비지도 학습

비지도 학습 (Unsupervised Learning)

- 미리 알려진 출력 값 또는 정보 없이 학습 알고리즘을 학습해야 하는 모든 종류의 머신러닝 기법
- 알고리즘은 입력 데이터만으로 데이터에서 지식을 추출할 수 있어야 함
- 대표적인 두 가지 비지도 학습은 변환과 군집

비지도 변환 (Unsupervised Transformation)

- 데이터를 새롭게 표현해서 사람 또는 다른 머신러닝 알고리즘이 원래 데이터보다 쉽게 해석할 수 있도록 만드는 것

- 대표적으로 차원 축소와 데이터 성분 찾기 등의 기법 포함

- 차원 축소
  - 특성이 많은 고차원 데이터에 대해 특성의 수를 줄이고 꼭 필요한 특징을 포함한 데이터로 표현하는 방법
  - 사례 → 시각화를 위해 데이터셋을 2차원으로 변경

- 데이터의 구성 단위 또는 성분 찾기
  - 사례 → 많은 텍스트에서 주제를 추출하는 작업

군집 (Clustring)

- 데이터를 비슷한 것끼리 그룹으로 묶는 것

- 사례
  - 소셜 미디어 사이트에 업로드된 사진 분류
  - 마케팅을 위한 시장 세분화
  - 소비자 분류

비지도 학습의 과제

- 가장 어려운 작업은 학습 결과에 대한 평가
  - 평가를 위해 직접 확인하는 것이 유일한 방법인 경우가 많음

- 데이터에 대한 이해 수준을 높이기 위해 수행하는 탐색적 분석 단계에서 많이 사용

- 또는 지도학습의 전처리 단계에서 많이 사용
  - 비지도 학습의 결과로 새롭게 표현된 데이터를 사용해 학습하면 지도 학습의 정확도가 좋아지기도 하며 메모리와 시간 절약 가능

차원 축소, 특성 추출, 매니폴드 학습

차원 감소

- 차원의 저주
  - 특성(차원)이 증가함에 따라 차원 내의 부피가 급격히 증가하지만 해당 공간에 놓일 데이터는 한정되어 있어서 빈 공간이 많아지는 상황
  - 데이터의 전반적인 구조는 바꾸지 않고 중복적인 정보를 가지는 차원을 줄이는 방법 필요

- 차원 감소 목표
  - 데이터의 구조는 최대한 살리면서 적은 수의 특징만으로 특정 현상을 설명

- 차원 감소 방법
  - 특성 추출 (feature selection) → 전체 특성 중 일부 특성만 선택
  - 차원 축소 (feature extraction) → 전체 특성 개수의 차원 벡터를 입력으로 적은 차원의 벡터를 출력하는 (선형) 함수 생성

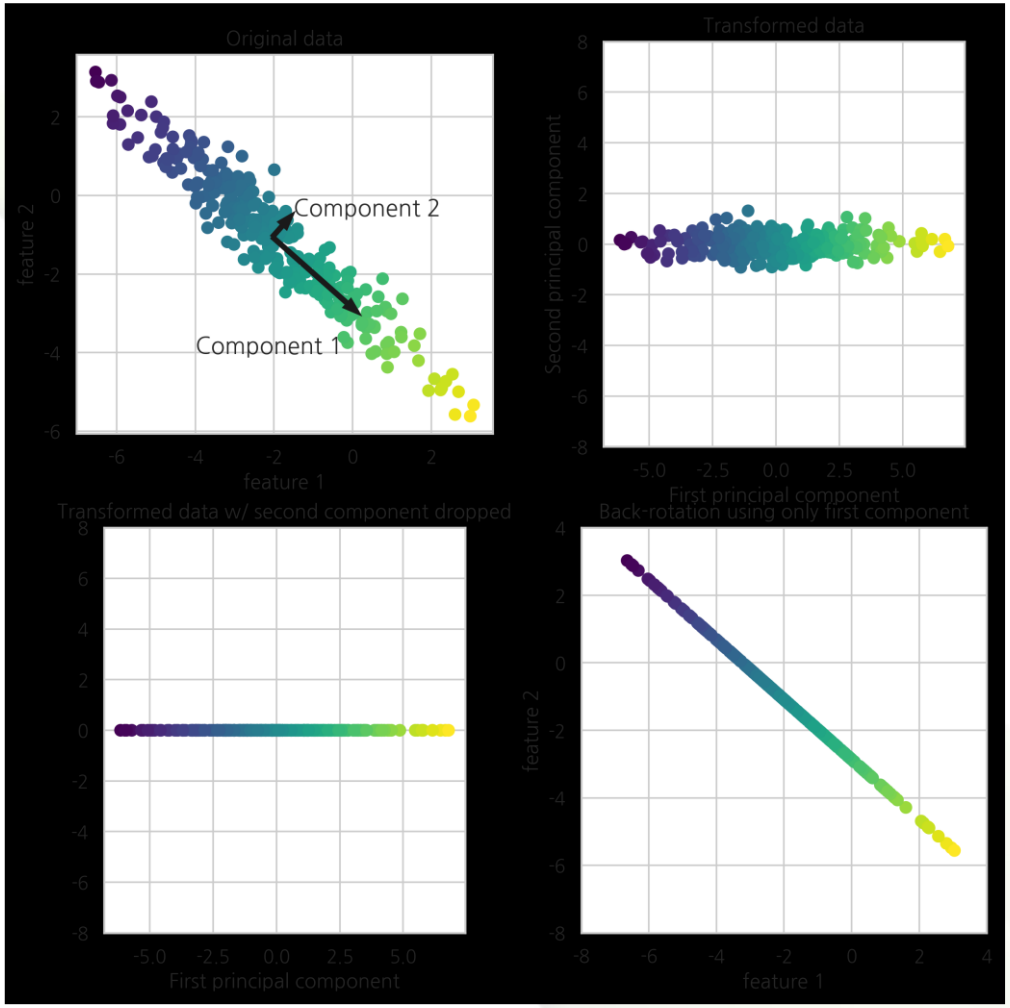
차원 축소, 특성 추출, 매니폴드 학습

- 비지도 학습을 통해 데이터를 변환하는 이유
  - 시각화
  - 데이터 압축
  - 추가적인 처리를 위해 정보가 더 잘 드러나는 표현 추적

- 종류
  - 주성분 분석 (Principal Component Analysis, PCA)
    » 가장 간단하고 일반적으로 사용되는 변환 기법
  - 비음수 행렬 분해 (Non-Negative Matrix Factorization, NMF)
    » 특성 추출에 널리 사용
  - t-Distributed Stochastic Neighbor Embedding, t-SNE
    » 2차원 산점도를 이용해 시각화

주성분 분석 (Principal Component Analysis, PCA)

- 특성들이 통계적으로 상관관계가 없도록 데이터 세트를 회전시키는 기술
- 회전한 뒤 데이터를 설명하는 중요도에 따라 새로운 특성의 일부만 선택
- 가장 널리 사용되는 분야는 고차원 데이터 세트의 시각화



주성분 분석 작업 단계

- 분산이 가장 큰 방향 찾기
  - 데이터에서 가장 많은 정보를 담고 있는 방향
  - 특성들의 상관관계가 가장 큰 방향

- 첫 번째 방향과 직각인 방향 중에서 가장 많은 정보를 담은 방향 찾기
  - 2차원에서는 직각 방향이 하나이지만 고차원에서는 더 많은 직각 방향 가능
  - 이렇게 발견한 방향이 주성분
  - 일반적으로 특성 개수만큼의 주성분 있음

- 축 회전
  - 주성분을 x축과 y축에 평행하게 회전이동 (이 때 평균을 빼서 중심을 원점에 일치하도록 한 후 이동)
  - 회전된 두 축은 연관되어 있지 않기 때문에 변환된 데이터의 상관관계 행렬은 대각선 방향을 제외하면 0

주성분 분석 작업 단계

- 필요하다면 주성분의 일부만 남기는 차원 축소 처리
  - 가장 유용한 방향을 찾아서 그 방향의 성분만 유지

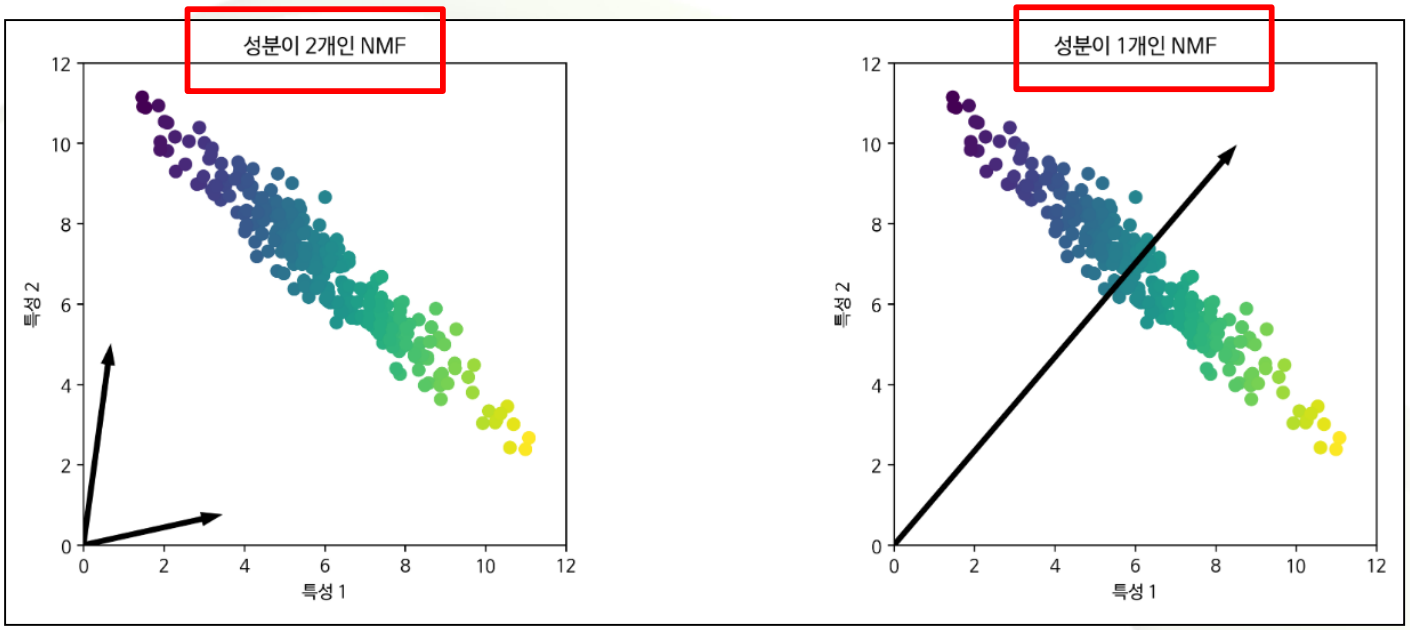
- 데이터에 다시 평균을 더해서 반대 방향으로 회전

비음수 행렬 분해

- 유용한 특성을 뽑아내기 위한 비지도 학습 알고리즘 중 하나
- PCA와 비슷하고 차원 축소에도 사용 가능
- 음수(-)가 아닌 성분과 계수 값 추적
  - 주성분과 계수가 모두 0 이상이어야 함
  - 음수가 아닌 특성을 가진 데이터에만 적용
- 독립된 소스를 추가해서 (덮어써서) 만든 데이터에 유용
  - 여러 사람의 목소리가 담긴 오디오 트랙
  - 여러 악기로 이루어진 음악
- 데이터를 인코딩하거나 재구성하는 용도보다는 데이터 있는 유용한 패턴을 발견하는데 활용

비음수 행렬 분해

- 비음수 행렬 분해 적용 결과 사례



- 하나의 성분을 사용하면 NMF는 데이터를 가장 잘 표현할 수 있는 평균으로 향하는 성분 생성
- 성분 개수를 줄이면 특정 방향이 제거되는 것뿐만 아니라 전체 성분이 완전히 변경됨
- 모든 성분은 동등하게 취급

t-SNE

- PCA는 종종 데이터 변환에 가장 먼저 시도해볼 만한 방법이지만 알고리즘의 태생상 유용성이 떨어짐

- 매니폴드 학습(manifold learning) 시각화 알고리즘은 훨씬 복잡한 매핑을 만들어 더 나은 시각화 제공 → 특별히 t-SNE 알고리즘을 많이 사용

- 시각화가 목적이기 때문에 3개 이상의 특성을 추출하는 경우가 많지 않음

- 훈련 데이터를 새로운 표현으로 변환하지만 테스트 데이터에는 적용할 수 없음
  - 탐색적 분석에는 유용하지만 지도 학습용으로는 거의 사용하지 않음

- t-SNE 알고리즘은 멀리 떨어진 포인트와 거리를 보존하는 것보다 가까이 있는 포인트에 더 많은 비중을 부여 → 이웃 데이터 포인트에 대한 정보를 보존하기 위해 노력

군집 (Clustring)

군집 (Clustring)

- 데이터 세트를 클러스터라는 그룹으로 나누는 작업
- 한 클러스터 안의 데이터 포인트는 매우 비슷하고 다른 클러스터의 데이터 포인트와 구분되도록 나누는 것이 목표
- 각 데이터 포인트가 어느 클러스터에 속하는지 할당 또는 예측

K-평균 군집 (K-Means)

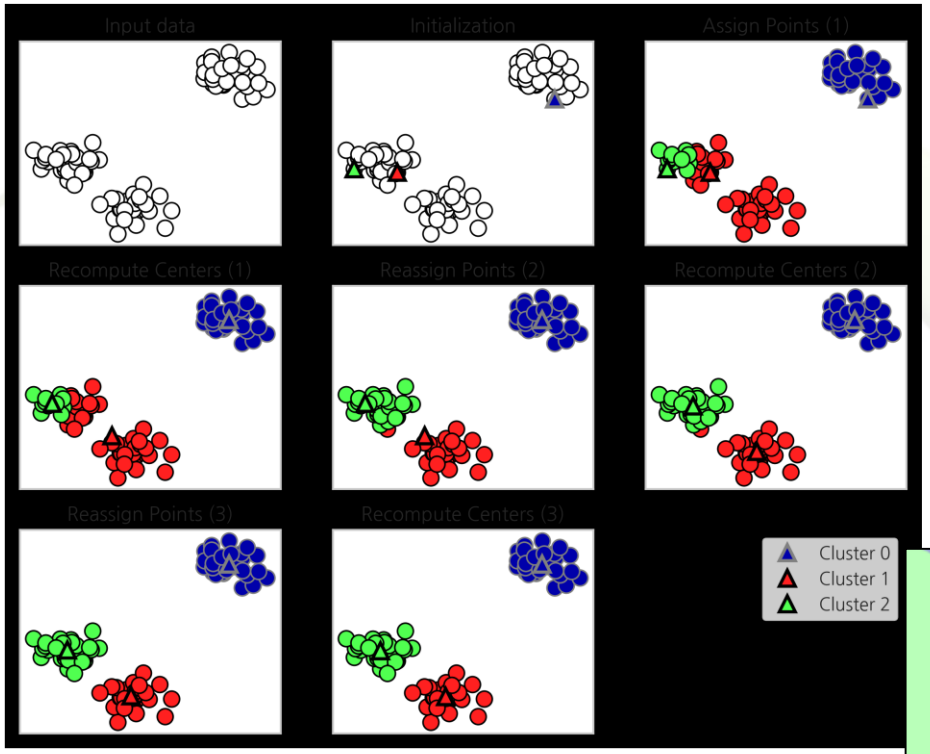
- 가장 간단하고 널리 사용하는 군집 알고리즘

- 데이터의 어떤 영역을 대표하는 클러스터 중심을 추적

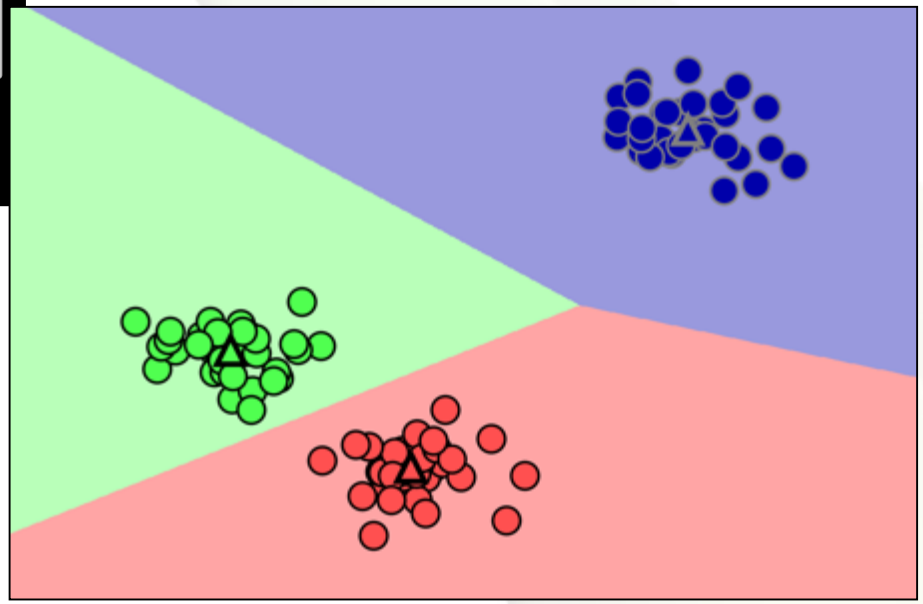
- 프로세스
  - 1. 데이터 포인트를 가장 가까운 클러스터 중심에 할당
  - 2. 클러스터에 할당된 데이터 포인트의 평균으로 클러스터 중심 이동
  - 1 ~ 2를 반복 → 클러스터에 할당되는 데이터 포인트에 변화가 없을 때 알고리즘 종료

- 각 데이터 포인트가 레이블을 가진다는 면에서 분류와 유사하지만,
  - 정답을 알지 못하고
  - 레이블 자체에 특별한 의미가 부여되지 않음

클러스터 할당 과정과 결정 경계



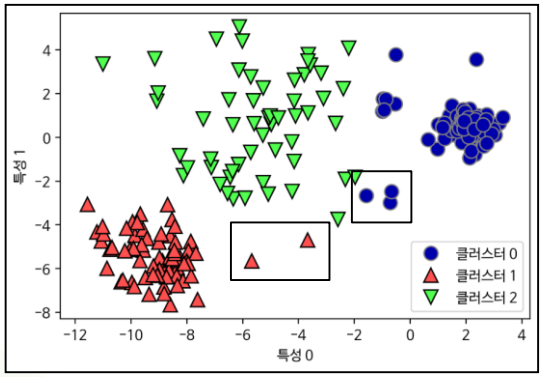
알고리즘 적용 과정



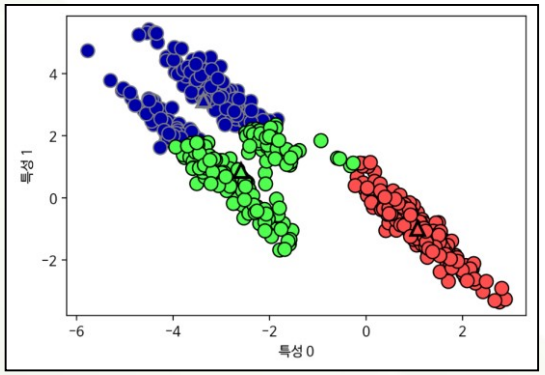
분류 결과

K-평균 군집의 분류 실패

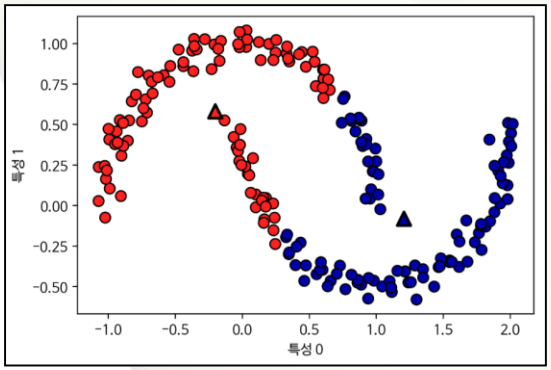
- 클러스터를 정의하는 것이 중심 하나뿐이기 때문에 클러스터는 둥근 형태로 나타나며 비교적 간단한 형태만 구분 가능 (밀도가 다른 그룹 분류 어려움)



- 모든 클러스터의 반경이 동일하다고 가정하기 때문에 범위가 다른 범주를 구분할 때 예상치 못한 결과 도출



- 모든 방향이 동일하게 중요하다고 가정하기 때문에 원형이 아닌 분포를 적절하게 판단하지 못함



k-평균 장단점

- 장점
  - 비교적 이해하기 쉽고
  - 구현도 쉬우며
  - 상대적으로 빠른 처리 속도

- 단점
  - 무작위 초기화를 사용하기 때문에 알고리즘 출력이 난수 초기값에 따라 달라짐
  - 클러스터의 모양을 가정하고 있기 때문에 활용 범위가 제한적
  - 클러스터의 개수를 사용자가 지정 (미지의 데이터)

병합 군집 (Agglomerative Clustring)

- 시작할 때 각 포인트를 하나의 클러스터로 지정하고 그 다음 어떤 종료 조건을 만족할 때까지 가장 비슷한 두 클러스터를 결합하는 과정을 반복

- scikit-learn 구현
  - 종료 조건은 클러스터 개수로 지정된 개수의 클러스터가 남을 때까지 비슷한 클러스터를 결합
  - linkage 옵션으로 가장 비슷한 클러스터를 측정하는 방법 지정
    - » ward : 클러스터 내의 분산을 가장 적게 증가시키는 두 클러스터 병합
    - » average : 클러스트 포인트 사이의 평균 거리가 가장 짧은 두 클러스터 병합
    - » complete : 클러스터 포인트 사이의 최대 거리가 가장 짧은 두 클러스터 병합

병합 군집 (Agglomerative Clustring)

- 새로운 데이터 포인트에 대해 예측 불가능 → predict 메서드 없음
  - 훈련 세트로 모델을 만들고 정보를 얻기 위해 fit_predict 메서드 사용

- 복잡한 데이터 형상을 구분하는데 한계 있음

DBSCAN

- Density-Based Spatial Clustering of Applications with Noise

- 장점
  - 클러스터 개수를 미리 지정할 필요 없음
  - 복잡한 형상도 구별 가능
  - 어떤 클래스에도 속하지 않는 포인트 식별 가능
  - 큰 데이터 세트에도 적용 가능 (k-Means, 병합군집 보다는 다소 느림)

- 원리
  - 데이터의 밀집 지역이 한 클러스터 구성
  - 비교적 비어 있는 지역을 경계로 다른 클러스터와 구분
  - 한 데이터 포인트에서 일정한 거리(eps) 안에 일정한 개수(min_samples) 이상의 데이터가 있으면 같은 클러스터로 병합

DBSCAN

- 동작
  - 무작위로 포인트 선택
  - 지정된 거리 안의 모든 포인트 검색
    - » 검색된 데이터 포인트의 수가 지정된 개수보다 적으면 잡음으로 처리
    - » 검색된 데이터 포인트의 수가 지정된 개수보다 많으면 핵심 샘플로 처리
  - 핵심 샘플로 지정된 데이터 포인트에 클러스터 레이블 할당
  - 핵심 샘플을 기준으로 지정된 거리 안에 있는 이웃 데이터 포인트를 탐색하고 클러스터 레이블 할당

- 포인트의 종류는 3가지 → 핵심 포인트, 경계 포인트, 잡음 포인트
  - 경계 포인트는 탐색 순서에 따라 달라질 수 있음

- 새로운 데이터 예측에 사용할 수 없으며 `fit_predict` 메서드를 사용해서 군집과 클러스터 레이블 계산