의 값을 설정해야 함
- 어떤 거리 척도가 분석에 적함한 지 불분명하며, 따라서 데이터의 특성에 맞는 거리측도를 임의로 선정해야 함
- 새로운 관측치와 각각의 학습 데이터 간 거리를 전부 측정해야 하므로, 계산시간이 오래 걸리는 단점이 있음

Weighted KNN
3NN 예측모델
거리에 따라 가중치를 다르게 둠

KNN요약
- KNN은 매우 단순한 접근방식으로 새로운 관측치를 분류 및 예측할 수 있는 방법임
- 선형 모델과 같이 학습 데이터로부터 특정 형태의 모델을 제시하는 것이 아니라, 학습 데이터 내 유사한 관측치들 만을 토대로 새로운 데이터의 예측을 수행함
- 일부 유사한 관측치의 반응변ㅅ의 조합을 통하여 예상된느 반응변수 값을 제공함
- Weighted KNN알고리즘으로 데이터의 가중치를 고려할 수 있으며, 이를 통해 보다 정확한 모델을 구축할 수 있음