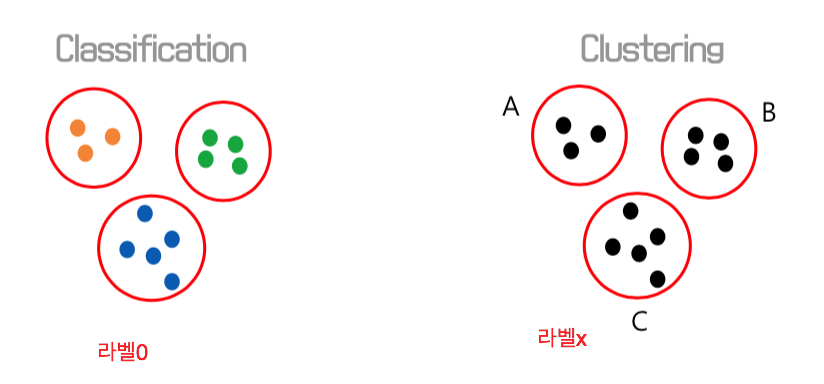
Classification / Regression / Clustering의 차이점

1. Classification (분류)
2. 지도 학습의 일종으로 기존에 존재하는 데이터의 Category 관계를 파악하고 새롭게 관측된 데이터의 Category를 스스로 판별하는 과정
3. 다중 분류는 비지도학습의 Clustering과 비슷하지만, 가장 큰 차이점은 Category의 도메인이 정의되있는가 그렇지 않은가
4. 지도 학습의 Classification은 이미 정해진 레이블 안에서 학습하고 분류한다.
5. 비지도 학습의 Clustering은 레이블을 원하는 만큼 생성하고, 분류한다.
6. 
7. 표로 정리

|  |  |  |
| --- | --- | --- |
|  | Classification (분류) | Clustering (군집화) |
| 소속집단에 대한 정보 | O | X |
| Label | O | X |
| 종류 | Supervised Learning | Unsupervised Learning |
| 공통점 | 데이터를 비슷한 집단으로 묶는 방법 | |

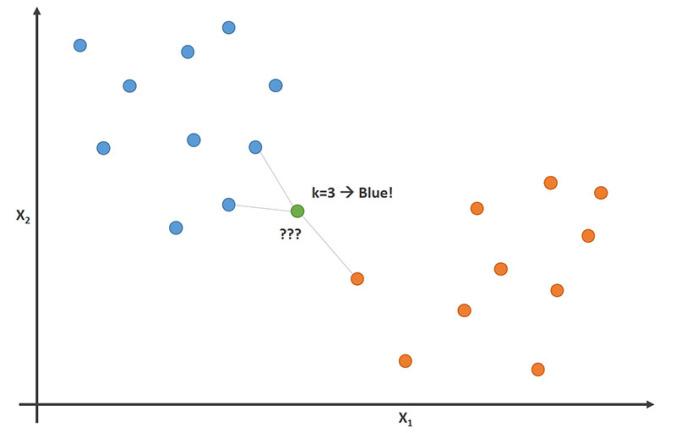
1. Classification의 종류
   1. KNN
   2. Decision Tree는 간략하게
   3. Random Forest
   4. Naive Bayes
   5. SVM
2. Quantitative한 데이터의 경우 Regression을 이용해 Fitting 하여 값을 추정한다
3. Qualitative한 데이터의 경우 Classification을 이용해 Class로 할당한다.
4. Regression은 보통 연속적인 숫자, 예측값이 float형태인 문제들을 해결하는 데 사용한다. 출력에 연속성이 있다
5. Regression은 현재 가지고 있는 data 특성을 통해서 어떠한 function을 얻었을 때, 미래의 data 또한 예측이 가능함.

K-NN

입력값과 K개의 가까운 점들이 있다고 생각하고 그 점들이 어떤 라벨과 가장 비슷한지를 판단하는 것

비슷하다는 것의 오차(거리)로 판단한다.

새로운 데이터 포인트의 카테고리를 결정할 때 K 개의 가장 가까운 포인트를 선점하고 그 중 가장 많이 선택된 포인트의 카테고리로 새로운 데이터를 분류하는 방법



k는 3일 때 가장 가까운 3개의 데이터를 본 뒤에 3개의 주변 데이터가 더 많이 포함되어 있는 파란색 범주로 분류하게 된다.

KNN은 훈련이 할 필요가 없다. Real -time 예측이 이루어 진다.

모델을 별로 구축하지 않아도 된다는 의미로 게으른 모델 (Lazy model)이라고도 부른다.

그래서 정확도는 SVM이나 선형회귀보다는 빠르다.

거리를 구해야 하는데 거리를 구하는 방법으로 여러가지 있다.

Euclidean Distance: 두 관측치 사이의 직선 최단거리를 의미한다.

Manhattan Distance:

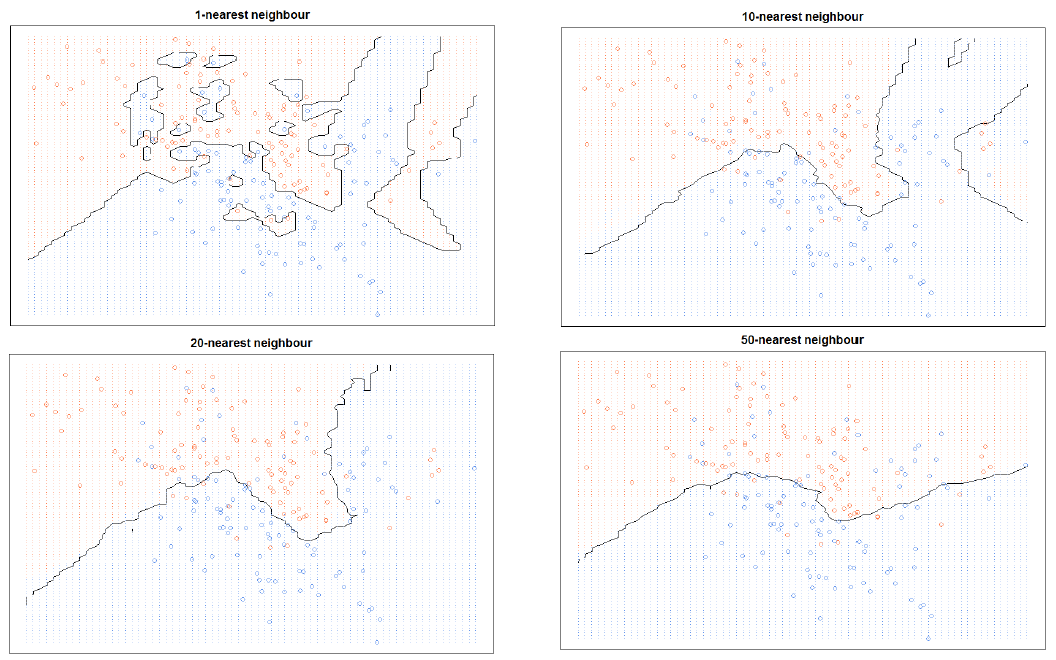
Mahalanobis Distance

변수 내 분산, 변수 간 공분산을 모두 반영하여 거리를 계산한다.

Correlation Distance

Rank Correlation Distance

그래서 이 모델의 하이퍼파라미터는 두가지, k의 수와 거리를 구하는 방법이다.



K의 수가 증가할 때마다 분류 경계면이 단순화되는 것을 알 수 있다

KNN 알고리즘은 새로운 데이터가 주어줬을 때 어느 정도 범주로 분류되는지 보기 좋게 시각화를 시켰기 때문에 작동을 한다.

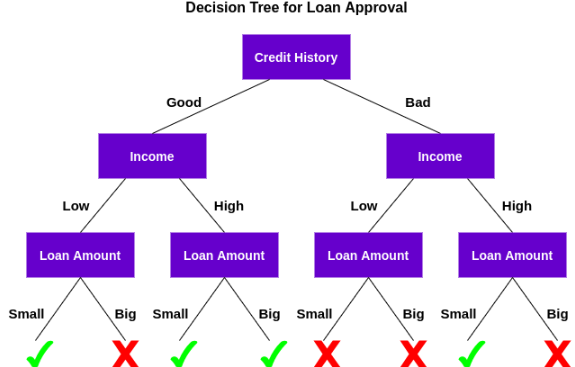
Decision Tree

의사 결정 트리

Decision Tree와 같은 도구를 이용해서 모델을 그래프로 나타낸다.

Root에서 적절한 node로 진행해 가고 최종노드를 선택하게 된다.

구조가 단순해서 누구나 이해할 수 있다.

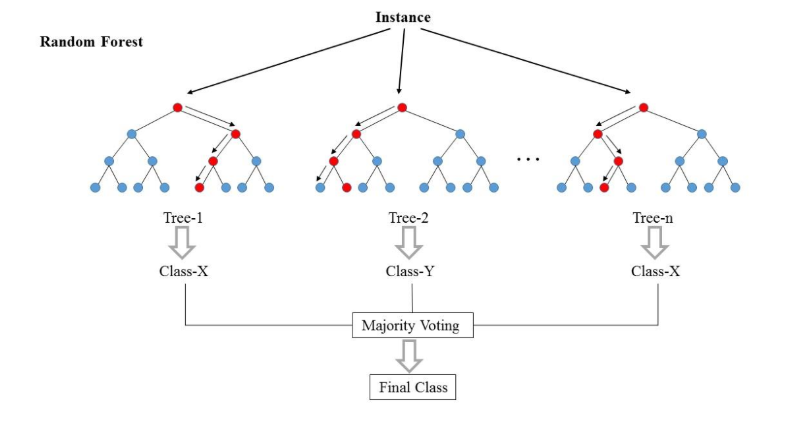


Random Forest

Decision Tree가 여러 개 모여 숲을 이루었다 해서 Forest가 들어갔다.

작은 Tree가 여러 개 모여 있고, 그 트리들의 결과를 합하여 더 많은 값을 최종결과로 본다.

새로운 데이터 포인트를 각 트리에 통과시키며, 각 트리가 분류한 결과에서 투표를 실시하여 가장 많이 득표한 결과를 최종 분류로 선택한다.



Random forest 는 맨처음 데이터를 받아들여 Baggin 이라는 과정을 거친다. Bagging은 트리를 만들 때 training set의 부분집합을 활용하여 생성하는 것이다.

