CHAPTER 7 비지도 학습 - 정지혜

레이블이 달린 데이터를 이용해 모델을 학습하는 과정 없이 데이터로부터 의미를 이끌어내는 통계적 기법

7.1 주성분분석

수치형 변수가 어떤 식으로 공변하는지 알아내는 기법

- 식당의 음식 값과 팁 → 가격
- 집의 면적, 방 개수, 화장실 개수 → 크기

주성분

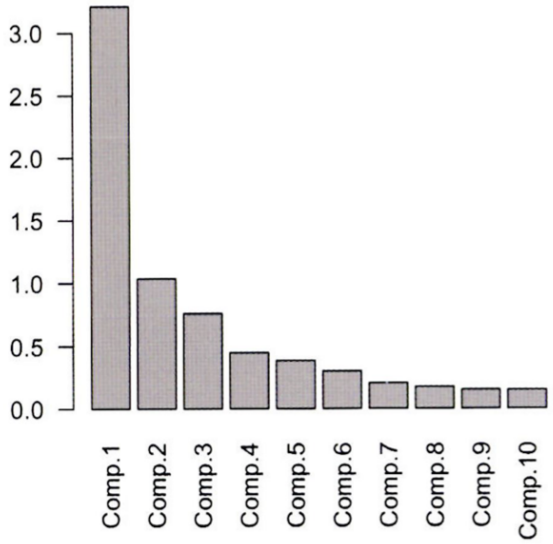
전체 변수들의 변동성을 대부분 설명할 수 있는 적은 수의 변수들의 집합

- 데이터의 차원 축소
- 새로운 변수들 → 기존 변수들에 가중치를 적용한 선형결합
- 가중치(부하) : 새로운 주성분에 대한 기존 변수들의 기여도
- 첫 주성분은 전체 변동성을 가장 잘 설명하는 선형결합이며, 추가 주성분들은 서로 수직이고 나머지 변동성을 설명한다.

주성분 해석

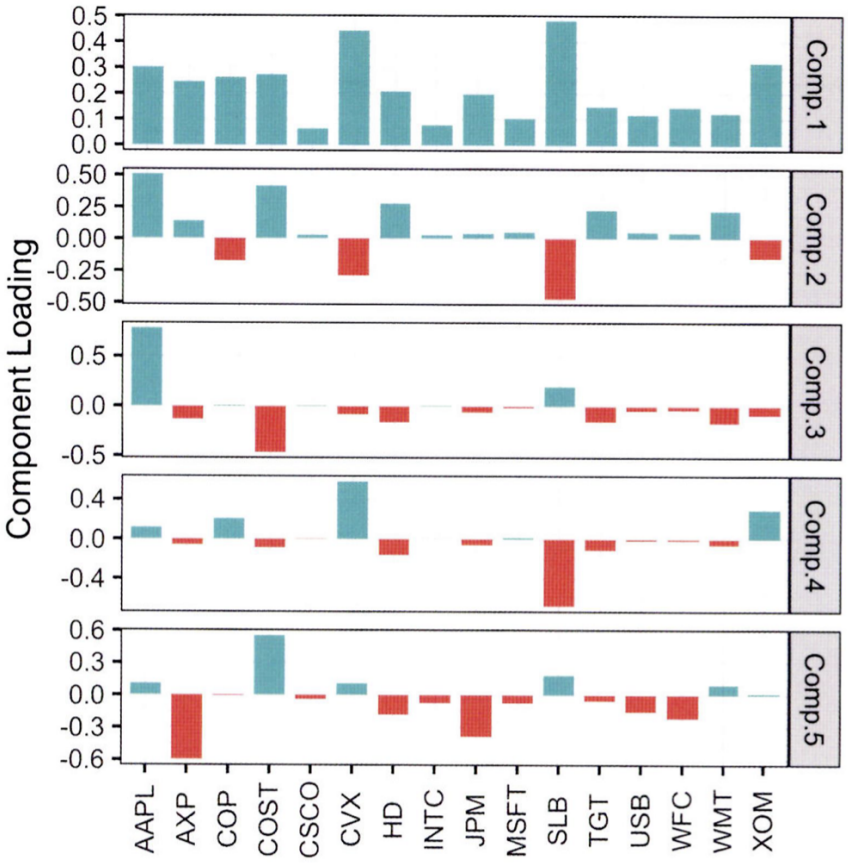
- 스크리그래프(screeplot)

  성분들의 변동성을 표시한 그림. explained variance ratio(설명 분산 비율)을 이용하여 성분들의 상대적인 중요도를 보여준다.



- 상위 주성분들의 가중치 표시

  상위 주성분들의 가중치를 표시해보는 것도 주성분을 이해하는 데 도움이 된다.



대응분석

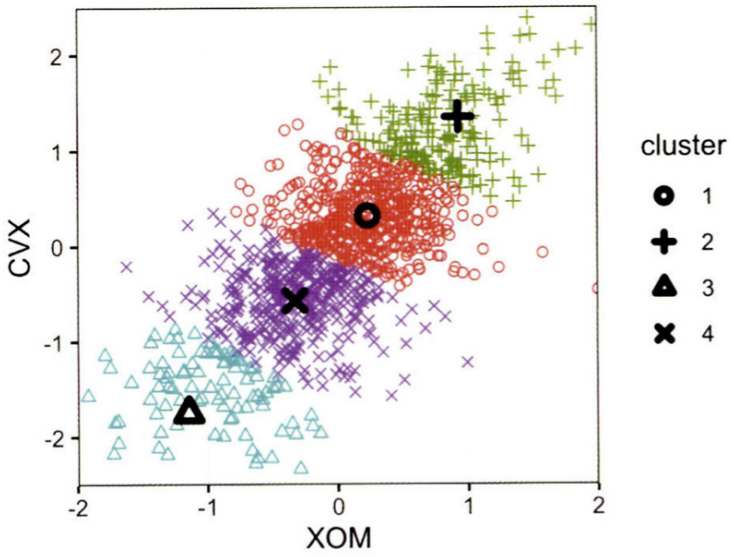
범주간 혹은 범주형 피쳐 간의 연관성 인식

- 주로 저차원 범주형 데이터의 그래프 분석에 이용

7.2 k-평균 클러스터링

서로 비슷한 데이터들끼리 k개의 그룹으로 분류

- 클러스터 : 서로 유사한 레코드들의 집합
- 할당된 클러스터의 평균과 포함된 데이터들의 거리 제곱합이 최소가 되도록 분류



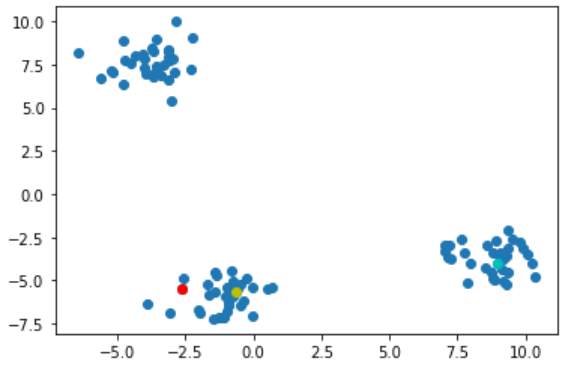
데이터들이 클러스터에 할당된 결과와 클러스터의 평균

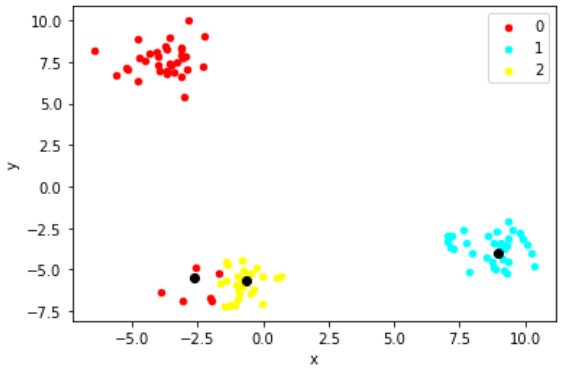
k-평균 알고리즘

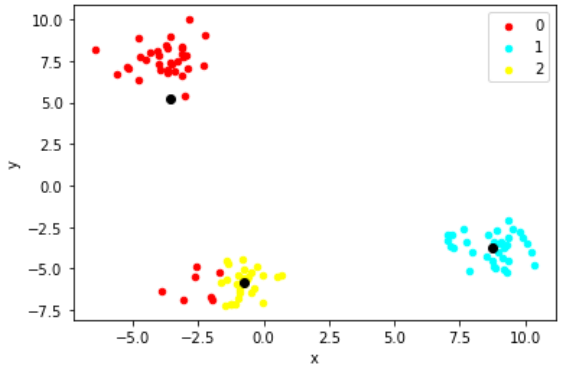
사용자가 미리 정해준 k 값과 랜덤하게 할당한 클러스터 평균의 초깃값으로 시작

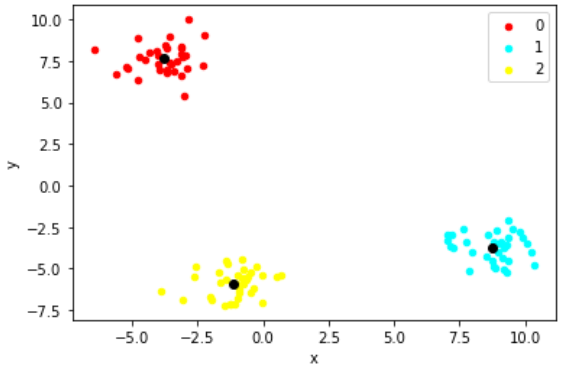
- 각 레코드를 거리가 가장 가까운 평균을 갖는 클러스터에 할당
- 새로 할당된 레코드들에 대하여 새로운 클러스터 평균 계산
- 클러스터에 유의미한 차이가 없을 때까지 반복

→ 평균의 초깃값에 따라 결과가 다를 수 있다. 항상 최적의 답을 준다는 보장이 없으므로 초깃값을 변화시켜 가며 알고리즘 반복 후 가장 좋은 성능을 보이는 결과 탐색





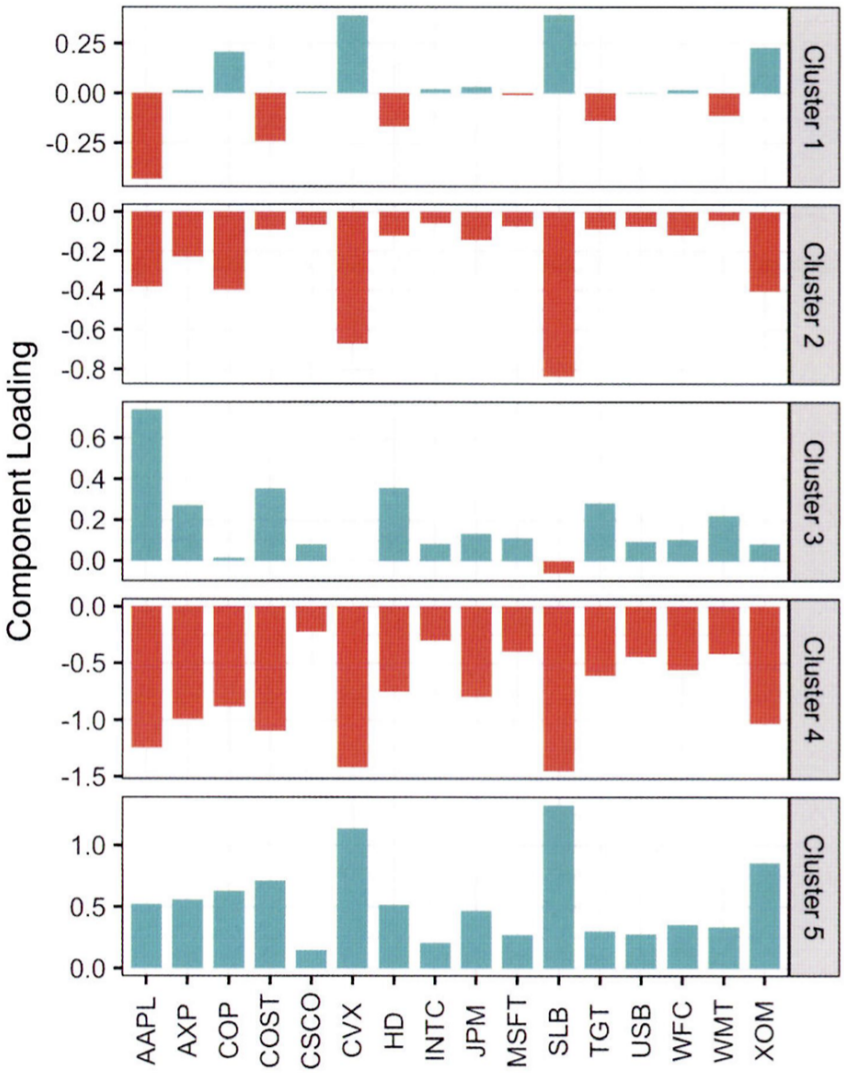




클러스터 해석

클러스터의 크기와 평균 중요

- 유난히 균형이 맞지 않는 클러스터가 존재한다면 멀리 떨어진 특이점 혹은 레코드 그룹 의미



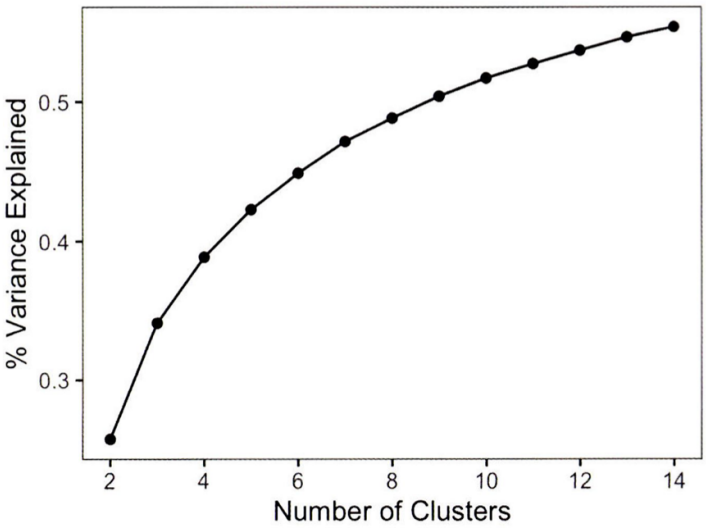
변수들의 평균(클러스터 평균 : 각 변수들의 평균으로 이루어진 벡터)

클러스터 개수 선정

- 실무적인 상황을 고려해 k 선택
- 통계적 접근 방식을 사용할 수 있으나, 최상의 클러스터 개수를 찾는 딱 한 가지 표준화된 방법은 없다.

  → 일반적으로 클러스터의 개수를 정확히 얻는 완벽한 방법은 없다.

  - elbow method : 언제 클러스터 세트가 데이터 분산의 '대부분'을 설명하는지를 알려준다.



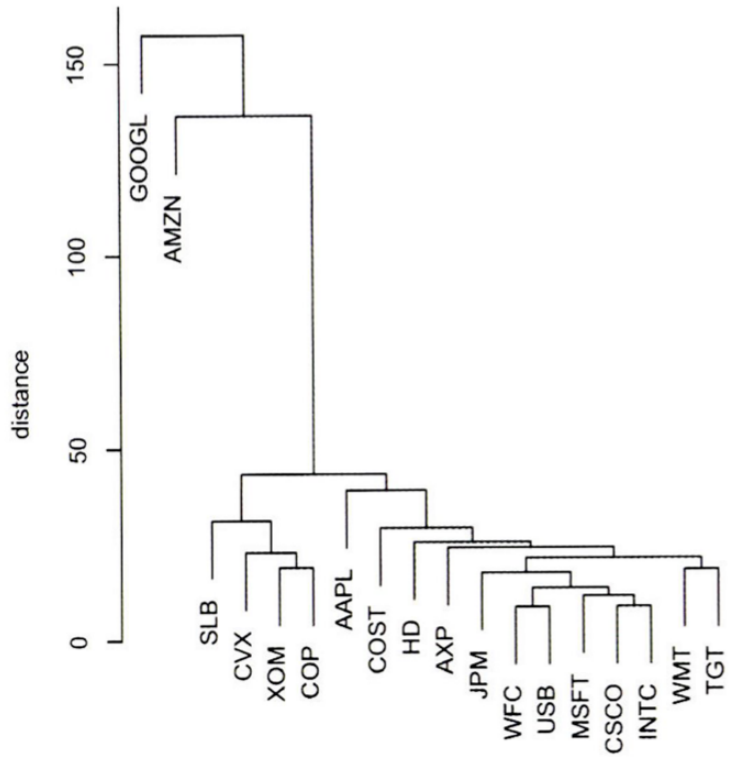
분산 증가율이 서서히 떨어지므로 눈에 띄는 팔꿈치 위치는 없다. 다만 데이터의 특성을 밝혀준다는 점에서 가치가 있다.

7.3 계층적 클러스터링

k-평균 대신 사용하는 클러스터링 방법으로, 서로 다른 수의 클러스터를 지정하는 효과를 시각화할 수 있다.

덴드로그램

레코드들, 그리고 레코드들이 속한 계층적 클러스터를 트리모델과 같이 시각적으로 표현



distance : 한 레코드가 다른 레코드들과 얼마나 가까운지를 보여주는 측정 지표

- 트리의 잎은 각 레코드를 의미한다.
- 트리의 가지 길이는 해당 클러스터 간의 차이 정도를 나타낸다.
- 사용자가 미리 클러스터의 수를 지정할 필요 없이 원하는 개수의 클러스터를 추출할 수 있다.

```
cutree(hcl, k=4)

GOOGL  AMZN  AAPL  MSFT  CSCO  INTC   CVX   XOM   SLB   COP   JPM   WFC
    1     2     3     3     3     3     4     4     4     4     3     3
  USB   AXP   WMT   TGT    HD  COST
    3     3     3     3     3     3
```

R에서 추출할 클러스터의 개수를 4로 설정한 예시

병합 알고리즘

유사한 클러스터들을 반복적으로 병합하는 역할

- 각 레코드 자체를 개별 클러스터로 설정하여 시작
- 모든 쌍의 클러스터 사이의 비유사도 계산

    - 비유사도 : 한 클러스터가 다른 클러스터들과 얼마나 가까운지를 보여주는 측정 지표
- 모든 레코드가 하나의 클러스터에 속할 때까지 가장 가까운 클러스터를 결합해 나가는 작업 반복
- 클러스터 간 거리는 모든 레코드 간 거리 정보를 사용하여 여러 가지 다른 방식으로 계산할 수 있다.

7.4 모델 기반 클러스터링

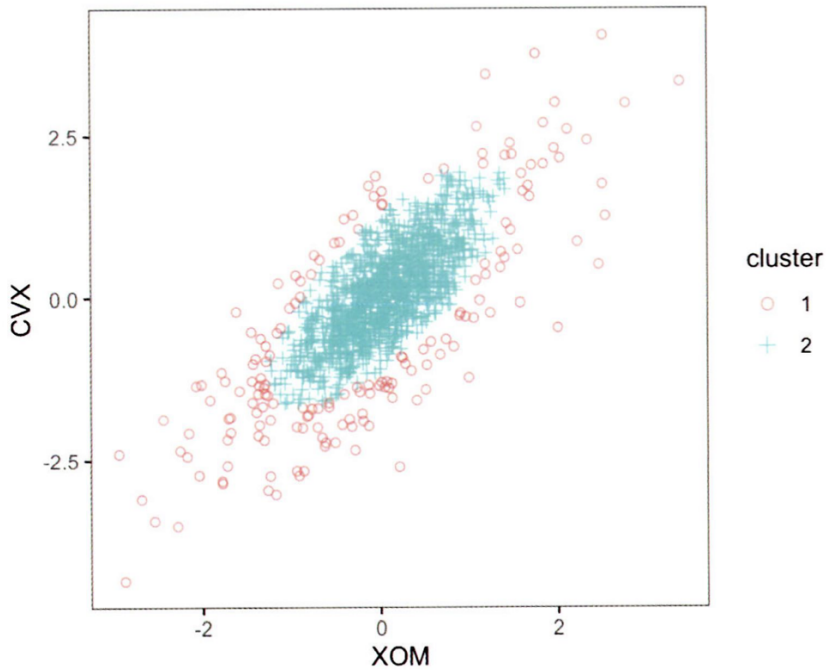
전반적으로는 서로 비슷하지만 모든 데이터가 반드시 서로 가까울 필요는 없는 그룹과, 서로 비슷하면서 데이터들이 아주 가까이에 있는 또 다른 그룹이 함께 있는 경우에 사용

다변량정규분포

- 가장 널리 사용되는 대부분의 모델 기반 클러스터링 방법은 모두 다변량정규분포를 따른다.
- 다변량정규분포는 p개의 변수 집합에 대해 정규분포를 일반화한 것이다.
- 분포는 평균과 공분산행렬로 정의된다.

정규혼합

- 모델 기반 클러스터링의 핵심 아이디어는 각 레코드가 k개의 다변량정규분포 중 하나로부터 발생했다고 가정하는 것이다.
- 각 분포는 서로 다른 평균과 공분산행렬을 갖는다.
- 모델 기반 클러스터링의 목적은 데이터를 가장 잘 설명하는 다변량정규분포를 찾는 것이다.



두 클러스터에서 하나는 데이터의 중심에 있고, 다른 하나는 외곽에 존재한다.

클러스터 개수 정하기

- 베이즈 정보기준 값이 가장 큰 클러스터의 개수를 자동으로 선택한다.

7.5 스케일링과 범주형 변수

- 비지도 학습에서는 데이터를 적절하게 스케일해야 한다.
  - 스케일링 : 데이터의 범위를 늘리거나 줄이는 방식으로 여러 변수들이 같은 스케일에 오도록 하는 것
  - 데이터의 크기가 조정되지 않으면 PCA, 클러스터링 방법은 큰 값을 갖는 변수들에 의해 결과가 좌우되고 작은 값을 갖는 변수들은 무시된다.

변수 스케일링

- 일반적인 스케일링 방법은 각 변수에서 평균을 빼고 표준편차로 나눠주는 정규화(표준화) 방법이다.

→ 클러스터들의 크기가 좀 더 균일하며 스케일에 큰 영향을 받지 않는다.

범주형 데이터와 고워 거리

범주형과 수치형 데이터가 혼합된 경우에는 고워 거리를 사용할 수 있다.

- 각 변수의 데이터 유형에 따라 거리 지표를 다르게 적용한다.

- 모든 변수를 0~1 범위로 스케일링 한다.

7.6 마치며

- k-평균은 매우 큰 데이터로 확장이 가능하고 이해하기 쉽다.
- 계층적 클러스터링은 수치형과 범주형이 혼합된 데이터 유형에 적용이 가능하며 직관적인 시각화 방법이 존재한다.
- 모델 기반 클러스터링은 통계 이론에 기초를 두고 있으며 더 엄밀한 접근 방식을 제시한다.

→ 데이터 크기나 응용 분야의 목표에 따라 사용되는 방법은 달라지게 된다.