K-Nearest-Neighbors Method(KNN)

➢ 통계적 학습 모델의 목적

: 데이터를 통해 값 예측 / 분류 / 데이터 구조 파악

➢ 통계적 학습 모델의 종류

- ✓ 지도학습(Supervised learning) : 예측과 분류를 진행하는 반응값(response)이 존재
  - ❖ 회귀(Regression) : 반응값이 연속형인 경우
  - ❖ 분류(Classification) : 반응값이 범주형인 경우
  - ❖ Parametric model : 반응값과 독립변수 사이에 특별한 모양을 가정 ex) 선형 회귀 모델
  - ❖ Non-Parametric model : 반응값과 독립변수 사이에 특별한 가정 x
- ✓ 비지도학습(Unsupervised learning) : 반응값이 없고 독립변수만 존재하는 학습 모델, 데이터 구조 파악이 주 목적

➢ HyperParameter

: 통계적 학습 모델에서 사용자가 직접 세팅해야 하는 값, 모델의 성능이 결정 되는 부분이지만

최적의 솔루션이 없기 때문에 사용자가 적절히 잘 선택해야 한다.

➢ HyperParameter에 따른 통계적 모델링의 변화

- ✓ Hyperparameter를 통해 모델의 복잡성을 조정할 수 있다.
- ✓ 모델의 복잡성이 높아지는 Hyperparameter를 사용하는 경우 모델의 편향(Bias)는 줄어들고 분산(Variance)는 증가
- ✓ 모델의 복잡성이 낮아지는 경우 모델의 편향(Bias)는 늘어나고 분산(Variance)는 증가

➢ 모델의 편향(Bias)과 분산(Variance)

- ✓ 모델의 편향(Bias)

  : 데이터의 모집단의 실제 모습과 model에서 예측하는 모습의 차이

- ✓ 모델의 분산(Variance)

  : 모집단에서 여러 표본을 뽑아 각 표본별로 model을 fitting했을 때 model간의 변동 정도

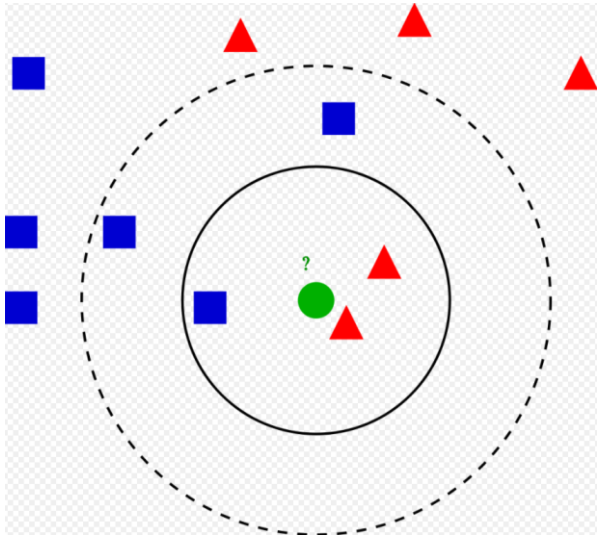
➢ K-Nearest Neighbor Method

: 지도학습, 분류, 회귀, non-parametric model

✓ HyperParameter : K(거리를 측정할 이웃 데이터의 개수)

✓ HyperParameter 이외에 이웃 데이터와의 거리 측정 방법에 따라 모델의 결과가 달라짐

✓ $\hat{Y}(X) = \frac{1}{k}\sum_{X_i \in \mathbb{N}_k(y)} y_i$



➢ 거리 측정 방법

- ✓ **유클리디안 거리(Euclidean Distance)**

  : $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

- ✓ **맨하탄 거리(Manhattan Distance)**

  : $\sum_{i=1}^{n}|x_i - y_i|$

- ✓ **마할라노비스 거리(Mahalanobis Distance)**

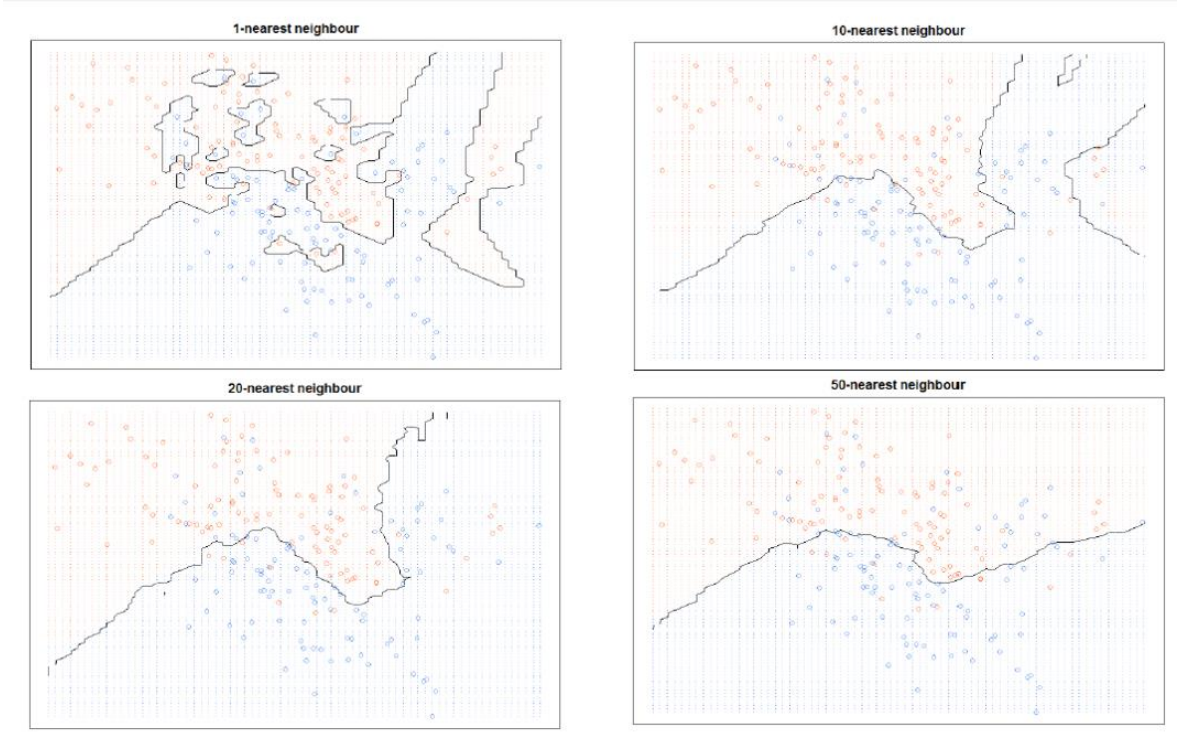
  : $\sqrt{\left(\vec{X} - \vec{Y}\right)^T \Sigma^{-1}\left(\vec{X} - \vec{Y}\right)}$

➢ 선택 기준 설정

: 사전 확률을 고려 할 것 , 기본적으로 KNN에서 결정 방식은 다수결(인접한 이웃 중 많은 범주)이지만
애초에 분류가 불균형한 데이터인 경우 이를 고려해서 선택 기준을 다양하게 설정해야 한다.

➢ K에 따른 모델성능의 변화

- ✓ K가 작을수록 불규칙한 경계 발생, 모델 편향 감소, 모델 분산 증가
- ✓ K가 클수록 경계의 불규칙성 감소, 모델 편향 증가, 모델 분산 감소
- ✓ K = 1일때 Training set에서 100%의 정확도를 보임
- ✓ K = N(데이터 개수)일때 표본에서 가장 수가 많은 군집으로만 분류



➢ 장점

- ✓ 학습 데이터의 오차(noise)에 크게 영향을 받지 않음
- ✓ 학습 데이터의 수가 많을 수록 정확도가 비교적 높은 효과적인 알고리즘
- ✓ 다양한 거리 공식을 활용하여 다양한 효과를 얻을 수 있음 ex) 마할라노비스 거리의 robust

➢ 단점

- ✓ HyperParameter인 K에 대한 최적값 선택이 불분명
- ✓ 선택 기준 결정 방식 또한 불분명(데이터의 특성에 맞게 사용자가 적당히..)
- ✓ 계산 시간이 오래 걸림(인접한 이웃을 찾기 위해 새로운 데이터와 기존의 모든 데이터의 거리 계산)
- ✓ 다양한 거리 공식을 활용하여 다양한 효과를 얻을 수 있음 ex) 마할라노비스 거리의 robust