**CS302 Introduction to machine learning**



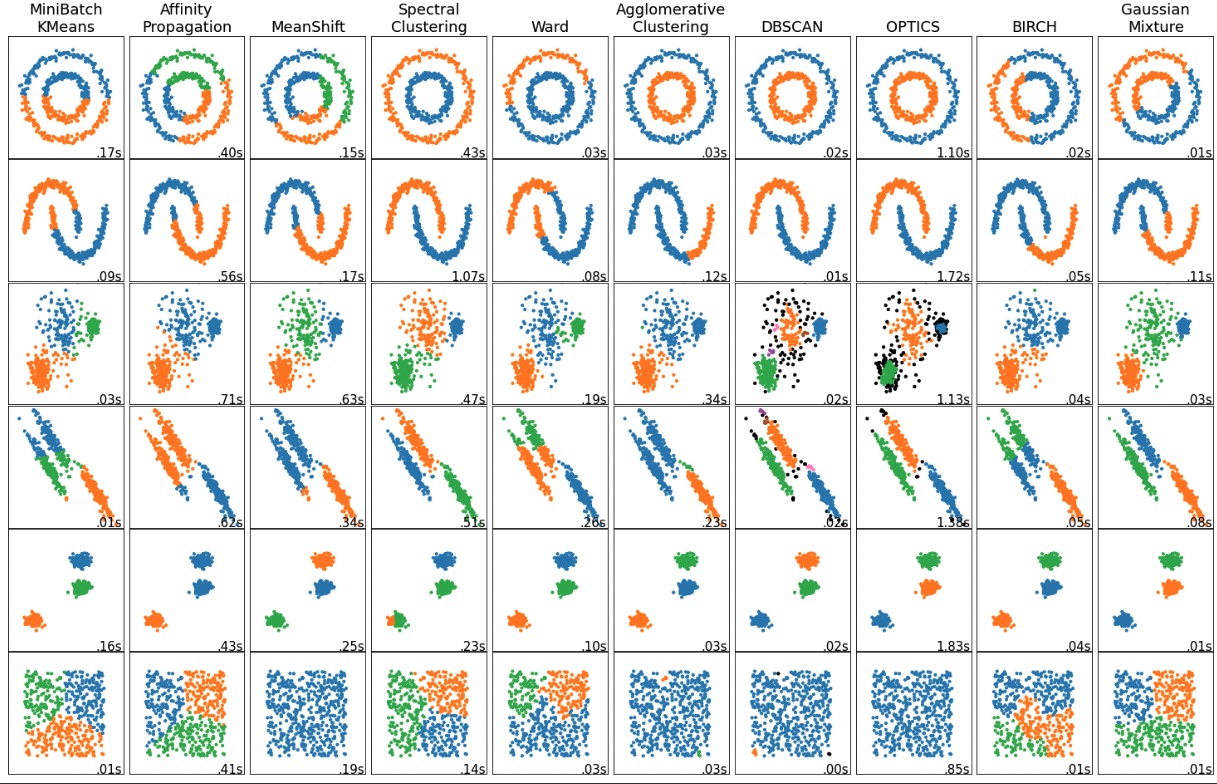
School of Undergraduate Studies, DGIST

201611161 정원균

Homework report 3.

2022. 05. 26

**Q1. Run the code below and arrange the results.**



Unsupervised Learning중 다양한 Clustering 결과를 실행한 결과를 나타낸다.

**Q2. Sample 100 images randomly for each class (total 1000 images) from the MNIST training data set.**

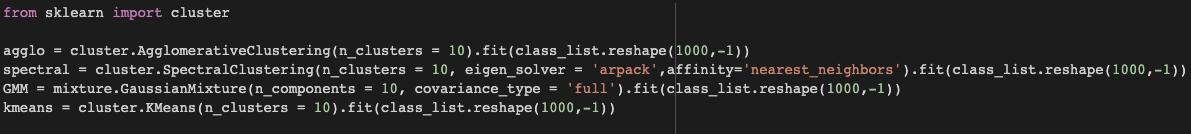
텍스트이(가) 표시된 사진

자동 생성된 설명

다음은 mnist 손글씨 데이터를 불러와서 X\_train , Y\_train, X\_test, Y\_test를 설정해주었고, 각각의 dimension을 맞춰주었다.

그 다음 X\_train과 y\_train이 mapping되는 60000개의 set을 분류하여 (e.g. 0으로 labeling되면 0번째 index로 분류)한 다음 100개의 샘플을 뽑는다. 이 과정에서 데이터의 편향을 막기 위해서 random.shuffle 메소드를 이용하여 각 index마다 셔플링을 한 다음 100개의 데이터를 추출하여 저장한다.

**Q3. For 1000 images, perform Agglomerative clustering, k-means clustering, Gaussian mixture model, Spectral clustering. (i.e., k = 10)**



1000개의 image들 (100 x 10)에 대해서 Agglomerative clustering, k-means clustering, Gaussian mixture model, Spectral clustering을 10개의 cluster로 clustering을 하였다.

**Q4. Based on the clustering results and the labels we know, compute “Rand index” and “mutual information-based score”. Explain your findings.**

4-1) Agglomerative Clustering

텍스트이(가) 표시된 사진

자동 생성된 설명

클러스터링의 성능을 평가하는 대표적인 기준으로 Rand\_index와 Mutual information score이 있는데 Rand\_index는 비교할 두 데이터의 순서쌍을 비교하여서 모든 데이터에 대해 일치하는 데이터 쌍의 갯수를 비율로 나타내며 1에 가까울 수록 좋은 성능의 클러스터링이라는 것을 알 수 있다. 하지만 cluster의 수가 증가할 수록 Rand\_index값이 증가하기 때문에 실제 정확도에 비해서 더 높은 정확도가 나오는 경향성이 있다.

또한 Mutual information score는 두 확률 변수의 의존도를 수치화한 점수이다. 이는 Correlation과 관계가 있는데 Correlation값이 높을 수록 두 확률 변수가 서로 의존도가 높고 0에 가까울 수록 두 확률변수가 독립적이라는 것을 의미한다. 역시 Cluster수가 증가하면 확률변수의 상호의존 정도가 낮아지기 때문에 실제보다 조금 더 낮은 값이 나올 수 있다.

다음은 agglomerative clustering을 진행하고 나온 image와 label을 추출하여 Rand\_index와 Mutual information score을 구해보았다.

약 0.889의 rand\_index와 0.643 normalized Mutual information score을 보임을 알 수 있다. 주어진 수치만 놓고 보았을 때 agglomerative clustering이 좋은 성능을 보이고 각 클러스터를 비교할 수 있을 정도의 상호 의존도를 보임을 알 수 있다.

4-2) Spectral Clustering

텍스트이(가) 표시된 사진

자동 생성된 설명

다음은 Spectral clustering을 진행하고 나온 image와 label을 추출하여 Rand\_index와 Mutual information score을 구해보았다.

약 0.875의 rand\_index와 0.625 normalized Mutual information score을 보임을 알 수 있다. 주어진 수치만 놓고 보았을 때 Spectral clustering이 좋은 성능을 보이고 각 클러스터를 비교할 수 있을 정도의 상호 의존도를 보임을 알 수 있다.

4-3) GMM

다음은 Gaussian Mixture model을 진행하고 나온 image와 label을 추출하여 Rand\_index와 Mutual information score을 구해보았다.

텍스트이(가) 표시된 사진

자동 생성된 설명

약 0.867의 rand\_index와 0.476 normalized Mutual information score을 보임을 알 수 있다. 주어진 수치만 놓고 보았을 때 Gaussian Mixture model이 좋은 성능을 보이고 각 클러스터를 비교할 수 있을 정도의 상호 의존도를 보임을 알 수 있다. 여기서 다른 모델들 보다 상호 의존도가 낮아서 좀 더 클러스터간의 비교하기에 용이한 것을 알 수 있다.

4-4) K-means Clustering

텍스트이(가) 표시된 사진

자동 생성된 설명

다음은 K-means clustering을 진행하고 나온 image와 label을 추출하여 Rand\_index와 Mutual information score을 구해보았다.

약 0.862의 rand\_index와 0.501 normalized Mutual information score을 보임을 알 수 있다. 주어진 수치만 놓고 보았을 때 K-means clustering이 좋은 성능을 보이고 각 클러스터를 비교할 수 있을 정도의 상호 의존도를 보임을 알 수 있다. 역시 앞의 두 모델들 보다 상호 의존도가 낮아서 좀 더 클러스터간의 비교하기에 용이한 것을 알 수 있다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster type | Agglomerative | Spectral | GMM | K-means |
| RI | 0.889 | 0.875 | 0.867 | 0.862 |
| MIS | 1.381 | 1.241 | 1.019 | 1.072 |
| NMIS | 0.643 | 0.625 | 0.476 | 0.501 |

**Q5. Based on the clustering results, you can get the center of each cluster. Classify the MNIST test data set using 1-NN classifier and provide accuracy. Explain your findings.**

5-1) Agglomerative Clustering

텍스트이(가) 표시된 사진

자동 생성된 설명

다음은 Agglomerative Clustering 을 처리한 다음에 0~9 까지 각 cluster의 center값을 찾아서 label과 train image값을 리스트에 넣어서 처리하였다. 그 다음 1-NN classifier을 이용하여 center 값들에 대해 fitting을 진행하였다. 학습이 완료된 1-NN classifier로 실제 Test image prediction을 하였고 y\_test 값과 Prediction한 값의 정확도를 구하였다.

그 결과 총 40%의 정확도가 나옴을 알 수 있는데, 총 6만개의 data 값들 중에 각 label당 100개씩의 data만 Sampling 하여clustering을 진행한 후 classifier로 predict를 진행하였기 때문에 적당한 정확도를 보이지만 더 많은 데이터를 sampling하여 clustering한 후 classifier로 예측을 진행하면 더 좋은 정확도가 나올 것으로 예상할 수 있다.

5-2) Spectral Clustering

텍스트이(가) 표시된 사진

자동 생성된 설명

다음은 Spectral Clustering 을 처리한 다음에 0~9 까지 각 cluster의 center값을 찾아서 label과 train image값을 리스트에 넣어서 처리하였다. 그 다음 1-NN classifier을 이용하여 center 값들에 대해 fitting을 진행하였다. 학습이 완료된 1-NN classifier로 실제 Test image prediction을 하였고 y\_test 값과 Prediction한 값의 정확도를 구하였다.

그 결과 총 46%의 정확도가 나옴을 알 수 있는데, 총 6만개의 data 값들 중에 각 label당 100개씩의 data만 Sampling 하여clustering을 진행한 후 classifier로 predict를 진행하였기 때문에 적당한 정확도를 보이지만 더 많은 데이터를 sampling하여 clustering한 후 classifier로 예측을 진행하면 더 좋은 정확도가 나올 것으로 예상할 수 있다.

5-3) GMM

다음은 Gaussian Mixture Model을 처리한 다음에 0~9 까지 각 cluster의 center값을 찾아서 label과 train image값을 리스트에 넣어서 처리하였다. 그 다음 1-NN classifier을 이용하여 center 값들에 대해 fitting을 진행하였다. 학습이 완료된 1-NN classifier로 실제 Test image prediction을 하였고 y\_test 값과 Prediction한 값의 정확도를 구하였다.

텍스트이(가) 표시된 사진

자동 생성된 설명

그 결과 총 45%의 정확도가 나옴을 알 수 있는데, 총 6만개의 data 값들 중에 각 label당 100개씩의 data만 Sampling 하여clustering을 진행한 후 classifier로 predict를 진행하였기 때문에 적당한 정확도를 보이지만 더 많은 데이터를 sampling하여 clustering한 후 classifier로 예측을 진행하면 더 좋은 정확도가 나올 것으로 예상할 수 있다.

5-4) K-means Clustering

텍스트이(가) 표시된 사진

자동 생성된 설명

다음은 K-means Clustering을 처리한 다음에 0~9 까지 각 cluster의 center값을 찾아서 label과 train image값을 리스트에 넣어서 처리하였다. 그 다음 1-NN classifier을 이용하여 center 값들에 대해 fitting을 진행하였다. 학습이 완료된 1-NN classifier로 실제 Test image prediction을 하였고 y\_test 값과 Prediction한 값의 정확도를 구하였다.

그 결과 총 45%의 정확도가 나옴을 알 수 있는데, 총 6만개의 data 값들 중에 각 label당 100개씩의 data만 Sampling 하여clustering을 진행한 후 classifier로 predict를 진행하였기 때문에 적당한 정확도를 보이지만 더 많은 데이터를 sampling하여 clustering한 후 classifier로 예측을 진행하면 더 좋은 정확도가 나올 것으로 예상할 수 있다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Clustering type | Agglomerative | Spectral | GMM | K-means |
| Accuracy | 40% | 46% | 45% | 45% |

최종 결과를 본다면 가장Naïve 한 Agglomerative Clustering이 가장 낮은 정확도를 보이고 나머지 3개의 Clustering은 비슷하게 조금 더 높은 정확도를 보임을 알 수 있다.