[knn 알고리즘]

* 거리기반 분류분석 모델
* 어떤 데이터에 대해 근접해 있는 k개의 데이터를 기반으로 해당 데이터를 분류하는 알고리즘
* 매우 단순하지만 실용적
* k값이 매우 작을 때 : 모델이 과적합이 될 가능성이 존재한다.
* k값이 매우 클 때 : 데이터의 구조를 파악하기 어려워진다. 기존에 더 많이 분류되어 있는 쪽으로 편향된다.
* 어림잡기(Rule of Thumb) : k = root(전체 n)

# k-최근접이웃 기반 이상탐지

거리기반 이상치 탐지 기법 중 가장 기본적 알고리즘

각 데이터에 대한 novelty score를 k개의 근접 이웃까지 거리를 이용하여 계산하는 것

이웃까지의 거리를 계산해봤더니 값이 다른 데이터들에 비해 크다면, 이는 이상치일 확률이 높다고 할 수 있다.

# Distance information

주변에서 가장 먼 이웃과의 거리 (maximum distance to the k-th nearest neighbor)

거리의 평균 (Average distance to the k-nearest neighbors)

이웃들 간 centroid를 구해서 거리 구함 (Distance to the mean of the k-nearest neighbors)

# Consider additional factor

기존 k-nn 계산 방법에 한가지 항을 더 추가. 이웃들의 convex hull까지 거리고 고려

Convex hull까지의 거리는 이웃들끼리 연결했을 때 그 안에 있으면 거리가 0, 그 밖에 있으면 거리가 0 이상.

즉, convex안에 있으면 penalty를 주지 않고 최대 2배까지 패널티를 주겠다는 말과 같다.

# 거리 계산법

* 유클리드 거리 : 두개의 서로 다른 점이 있다. 각 요소 간의 차이의 제곱을 다 더하고 제곱근을 취한 값
* 코사인 유사도 : 두개의 실수값을 갖는 속성벡터간에 사용될 수 있다. 값이 1이면 완전동일, 0이면 상호독립, -1이면 완전 반대
* 자커드 거리 : 객체집한 간의 거리를 표현, 두 집합이 얼마나 유사한지 표시
* 마할라노비스 거리 : 2개의 실수벡터 간에 사용. 유클리드 거리에 비해 상관을 고려할 수 있고 단위와 무관한 거리
* 해밍 거리 : 2개의 문자열, 같은 길이의 DNA 서열 간의 거리를 구하는 데 사용. 문자의 위치를 옮겨가며 그 위치에 있는 문자들이 서로 같은지 보고 다르면 1씩 늘려간다.
* 맨해튼 거리 : 두개의 k-차원 실수벡터간 거리. 격자 모양인 맨해튼 도심거리를 다니는 택시를 머릿속으로 그려보라.
* 이외, 민코우스키 거리, 캔버라 거리, 체비셰프 거리 등
* 데이터 형태에 따라 사용하는 거리 척도가 있다. 적용을 통해 최적의 거리 계산법 선택

## 최근접 거리 (Nearest neighbor) 기반 이상 탐지 기법

* ‘정상값들은 어떤 근방(들) (neighbor)에 밀집되어 있고, 이상값은 각 근방에서 멀리 떨어져 있다’고 가정
* NN기반 기법을 쓰려면 두 개체 사이 거리의 개념이 정의되어야 한다. 거리는 여러가지 방법으로 정의할 수 있다.
* 연속형 변수에 대해서는 유클리드 거리가 일반적인 선택이고, 다른 척도를 사용할 수도 있다.
* 범주형 변수에 대해서는 단순 일치 계수가 자주 쓰이며 마찬가지로 더 복잡한 척도가 존재한다.
* # k번째로 가까운 개체와의 거리 이용
* 이상점수를 k번째로 가까운 개체와의 거리로 정의
* 일반적으로 이상 점수의 경곗값(threshold)을 설정하지만, ramaswamyet은 이상 점수를 기준으로 정렬시킨 뒤 이상 점수가 가장 큰 m개를 이상값으로 보았고, 여기서 m은 분석자가 선택하는 상수이다.
* 이상 점수를 구하는 다른 방법으로 한 개체에서 일정거리(d) 이내에 있는 개체의 수(n)을 세는 방법이다.
* # 상대 밀도 이용
* 밀도 기반 기법은 각 개체의 근방의 밀도를 추정한다. 근방의 밀도가 낮은 개체는 이상값이라 판단한다.
* 이러한 기법은 영역에 따라 밀도가 다를 때 취약하다.
* ## NN기반 이상 탐지 기법의 장단점
* 비지도 학습 기반이며, 자료에 대해 어떠한 가정도 필요하지 않다.
* 이상값이 학습 집합에서 가까운 근방을 형성할 확률이 매우 낮아 준지도 기법이 비지도 기법보다 이상값을 잘 찾아내는 면에서는 더 우수하다.
* 거리만 잘 정의되어 있으면 자료의 형태에 구애받지 않는다.
* 정상값이 가까운 이웃이 없다거나, 이상값이 가까운 이웃이 있는 상황에서 비지도 기법은 이상을 정상으로 보는 오류를 범한다.
* 준지도 기법을 적용할 때 학습 자료에서 없었던 패턴의 정상값이 테스트 집합에서 나타나면 이상으로 처리할 가능성이 높다.
* 성능이 거리 척도의 영향을 많이 받으며, 그래프, 순차 자료같이 형태가 복잡하면 거리를 정의하기 까다로울 수 있다.
* 일반적인 유클리드 거리를 사용할 경우, 각 변수별로 척도의 차이가 존재하기 때문에 왜곡이 일어날 수 있으며, 표준화를 통해 이를 바로 잡더라도 범주형 변수가 존재하는 경우에는 가변수를 생성하는 등 다른 방식으로 문제를 해결해야 한다.
* 자료의 차원이 높아질수록 거리를 계산하는 데 있어 어려움이 많이 따른다.