혼자 공부하는 머신러닝 + 딥러닝

CH06. 비지도 학습

인공지능공학과 12223557 여예진

비지도 학습 (Unsupervised Learning)

= 타깃이 없는 입력 데이터만을 학습하는 방법

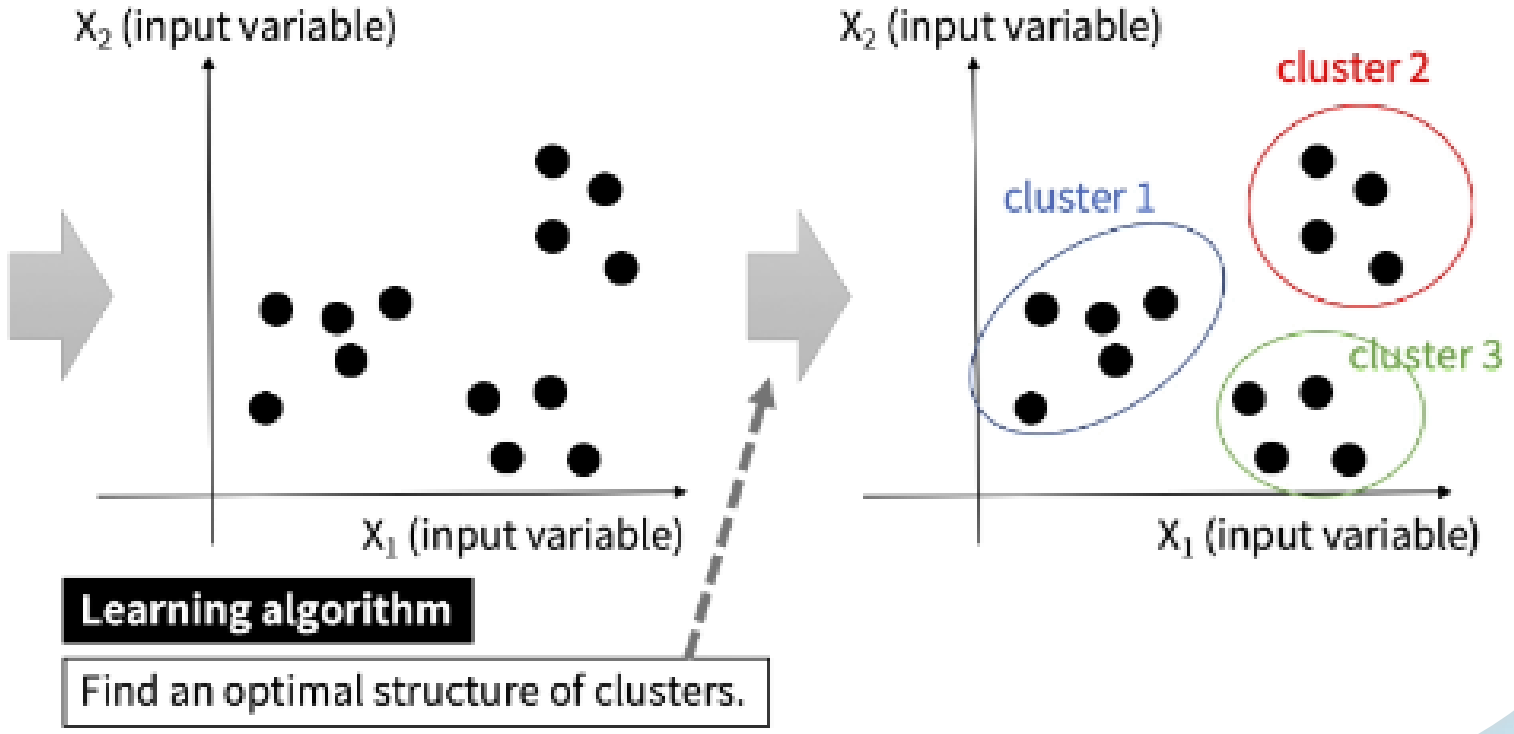
- 군집화
- 차원 축소
- 이상치 탐지
- 밀도 추정
- 잠재 변수 모델

군집화 (Clustering)

= 주어진 데이터 집합을 비슷한 특성을 가진 데이터들의 그룹으로 나누는 작업

- 군집(cluster) = 군집화로 나누어진 유사한 데이터의 그룹
- ex) 고객 데이터 군집화 -> 비슷한 행동 패턴이나 관심사를 가진 그룹 형성 가능

ID	X_1	X_2
1	128	142
2	...	...
3	...	...
4	...	...
...	...	...



Q. 군집화와 분류는 다른 개념인가? 무엇이 다른가?

clustering	classification
데이터 간의 유사도를 정의하고 그 유사도에 가까운 것부터 순서대로 합쳐가는 방법	기존에 존재하는 데이터의 category 관계를 파악하고, 새롭게 관측된 데이터의 category를 스스로 판별하는 과정
비지도 학습	지도 학습
Label(category) 없을 때	Label(category) 있을 때
데이터 자체의 특성을 알고자 하는 목적	새로운 데이터의 그룹을 예측하기 위한 목적

K-means 알고리즘

= 최적의 군집을 구성하는 대표적 군집 알고리즘

1) 군집의 개수(K) 설정

⇩

2) 초기 중심점 설정

⇩

3) 데이터를 군집에 할당(배정)

⇩

4) 중심점 재설정(갱신)

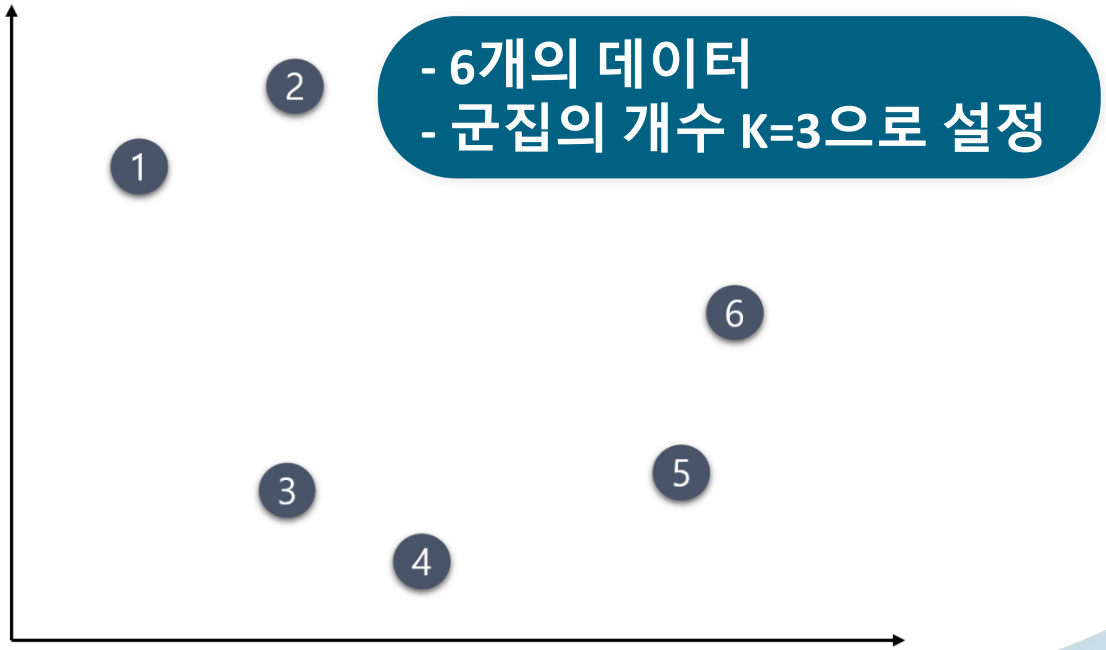
⇩

5) 데이터를 군집에 재할당(배정)

중심점의 위치가 더 이상 변하지 않을 때까지 반복

Step1) 군집의 개수(K) 설정

1) Rule of thumb
2) Elbow Method
3) Information Criterion Approach (정보 기준 접근법)



K-means 알고리즘

= 최적의 군집을 구성하는 대표적 군집 알고리즘

1) 군집의 개수(K) 설정

2) 초기 중심점 설정

3) 데이터를 군집에 할당(배정)

4) 중심점 재설정(갱신)

5) 데이터를 군집에 재할당(배정)

Step2) 초기 중심점(클러스터 중심) 설정하기

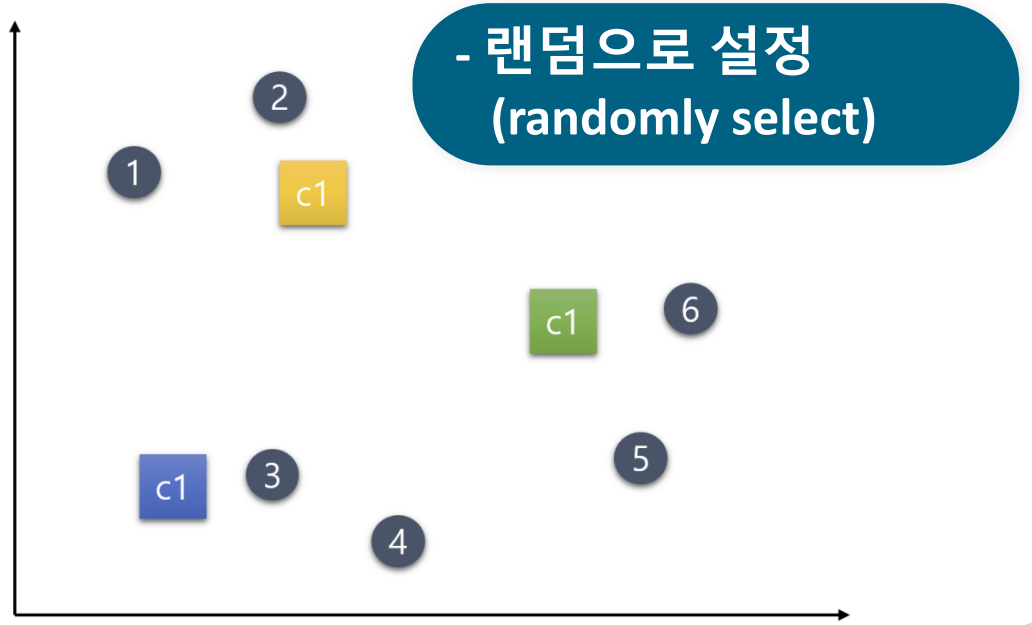
- 클러스터 중심(Center of Cluster, Centroid) = k-means 알고리즘이 만든 클러스터에 속한 샘플의 특성 평균값
- 초기 중심점으로 어떤 값을 선택하는가에 따라 성능이 크게 달라짐

1) Randomly select

2) Manually assign

3) K-means++

중심점의 위치가 더 이상 변하지 않을 때까지 반복



K-means 알고리즘

= 최적의 군집을 구성하는 대표적 군집 알고리즘

1) 군집의 개수(K) 설정

⇩

2) 초기 중심점 설정

⇩

3) 데이터를 군집에 할당(배정)

⇩

4) 중심점 재설정(갱신)

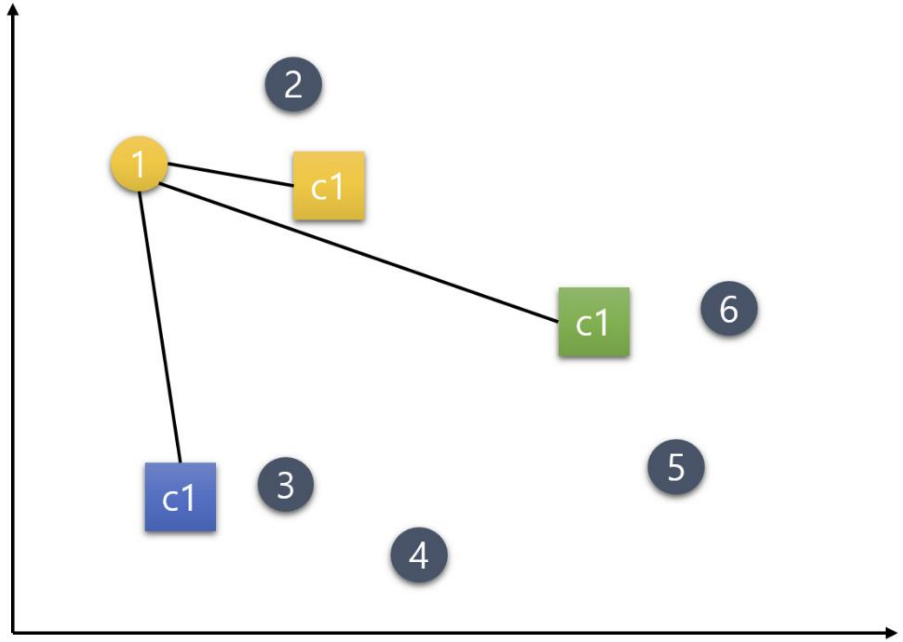
⇩

5) 데이터를 군집에 재할당(배정)

중심점의 위치가 더 이상 변하지 않을 때까지 반복

Step3) 데이터를 군집에 할당(배정)하기

- 거리 상 가장 가까운 군집(중심점)으로 주어진 모든 데이터를 할당 또는 배정함
- 거리 측정 방법 : (일반적으로) 유클리드 거리



K-means 알고리즘

= 최적의 군집을 구성하는 대표적 군집 알고리즘

1) 군집의 개수(K) 설정

2) 초기 중심점 설정

3) 데이터를 군집에 할당(배정)

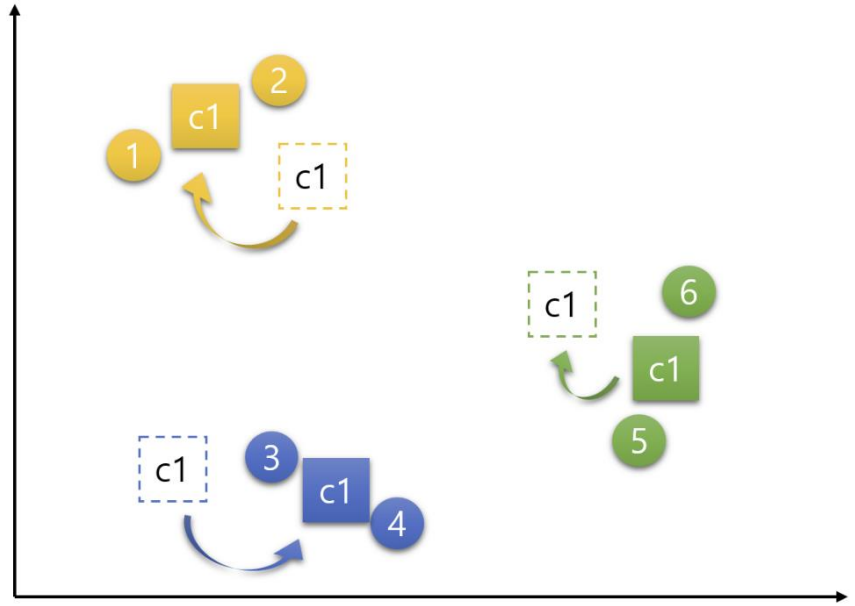
4) 중심점 재설정(갱신)

5) 데이터를 군집에 재할당(배정)

중심점의 위치가 더 이상 변하지 않을 때까지 반복

Step4) 중심점 재설정(갱신)하기

- 각각의 중심점은 그 군집의 속하는 데이터들의 가장 중간(평균)에 위치한 지점으로 재설정함
- 중심점들은 각 데이터들의 평균인 지점으로 갱신됨



K-means 알고리즘

= 최적의 군집을 구성하는 대표적 군집 알고리즘

1) 군집의 개수(K) 설정

2) 초기 중심점 설정

3) 데이터를 군집에 할당(배정)

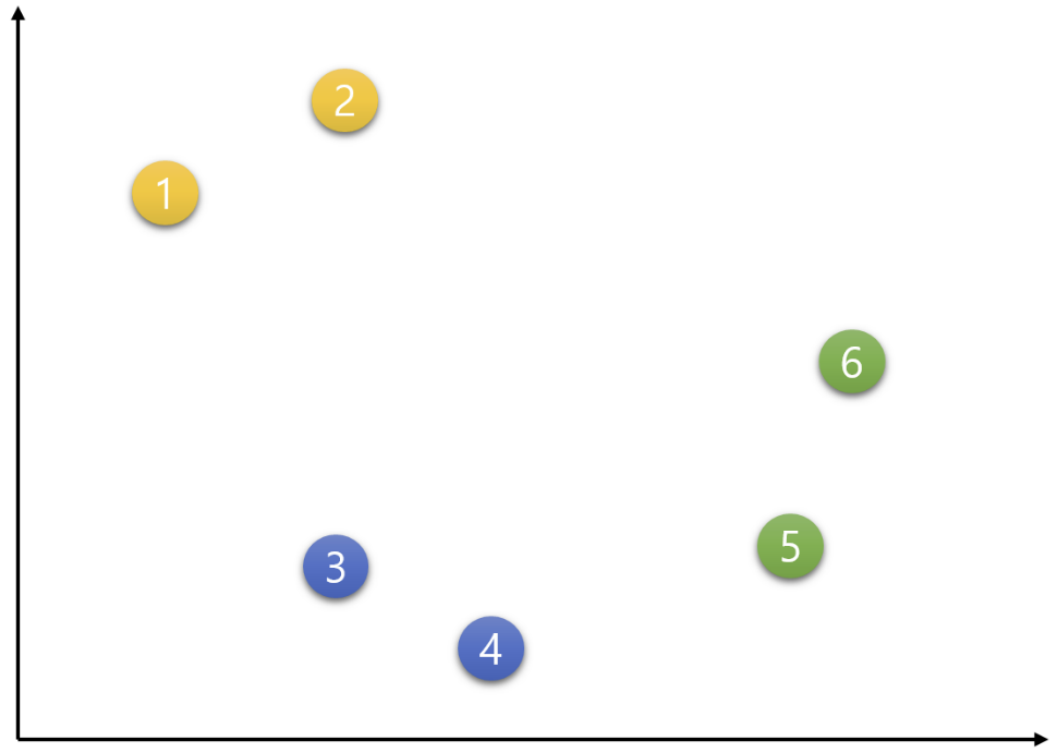
4) 중심점 재설정(갱신)

5) 데이터를 군집에 재할당(배정)

중심점의 위치가 더 이상 변하지 않을 때까지 반복

Step5) 데이터를 군집에 재할당(배정)하기

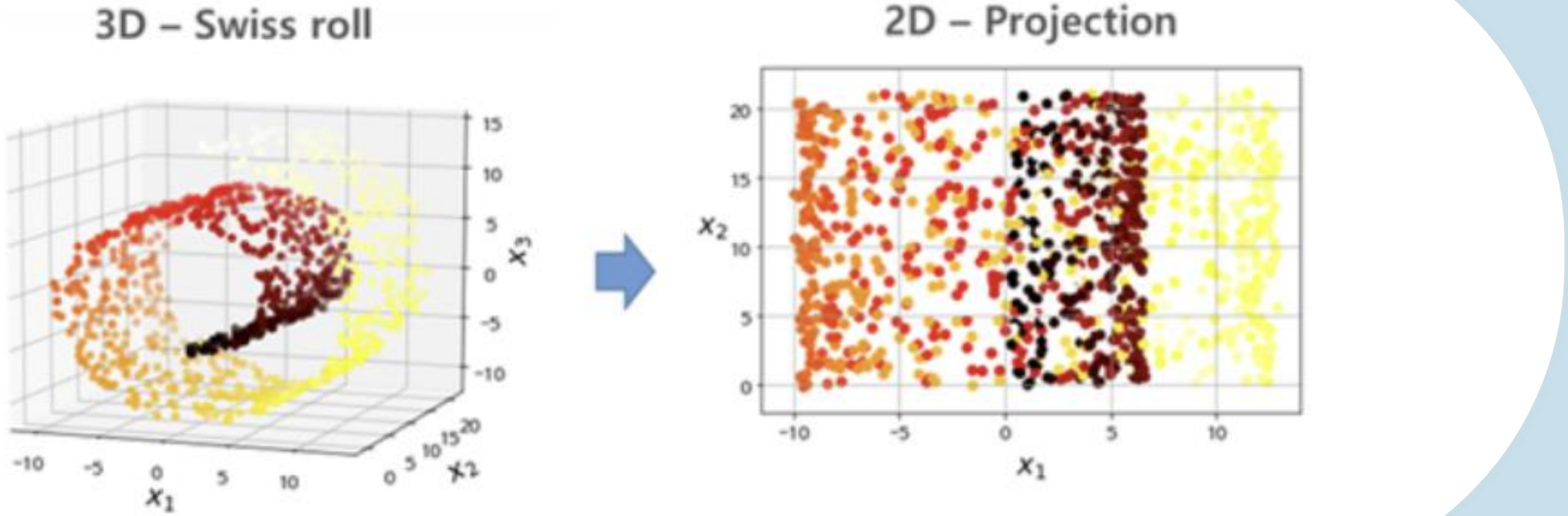
- 더 이상 중심점의 이동이 없을 때까지 step4)와 step5) 반복



차원 축소 (Demension Reduction)

= 원본 데이터의 특성을 적은 수의 새로운 특성으로 변환하는 비지도 학습

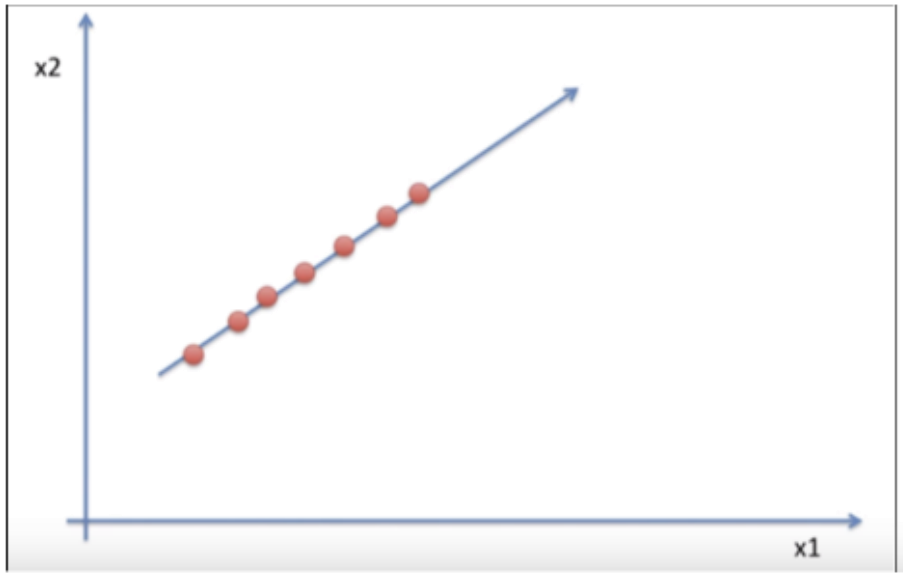
- 노이즈 데이터를 제거하기 위해 특성 전처리 단계에서 종종 적용함
- 데이터가 복잡하고 높은 차원을 가져 시각화하기 어려울 때 2,3차원으로 표현함

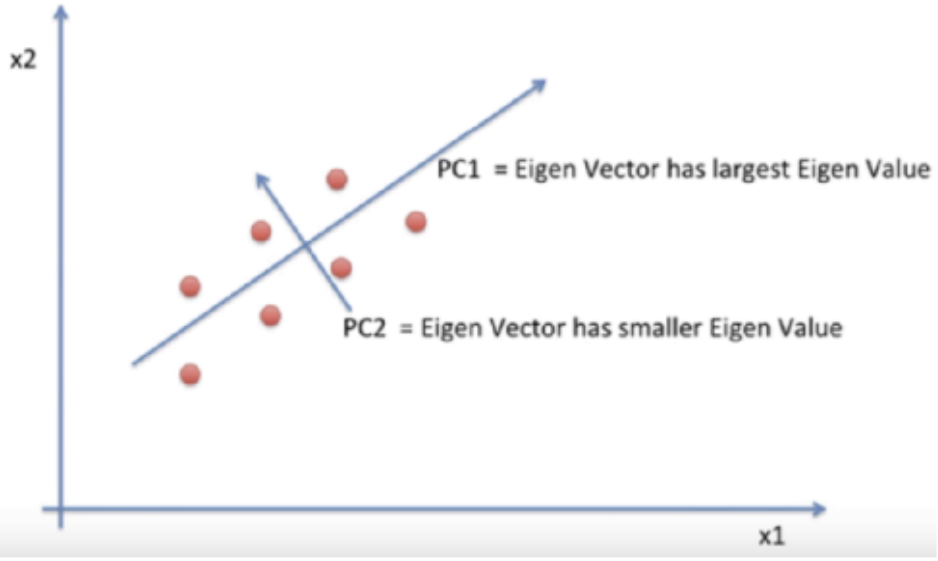


주성분 분석(PCA)

= 데이터에서 가장 분산이 큰 방향을 찾는 방법 / 차원 축소 알고리즘

- 주성분(PC) = 전체 데이터(독립변수들)의 분산을 가장 잘 설명하는 성분
- 변수의 개수 = 차원의 개수
- 변수가 너무 많아 기존 변수를 조합해 새로운 변수를 가지고 모델링을 하려고 할 때 주로 PCA를 사용함





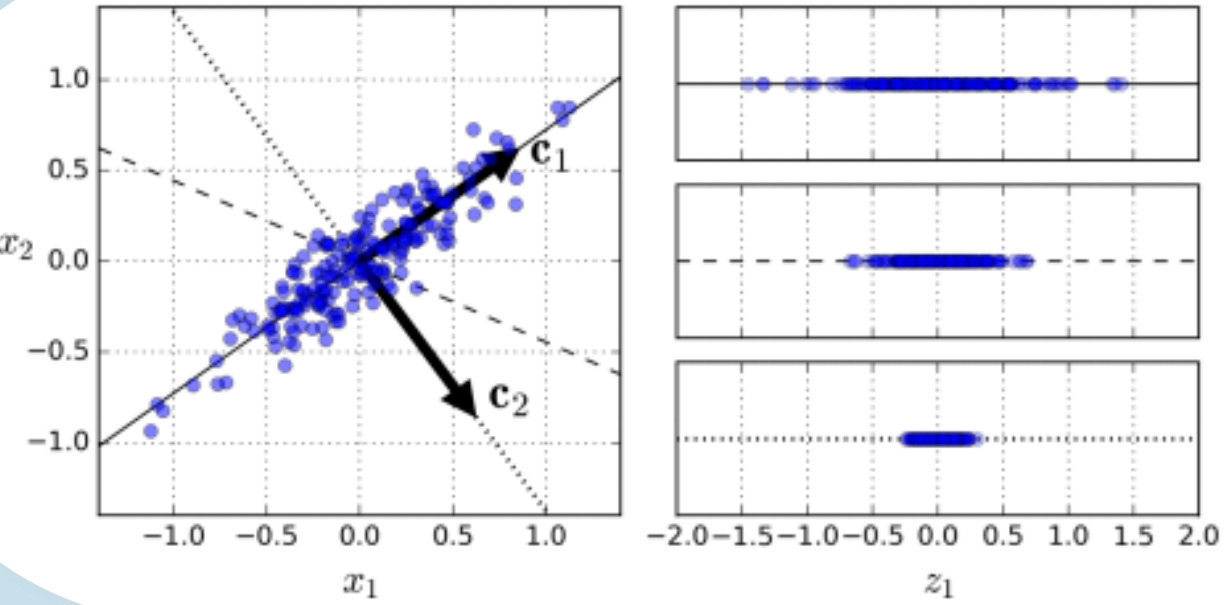
Q. 왜 분산을 최대로 보존할 수 있는 축을 선택하는가?

- 분산이 커져야 데이터들 사이의 차이점의 명확해짐
- PCA에서는

  첫번째 축 : 분산이 최대인 축

  두번째 축 : 첫번째 축에 직교하고 남은 분산을 최대한 보존하는 축

2차원 - 두번째 축에 대한 선택의 여지가 없음
고차원 – 여러 방향의 직교하는 축을 찾을 수 있음

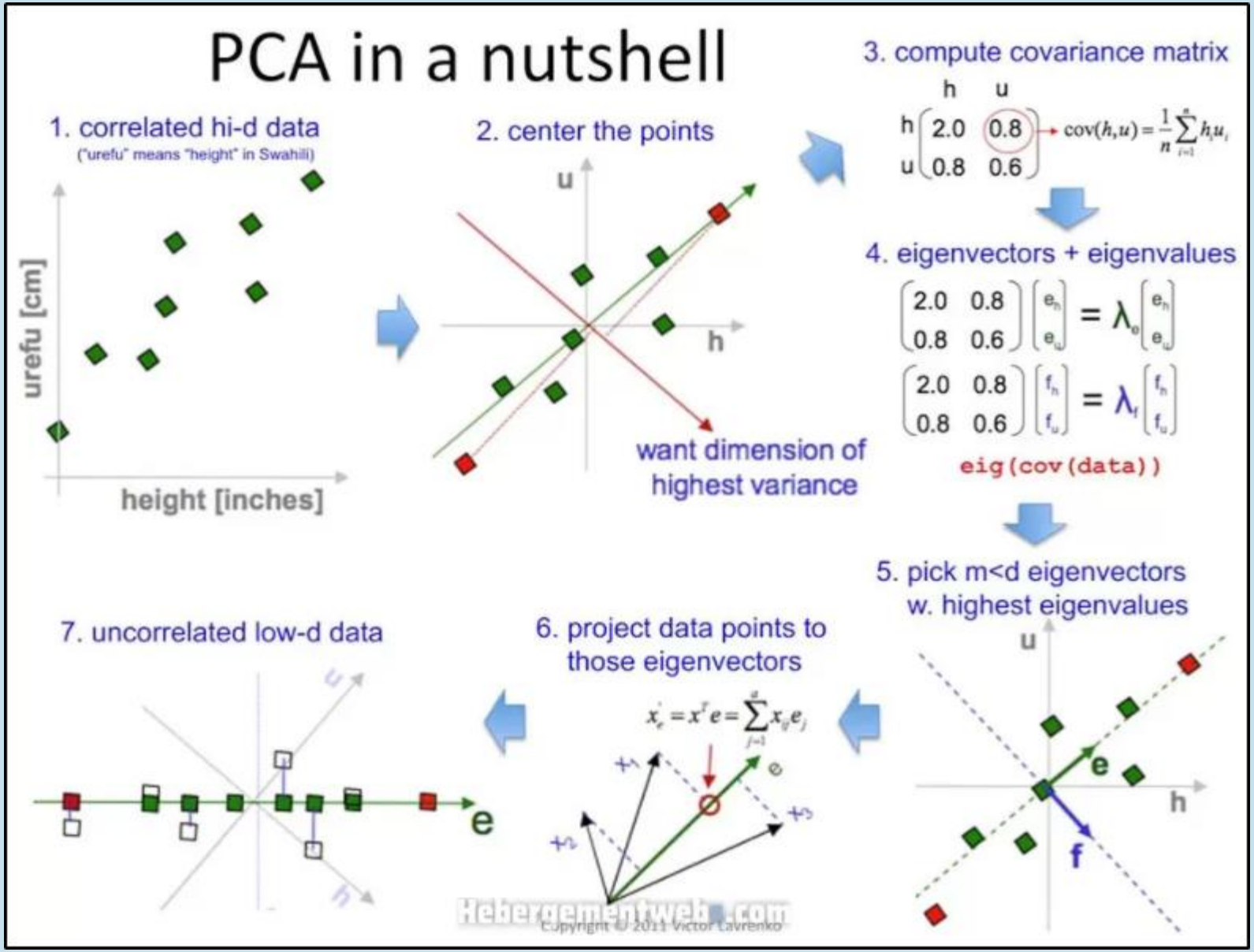


2D 데이터셋이 있을 때, 세 개의 축 후보

- 실선 선택 -> 분산 최대로 보존하는 것
- C2의 점선 선택 -> 분산을 적게 만들어버리는 것
- 첫번째 PC : c1 / 두번째 PC : c2

PCA 수행 과정

1. Mean centering
2. SVD(특잇값 분해) 수행하기
3. PC score 구하기
4. PC score를 설명변수로 활용하여 분석 진행하기



Thank You