4장. 정형 데이터의 전처리(2)

어떤 한 은행에서 정기예금 계약을 위한 텔레마케팅을 실시하고 있습니다. 오퍼레이터가 과거의 경험에 의거해 잠재고객에게 전화를 걸고 있습니다만 최근 수 년간은 신규고객의 계약 건수가 주춤하고 있고 실패 건수가 눈에 띄는 상태입니다. 전화를 걸면 걸수록 인건비만 늘어나므로 고객수를 늘리기 위해서 무언가 특단의 대책이 필요하다고 생각하고 있습니다.

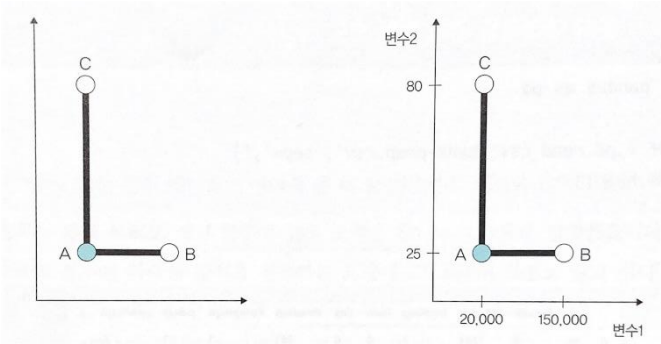
3장의 분석 목표 : 계약 가능성이 높은 고객을 찾아내는 것

⇨ 지도 학습 : 정답이 있는 데이터를 활용하여 데이터 학습

분석 목표: 고객의 특성을 한 가지 이상 찾아내는 것

⇨ 비지도 학습 : 정답 label이 없는 데이터를 비슷한 특징끼리 군집화하여 새로운 데이터에 대한 결과를 예측하는 방법

<데이터 정규화>

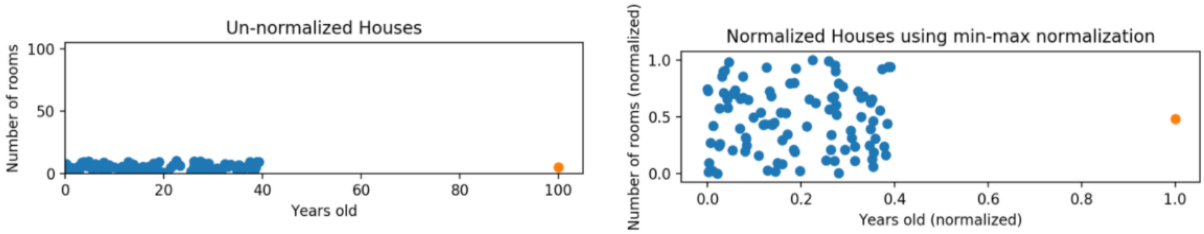


⇨ 변수의 척도가 중요

- 범위 변환 : 정규화 후 변수의 최솟값을 0, 최댓값을 1로 했을 때, 값들이 이 사이에 머물도록 함(Min-Max)

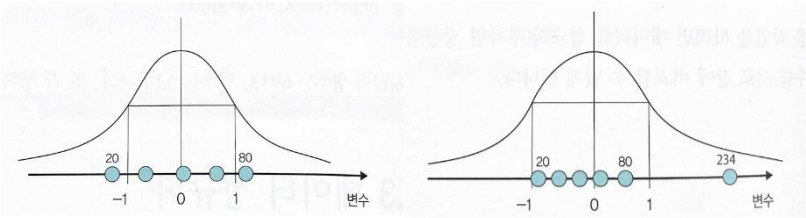
⇨ 이상치(outlier)에 많은 영향을 받음

⇨ 100개의 값들 중 99개는 0-44사이에 있고, 나머지 하나만 100인 경우

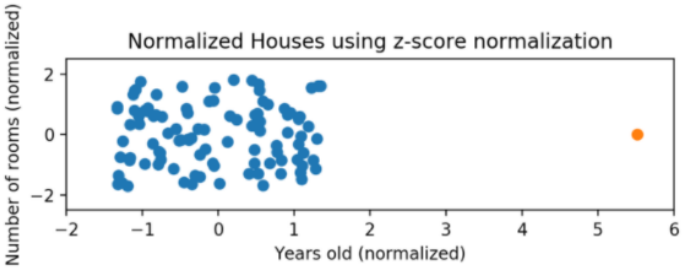


- Z변환 : 정규화 후 변수 평균값이 0, 표준편차가 1이 되도록 값을 변환

➔ (X – 평균) / 표준편차



⇨ 이상치(outlier) 문제를 피하는 데이터 정규화 전략



⇨ 단, 정확히 동일한 척도로 정규화 된 데이터를 생성하지는 않음

<그룹화(군집화)>

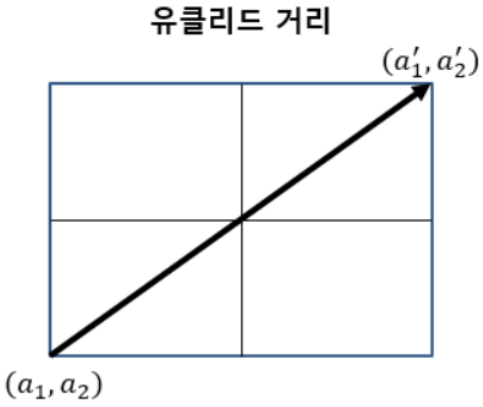
- 계층형 클러스터링
- 비계층형 클러스터링

데이터 간의 거리가 가까운 것들이 모여서 형성

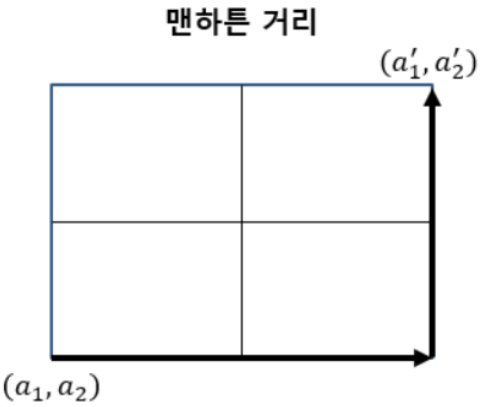
데이터 사이에는 스케일과 같은 여러 요소가 포함되어 있어 단순하게 비교하기는 힘듬

➔ 데이터 간의 거리 측정에 사용하는 대표적인 거리 함수

■ 유클리드 거리(Euclidean Distance) : 피타고라스 정리



■ 맨하튼 거리(Manhattan Distance) : 축을 따라 진행했을 때의 경로 길이

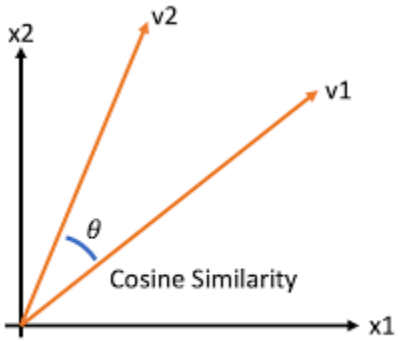


■ 코사인 유사도(Cosine Similarity) : 변수에 포함되는 값의 유사도를 나타냄

⇨ 문서를 분류할 때 자주 사용

⇨ 유사도는 두 개의 벡터 각도에 의해 결정됨

(작을수록 내용이 비슷한 문서)



<계층형 클러스터링(Hierarchical Clustering)>

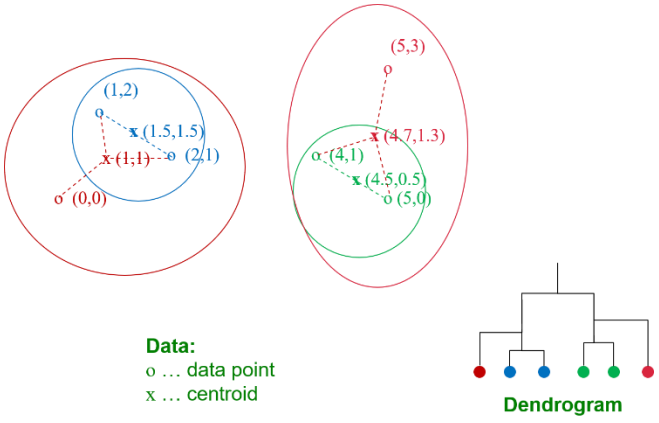
거리가 가까운 데이터부터 순서대로 병합해 그룹을 형성

가장 중요한 과정? 반복적으로 두 개의 "가까운" 클러스터를 찾는 것

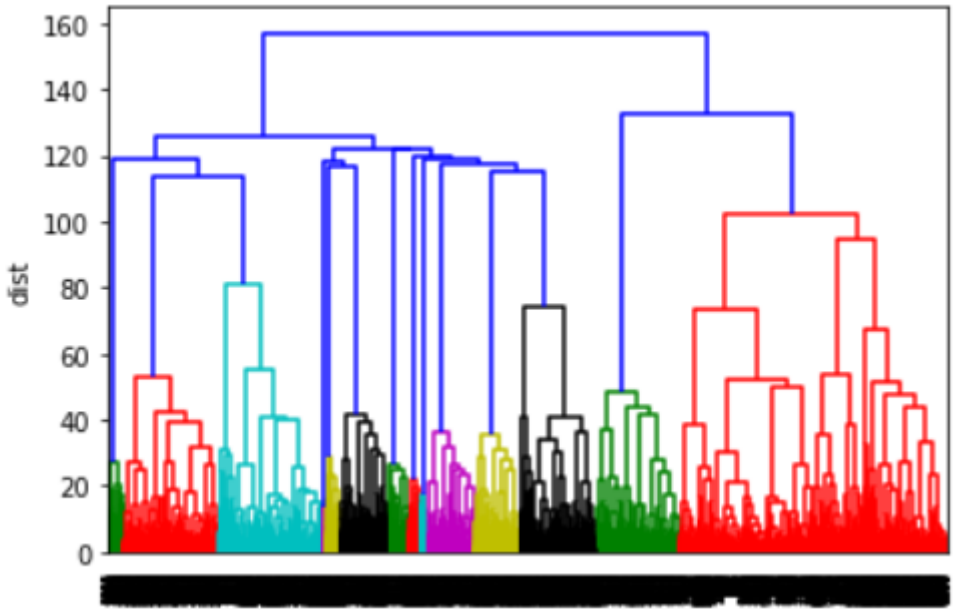
결과는 덴드로그램(Dendrogram)으로 표현

워드법(Ward's Method) : 데이터 병합 방법 중 하나

1. 병합하고 싶은 두 개의 그룹 L과 M에 대해서 각각의 그룹 내에서 얼마나 흩어져 있는지 그 정도(중심과 각 데이터 간 거리의 제곱 합)를 계산
2. 병합했을 때의 중심과 각 데이터 간에 흩어져 있는 정도를 계산
3. 세 가지의 흩어져 있는 정도 차이를 계산



4. 차이가 최소가 되는 그룹을 결합

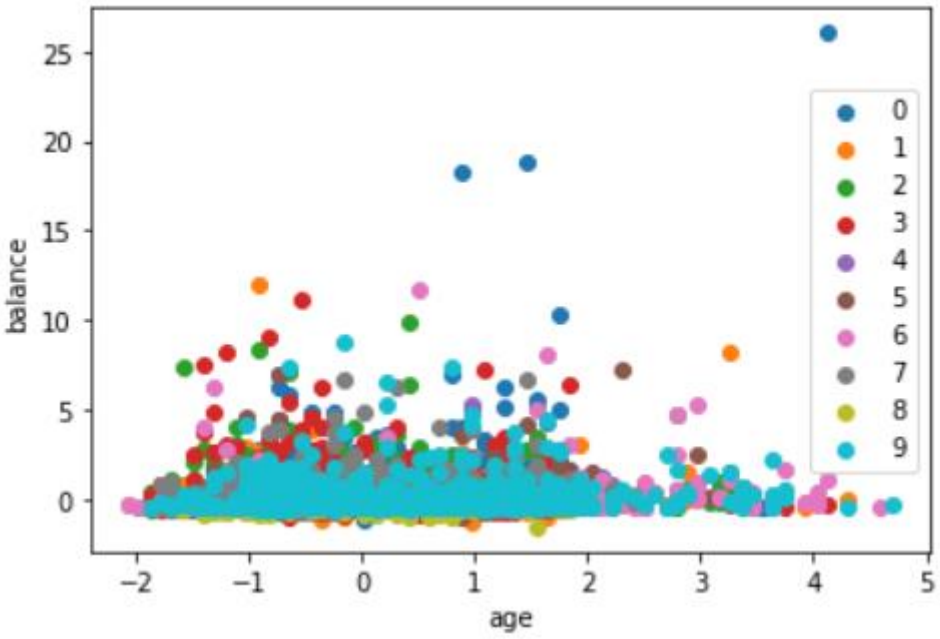


⇨ 가로축에 고객(index), 세로축에 거리를 표시하는 덴드로그램

<비계층형 클러스터링>

k-Means : 데이터의 덩어리를 성질이 비슷한 k개의 덩어리로 분할해 그룹을 형성하는 방법

1. 데이터를 몇 개의 그룹으로 분할할지 개수를 정함
2. 그 개수만큼 임의의 위치로 점을 배치하고 각 그룹의 중심으로 설정
3. 각각의 데이터와 각 중심과의 거리를 계산하고 데이터 점과 소속되는 그룹을 정함
4. 그룹 내의 데이터 점의 중심 위치를 계산하고 원래 중심을 그 위치로 이동시킴
5. 중심 위치가 움직이지 않을 때까지 반복



<주성분 분석(Principal Component Analysis)>

데이터 샘플 수에 비해 특성의 수가 훨씬 많다면? 예측 성능 악화

⇨ 특성이 너무 많아지면 어떤 특성이 타깃에 어떤 영향을 미치는지 인과관계를 파악하기 어려워지기 때문에 발생

⇨ 해결방법?

각각의 특성이 미치는 영향을 파악할 수 있을 정도로 데이터 샘플을 충분히 모으기 or 특성의 수를 줄이는 것(특성 추출)

⇨ PCA : 기존의 변수를 조합해 다른 변수를 새롭게 작성하는 방법

1. 데이터 샘플의 분포에서 분산이 가장 큰 방향으로 새로운 축을 설정(제 1주성분)
2. 분산이 두 번째로 커지는 방향으로, 제 1주성분과 직교하도록 제 2주성분을 설정
3. 누적 기여율(Cumulative Proportion)이 70~80%에 도달할 때까지의 주성분을 채용

기여율 : 각 주성분이 가지고 있는 정보가 데이터에 대해서 얼마나 영향을 미치는지를 나타내는 지수