데이터 과학

L09: Clustering

Kookmin University

지도학습 Supervised Learning

훈련 데이터(Training Data)로부터 하나의 함수를 유추해내기 위한 기계 학습(Machine Learning)의 한 방법

Training Data

```
[1.2,  3.8, -1.4, …,  4.1] →  1.1
[3.2, -1.2, -0.2, …,  2.1] →  2.7
[2.8, -1.4, -0.3, …,  2.3] →  2.8
[1.2,  3.4, -1.5, …,  4.2] →  0.9
[4.2,  2.1,  2.8, …, -0.5] → -0.1
                    …
[3.2,  2.2,  2.2, …, -0.4] → -0.2
```

Test

```
[1.3,  3.2, -1.5, …,  4.1] → ?
```

비지도학습 Unsupervised Learning

“데이터 패턴 학습”

기계 학습의 일종으로, 데이터가 어떻게 구성되었는지를 알아내는 문제의 범주에 속함

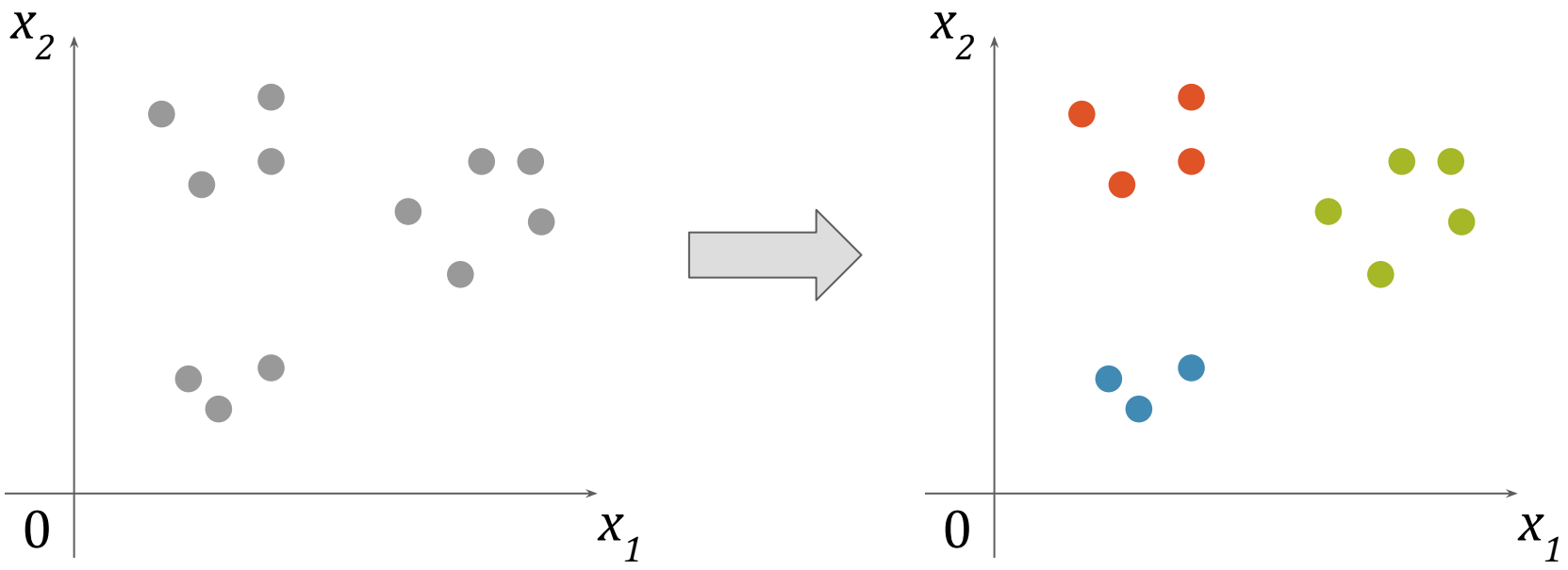
- Clustering
- Dimensionality Reduction
- Association Analysis

Data

```
[1.2,  3.8, -1.4, …,  4.1] →  ?
[3.2, -1.2, -0.2, …,  2.1] →  ?
[2.8, -1.4, -0.3, …,  2.3] →  ?
[1.2,  3.4, -1.5, …,  4.2] →  ?
[4.2,  2.1,  2.8, …, -0.5] →  ?
                    …
[3.2,  2.2,  2.2, …, -0.4] →  ?
```

클러스터 분석 Clustering

다차원 공간에서 여러개의 점들이 존재할 때,
서로 가까이 있는 점들을 서로 연관시키는 문제



클러스터 분석 활용 예

- 인물 사진 분류
  - 인물사진들 중에 닮은 사진 모으기

같은 사람인가요, 다른 사람인가요?

Google 포토

출처: http://www.gearbax.com/19361

클러스터 분석 활용 예

- 비슷한 뉴스 모으기
- 스팸메일 분류
- 비슷한 성향의 사용자/영화 모으기
- 사진 압축

출처: http://norman3.github.io/prml/docs/chapter09/1.html

K-Means Clustering

반복적인 연산을 통해 데이터를 k 개의 클러스터로 분할하는 알고리즘

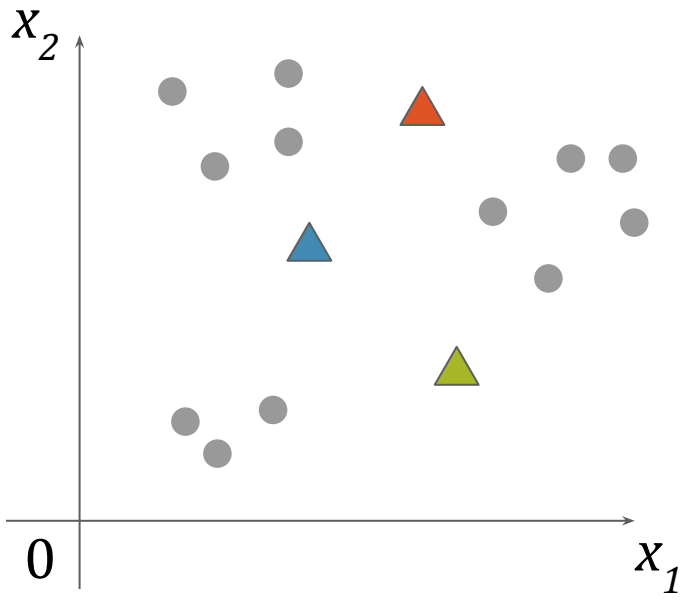
- 클러스터 분석 알고리즘
- 분할법 (partitioning)
- 클러스터 개수 (k) 지정 필요
- 반복연산 (iterative process)

K-Means Clustering

1. 임의로 k 개의 중심점(centroid)을 생성
2. 각각의 점을 가장 가까운 중심점의 클러스터에 포함시킴
3. 각 클러스터에 포함된 점들을 평균내어 새로운 중심점을 계산
4. 2-3 을 반복하다가 클러스터에 변화가 없으면 종료

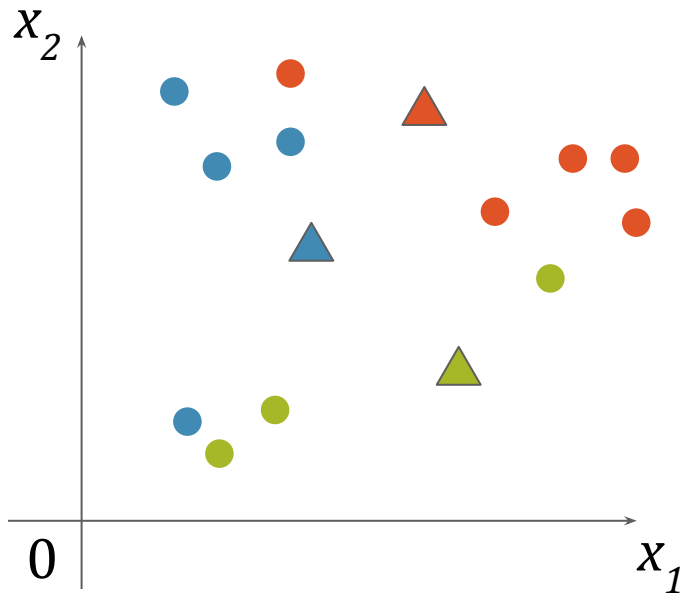
K-Means Clustering

1. **임의로 k 개의 중심점(centroid)을 생성**
2. 각각의 점을 가장 가까운 중심점의 클러스터에 포함시킴
3. 각 클러스터에 포함된 점들을 평균내어 새로운 중심점을 계산
4. 클러스터에 변화가 없으면 종료



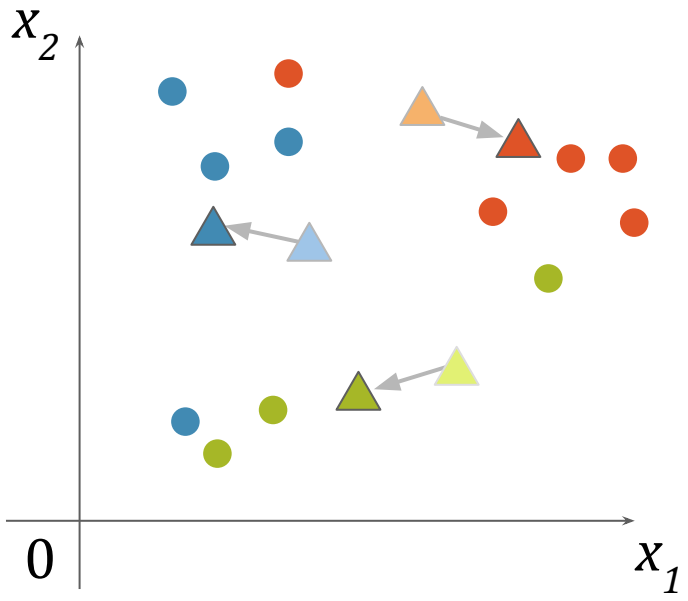
K-Means Clustering

1. 임의로 k 개의 중심점(centroid)을 생성
2. **각각의 점을 가장 가까운 중심점의 클러스터에 포함시킴**
3. 각 클러스터에 포함된 점들을 평균내어 새로운 중심점을 계산
4. 클러스터에 변화가 없으면 종료



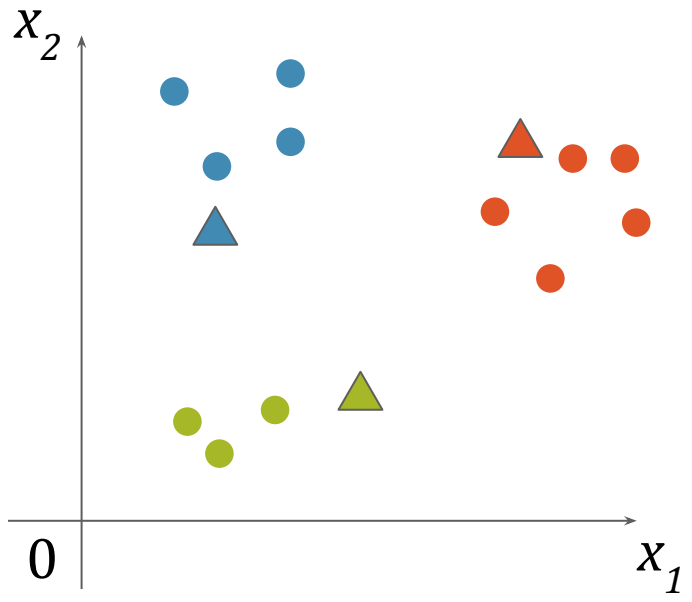
K-Means Clustering

1. 임의로 k 개의 중심점(centroid)을 생성
2. 각각의 점을 가장 가까운 중심점의 클러스터에 포함시킴
3. **각 클러스터에 포함된 점들을 평균내어 새로운 중심점을 계산**
4. 클러스터에 변화가 없으면 종료



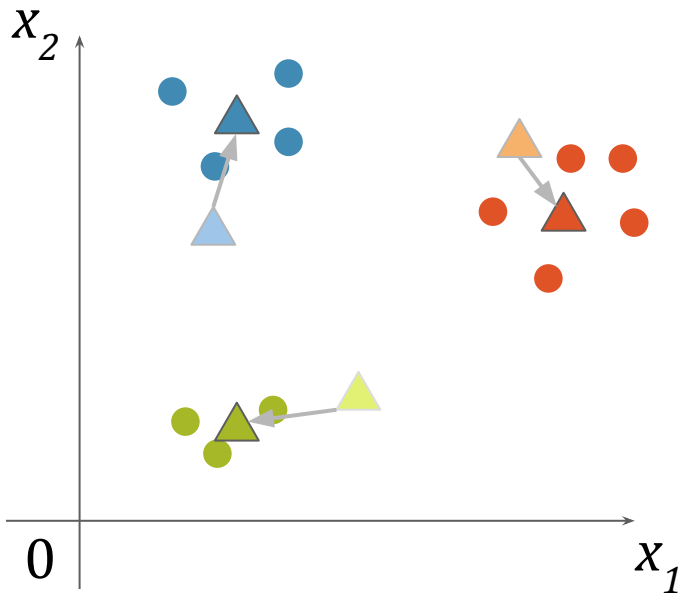
K-Means Clustering

1. 임의로 k 개의 중심점(centroid)을 생성
2. **각각의 점을 가장 가까운 중심점의 클러스터에 포함시킴**
3. 각 클러스터에 포함된 점들을 평균내어 새로운 중심점을 계산
4. 클러스터에 변화가 없으면 종료



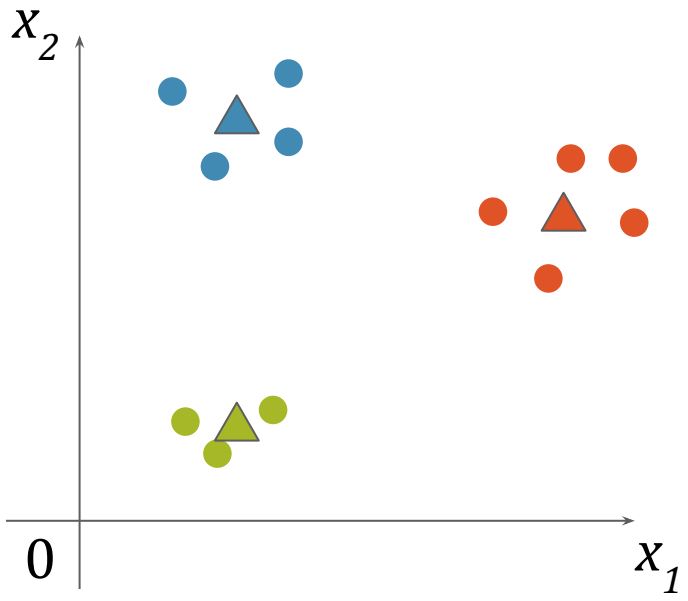
K-Means Clustering

1. 임의로 k 개의 중심점(centroid)을 생성
2. 각각의 점을 가장 가까운 중심점의 클러스터에 포함시킴
3. **각 클러스터에 포함된 점들을 평균내어 새로운 중심점을 계산**
4. 클러스터에 변화가 없으면 종료



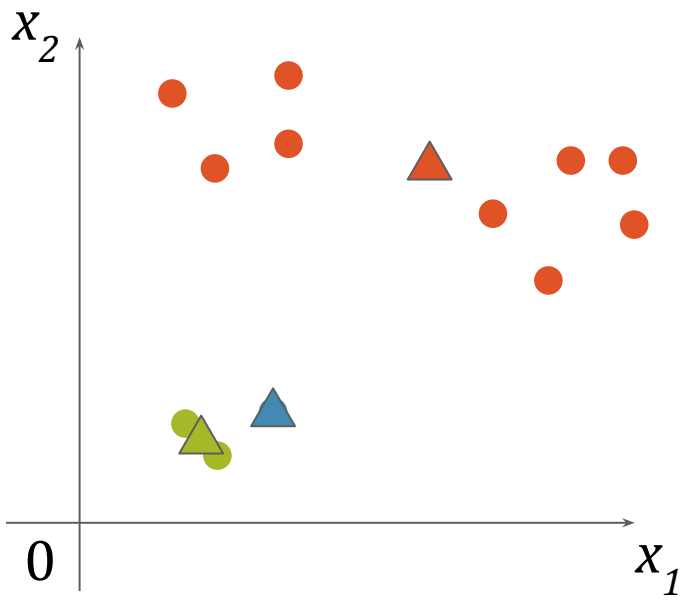
K-Means Clustering

1. 임의로 k 개의 중심점(centroid)을 생성
2. 각각의 점을 가장 가까운 중심점의 클러스터에 포함시킴
3. 각 클러스터에 포함된 점들을 평균내어 새로운 중심점을 계산
4. **클러스터에 변화가 없으면 종료**



이상한 경우...

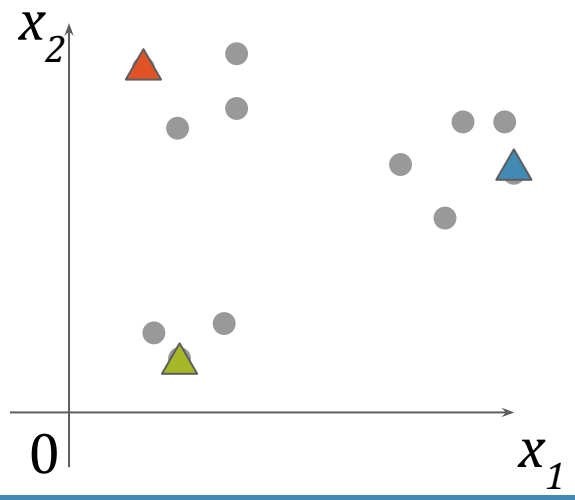
K-Means 알고리즘은 Local Optimum에 빠질 수 있다



중심점 초기화

중심점 초기화 방법에 따라 결과가 달라질 수 있다

- 임의의 벡터로 중심점 초기화
  - 여러번 반복하여 가장 좋은(?) 결과 선택
- Forgy: 데이터 점 들 중 임의로 선택
- 직접 중심점 지정하기 → 데이터를 얼추 알고 있을 때
- K-Means++: 멀리 떨어진 점들을 초기 중심점으로 사용



k 값 선택하기

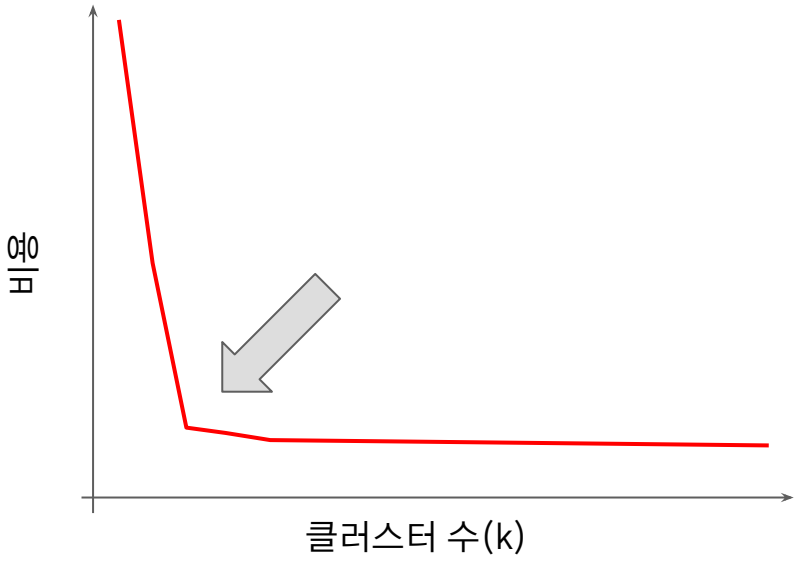
- 좋은 클러스터? → 분산이 낮다 → 비용이 낮다
- 목표함수, 비용:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

S: 데이터 집합
k: 클러스터 수
S_i:i 번째 클러스터에 속한 데이터 집합
μ_i:i 번째 클러스터의 centroid (데이터 평균)

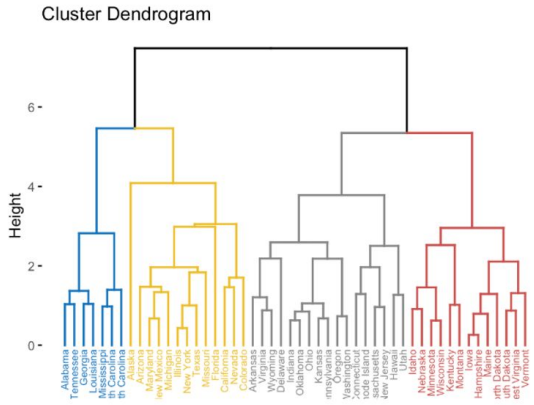
k 값 선택하기

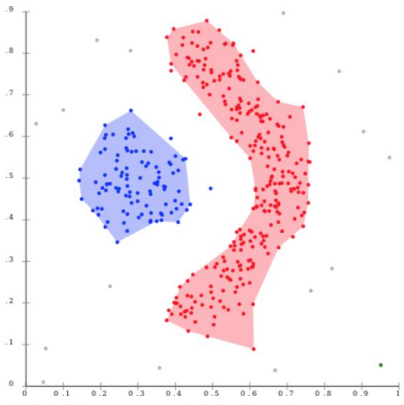
- k값을 1부터 증가시켜가며 비용을 분석
- 비용의 감소가 급격히 줄어드는 지점 선택

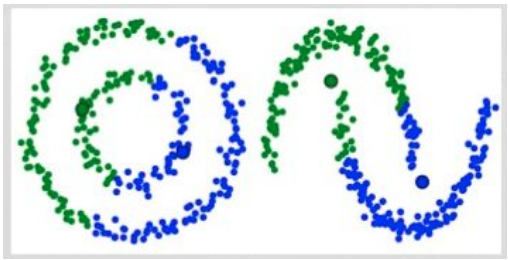


다른 클러스터링 방법

- k-medoids: k-means 알고리즘이 이상치에 민감한 문제를 보완
- 계층적 클러스터링
- DBSCAN: 밀도기반 클러스터링
- 등등...







Questions?