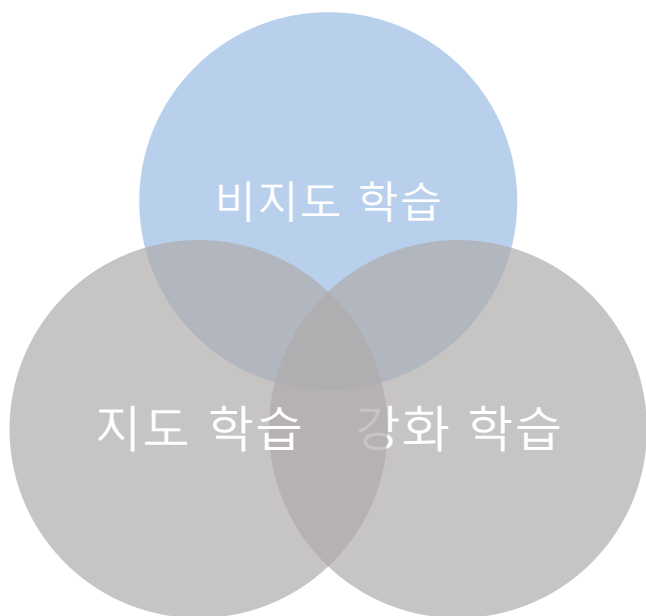


## 비지도학습(unsupervised learning)

---

(= 타깃이 없는 데이터 훈련)

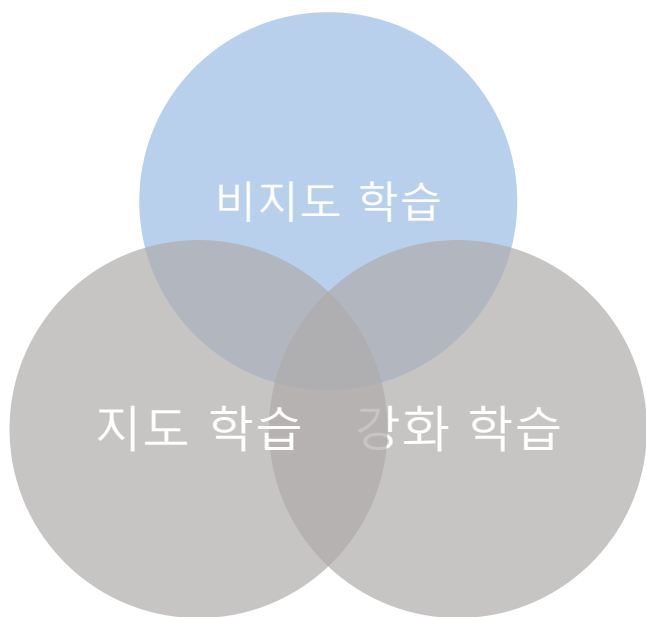
- 비지도 학습은 지도학습과 반대로 타깃 데이터가 없다.
- 많은 자료들을 비슷한 것으로 분류
- 쏟아지는 각종 소식들을 비슷한 주제에 따라 분류해주고 기업에서는 비슷한 성향의 고객을 묶어 관리할 수 있게 해준다.
- 그 대신 방대한 자료가 필요하다.



## 비지도학습 동작 방식

---

- 비지도 학습은 패턴을 찾고 데이터의 추세를 이해하여 결과를 발견합니다.
- 따라서 모델은 입력 데이터의 기능을 기반으로 데이터에 레이블을 지정하려고 합니다.
- 비지도 학습 기술에 사용되는 훈련 프로세스는 모델을 구축하기 위해 감독이 필요하지 않습니다.
- 스스로 학습하고 결과를 예측합니다.



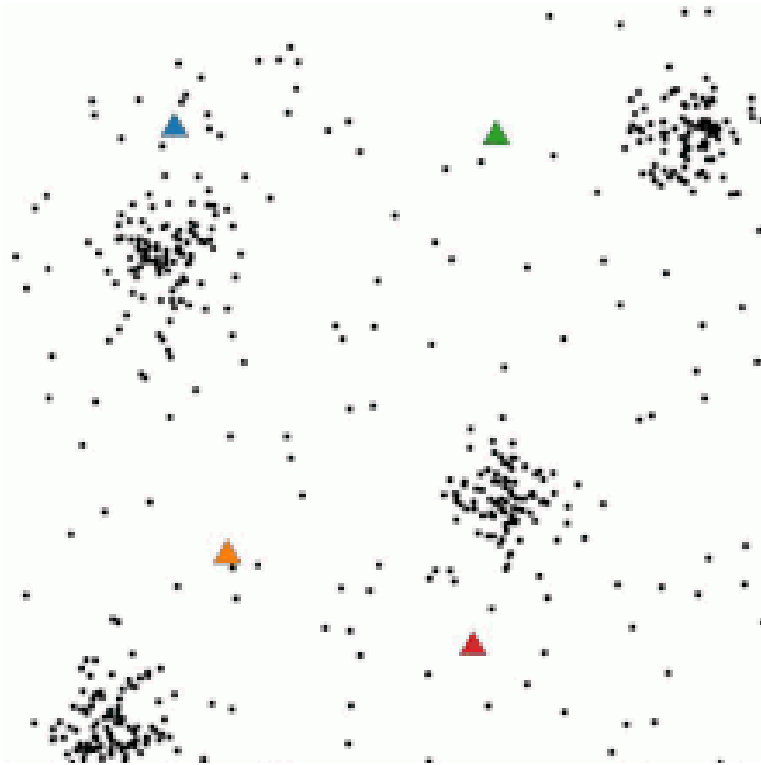
## 지도학습에서 많이 사용되는 알고리즘

---

- K - means clustering
- Hierarchical clustering
- Apriori algorithm
- DBSCAN DBSCAN
- Principal Component Analysis

## 비지도학습 알고리즘 : K - means clustering

---



- 클러스터링의 대표적인 알고리즘은 K-means 알고리즘입니다.
- 알고리즘은 아래와 같이 4단계로 구성됩니다.
- 1. 초기 k 평균 값(위의 경우  $k = 3$ )은 데이터 오브젝트 중에서 무작위로 뽑습니다.
- 2. k 각 데이터 오브젝트들은 가장 가까이 있는 평균값을 기준으로 묶으며 평균값을 기준으로 분할된 영역은 보로노이 다이어그램으로 표시합니다.
- 3. k의 클러스터의 중심점을 기준으로 평균값이 재 조정됩니다.
- 4. 수렴할 때까지 2, 3 과정을 반복합니다.

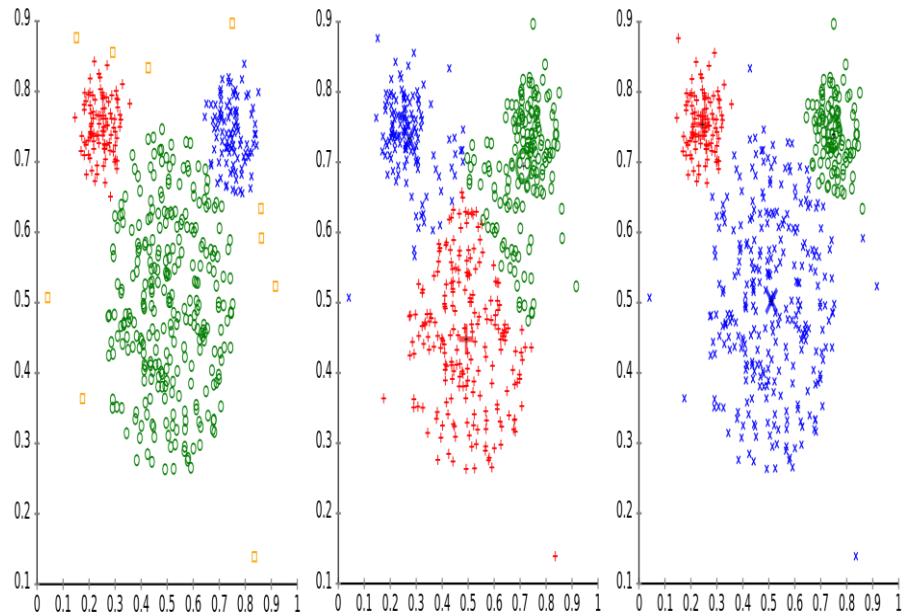
## 비지도학습 알고리즘 : K - means clustering 한계점

Different cluster analysis results on "mouse" data set:

Original Data

k-Means Clustering

EM Clustering



- K - 평균 알고리즘은 몇가지 한계점을 가지고 있는데 이러한 단점을 극복하기 위한 개선 알고리즘이 존재합니다.
- 1. 클러스터 개수 k값을 입력 파라미터로 지정해주어야 합니다.
- 2. 알고리즘의 에러 수렴이 전역 최솟값이 아닌 지역 최솟값으로 수렴할 가능성이 있습니다.
- 3. 이상 값에 민감합니다.
- 4. 구형이 아닌 클러스터를 찾는 데에는 적절하지 않습니다.

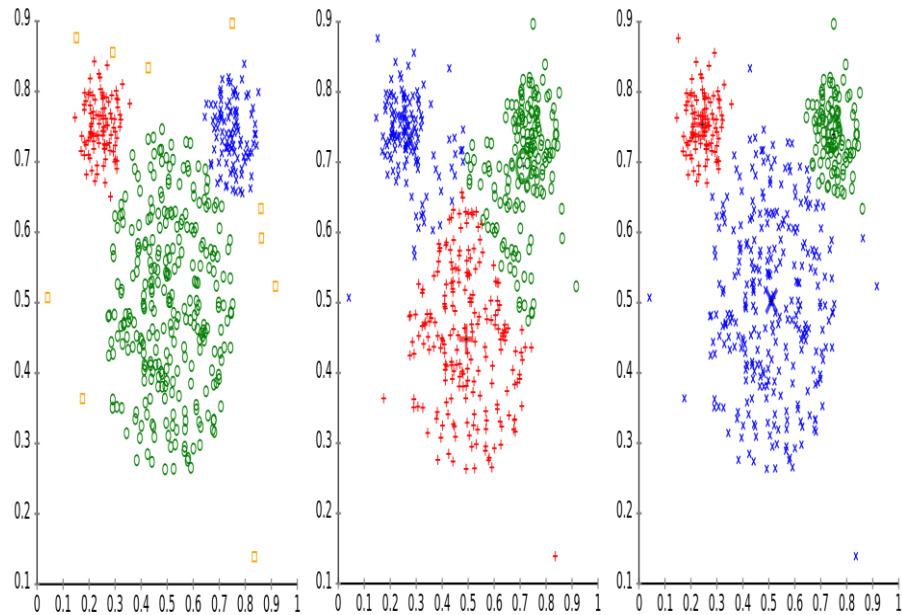
## 비지도학습 알고리즘 : K - means clustering 알고리즘 구현

Different cluster analysis results on "mouse" data set:

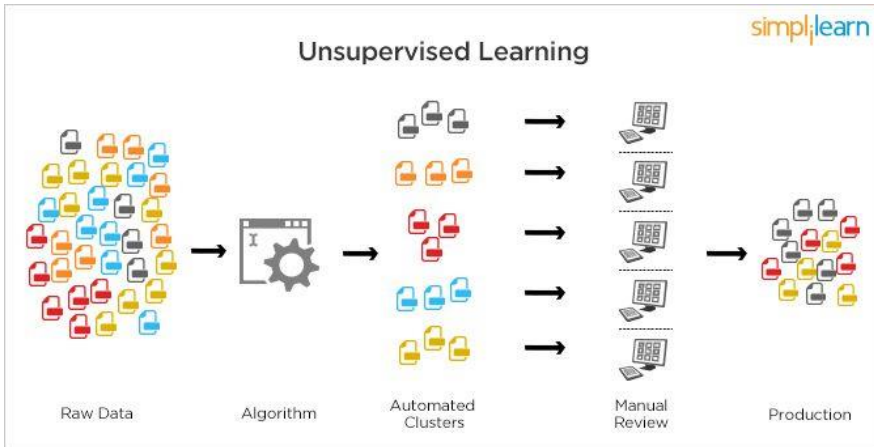
Original Data

k-Means Clustering

EM Clustering



– <https://kibua20.tistory.com/214>



## - 비지도학습 알고리즘 : Apriori algorithm

- Apriori algorithm은 association rule learning을 하기 위해서 자주 쓰이는 알고리즘 입니다.
- 예를 들자면 물건을 구매한 사람이 또 구매하게 된다 입니다.
- 연관성을 찾아서 그 뒤에 사람들이 어떤 물건을 살 가능성이 높은 지 분석하는 것입니다.
- 빈도 -> 척도 -> 연관을 통해 데이터에 대한 발생빈도를 기반으로 각 데이터 간의 연관 관계를 밝히기 위한 수학적 접근 알고리즘 입니다.
- 이런 알고리즘은 통신회사에서 이반 고객 예상이나 기지국 위치 선정, 유통업체에서 매장 진열 방법이나 고객 장바구니 분석, 의료에서는 환자의 질병 예측이나 약품 부작용 분석, 카드회사에서는 대출 심사나 카드 연체 고객 예상 등에서 활용되고 있습니다.

Movie recommendation:

$$\text{Support (M)} = \frac{\text{\# user watchlists containing M}}{\text{\# user watchlists}}$$

Market Basket Optimisation:

$$\text{Support (I)} = \frac{\text{\# transaction containing I}}{\text{\# transactions}}$$

Movie recommendation:

$$\text{Confidence (M1} \rightarrow \text{M2)} = \frac{\text{\# user watchlists containing M1 and M2}}{\text{\# user watchlists containing M1}}$$

Market Basket Optimisation:

$$\text{Confidence (I1} \rightarrow \text{I2)} = \frac{\text{\# transaction containing I1 and I2}}{\text{\# transactions containing I1}}$$

Movie recommendation:

$$\text{Lift (M1} \rightarrow \text{M2)} = \frac{\text{confidence (M1} \rightarrow \text{M2)}}{\text{support (M2)}}$$

Market Basket Optimisation:

$$\text{Lift (I1} \rightarrow \text{I2)} = \frac{\text{confidence (I1} \rightarrow \text{I2)}}{\text{support (I2)}}$$

## - 비지도학습 알고리즘 : Apriori algorithm 분석 세가지 단계

### - Support 단계

- -> 100명의 사람들 중에 가디언스 오브 갤럭시를 본 사람은 10명이라고 하면 이때  $\text{Support} = 10 / 100$  으로 10%가 됩니다.

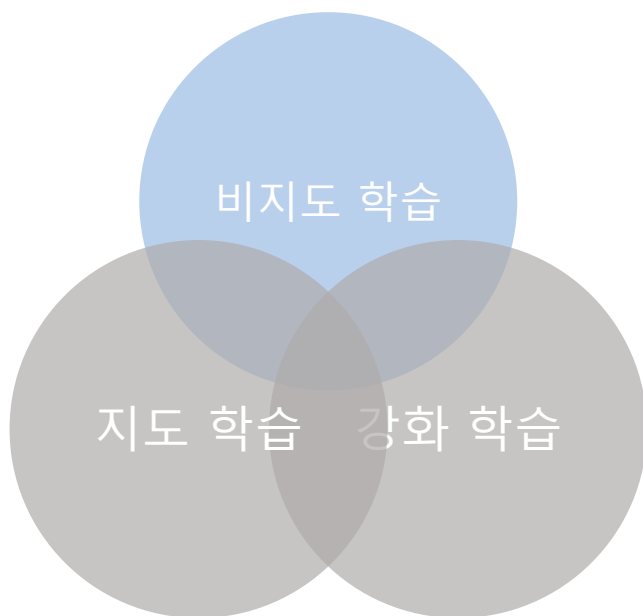
### - Confidence 단계

- -> 100명의 사람들 중에 스타워즈를 본 사람은 40명이라고 하면 이 40명 중에서 갤럭시 오브 가디언스를 본 사람이 7명 있다고 하면 이때  $\text{Confidence} = 7 / 40$  으로 17.5%가 됩니다.

### - Lift 단계

- -> 무작위 100명의 사람들 중에 가디언스 오브 갤럭시를 보라고 권유했을 때 보게 될 가능성은 10명이 됩니다. 하지만 스타워즈를 본 40명에게 가디언스 오브 갤럭시를 보라고 권유했을 경우 40명 중에서 보게될 가능성은 7명이 됩니다.



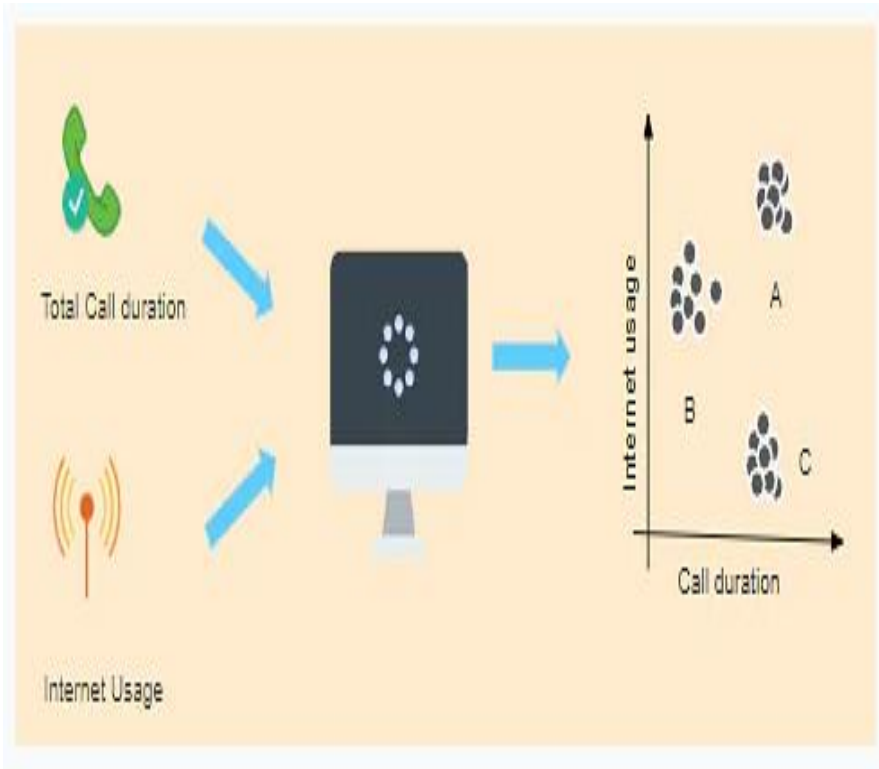


## 비지도학습의 두가지 형태 : 클러스터링

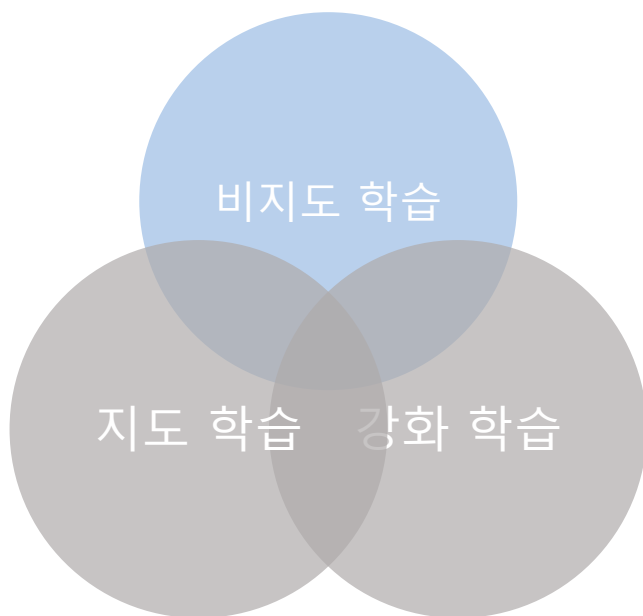
---

- 클러스터링은 객체를 유사한 클러스터로 나누고 다른 클러스터에 속하는 객체와 유사하지 않은 클러스터로 나누는 방법입니다.
- 예를 들어, 유사한 제품을 구매한 고객을 찾는 것입니다.

# 비지도학습의 두가지 형태 : 클러스tring 예시



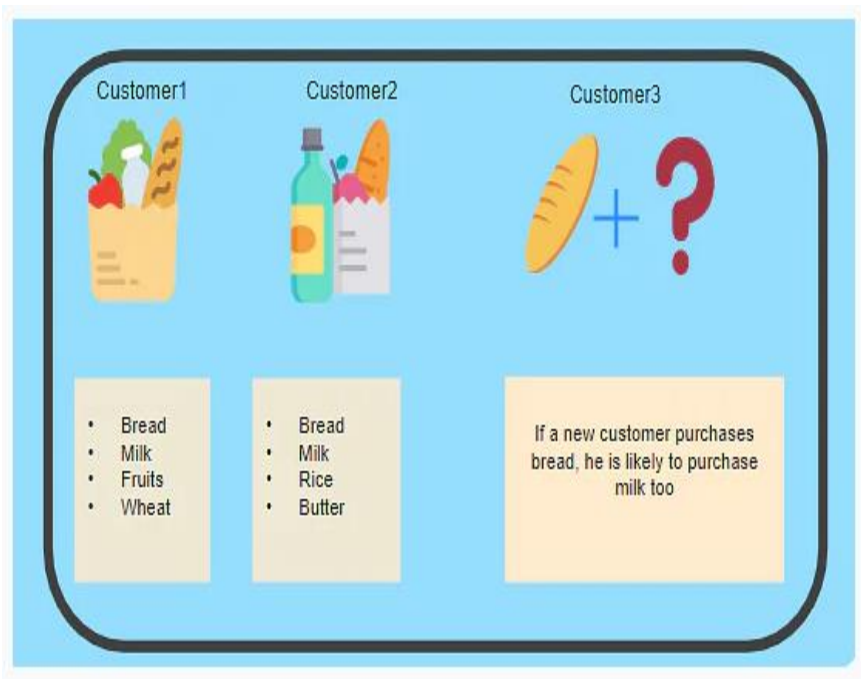
- 통신 회사가 개인 통화 및 데이터 요금제를 제공하여 고객 이탈률을 줄이려고 한다고 가정해 보겠습니다.
- 적절한 프로모션과 캠페인을 통해 해지율을 최소화하고 수익을 극대화 하기 위해서 여러 전략이 채택됩니다.
- 이미지 오른쪽에서 고객을 그룹화한 그래프를 볼 수 있습니다.
- 그룹 A고객은 더 많은 데이터를 사용하고 또한 긴 통화 시간을 가집니다.
- 그룹 B고객은 인터넷 사용자가 많고 그룹 C고객은 통화 시간이 많습니다.
- 따라서 그룹 B에는 더 많은 데이터혜택 플랜이 제공되고 그룹 C에는 더 저렴한 통화 요금 플랜이 제공되고 그룹 A에는 두 가지 혜택 모두 제공됩니다.



## 비지도학습의 두가지 형태 : 연관

---

- 연관은 컬렉션에서 항목이 동시에 발생할 확률을 발견하는 규칙 기반 머신 러닝입니다.
- 예를 들어 함께 구매한 제품을 찾는 것입니다.



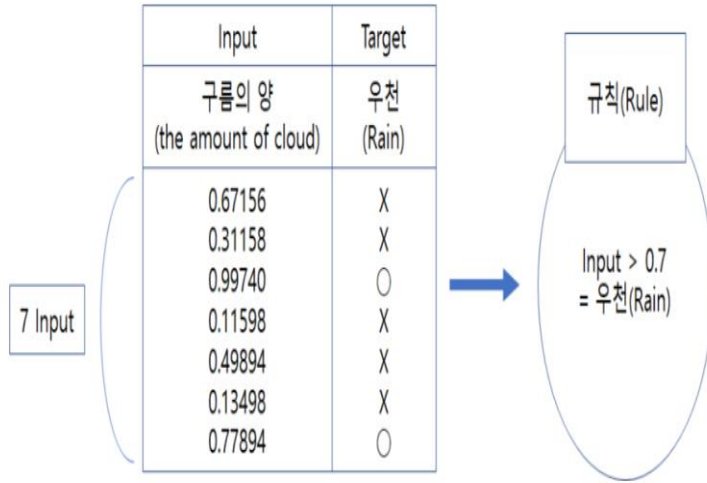
## 비지도학습의 두가지 형태 : 연관 예시

- 고객이 슈퍼마켓에 가서 빵, 우유, 과일, 밀을 산다고 가정해 보겠습니다.
- 또 다른 손님이 와서 빵, 우유, 쌀, 버터를 삽니다.
- 이제 다른 고객이 방문 했을 때 빵을 사면 우유도 사게 될 가능성이 큼니다.
- 따라서 고객 행동을 기반으로 관계가 설정되고 권장 사항이 만들어 집니다.

# 비지도학습의 예시1 : 군집화

(= 그룹분류)

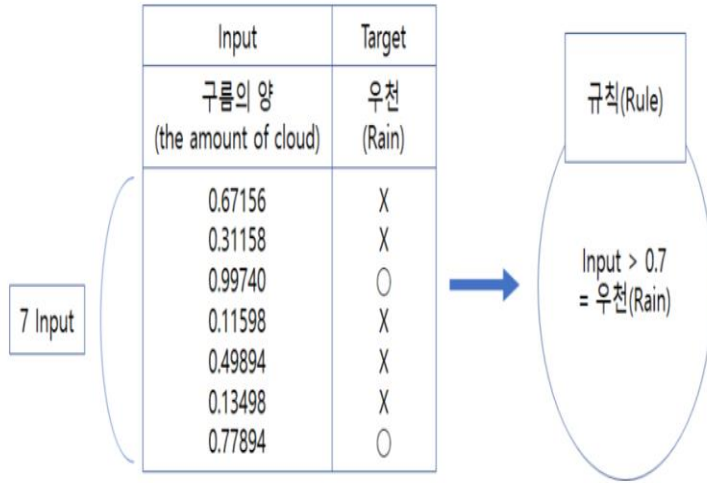
- 그룹을 만들기 전까지는 어떤 그룹도 존재하지 않고 어떤 그룹이 있을지도 모르기 때문에 타깃이 없는 것과 같다.
- 타깃이 없는 상태에서 프로그램은 해당 입력들을 특성에 맞게 분류하여 군집을 만든다.
- 군집화는 훈련 데이터 집합에서 서로 유사한 것들끼리 묶어서 군집을 형성할 함으로서 데이터의 집합을 분할하는 것을 말합니다.
- 이 유사성은 데이터 간의 거리를 가지고 판단할 수 있습니다.



# 비지도학습의 예시1 : 군집화

(= 그룹분류)

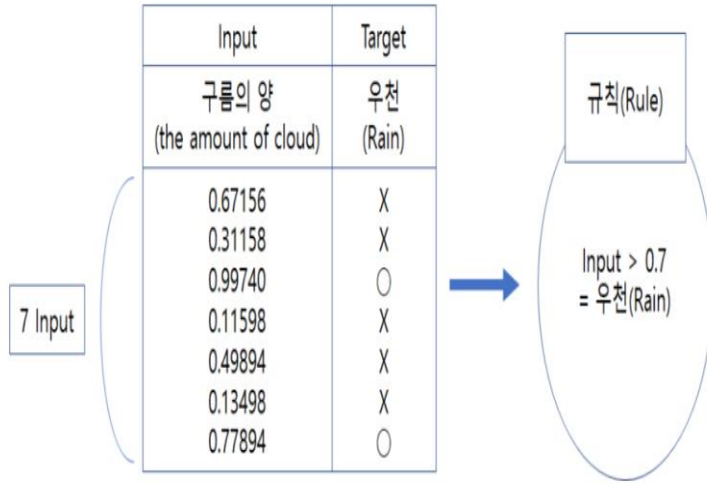
- 군집화 알고리즘의 기본 틀은 같은 군집에 속한 데이터 간의 거리는 최소로 줄이고 다른 군집의 데이터와의 거리는 최대로 늘리기 위하여 군집 소속을 바꿔가면서 최적의 군집을 찾는 것입니다.
- 소비자의 구입이력을 사용하여 새로운 상품을 추천하는 데에 이 군집화 방법이 사용됩니다.
- 소비자 A와 B가 유사한 구매패턴을 가지고 있다는 것을 학습하였다면 소비자 A가 구입한 물건은 B도 구매할 확률이 높다는 전제로 추천을 수행하는 것입니다.



# 비지도학습의 예시1 : 군집화

(= 그룹분류)

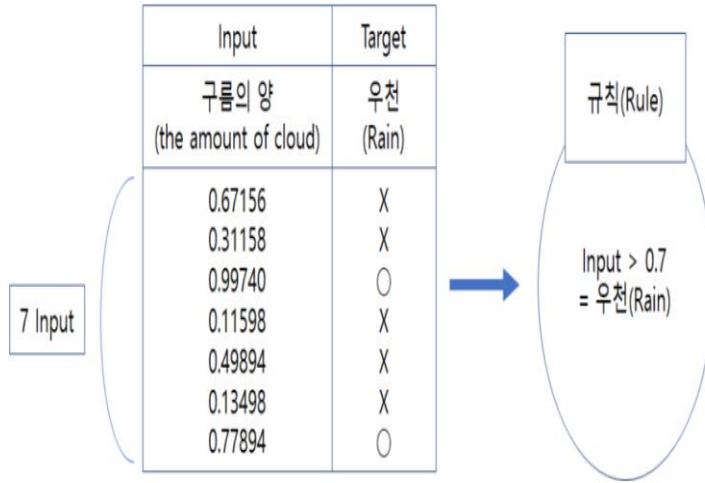
- 클러스터링 또는 군집 분석 알고리즘인 K-means다.
- 데이터들이 아래 그림과 같이 분포되어 있다고 할 때 자동으로 그룹화하기 위해 사용한다,
- 그래프 상에 임의의 중심점 몇 개를 찍은 후 각 점을 중심으로 가장 가까운 점부터 그룹에 편입시킨다.



# 비지도학습의 예시1 : 군집화

(= 그룹분류)

- 이렇게 만들어진 그룹의 평균값을 구하고 다시 새로운 중심점을 설정하면서 그룹들이 만들어진다.
- 이렇게 만들어진 그룹의 평균값을 구하고 다시 새로운 중심점을 설정하면서 그룹들이 만들어진다.
- 이런 과정을 반복하면서 데이터 셋이 갖는 고유한 군집 특성이 차츰 드러나게 된다.

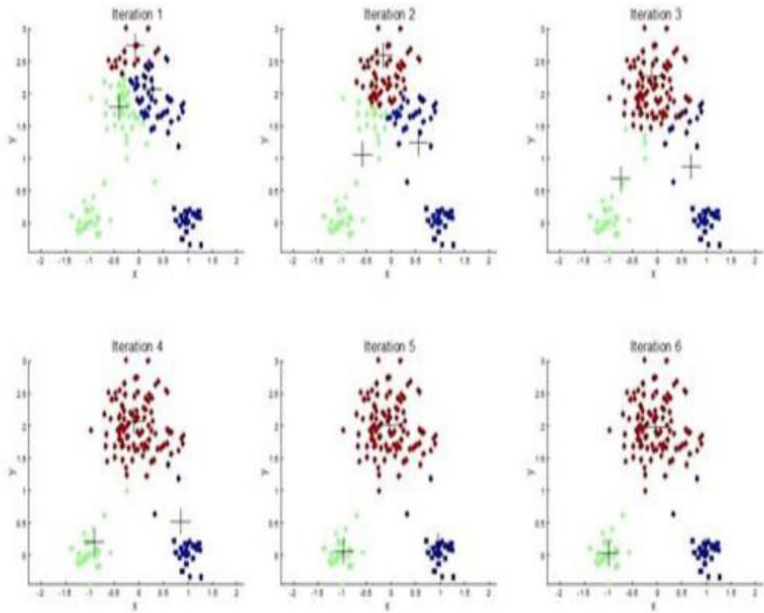




# 비지도학습의 예시2 : 차원감소

(Dimensionality Reduction)

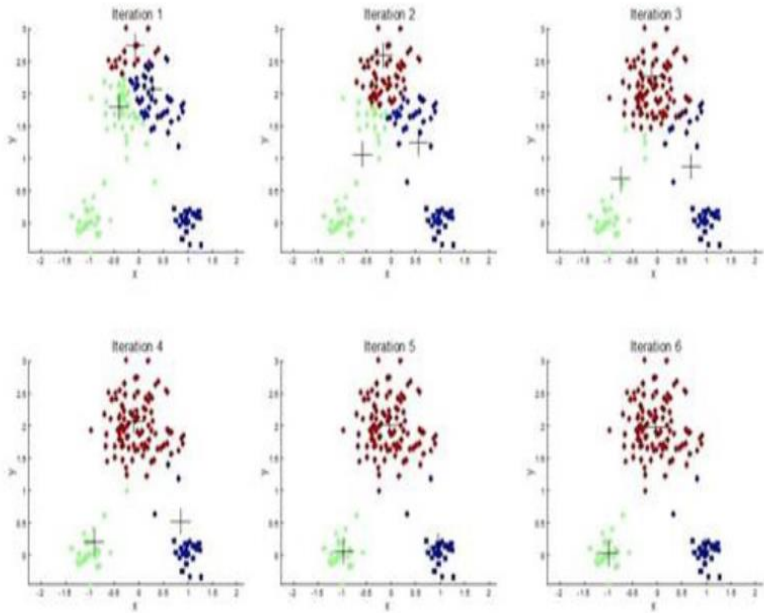
- 차원이란 데이터 특성의 개수를 말합니다.
- 너무 많은 특성으로 데이터를 표현하면 찾으려고 하는 해당 데이터의 본질을 찾지 못하기 때문에 꼭 필요한 특성으로 간추려서 표현하는 방법을 차원감소라고 합니다.
- 예를 들어 여러 각도에서 찍은 얼굴 사진을 따로 분류 위해 2차원 공간으로 공통된 특징만을 매핑하는 것 등이죠.



# 비지도학습의 예시2 : 차원감소

(Dimensionality Reduction)

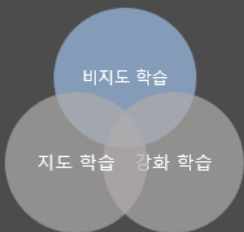
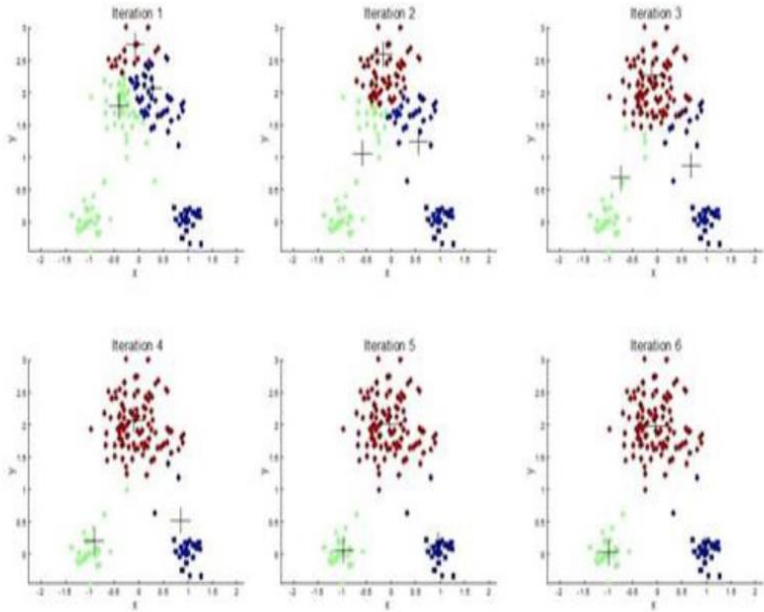
- 변환 과정에서 명암 배치, 그림자 위치, 점의 개수 등 덜 중요한 정보를 잃어버리긴 하겠지만 눈, 코, 입의 배열, 얼굴형, 머리카락 등의 중요한 특징은 유지시켜서 데이터 분류를 쉽게 하는 것이다.
- 주성분 분석은 높은 차원의 공간을 낮은 공간으로 축소하는 분석방법이다. 변환 과정에서 정보를 잃어버리긴 하지만 중요한 특징은 유지되고 데이터 분별이 쉬워질 것으로 기대한다.



# 비지도학습의 예시2 : 차원감소

(Dimensionality Reduction)

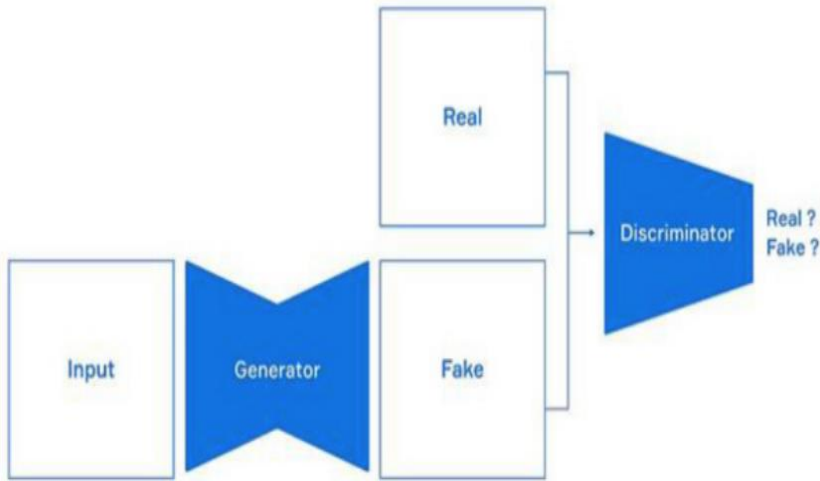
- 오토엔 코더는 데이터 차수 감소를 목적으로 자주 사용되는 인공 신경망 기법이다. 라벨이 없는 학습데이터에 대해 입력과 동일한 출력을 내도록 지도 학습으로 훈련시킨다는 아이디어에서 출발한다.
- 비선형 변환이 가능하고 강력한 특성을 추출하기 위한 의미적 변환으로 볼 수 있다.



# 비지도학습의 예시3 : GAN

(Generative Adversarial Network)

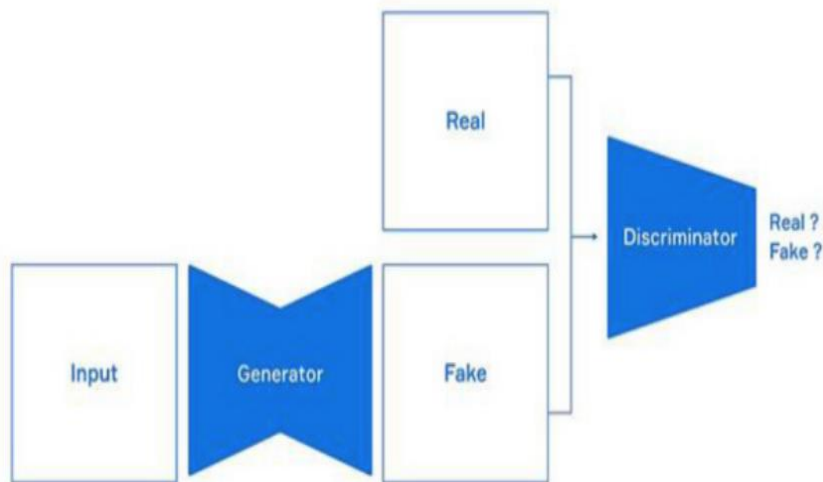
- 진짜 같은 가짜를 생성하는 알고리즘입니다.
- 진짜 데이터를 학습하여 진짜와 유사한 완전한 가짜 데이터를 생성하는 제너레이터(Generator)와 진짜 데이터와 가짜 데이터를 판별하는 디스크리미네이터(Discriminator)를 두어 서로 경쟁하도록 하였습니다.



# 비지도학습의 예시3 : GAN

(Generative Adversarial Network)

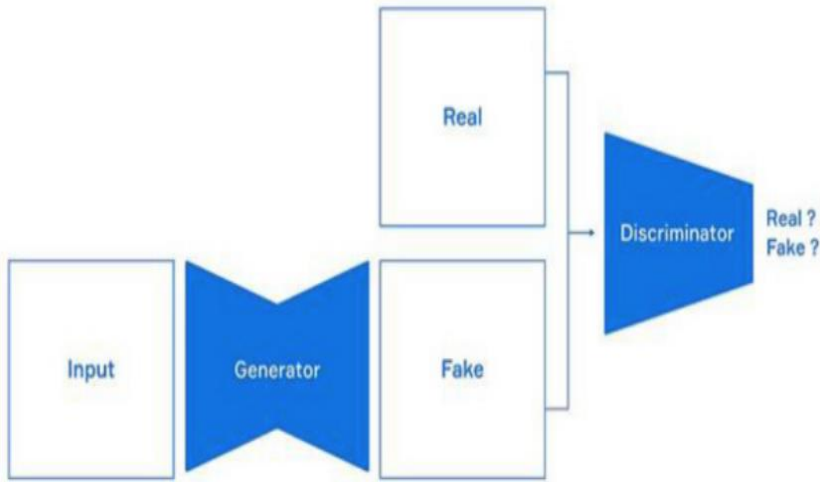
- 제너레이터는 디스크리미네이터가 틀린 데이터의 특징 또는 맞춘 데이터의 특징을 계속 학습하여 감별자가 맞추기 어렵게 계속 진화합니다.
- 디스크리미네이터는 제너레이터가 만든 가짜의 특징을 계속 학습하여 판별 능력을 향상합니다.
- 결국 서로 경쟁하는 과정을 통해 진짜와 가짜를 구분할 수 없는 수준의 산출물을 만들게 되어, 디스크리미네이터가 맞추는 확률이 50%가 되면 학습이 마무리 됩니다.

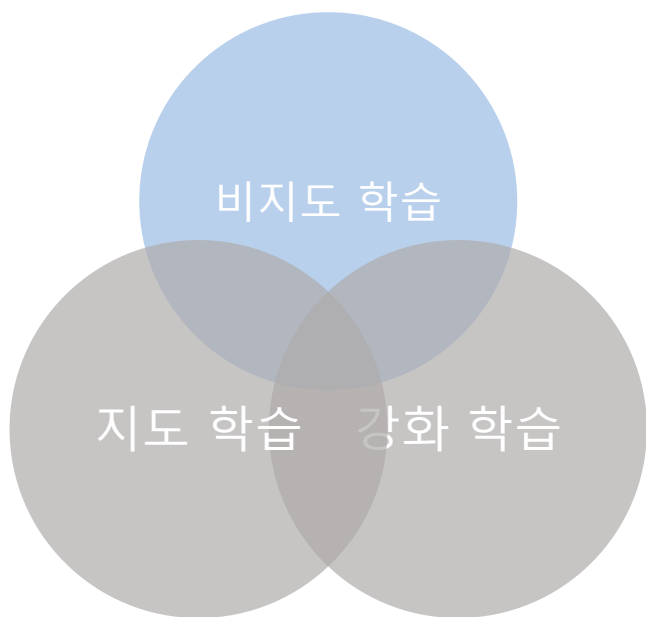


# 비지도학습의 예시3 : GAN

(Generative Adversarial Network)

- GAN은 저해상도 이미지를 고해상도로 개선하거나 음성 복원, 게임 배경화면 생성, 예술 창작활동 등 다양한 분야에 쓰일 것으로 기대되는 기술입니다.

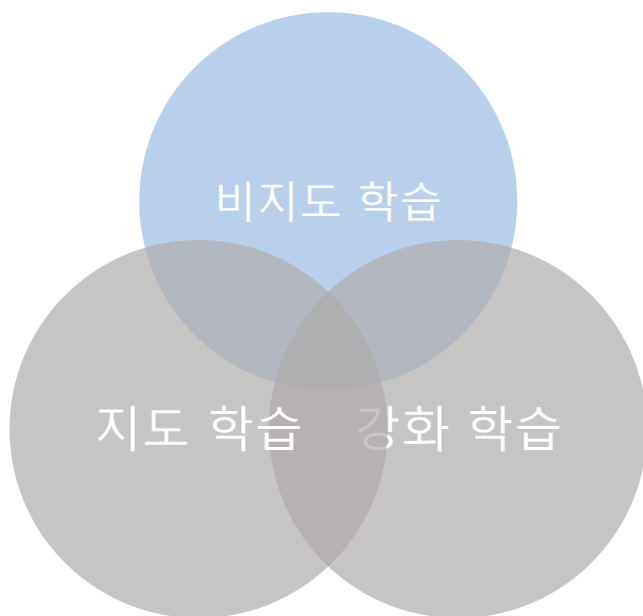




## 비지도학습의 실제 적용

---

- 시장 바구니 분석 (Market Basket Analysis)
  - > 특정 그룹의 품목을 구매하면 다른 그룹의 품목을 구매할 가능성이 적거나 더 높아진다는 알고리즘을 기반으로 하는 머신 러닝 모델입니다.
- 시맨틱 클러스터링(Semantic Clustering)
  - > 의미적으로 유사한 단어는 비슷한 맥락을 공유합니다. 사람들은 각자의 방식으로 웹사이트에 검색어를 게시합니다. 시맨틱 클러스터링은 이러한 모든 응답을 클러스터에서 동일한 의미로 그룹화하여 고객이 원하는 정보를 빠르고 쉽게 찾을 수 있도록 합니다. 정보 검색, 멋진 브라우징 경험 및 이해에 중요한 역할을 합니다.



## 비지도학습의 실제 적용

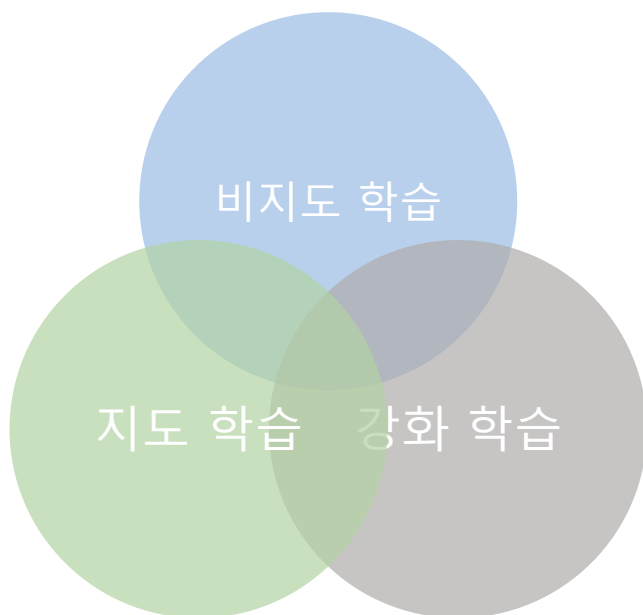
---

- 배달 매장 최적화 (Delivery Store Optimization)
  - > 머신 러닝 모델은 수요를 예측하고 공급을 따라가는 데 사용됩니다. 또한 수요가 더 높은 매장을 열고 과거 데이터 및 행동에 따라 보다 효율적인 배송을 위해 경로를 최적화하는 데 사용됩니다.
- 사고취약지역 파악 (Identifying Accident Prone Areas)
  - > 비지도 머신 러닝 모델을 사용하여 사고가 발생하기 쉬운 영역을 식별하고 해당 사고의 강도에 따라 안전 조치를 도입할 수 있습니다.



# 지도학습과 비지도 학습 차이점

---



- 기계학습 알고리즘은 훈련데이터의 쌍과 그 사용방법에 따라 지도학습과 비지도 학습으로 나뉩니다.
- 지도학습은 입력 값으로 알려진 라벨링 된 데이터를 사용하지만 비지도 학습은 입력으로 라벨링 되지 않은 데이터를 사용합니다.
- 지도학습은 피드백 메커니즘을 가지고 있지만 비지도 학습은 피드백 메커니즘이 없다.