Assignment 5

2015313254 노인호

December 1th 2019

1 Progress

1.1 Problem

2개의 초승달 데이터를 비지도 학습을 사용해서 군집화(clustering)하여 시각화하는 문제이다. 반원이 2개 생기는데 DBSCAN이 가장 적합하다고 생각해서 DBSCAN을 이용하여 시각화하였다.

1.2 Data Load

scikit-learn에 있는 make_moon 데이터를 'load_iris()'를 이용하여 로드한다. 샘플수는 100개로 주고 약간의 노이즈를 주었다.

1.3 Data preprocessing

- data scaling은 데이터를 정규화 시켜주는 StandardScaler함수를 사용하였다.

1.4 DBSCAN

Scikit-learn의 skleran.clustering DBSCAN 함수를 사용하였다. hyperparameter는 default 값으로 설정하였다.

1.5 Visualization

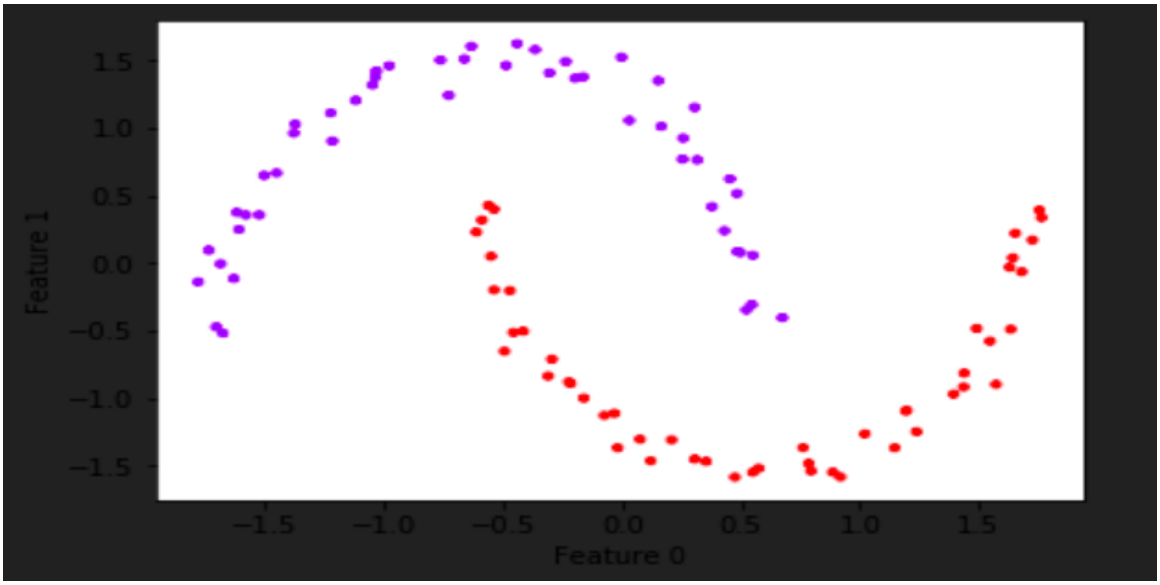

Figure 1: Clustering of make_moon data using DBSCAN

2 Conclusion

군집화가 잘 되었는지 평가해주는 지표인 ARI(adjusted rand index)를 계산해본 결과 1로 나름 잘 됐다고 평가할 수 있다.

3 Python code in jupyter notebook

웹사이트 https://nbviewer.jupyter.org/ 에서 다음의 gist 주소를 입력하면

- https://gist.github.com/nosy0411/a55f9ec2c145b2af93cef71bb44263c3

assignment5.ipynb 파일의 코드를 볼 수 있다.

```
from sklearn.cluster import DBSCAN
from sklearn.datasets import make_moons
from sklearn.preprocessing import StandardScaler

X, y = make_moons(n_samples=100, noise=0.05, random_state=0)

scaler = StandardScaler()
scaler.fit(X)
X_scaled = scaler.transform(X)

dbscan = DBSCAN(eps=0.5, min_samples=5, metric='euclidean')
clusters = dbscan.fit_predict(X_scaled)

X_train_scaled=dbscan.labels_
print(X_train_scaled)
print(clusters)

import matplotlib.pyplot as plt
plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=clusters, cmap=plt.cm.rainbow, s=10)
plt.xlabel("Feature 0")
plt.ylabel("Feature 1")

from sklearn.metrics.cluster import adjusted_rand_score
print("ARI: {:.2f}".format(adjusted_rand_score(y, clusters)))
```