이해동
정보과학과
202584-010115

머신러닝특론 과제3 리포트

서론

본 보고서는 K-means 군집화 알고리즘과 EM 알고리즘을 통하여 가우시안 혼합모델을 통하여 분류기를 만들고 그 결과를 검증합니다.

문제 1 K-means 군집화 알고리즘

문제 1.1 산점도 그리기

본 문제에서는 다음과 같은 평균과 공분산 행렬을 갖는 가우시안 분포를 따르는 세 개의 클래스 데이터를 각각 100개씩 생성하였습니다.

$$\mu1=\begin{pmatrix}0\\4\end{pmatrix}, \mu2=\begin{pmatrix}4\\4\end{pmatrix}, \mu3=\begin{pmatrix}2\\0\end{pmatrix} \quad \Sigma1 = \Sigma2 = \Sigma3 = \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$$

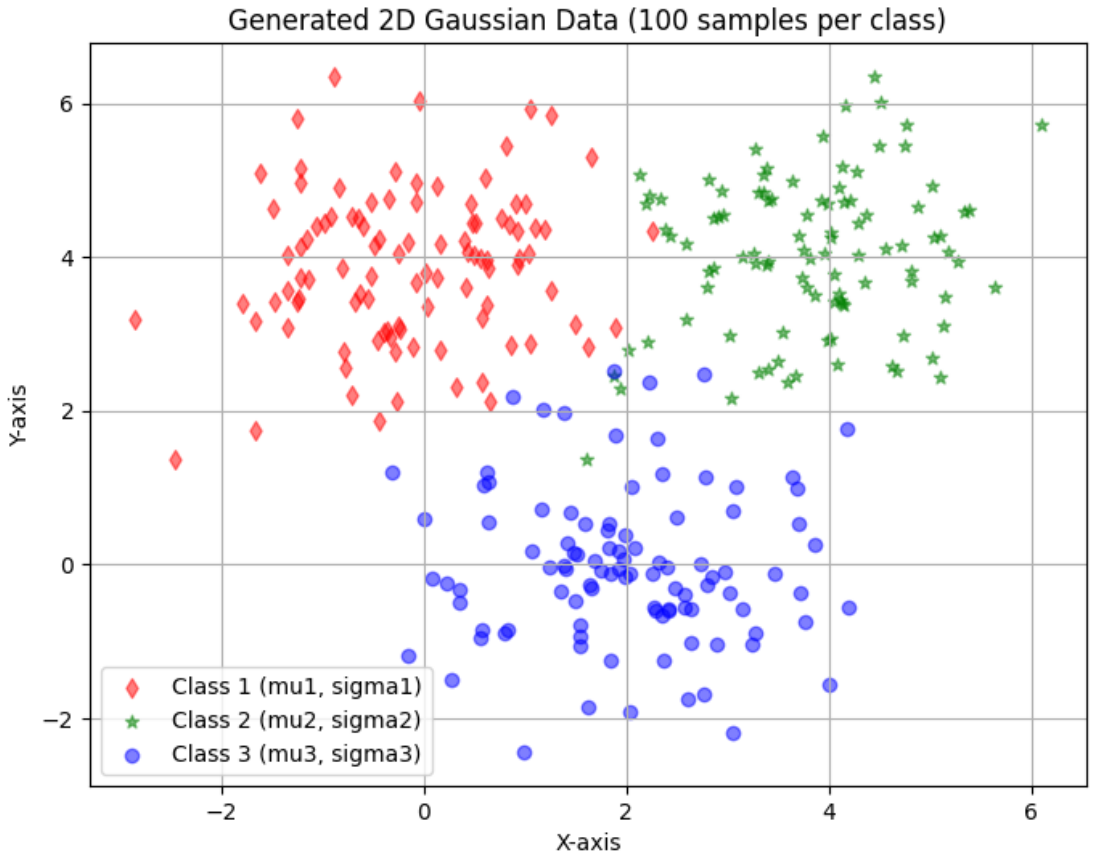
생성된 데이터의 산점도는 <그림 1>과 같습니다.

산점도는 python 의 numpy 라이브러의 np.random.multivariate_normal() 함수를 통하여 주어진 평균과 공분산을 가진 가우시안 분포를 따르는 데이터를 생성하였습니다.

Source code : K_Means1_1.py

결과

- m1 은 Class 1 로 붉은색 다이아몬드로 표현됨
- m2 은 Class 2 로 녹색 별로 표현됨
- m3 은 Class 3 로 파란색 원으로 표현됨



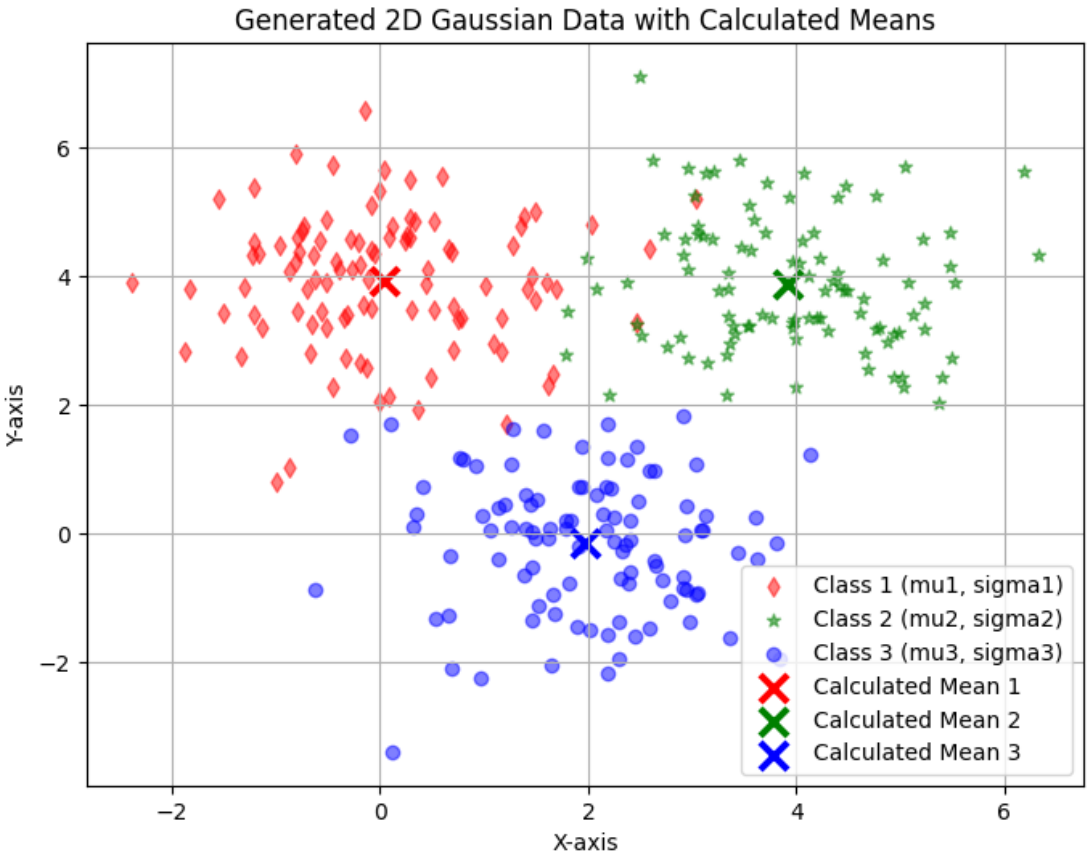
<그림 1>

문제 1.2 각 클래스의 평균을 계산하여 표시

앞서 생성한 산점도 데이터를 기반으로 각 클래스의 데이터의 평균을 계산하여 산점도 상에 X 자로 표시 하였습니다. 평균의 계산은 numpy 라이브러리의 np.mean() 함수를 사용하였습니다.

Source code : K_Means1_2.py

결과



<그림 2>

문제 1.3 K-means 알고리즘을 적용하여 군집화를 수행하고 그 변화 양상을 표시

K-means 알고리즘을 사용한 군집화는 반복을 통하여 이루어 집니다.

처음 <그림3>의 데이터는 분류가 되지않은 데이터를 보여줍니다. 검은색 사각형은 무작위로 선택된 군집의 중심점 입니다.

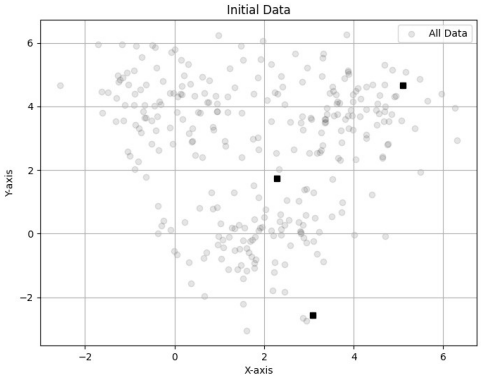
<그림4> 는 첫번째 반복이 이루어진 결과를 보여줍니다. 붉은색으로 많은 데이터가 분류되었습니다.

<그림7> 에서 더이상 반복을 해도 분류의 변화가 없음을 알 수 있습니다.

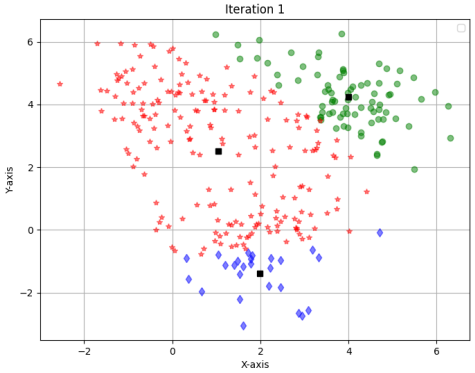
<그림8> 에서는 K-means 알고리즘을 통하여 분류한 데이터들을 계산한 중심점과 원래 데이터를 더하여 평균을 구한 점 (x 표시) 를 함께 표시하였습니다.

Source code : K_Means1_3.py

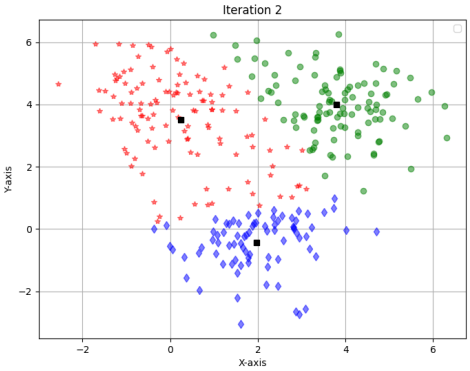
결과



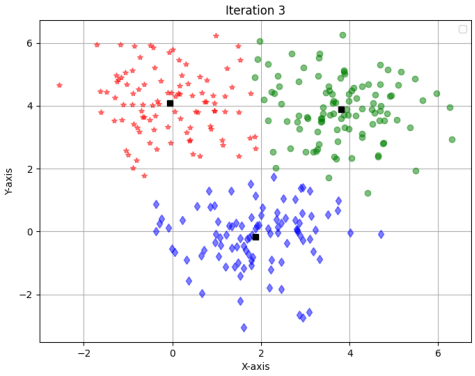
<그림 3>



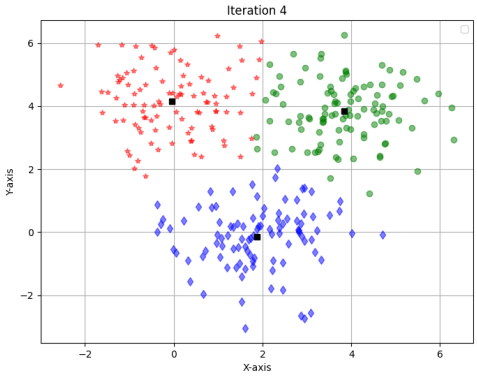
<그림 4>



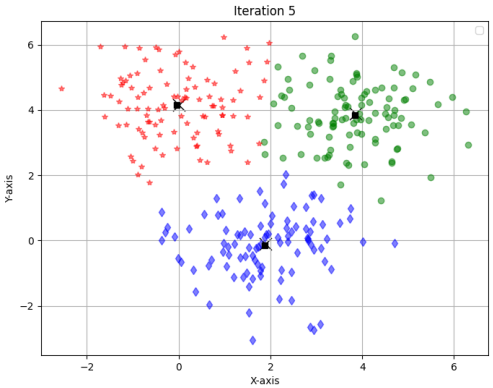
<그림 5>



<그림 6>



<그림 7>



<그림 8>

문제 2 EM 알고리즘과 가우시안 혼합 모델을 사용한 군집화

2.1 산점도 표시

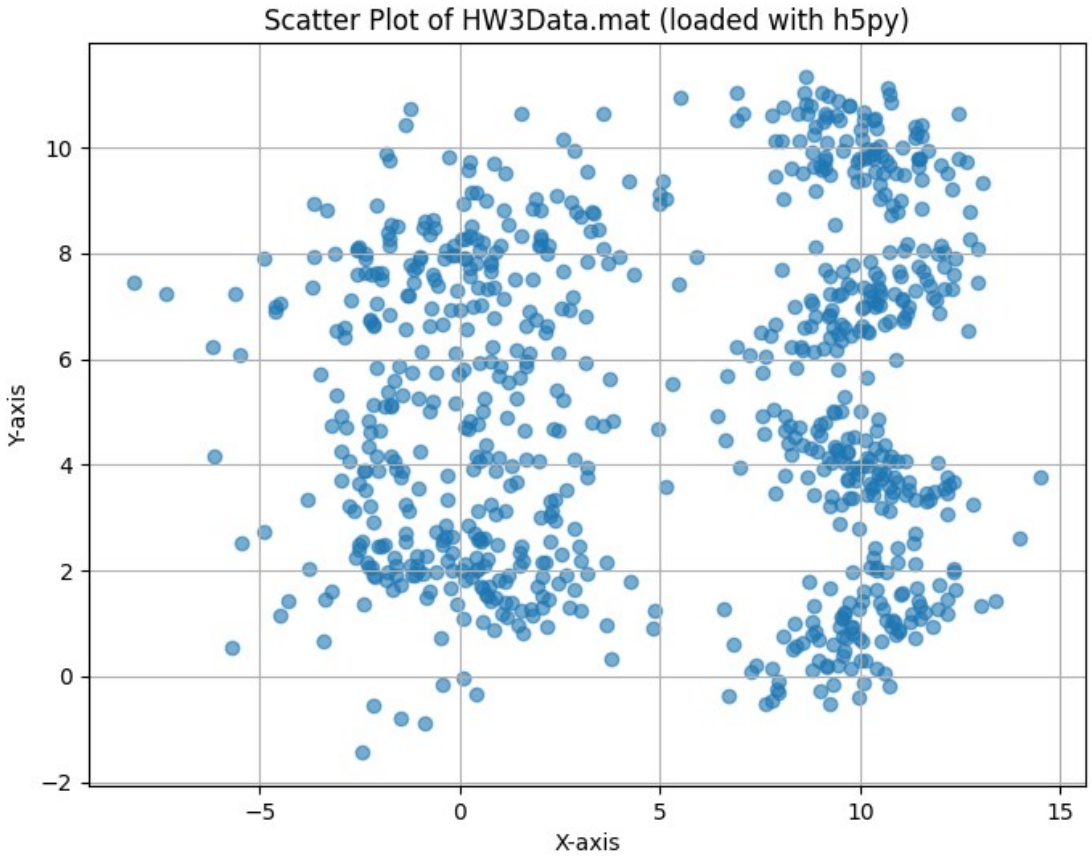
HW3Data.mat 파일을 읽어서 산점도를 표시 합니다. HW3Data.mat 에는 800 개의 데이터가 들어 있습니다.

Python 에서는 h5py 라이브러리를 사용하여 matlab 파일을 읽어들입니다.

<그림9> 에 결과를 표시합니다.

Source code : Gaussian_Mixture2_1.py

결과



<그림 9>

2.2 가우시안 혼합 모델을 적용하여 분석 수행

성분수가 2인 경우 빠르게 군집을 찾고, 반복을 하여도 큰 변화가 없었습니다.

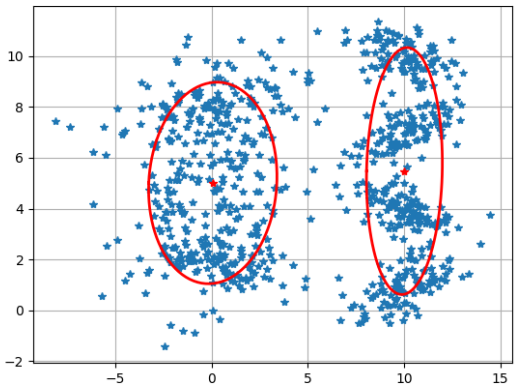
성분수가 6, 10인 경우는 성분수가 2인 경우 보다는 최적의 군집을 찾는데 더 많은 반복을 필요로 하였습니다.

성분수가 늘어날 수록 더 많은 반복을 필요로 하는 것으로 파악됩니다.

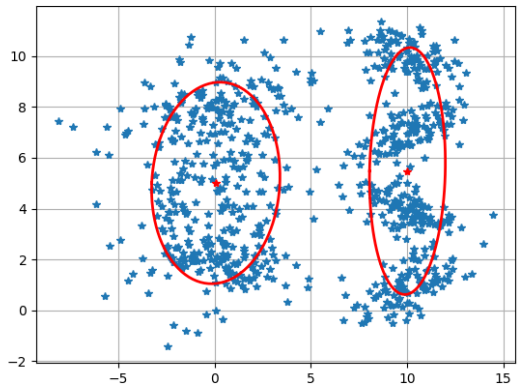
Source code : Gaussian_Mixture2_2.py

결과

- **성분수 2**

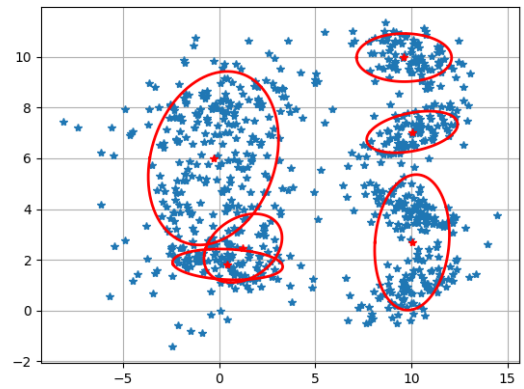


시작

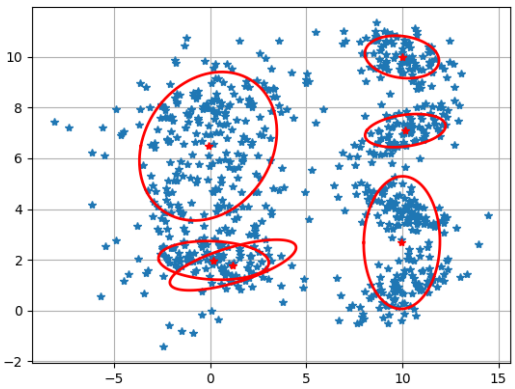


반복 종료

- **성분수 6**

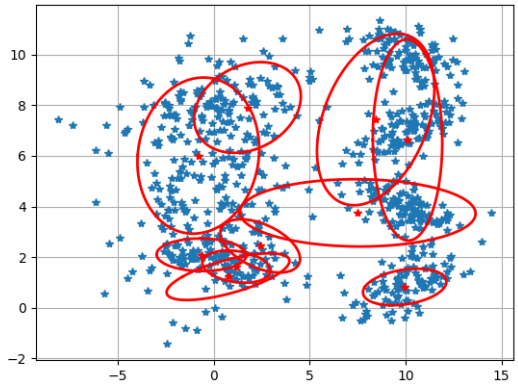


시작

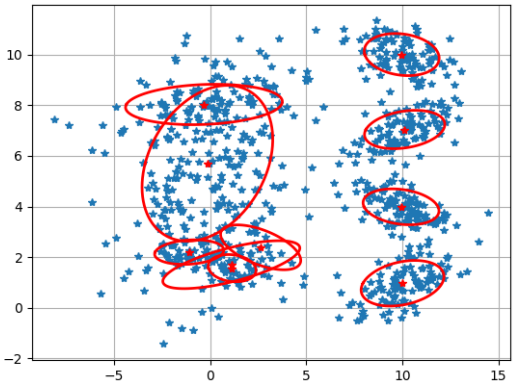


반복 종료

- **성분수 10**



시작



반복 종료