노트북: 첫 번째 노트북
만든 날짜: 2021-01-11 오후 10:26
URL: https://continuous-development.tistory.com/208?category=736685

ML,DL

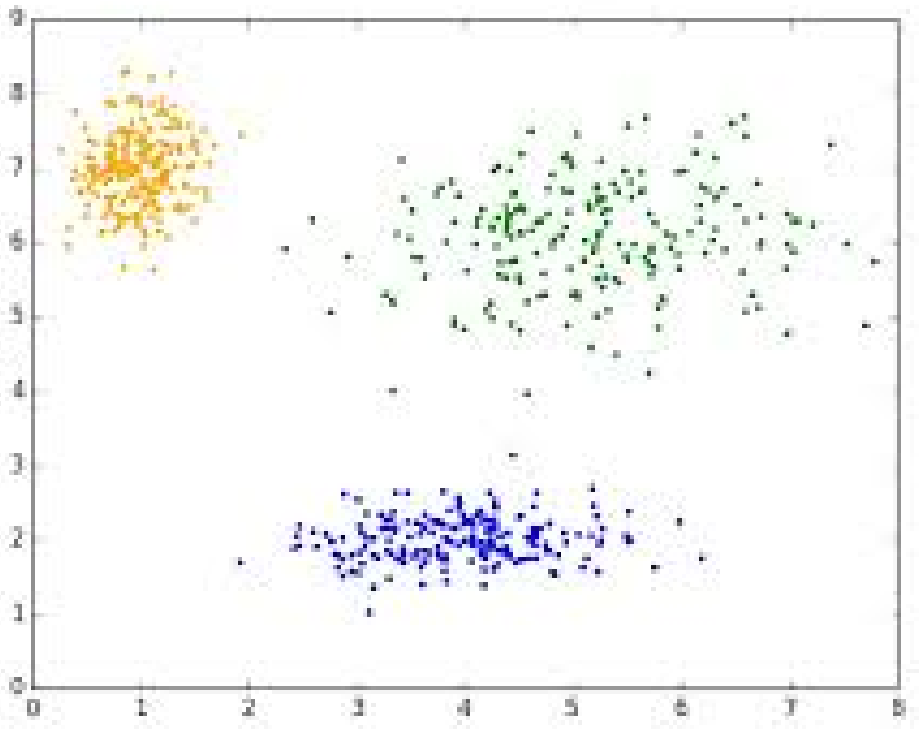
[ML/DL] 군집화의 정의와 종류 및 구현

2021. 1. 7. 05:09 수정 삭제 공개

1.군집화

1-1.군집화란?

데이터들의 특성을 고려해 데이터 집단을 정의하고 데이터 집단의 대표할 수 있는 대표점을 찾는 것
비지도 학습의 종류중 하나로서 **답**을 알지 못하는 상태에서 데이터들 간의 **분할**을 진행하는 방법이다.
그래서 결론은 답없는 애들을 비슷한 애들끼리 끼리끼리 모아 놓는 느낌이다



여기서는 지금 3가지로 분류를 하였는데 저 색깔별로 중심에 점을 두고 **거리**를 계산해서 가까운 곳에 있는 애들을 같은 색으로 칠한다. 그런느낌이다.
군집화를 어떻게 할 것인가에 대한 기준은 거리다! 거리 척도 유형에는 두 가지가 있다.

1-2.거리척도유형

1.유클리디안 거리(Euclidean Distance)

점 $\mathbf{p} = (p_1, p_2, \ldots, p_n)$와 $\mathbf{q} = (q_1, q_2, \ldots, q_n)$가 있을때,

$$\|\mathbf{p} - \mathbf{q}\| = \sqrt{(\mathbf{p} - \mathbf{q}) \cdot (\mathbf{p} - \mathbf{q})} = \sqrt{\|\mathbf{p}\|^2 + \|\mathbf{q}\|^2 - 2\mathbf{p} \cdot \mathbf{q}}.$$

2.맨하탄 거리(Manhattan distance)

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|$$

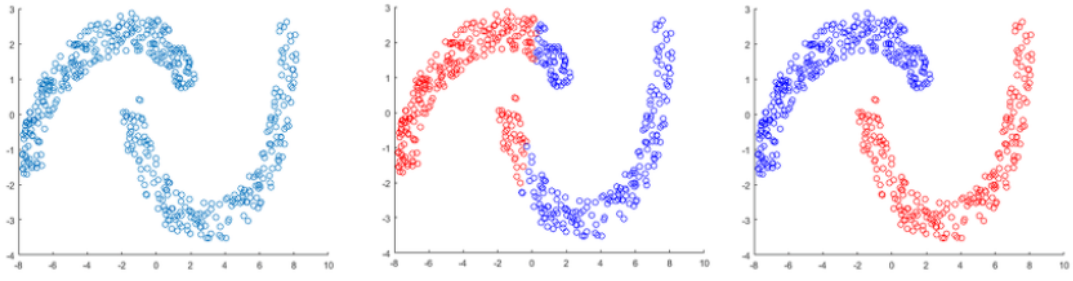
1-3.군집분석의 유형

1.분리형(비계층적) 군집화(Partitioning Clustering)

:사전에 군집의 수를 정해주어 대상들이 군집에 할당되도록 하는 것
ex) K-Means algorithm

분리형에서도 종류가 있다. 여기서도 두가지로 나눠진다.

- 중심기반(Center-based clustering) - 프로토타입 기반이라고도 하며 동일한 군집에 속하는 데이터는 어떠한 중심을 기준으로 분포 할 것 이라는 가정을 기반으로 한다. (K-Means algorithm)
- 밀도기반(Density-based clustering) - 동일한 군집에 속하는 데이터는 서로 근접하게 분포 할것이라는 가정(DBSCAN algorithm)



(a) 원본 데이터 (b) k-means clustering의 결과 (c) DBSCAN의 결과

2.계층적 군집화(Hierarchical Clustering)

: 각 객체가 n개(객체의 수)의 독립적인 각각의 군집에서 출발하여 점차 거리가 가까운 대상과 군집을 이루어 가는 것

1-4.사용 예시

- **고객, 마켓 브랜드, 사회경제 활동 세분화**
- **image 검출, 세분화, 트랙킹**
- **이상 검출 등등**

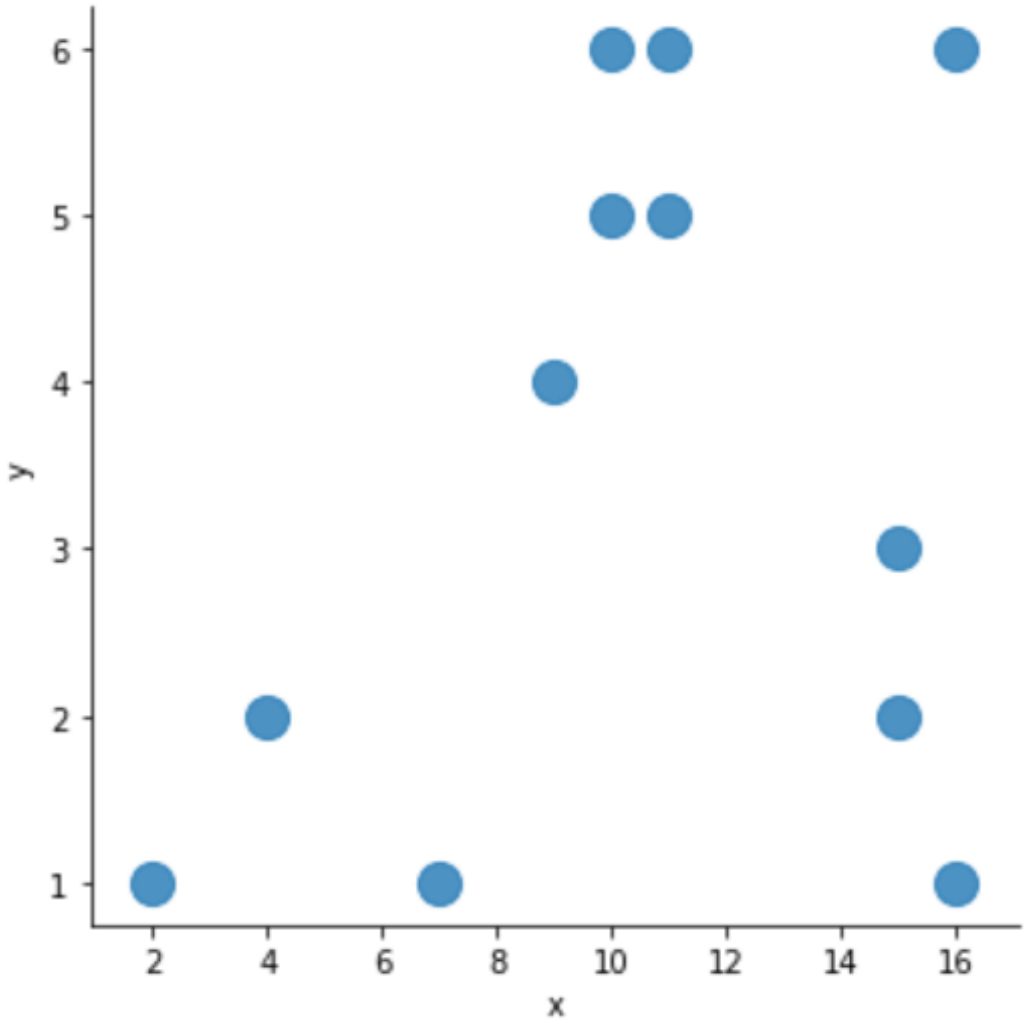
1-5.구현

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

import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline


df = pd.DataFrame(columns=('x','y'))
df.loc[0] = [7,1]
df.loc[1] = [2,1]
df.loc[2] = [4,2]
df.loc[3] = [9,4]
df.loc[4] = [10,5]
df.loc[5] = [10,6]
df.loc[6] = [11,5]
df.loc[7] = [11,6]
df.loc[8] = [15,3]
df.loc[9] = [15,2]
df.loc[10] = [16,6]
df.loc[11] = [16,1]


sns.lmplot('x','y',data=df,fit_reg=False,scatter_kws={'s':200})
```



이렇게 하나하나의 점을 찍는다.

```
data_points = df.values
data_points
```

```
array([[7, 1],
       [2, 1],
       [4, 2],
       [9, 4],
       [10, 5],
       [10, 6],
       [11, 5],
       [11, 6],
       [15, 3],
       [15, 2],
       [16, 6],
       [16, 1]], dtype=object)
```

```
#KMeans 는 가까운 애들로 군집화하는 것이다. 여기서는 3개의 클러스터링으로 만들겠다는것이다.
kmeans = KMeans(n_clusters=3).fit(data_points)
kmeans
```

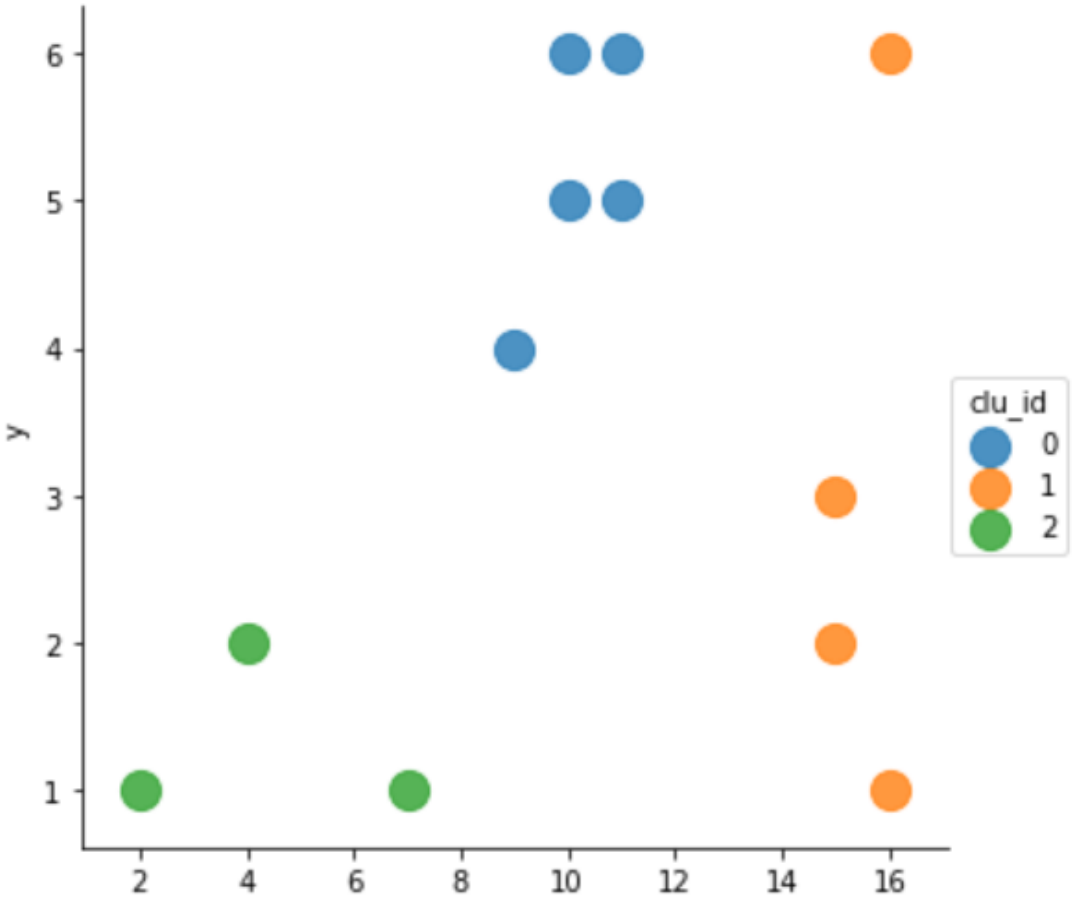
```
array([2, 2, 2, 0, 0, 0, 0, 0, 1, 1, 1, 1])
```

```
# clu_id 라는 컬럼을 만들고 만든 값의 labels를 넣는다.
df['clu_id']= kmeans.labels_
df
```

	x	y	clu_id
0	7	1	2
1	2	1	2
2	4	2	2
3	9	4	0
4	10	5	0
5	10	6	0
6	11	5	0
7	11	6	0
8	15	3	1
9	15	2	1

10	16	6	1
11	16	1	1

```
# hue라는 것을 통해 그룹화 해준다.
# visualzation data point
sns.lmplot('x','y',data=df,fit_reg=False,scatter_kws={'s':200}
           ,hue='clu_id')
```



'ML,DL' 카테고리의 다른 글□

[ML/DL] 회귀(Regression)의 정의와 구현□

[ML/DL] 군집화의 정의와 종류 및 구현□

[ML/DL] XGboost의 정의와 구현 및 hyper parameter 설정□

[ML/DL] 앙상블 학습 (Ensemble Learning): 3.Boosting(부스팅)이란?□

[ML/DL] 앙상블 학습 (Ensemble Learning): 2. Voting(보팅)이란?□

[ML/DL] 앙상블 학습 (Ensemble Learning): 1. bagging(배깅)이란?□

Clustering 군집화 군집화란 군집화의 종류 클러스터링

나아무늘보

혼자 끄적끄적하는 블로그 입니다.