군집 모델 분석

충북대학교 소프트웨어학과
류관희

목 차

- ❖ Part 1. 군집분석 개념 및 k-Means 군집 분석
  - ▪ 개념
  - ▪ K-Means 군집 분석
- ❖ Part 2. DBSCAN 군집 분석
  - ▪ 개념
  - ▪ 군집 분석
  - ▪ 파이션 구현
- ❖ Part 3. 병렬 군집 분석 및 군집 분석 평가
  - ▪ 병렬 군집 분석
  - ▪ 군집 분석 평가

01

군집분석 개념 및 K-MEANS 군집 분석

- 군집분석 개념
- K-MEANS 군집 분석

02

DBSCAN 군집 분석

- 개요
- 군집 분석 설명
- 파이션 구현

03

병합 군집 분석 및 군집 분석 성능 평가

- 병합 군집 분석
- 군집분석 성능 평가

학습목표

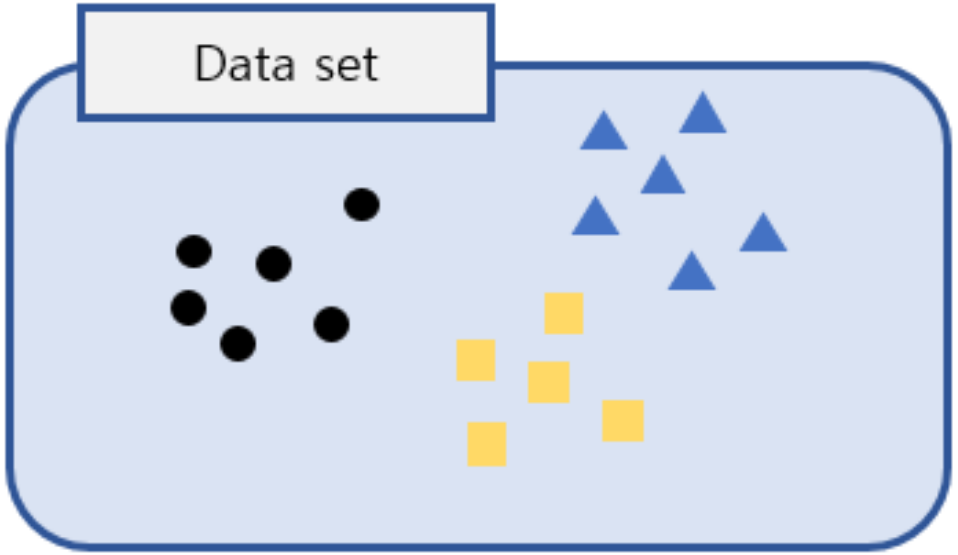
군집 분석에 대한 개념을 이해 한다.

군집 분석을 사용하는 응용 문제를 경험한다.

K-MEANS 군집 분석 개념을 이해한다.

클러스팅 개념

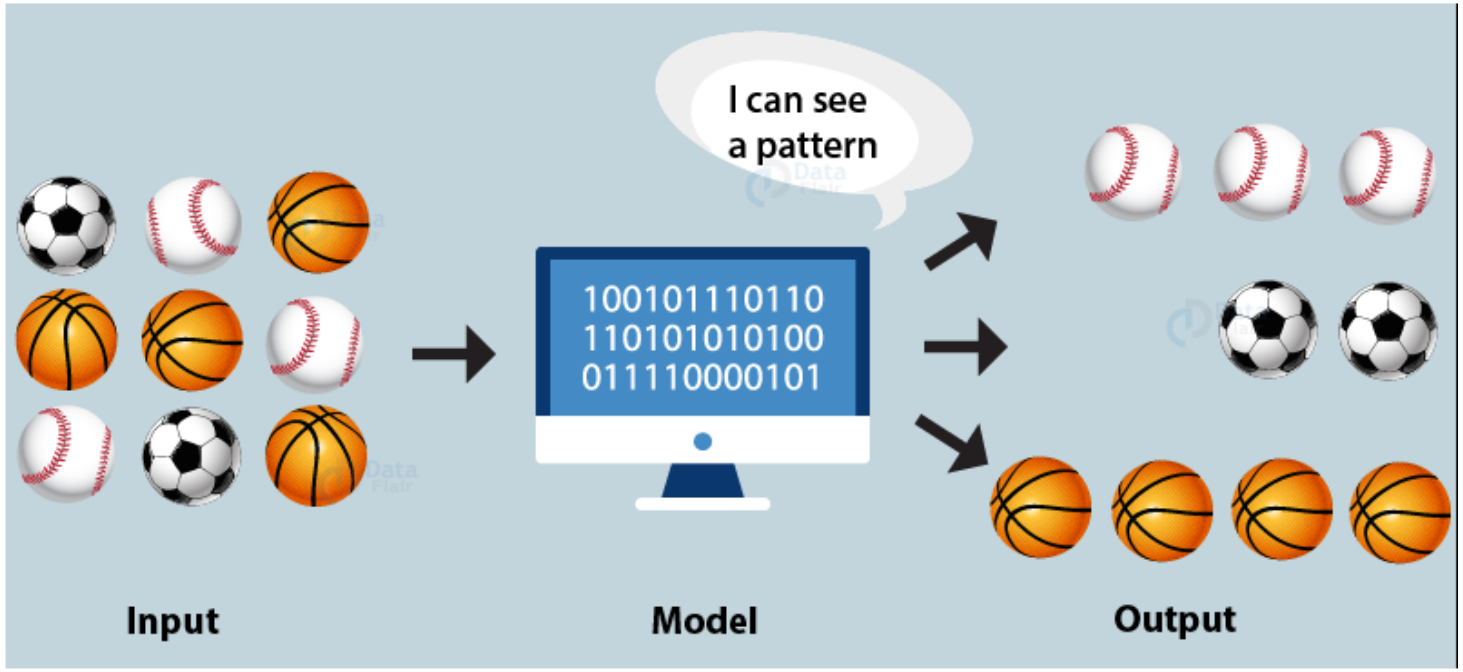
- 클러스터링이 무엇인가?
  - 데이터 개체 집합을 하위 집합으로 분할하는 과정
    - 하위 집합을 클러스터라고 함



- 예측보다는 지식 추출에 사용

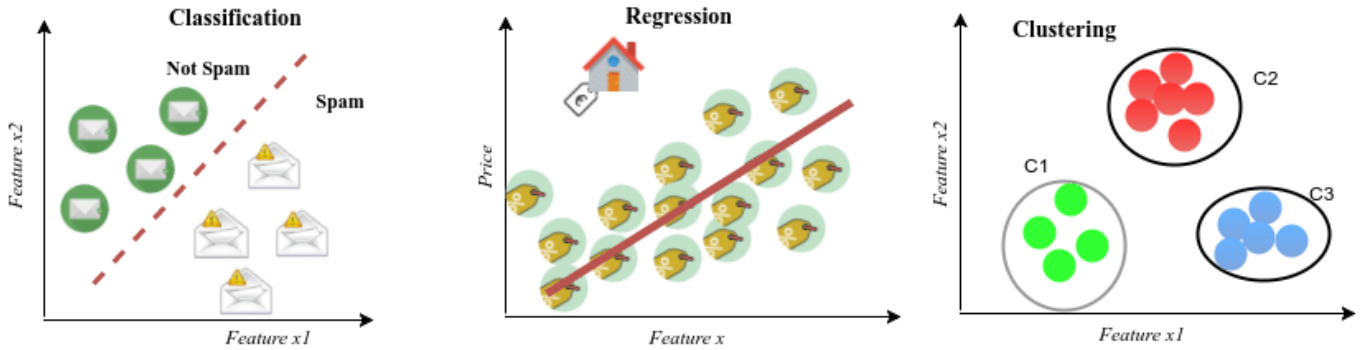
클러스팅 개념

- 클러스터링 예



클러스터링 개념

- 차이점
  - 분류: KNN, Decision Tree
  - 회귀: Linear and Logistic regression
  - 클러스터링: K-Means, Agglomerative Filtering, DBSCAN

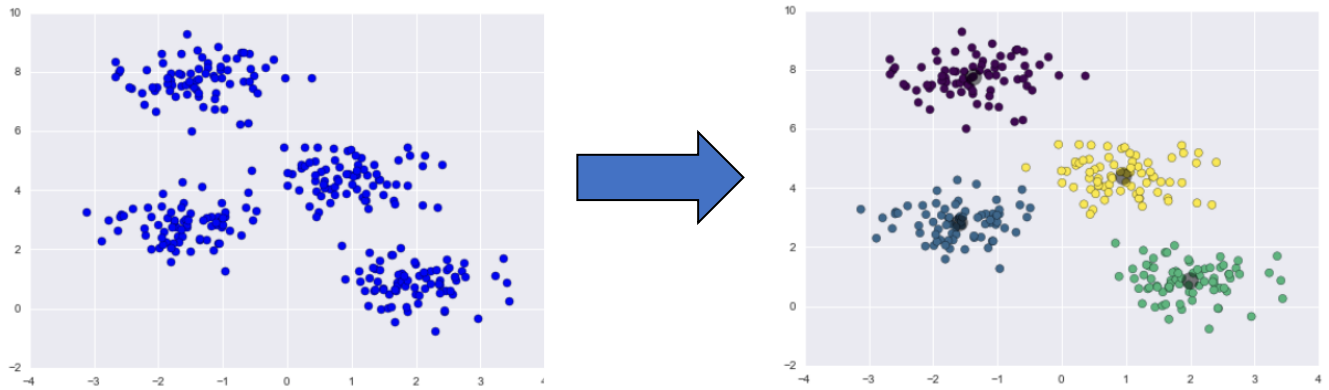


클러스터링 개념

- 적용
  - 타겟 마케팅 캠페인을 위해 유사한 인구통계나 구매 패턴을 가진 그룹으로 고객 세분화
  - 비정상 행동 탐지
    - 알려진 클러스터를 벗어나는 사용 패턴을 식별하여 무단 네트워크 침입
  - 매우 큰 데이터셋 단순화
    - 비슷한 값을 가진 특징을 더 적은 수의 동종 범주로 그룹화

K-Means 클러스터링

- K-Means
  - K는 클러스터의 수
  - 중심 기반 기술
    - 중심은 각 클러스터에 속한 개체의 평균
  - 가장 가까운 중심으로 각 개체를 그룹화



K-Means 클러스터링

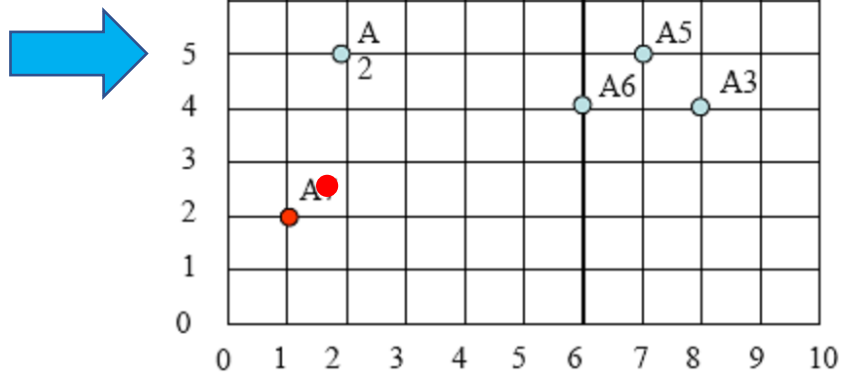
K-Means 절차

1. Step 1: 매개 변수(parameter) k 결정 (k > 0)
2. Step 2: 중심점을 시작하기 위해 k개의 점을 무작위로 선택
3. Step 3: 모든 점을 가장 가까운 중심에 할당하여 k 클러스터 형성
4. Step 4: 각 클러스터의 중심을 다시 계산 (각 클러스터의 평균 계산)
5. Step 5: 중심이 변하지 않을 때까지 step 3를 반복

K-Means 클러스터링

- K-Means 예
  - Step 1: 매개 변수 k 결정 (k > 0)
  - Step 2: 중심점을 시작하기 위해 k개의 점을 무작위로 선택

		Point
C1	A1	(2, 10)
	A2	(2, 5)
	A3	(8, 4)
C2	A4	(5, 8)
	A5	(7, 5)
	A6	(6, 4)
C3	A7	(1, 2)
	A8	(4, 9)



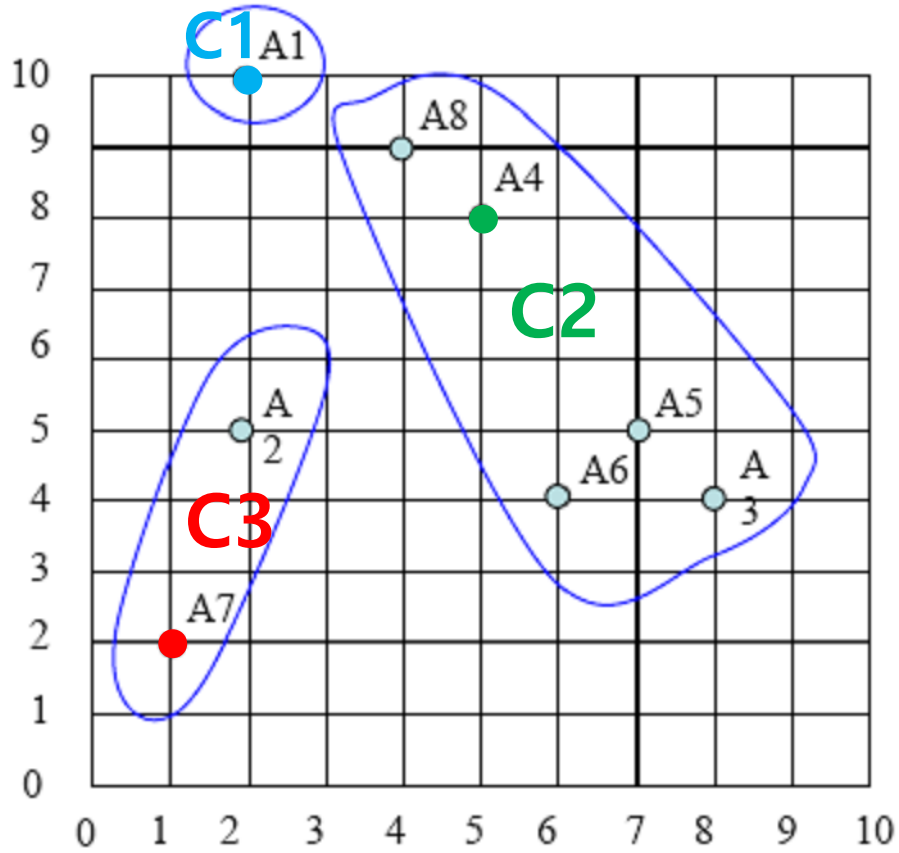
K-Means 클러스터링

- K-Means 예
  - Step 3: 모든 점을 가장 가까운 중심에 할당하여 k 클러스터 형성

	Point	C1(2, 10)	C2(5, 8)	C3(1, 2)	Cluster
A1	(2, 10)	0	3.60	8.06	C1
A2	(2, 5)	5	4.24	3.16	C3
A3	(8, 4)	8.48	5	7.28	C2
A4	(5, 8)	3.60	0	7.21	C2
A5	(7, 5)	7.07	3.60	6.70	C2
A6	(6, 4)	7.21	4.12	5.38	C2
A7	(1, 2)	8.06	7.21	0	C3
A8	(4, 9)	2.23	1.41	7.61	C2

K-Means 클러스터링

- K-Means 예
  - Step 3: 모든 점을 가장 가까운 중심에 할당하여 k 클러스터 형성



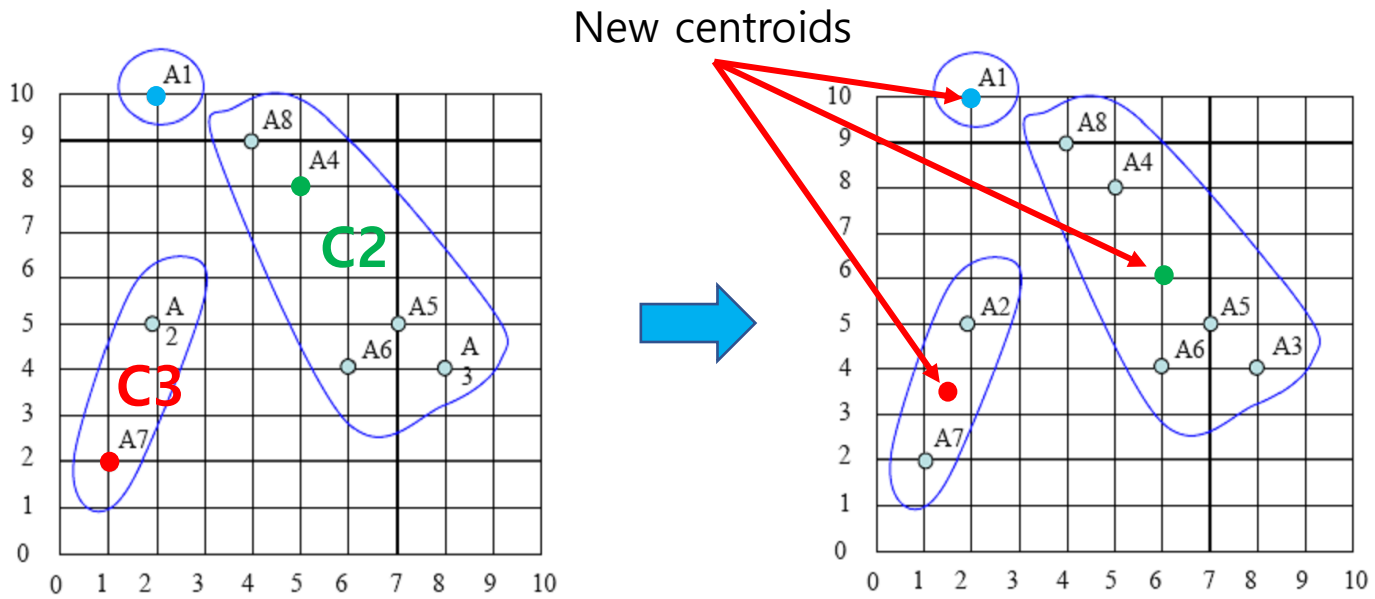
K-Means 클러스터링

- K-Means 예
  - Step 4: 각 클러스터의 중심을 다시 계산 (각 클러스터의 평균 계산)

	Point	Cluster	Mean of C#
A1	(2, 10)	C1	(2, 10)
A3	(8, 4)	C2	(6, 6)
A4	(5, 8)	C2	
A5	(7, 5)	C2	
A6	(6, 4)	C2	
A8	(4, 9)	C2	
A7	(1, 2)	C3	(1.5, 3.5)
A2	(2, 5)	C3	

K-Means 클러스터링

- K-Means 예
  - Step 4: 각 클러스터의 중심을 다시 계산
    (각 클러스터의 평균 계산)



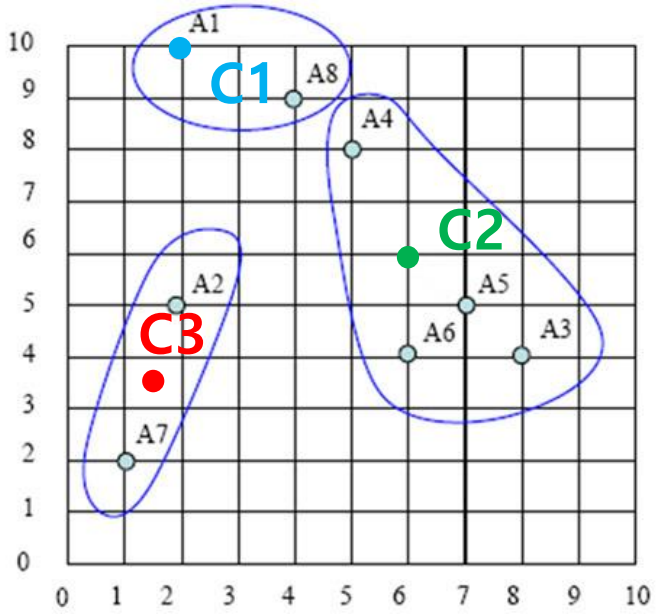
K-Means 클러스터링

- K-Means 예
  - Step 5: 중심이 변하지 않을 때까지 step 3를 반복
  - Step 3: 모든 점을 가장 가까운 중심에 할당하여 k 클러스터 형성

	Point	C1(2, 10)	C2(6, 6)	C3(1.5, 3.5)	Cluster
A1	(2, 10)	0	5.65	6.51	C1
A2	(2, 5)	5	4.12	1.58	C3
A3	(8, 4)	8.48	2.82	6.51	C2
A4	(5, 8)	3.60	2.23	5.70	C2
A5	(7, 5)	7.07	1.41	5.70	C2
A6	(6, 4)	7.21	2	4.52	C2
A7	(1, 2)	8.06	6.40	1.58	C3
A8	(4, 9)	2.23	3.60	6.04	C1

K-Means 클러스터링

- K-Means 예
  - Step 5: 중심이 변하지 않을 때까지 step 3를 반복
  - Step 3: 모든 점을 가장 가까운 중심에 할당하여 k 클러스터 형성



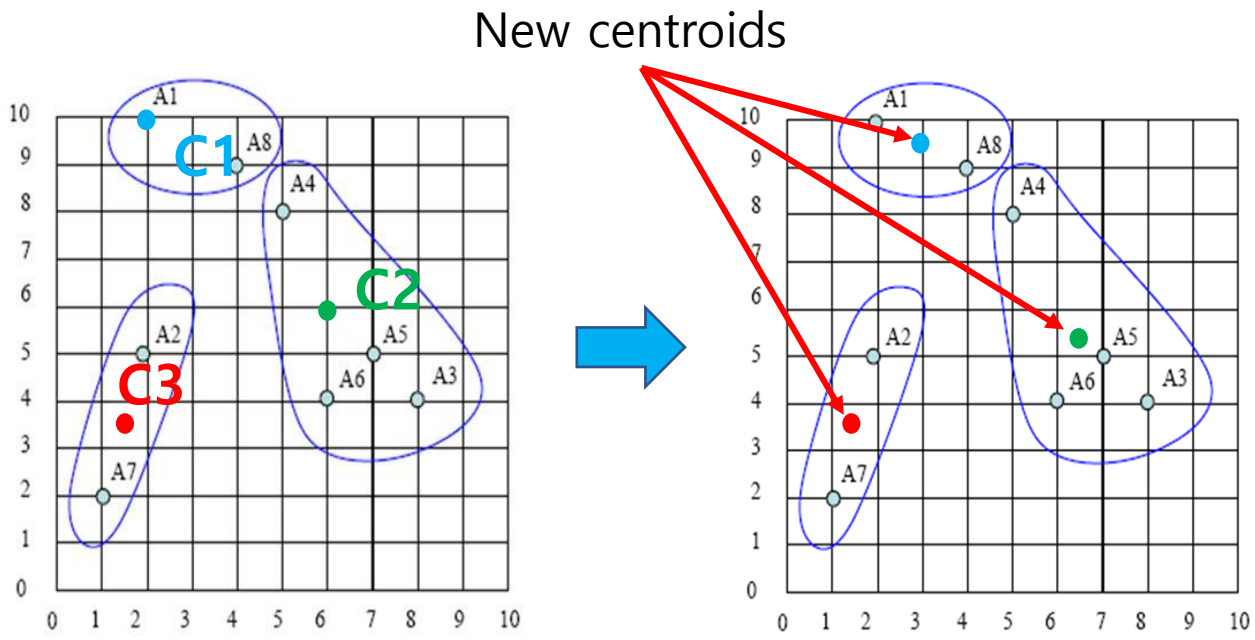
K-Means 클러스터링

- K-Means 예
  - Step 4: 각 클러스터의 중심을 다시 계산 (각 클러스터의 평균 계산)

	Point	Cluster	Mean of C#
A1	(2, 10)	C1	(3, 9.5)
A8	(4, 9)	C1	
A3	(8, 4)	C2	(6.5, 5.25)
A4	(5, 8)	C2	
A5	(7, 5)	C2	
A6	(6, 4)	C2	
A7	(1, 2)	C3	(1.5, 3.5)
A2	(2, 5)	C3	

K-Means 클러스터링

- K-Means 예
  - Step 4: 각 클러스터의 중심을 다시 계산 (각 클러스터의 평균 계산)



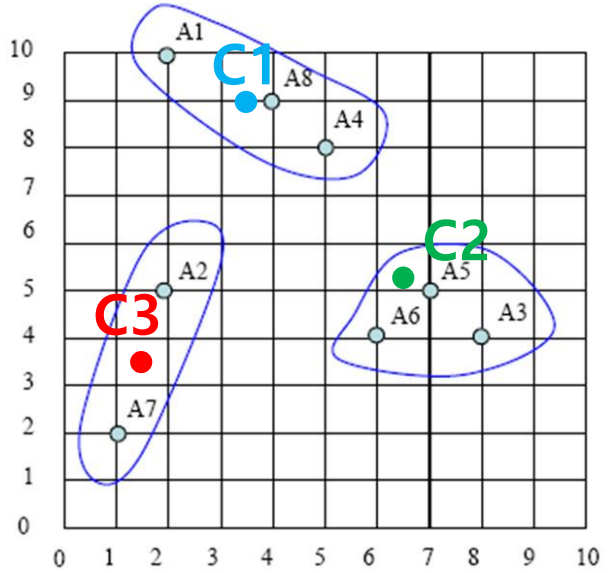
K-Means 클러스터링

- K-Means 예
  - Step 5: 중심이 변하지 않을 때까지 step 3를 반복
  - Step 3: 모든 점을 가장 가까운 중심에 할당하여 k 클러스터 형성

	Point	C1(3, 9.5)	C2(6.5, 5.25)	C3(1.5, 3.5)	Cluster
A1	(2, 10)	1.11	6.54	6.51	C1
A2	(2, 5)	4.60	4.5	1.58	C3
A3	(8, 4)	7.43	1.95	6.51	C2
A4	(5, 8)	2.5	3.13	5.70	C1
A5	(7, 5)	6.02	0.55	5.70	C2
A6	(6, 4)	6.26	1.34	4.52	C2
A7	(1, 2)	7.76	6.38	1.58	C3
A8	(4, 9)	1.11	4.5	6.04	C1

K-Means 클러스터링

- K-Means 예
  - Step 5: 중심이 변하지 않을 때까지 step 3를 반복
  - Step 3: 모든 점을 가장 가까운 중심에 할당하여 k 클러스터 형성



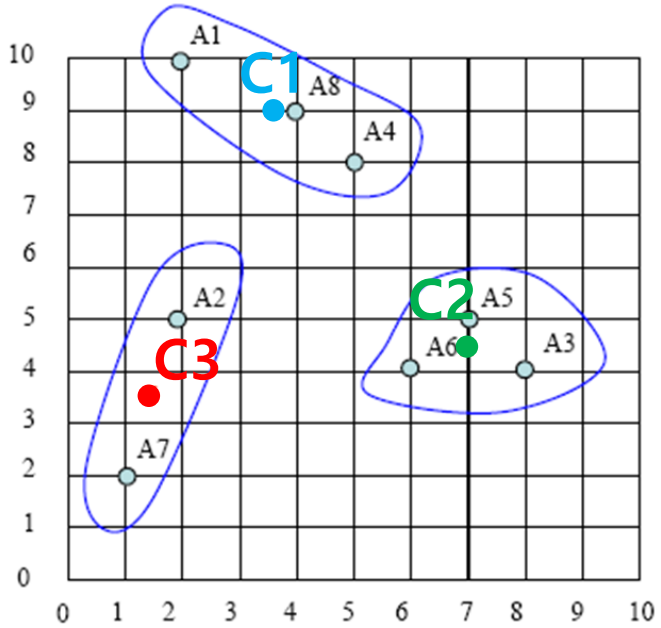
K-Means 클러스터링

- K-Means 예
  - Step 4: 각 클러스터의 중심을 다시 계산 (각 클러스터의 평균 계산)

	Point	Cluster	Mean of C#
A1	(2, 10)	C1	(3.66, 9)
A8	(4, 9)	C1	
A4	(5, 8)	C1	
A3	(8, 4)	C2	(7, 4.33)
A5	(7, 5)	C2	
A6	(6, 4)	C2	
A7	(1, 2)	C3	(1.5, 3.5)
A2	(2, 5)	C3	

K-Means 클러스터링

- K-Means 예
  - 최종 클러스터 결과



K-Means 클러스터링

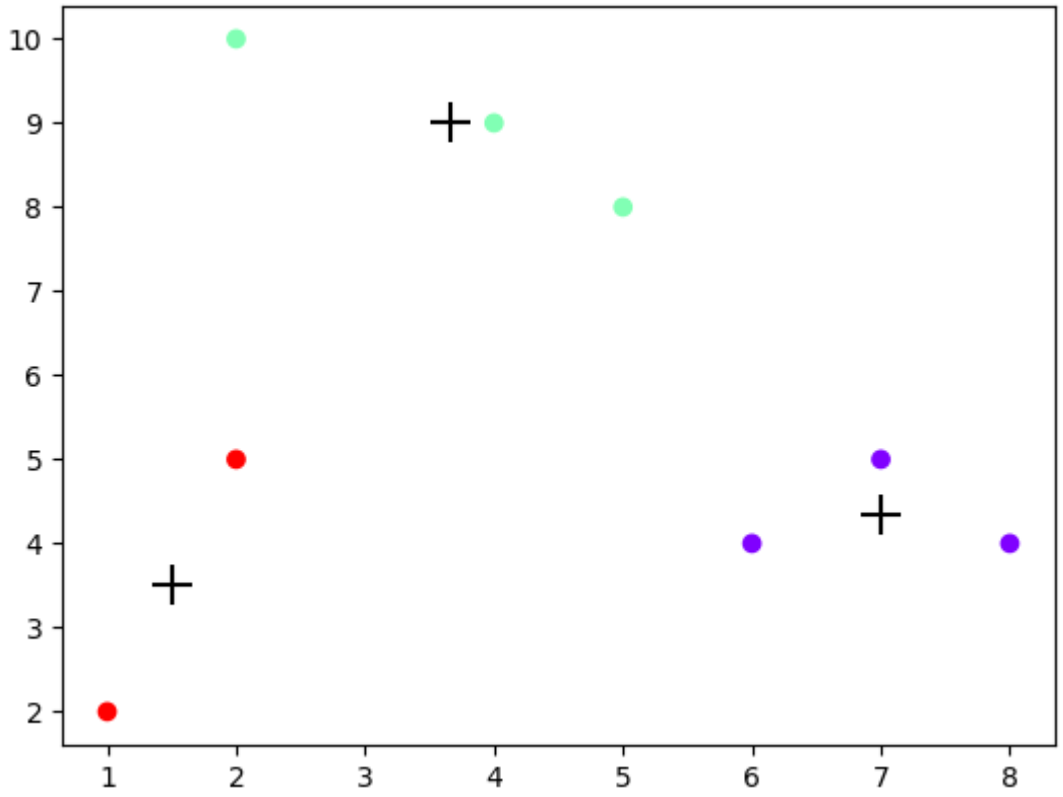
- Python에서의 K-Means

```
from sklearn.cluster import KMeans
import numpy as np

X = np.array([[2, 10], [2, 5], [8, 4],[5, 8], [7, 5], [6, 4], [1, 2], [4, 9
]])

kmeans = KMeans(n_clusters=3).fit(X)

print("Labels: ", kmeans.labels_)

print("Cluster Centers: ", kmeans.cluster_centers_)

print("Predict Values: ", kmeans.predict([[1, 1]]))
```

```
Labels:          [2 1 0 2 0 0 1 2]
Cluster Centers: [[7.         4.33333333]
                  [1.5        3.5       ]
                  [3.66666667 9.        ]]
Predict Values:  [1]
```

K-Means 클러스터링

- Python에서의 K-Means 시각화



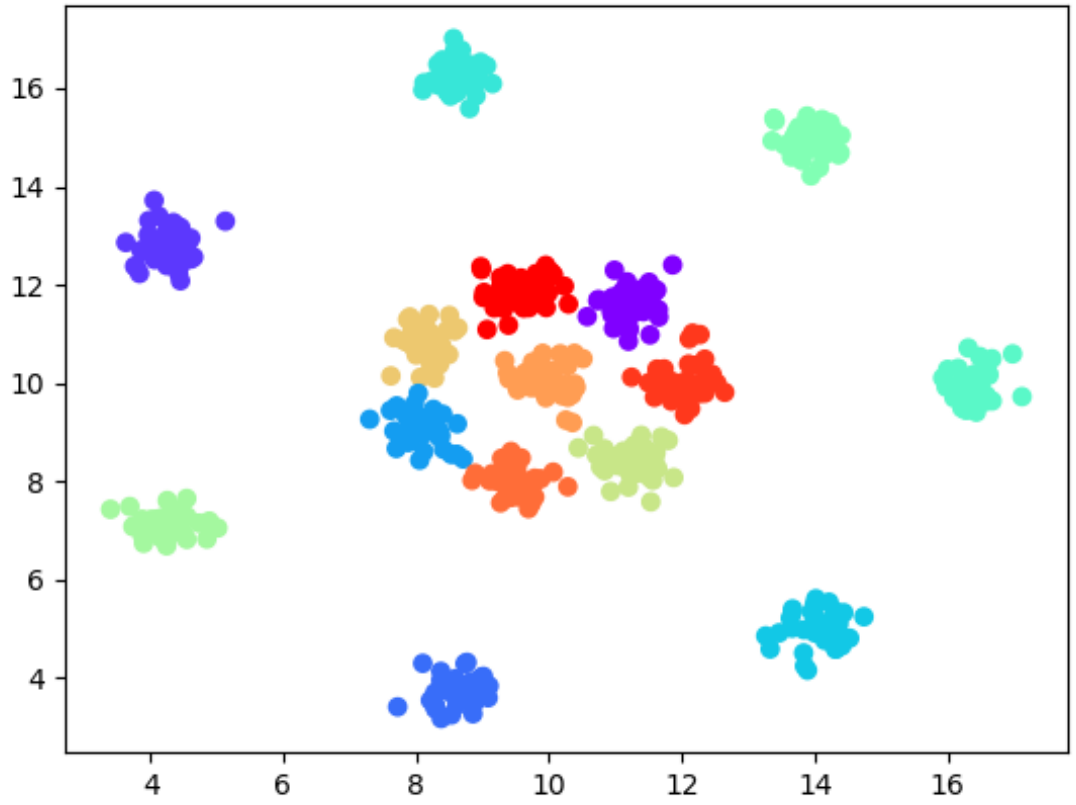
K-Means 클러스터링

- Python에서의 K-Means 시각화
  - r15 dataset (r15.csv) -> n_clusters=15

```
from sklearn.cluster import KMeans
import pandas as pd
import matplotlib.pyplot as plt

sample_df = pd.read_csv("r15.csv")

training_points = sample_df[["col1", "col2"]]
training_labels = sample_df["target"]

kmeans = KMeans(n_clusters=15).fit(training_points)

plt.scatter(training_points["col1"], training_points["col2"], c=kmeans.labels_,
                                cmap='rainbow')

plt.show()
```

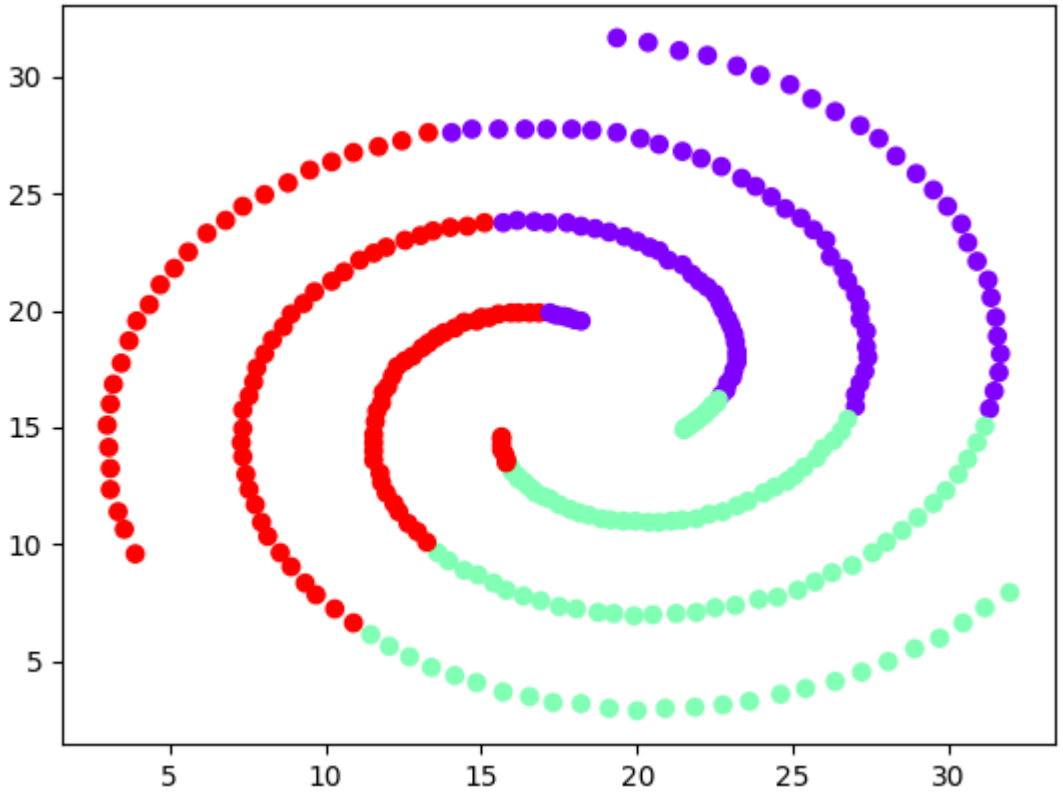
K-Means 클러스터링

- Python에서의 K-Means 시각화
  - r15 dataset (r15.csv) -> n_clusters=15



K-Means 클러스터링

- Python에서의 K-Means 시각화
  - Spiral (spiral.csv) -> n_clusters=3



문제풀이

- 군집 분석의 개념을 설명하시오.
- K-Means 군집 분석의 절차를 순서적으로 나열하시오.

요약

- 군집 분석의 개념을 이해하였음.
- K-Means 군집 분석의 과정을 공부하였고, K-Means 기법을 이용하여 군집 분석을 할 수 있는 응용 분야를 공부하였음.

01

군집분석 개념 및 K-MEANS 군집 분석

- 군집분석 개념
- K-MEANS 군집 분석

02

DBSCAN 군집 분석

- 개요
- 군집 분석 설명
- 파이션 구현

03

병합 군집 분석 및 군집 분석 성능 평가

- 병합 군집 분석
- 군집분석 성능 평가

학습목표

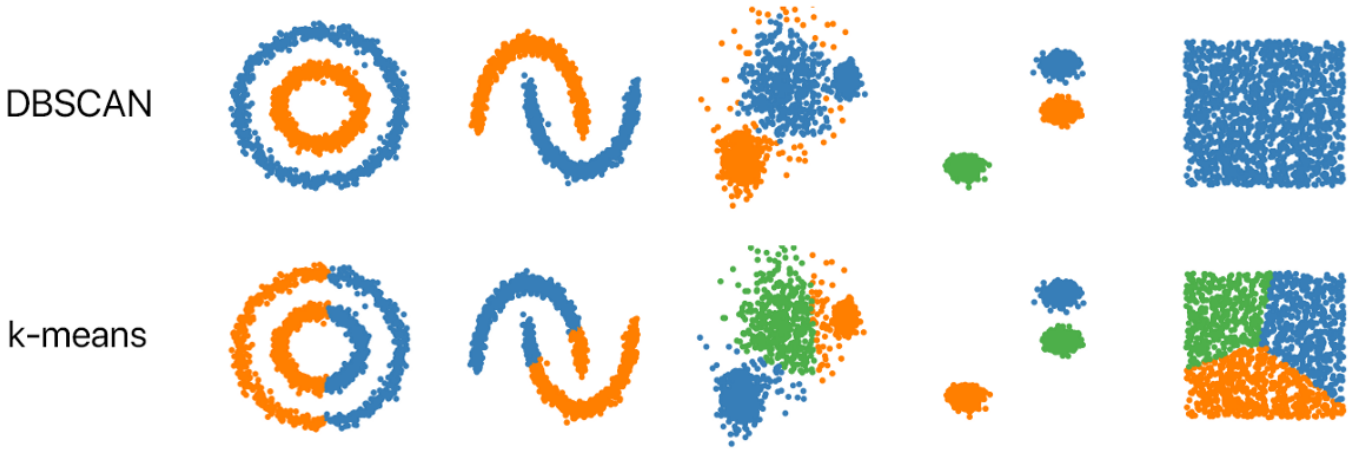
DBSCAN 개념을 이해한다.

DBSCAN 군집화 과정을 이해한다.

DBSCAN을 위한 파이션 프로그램을 만들 수 있다.

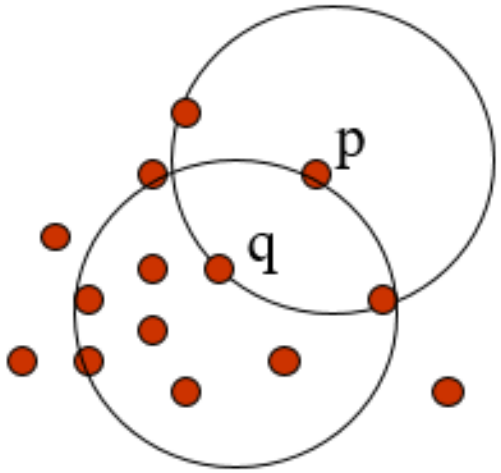
DBSCAN을 이용한 클러스터링

- DBSCAN이 무엇인가?
  - Density-based spatial clustering of applications with noise
  - 연결된 점을 기반으로 하는 클러스터링
  - “이웃”의 밀도가 일부 임계값을 초과하는 한 주어진 클러스터를 계속 확장



DBSCAN을 이용한 클러스터링

- DBSCAN 매개 변수
  - Epsilon (ε)
    - 이웃의 최대 반지름
  - minPts
    - 해당 지점의 Eps-이웃에 있는 최소 점의 수

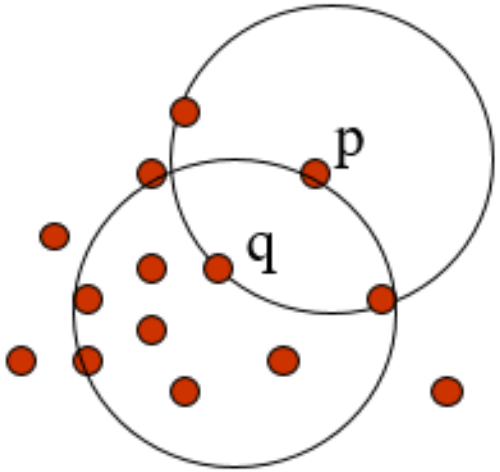


MinPts = 5

Eps = 1 cm

DBSCAN을 이용한 클러스터링

- DBSCAN 매개 변수
  - Core 개체는 ε-이웃의 최소 점의 수를 충족하는 개체임
  - 개체 q 는 core 개체

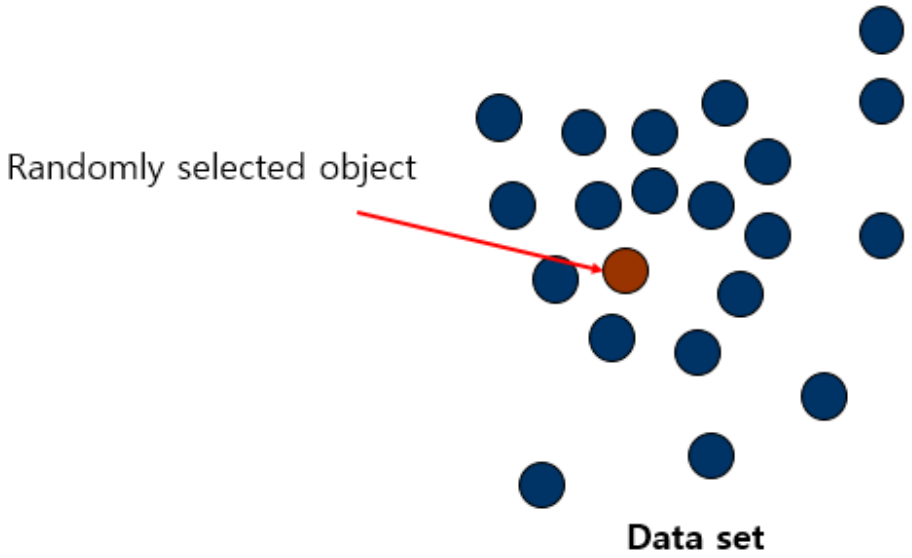


MinPts = 5

Eps = 1 cm

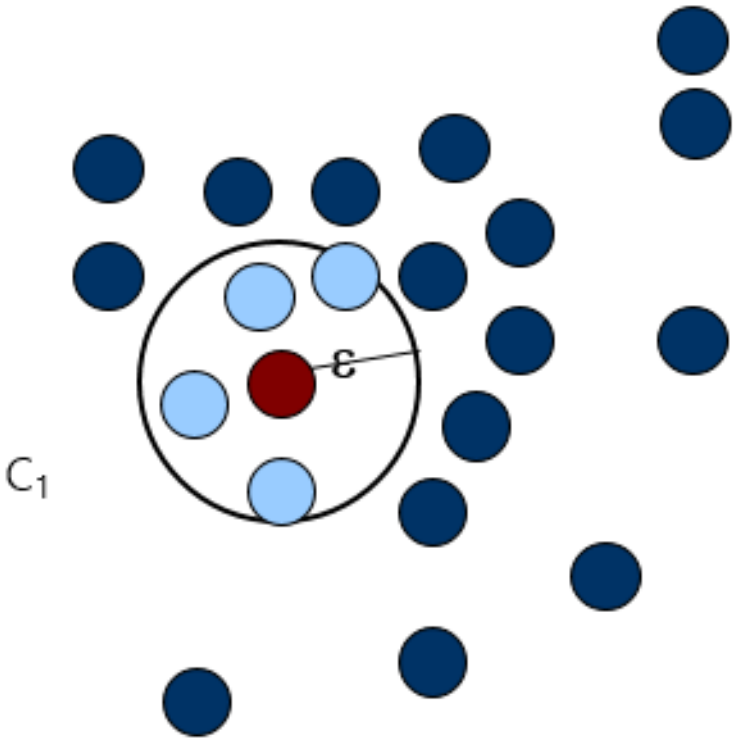
DBSCAN을 이용한 클러스터링

- DBSCAN 예



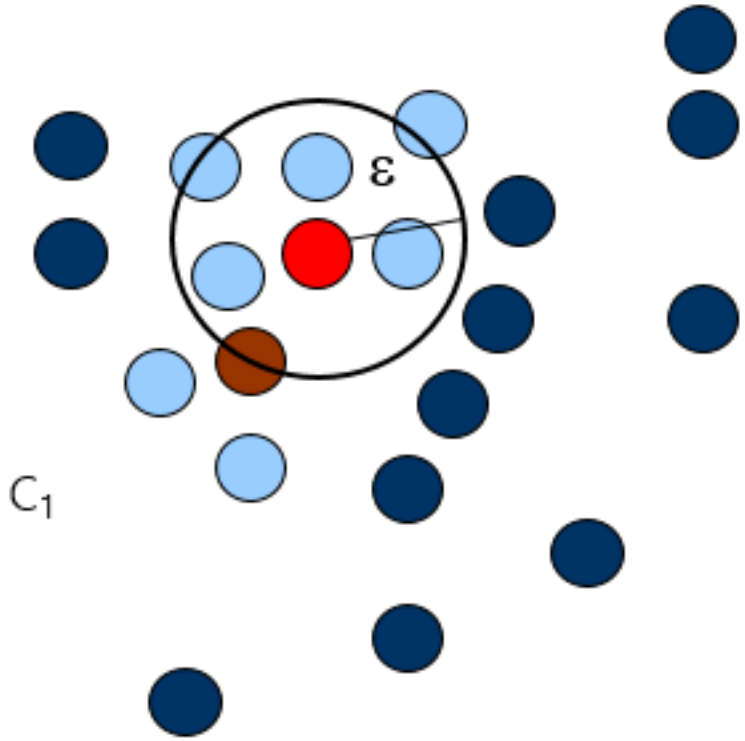
DBSCAN을 이용한 클러스터링

- DBSCAN 예



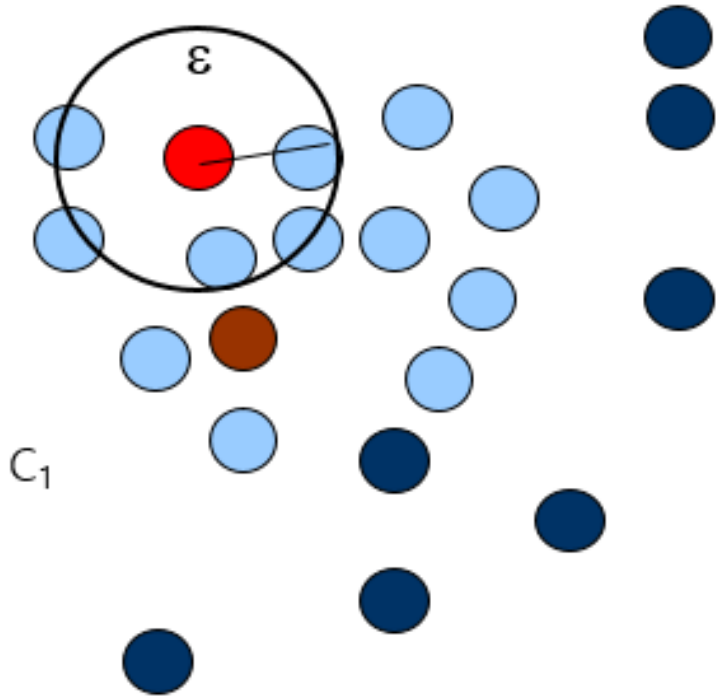
DBSCAN을 이용한 클러스터링

- DBSCAN 예



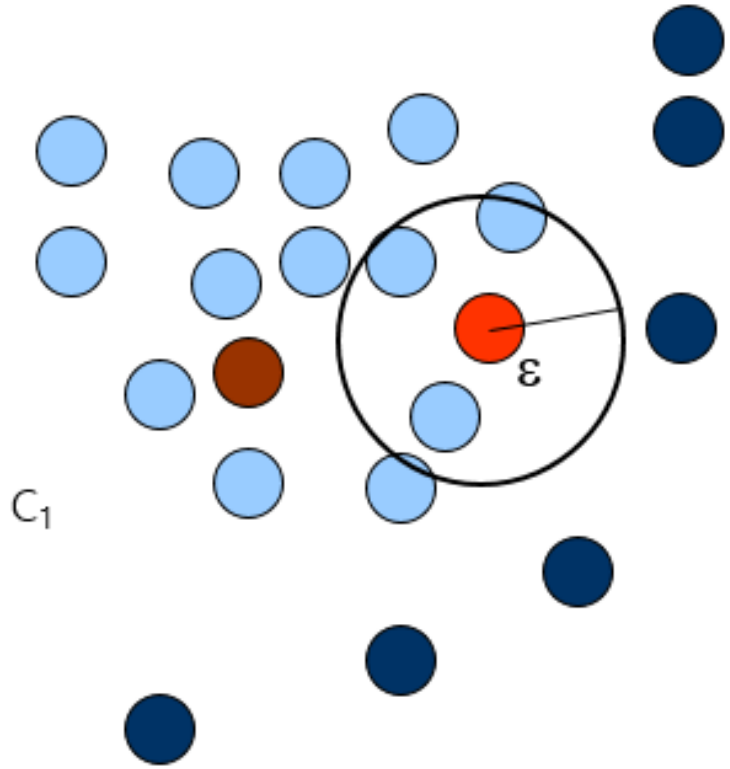
DBSCAN을 이용한 클러스터링

- DBSCAN 예



DBSCAN을 이용한 클러스터링

- DBSCAN 예



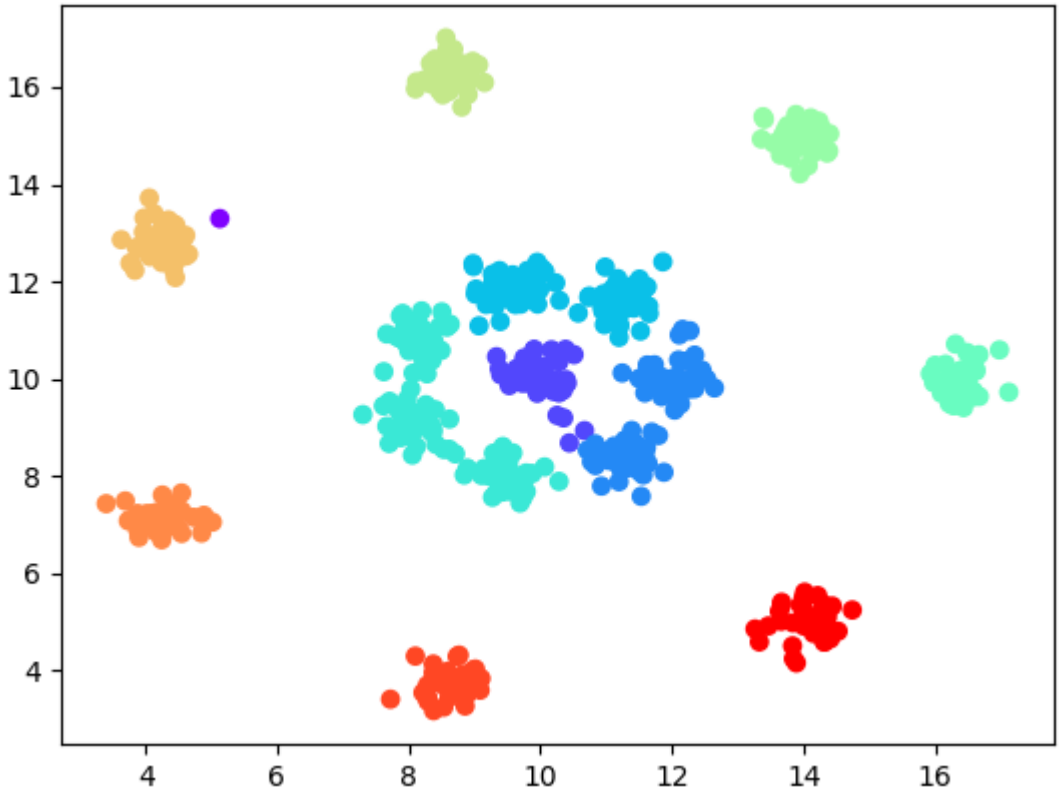
DBSCAN을 이용한 클러스터링

- Python에서의 DBSCAN
  - r15 dataset (r15.csv)

```
from sklearn.cluster import DBSCAN
import pandas as pd
import matplotlib.pyplot as plt

sample_df = pd.read_csv("r15.csv")

training_points = sample_df[["col1", "col2"]]
training_labels = sample_df["target"]

dbscan = DBSCAN(eps=0.6, min_samples=10).fit(training_points)

plt.scatter(training_points["col1"], training_points["col2"], c=dbscan.labels_,
            cmap='rainbow')

plt.show()
```

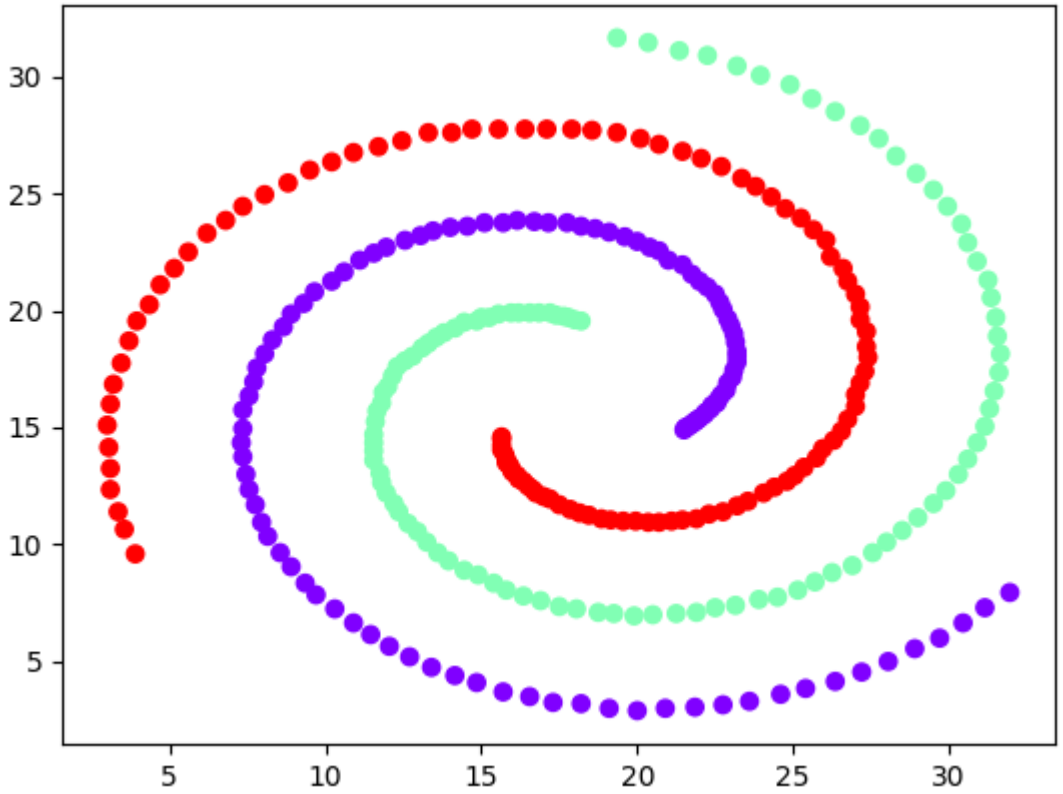
DBSCAN을 이용한 클러스터링

- Python에서의 DBSCAN



DBSCAN을 이용한 클러스터링

- Python에서의 DBSCAN



문제풀이

- DBSCAN 군집 분석의 개념을 설명하시오.
- DBSCAN 과 K-Means 군집 분석의 차이점을 설명하시오.
- DBSCAN을 적용하여 군집분석하기 좋은 분야를 제시하시오.

요약

- DBSCAN 군집 분석의 개념을 설명하시오.
- DBSCAN 군집 분석의 절차를 순서적으로 나열하시오.

01

군집분석 개념 및 K-MEANS 군집 분석

- 군집분석 개념
- K-MEANS 군집 분석

02

DBSCAN 군집 분석

- 개요
- 군집 분석 설명
- 파이션 구현

03

병합 군집 분석 및 군집 분석 성능 평가

- 병합 군집 분석
- 군집분석 성능 평가

학습목표

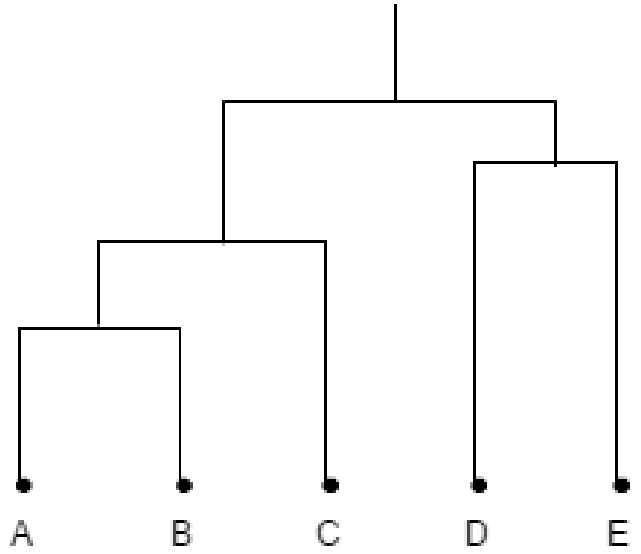
병합 군집 분석 개념을 이해한다.

병합 군집화 과정을 이해한다.

군집 분석 평가 방법을 이해 한다.

병합 클러스터링

- 병합 클러스터링
- 데이터 개체를 클러스터의 계층 또는 " 트리 " 로 그룹화
- Dendrgram은 계층적 클러스터링 과정을 나타내는데 사용



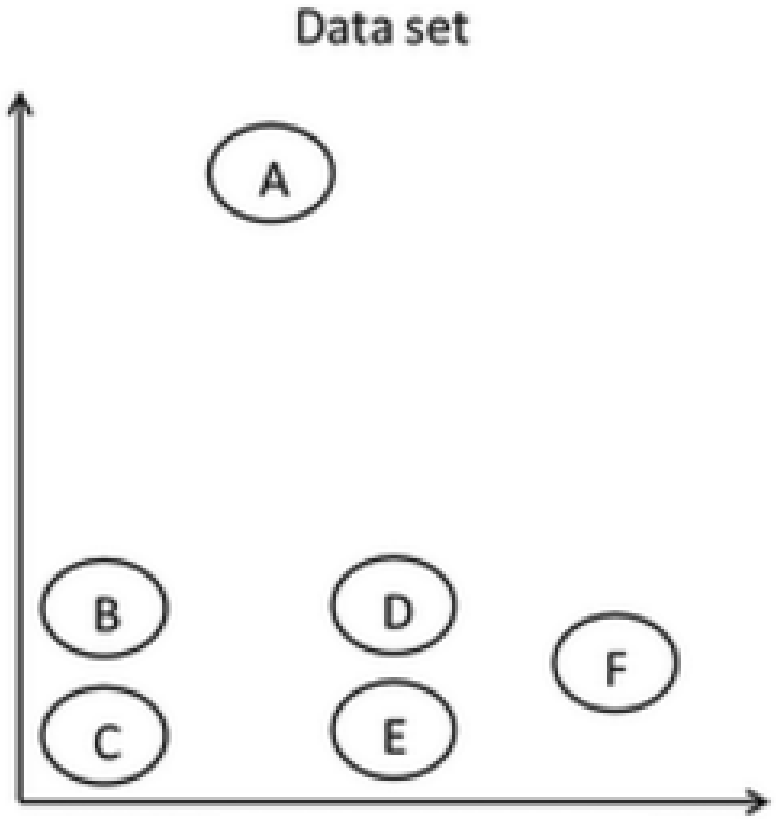
병합 클러스터링

- 병합 클러스터링
  - Bottom-up 전략
  - 각 개체가 자체 클러스터를 형성하도록 하는 것으로 시작
  - 클러스터를 더 큰 클러스터로 반복적으로 병합
    - 모든 개체가 단일 클러스터에 있을 때까지

- 병합 계층적 클러스터링 : 절차
  1. 각 개체는 하나의 클러스터를 형성
  2. 가장 낮은 수준의 가장 가까운 (유사한) 두 클러스터를 하나의 클러스터로 병합
  3. 단일 클러스터가 될 때까지 2 단계를 반복

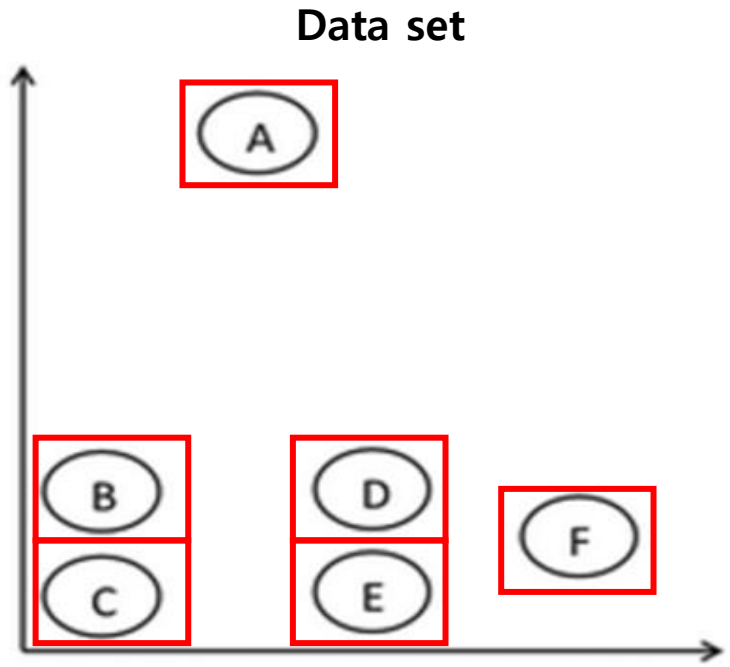
병합 클러스터링

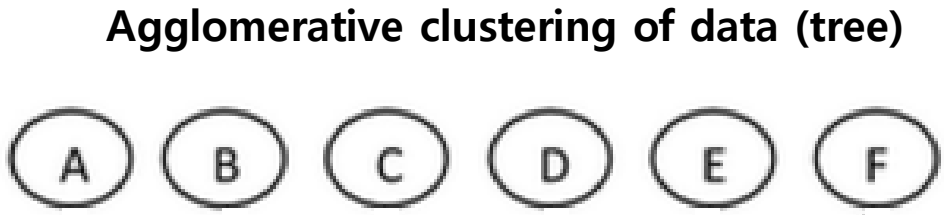
- 병함 클러스터링 예



병합 클러스터링

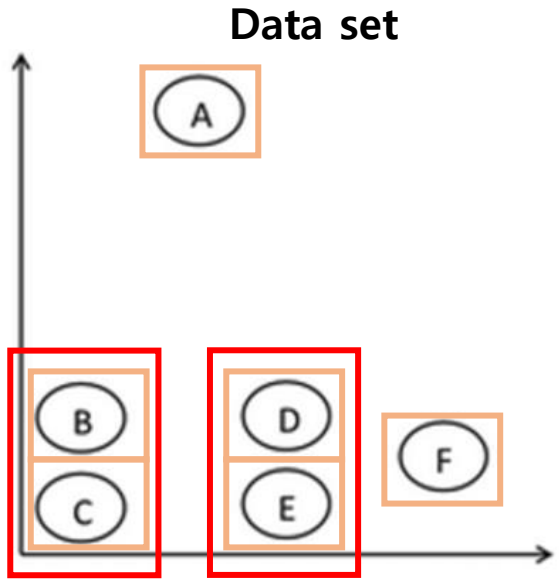
- 병합 클러스터링 예



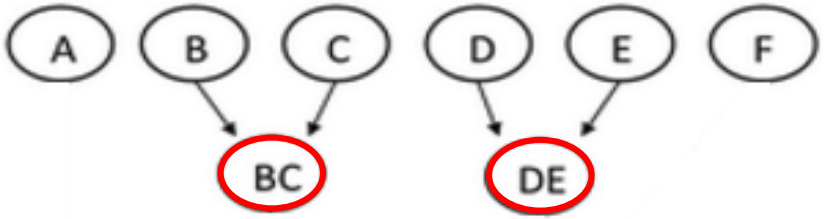


병합 클러스터링

- 병합 클러스터링 예

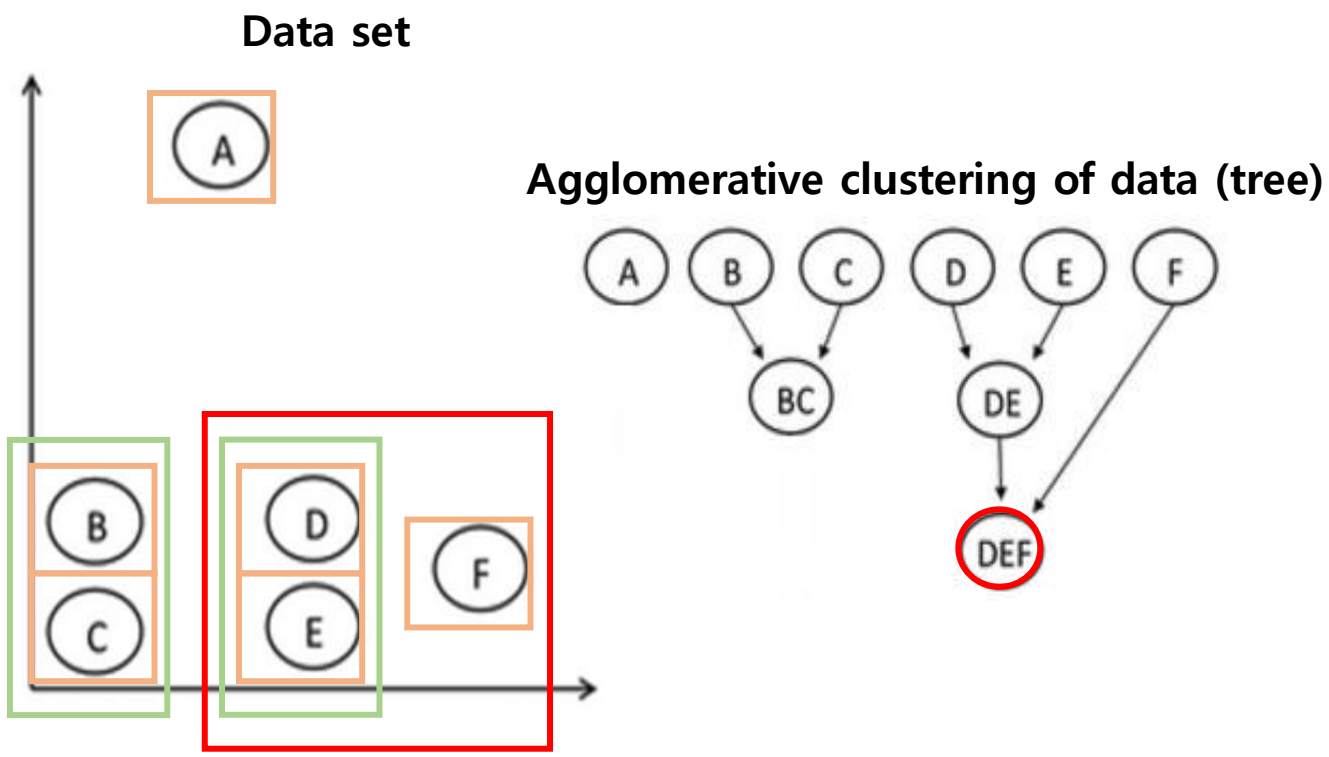


Agglomerative clustering of data (tree)



병합 클러스터링

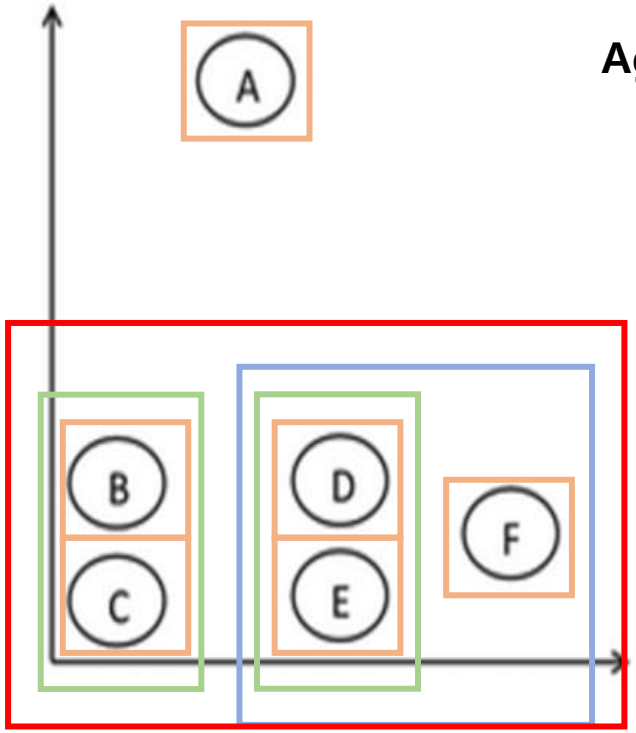
- 병합 클러스터링 예



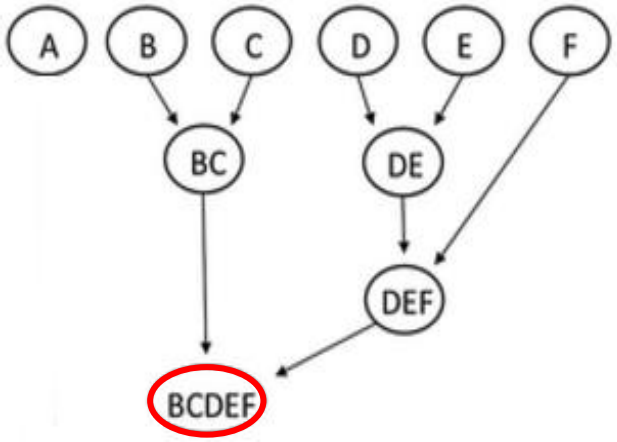
병합 클러스터링

- 병합 클러스터링 예

Data set



Agglomerative clustering of data (tree)



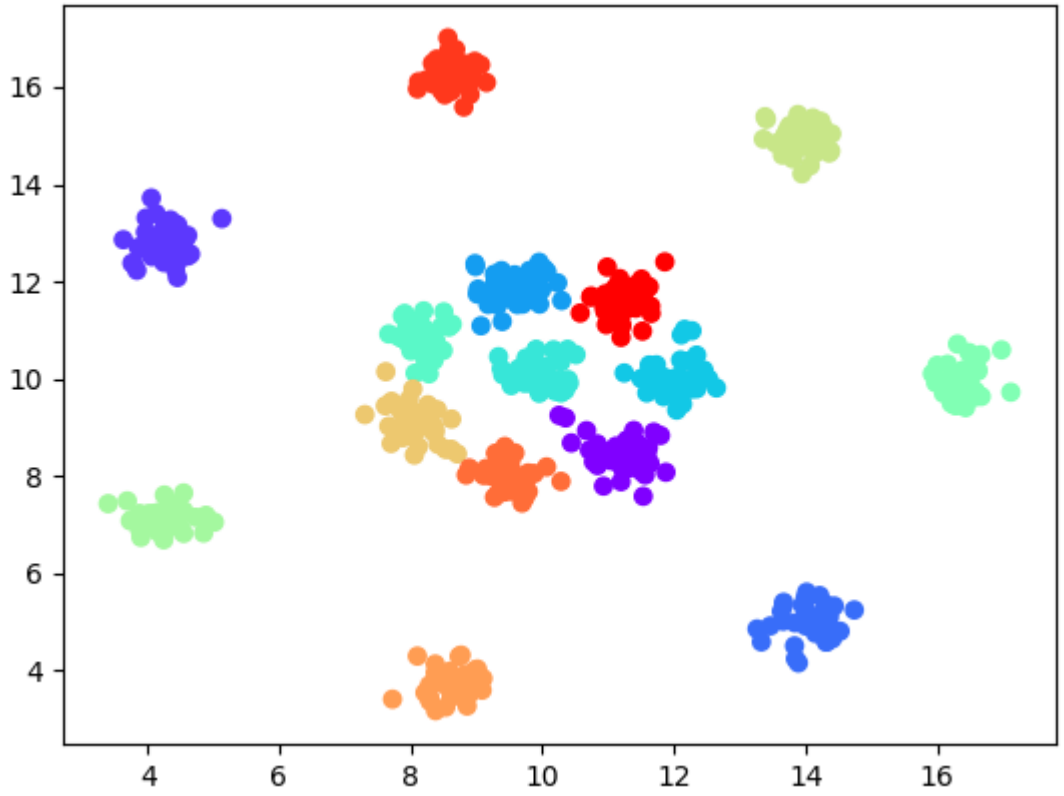
병합 클러스터링

- Python에서의 병합 클러스터링
  - r15 dataset (r15.csv) -> n_clusters=15

```
from sklearn.cluster import AgglomerativeClustering
import pandas as pd
import matplotlib.pyplot as plt

sample_df = pd.read_csv("r15.csv")

training_points = sample_df[["col1", "col2"]]
training_labels = sample_df["target"]

agglo = AgglomerativeClustering(n_clusters=15).fit(training_points)

plt.scatter(training_points["col1"], training_points["col2"], c=agglo.labels_,
                                cmap='rainbow')

plt.show()
```

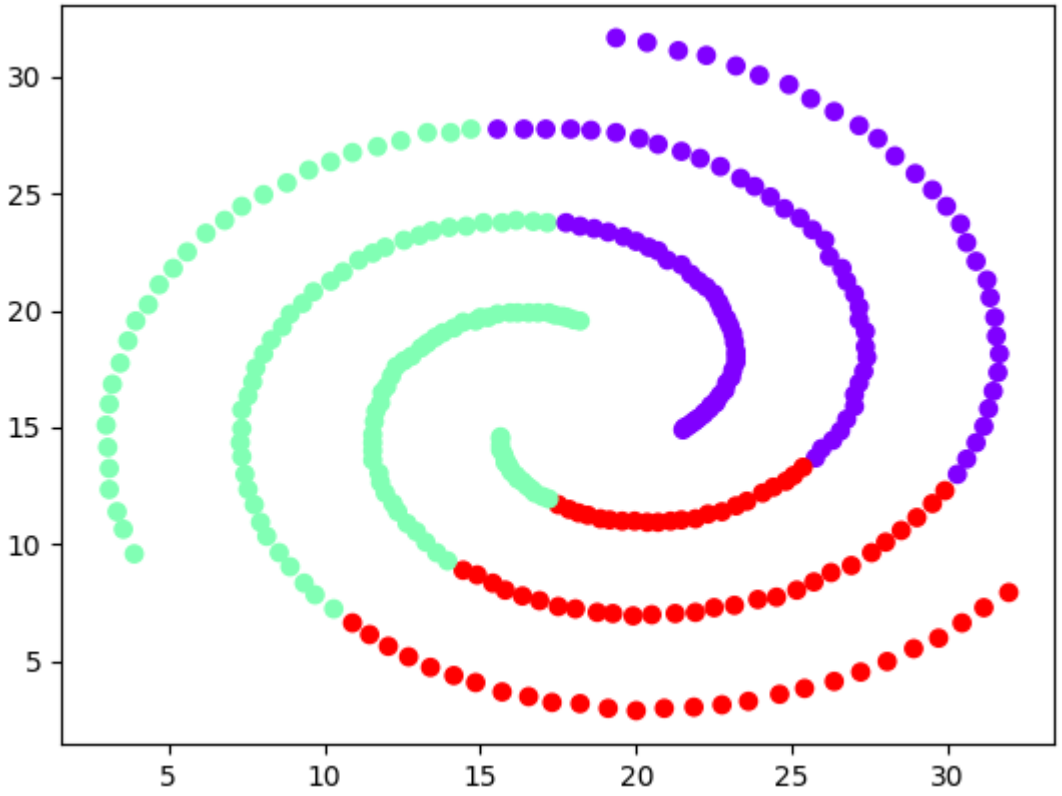
병합 클러스터링

- Python에서의 병합 클러스터링
  - r15 dataset (r15.csv) -> n_clusters=15



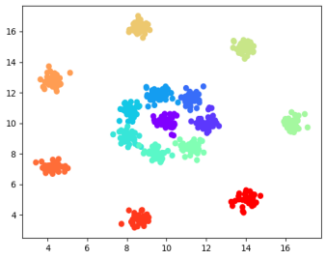
병합 클러스터링

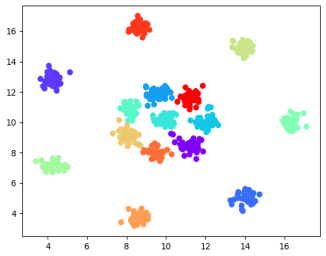
- Python에서의 병합 클러스터링
  - spiral dataset (spiral.csv) -> n_clusters=3

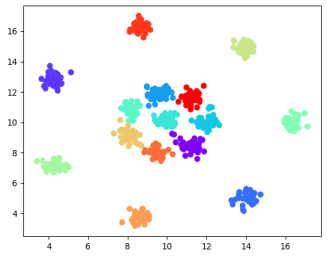


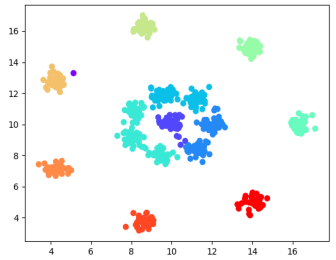
클러스터링 평가

- 평가하는 이유는 무엇인가?
  - r15에서의 비교

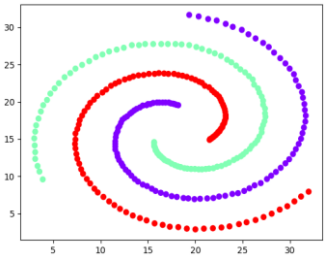

<Ground Truth>

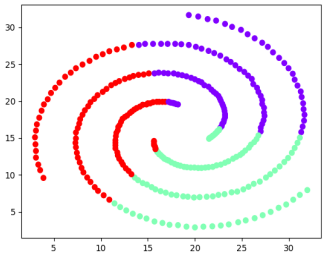

<K-Means>

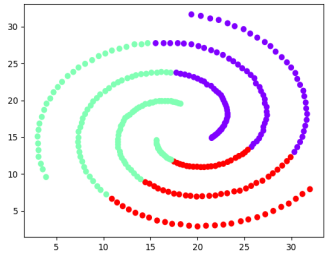

<Agglomerative>

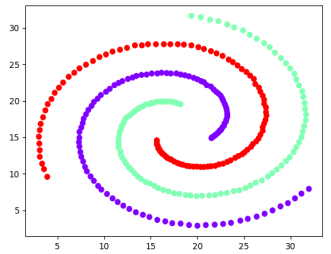

<DBSCAN>

  - 나선 비교


<Ground Truth>


<K-Means>


<Agglomerative>


<DBSCAN>

클러스터링 평가

- 조정 랜드 지수
  - 두 클러스터링 간의 유사성 측정 값을 계산

- 다음 공식을 사용하여 계산:

$$Adjusted\ RI = (RI - Expected\_RI) / (max\,(RI) - Expected\_RI)$$

- 두 개의 매개 변수
  - labels_true
    - Ground truth class labels
  - labels_pred
    - 평가하려는 clusters label

클러스터링 평가

- K-Means를 위한 조정 랜드 지수
  - r15 dataset (r15.csv)

```
from sklearn.cluster import KMeans
import pandas as pd
from sklearn.metrics.cluster import adjusted_rand_score

sample_df = pd.read_csv("D:/r15.csv")

training_points = sample_df[["col1", "col2"]]
training_labels = sample_df["target"]

kmeans = KMeans(n_clusters=15).fit(training_points)

arc = adjusted_rand_score(training_labels, kmeans.labels_
)

print(arc)
```

0.9927781994136302

클러스터링 평가

- DBSCAN을 위한 조정 랜드 지수
  - spiral dataset (spiral.csv)

```
from sklearn.cluster import DBSCAN
import pandas as pd
from sklearn.metrics.cluster import adjusted_rand_score

sample_df = pd.read_csv("spiral.csv")

training_points = sample_df[["col1", "col2"]]
training_labels = sample_df["target"]

dbscan = DBSCAN(eps=3, min_samples=2).fit(training_points)

arc = adjusted_rand_score(training_labels, dbscan.labels_)

print(arc)
```

```
1.0
```

문제풀이

- 병합 군집 분석의 개념을 설명하시오.
- 병합 군집 분석 과정을 설명하시오.
- 군집분석 평가를 위한 조정랜드 지수를 설명하시오.

요약

- 병합 군집 분석의 개념을 공부하였음.
- 병합 군집 분석 과정을 공부하였음.
- 군집분석 평가를 위한 조정랜드 지수를 이해하였음.