군집분석 성능평가

```
import warnings
warnings.filterwarnings('ignore')
```

I. Import Packages and Lead Dataset

1) Import Packages

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
```

2) Load Dataset

- Load iris Dataset

```
from sklearn.datasets import load_iris

iris = load_iris()
```

- pandas DataFrame

```
DF = pd.DataFrame(data = iris.data,
                  columns = ['sepal_length',
                             'sepal_width',
                             'petal_length',
                             'petal_width'])

DF.head(3)
```

	sepal_length	sepal_width	petal_length	petal_width
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2

II. K-means Modeling

1) Modeling

- n_clusters : 군집 개수 지정
- init : 초기 중심 설정 방식(기본값)
- max_iter : 최대 반복 횟수

```
from sklearn.cluster import KMeans

kmeans_3 = KMeans(n_clusters = 3,
                  init ='k-means++',
                  max_iter = 15,
                  random_state = 2045)

kmeans_3.fit(DF)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=15,
       n_clusters=3, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=2045, tol=0.0001, verbose=0)
```

III. Silhouette Analysis

- 실수엣 계수(Silhouette Coefficient) 측정지표
  - 개별 데이터포인트가 가지는 군집화 지표
  - 데이터포인트가 같은 군집 내의 다른 데이터포인트와 얼나마 가깝게 군집되어 있고
  - 다른 군집에 있는 데이터포인트와 얼마나 멀게 분리되어 있는지 나타내는 지표

- 각 군집 간의 거리가 얼마나 효율적으로 분리되었는지 평가
  - 다른 군집과의 거리는 멀고, 군집 내 데이터포인트 간의 거리는 가깝게 형성

1) DF에 'Clustering' 추가

- 3개로 군집분석한 결과 사용

```
DF['Clustering'] = kmeans_3.labels_
```

```
DF.head(3)
```

	sepal_length	sepal_width	petal_length	petal_width	Clustering
0	5.1	3.5	1.4	0.2	1
1	4.9	3.0	1.4	0.2	1
2	4.7	3.2	1.3	0.2	1

2) 실수엣 계수값

- 개별 데이터포인터들의 실수엣 계수값 계산
  - 'Clustering' 정보 사용

- 실수엣 계수는 -1 ~ 1 사이의 값을 가짐
  - 1에 가까울 수록 근접한 다른 군집과 거리가 멀리 떨어져 있음을 의미
  - 0에 가까울 수록 근접한 다른 군집과 거리가 가까운 것을 의미
  - -1값은 전혀 다른 군집에 데이터포인트가 할당 되었음을 의미

- silhouette_samples( )

```
from sklearn.metrics import silhouette_samples

silhouette_samples(iris.data, DF['Clustering'])
```

```
array([0.85295506, 0.81549476, 0.8293151 , 0.80501395, 0.8493016 ,
       0.74828037, 0.82165093, 0.85390505, 0.75215011, 0.825294  ,
       0.80310303, 0.83591262, 0.81056389, 0.74615046, 0.70259371,
       0.64377156, 0.77568391, 0.85101831, 0.70685782, 0.82030124,
       0.78418399, 0.82590584, 0.79297218, 0.7941134 , 0.77503635,
       0.79865509, 0.83346695, 0.84201773, 0.84364429, 0.81784646,
       0.81518962, 0.79899235, 0.76272528, 0.72224615, 0.82877171,
       0.83224831, 0.79415322, 0.84188954, 0.76856774, 0.85033231,
       0.84941579, 0.63900017, 0.78657771, 0.80023815, 0.74698726,
       0.80977534, 0.81340268, 0.81902059, 0.8182324 , 0.85209835,
       0.02672203, 0.38118643, 0.05340075, 0.59294381, 0.36885321,
       0.59221025, 0.28232583, 0.26525405, 0.34419223, 0.57829491,
       0.37478707, 0.58710354, 0.55107857, 0.48216686, 0.56310057,
       0.32459291, 0.55751057, 0.61072967, 0.46149897, 0.6115753 ,
       0.32909528, 0.58968904, 0.31046301, 0.49424779, 0.5000461 ,
       0.38548959, 0.12629433, 0.11798213, 0.55293611, 0.5069822 ,
       0.59466094, 0.5607585 , 0.61972579, 0.26087292, 0.54077013,
       0.41598629, 0.16655431, 0.48935747, 0.60716023, 0.61436443,
       0.59560929, 0.50352722, 0.62444848, 0.29362234, 0.62754454,
       0.60657448, 0.62205599, 0.55780204, 0.14131742, 0.63064081,
       0.49927538, 0.23225278, 0.61193633, 0.36075942, 0.5577792 ,
       0.54384277, 0.46682151, 0.55917348, 0.44076207, 0.56152256,
       0.26062588, 0.22965423, 0.55509948, 0.28503067, 0.02635881,
       0.39825264, 0.42110831, 0.49486598, 0.48341063, 0.32868889,
       0.6070348 , 0.33355947, 0.51237366, 0.20297372, 0.580154  ,
       0.57818326, 0.30904249, 0.25226992, 0.45434264, 0.51608826,
```

```
0.56017398, 0.48442397, 0.46255248, 0.13900039, 0.05328614,
0.55186784, 0.45549975, 0.3887791 , 0.35124673, 0.53444618,
0.5702338 , 0.41025549, 0.23225278, 0.61324746, 0.5670778 ,
0.42513648, 0.10417086, 0.31493016, 0.35245379, 0.18544229])
```

- 데이터포인트 별 실수엣 계수값 추가

```
DF['Silh_Coef'] = silhouette_samples(iris.data, DF['Clustering'])
```

```
DF.head(3)
```

	sepal_length	sepal_width	petal_length	petal_width	Clustering	Silh_Coef
0	5.1	3.5	1.4	0.2	1	0.852955
1	4.9	3.0	1.4	0.2	1	0.815495
2	4.7	3.2	1.3	0.2	1	0.829315

3) 실루엣 점수(Silhouette Score)

- 데이터포인트들의 실루엣 계수값의 평균

- 권장 실루엣 점수값
  - 전체 실수엣 계수 평균이 0 ~ 1 사이의 값을 가지며, 1에 가까운 경우
  - 개별 군집의 실루엣 계수 평균들이 전체 실루엣 계수 평균과 크게 차이나지 않는 경우

- silhouette_score( )

```
from sklearn.metrics import silhouette_score

silhouette_score(iris.data, DF['Clustering'])
```

```
0.5528190123564091
```

- 3개 군집의 실수엣 점수와 전체 실수엣 점수 비교

```
DF.groupby('Clustering')['Silh_Coef'].mean()
```

```
Clustering
0    0.417320
1    0.798140
2    0.451105
Name: Silh_Coef, dtype: float64
```

#

#

#

The End

#

#

#