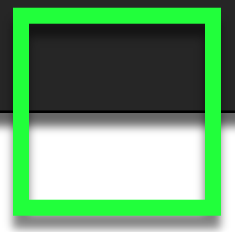
#수치 해석
#프로젝트 2

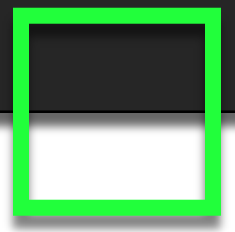
Pattern Recognition

컴퓨터소프트웨어학부
2018008395 박정호



#1

구현



데이터의 분포

이번 과제는 필자가 임의로 설정한 데이터의 분포에 대해서 K-Means Clustering Method를 적용해서 어떤 데이터가 어느 cluster에 속하는지를 인식하는 모델링을 하는 것이었다. 따라서, 우선 학습 및 테스트 데이터의 분포에 대해서 미리 언급하려 한다. Cluster 6은 모델의 테스트 과정에서 사용하는 학습된 데이터 외의 클러스터이다. 분석 과정에서도 이 분포를 기반으로 약간의 변경을 더할 계획이다.

	m_x	σ_x^2	m_y	σ_y^2	m_z	σ_z^2
Cluster 1	0	1	0	1	0	1
Cluster 2	5	2	6	0.8	7	1.5
Cluster 3	1	0.8	2	0.5	3	0.8
Cluster 4	10	2	7	3	1	2
Cluster 5	3	1.5	12	1	7	0.5
Cluster 6	-1	4	-3	0.5	0.7	0.5

모델링 – maximum distance의 설정

기본적으로 pattern recognition의 criterion은 K-Means Clustering으로 인해 만들어진 각 cluster의 mean vector 이다. 이 mean vector 와 vector distance를 구해서 가장 작은 거리를 가진 cluster를 주어진 데이터의 cluster로 보는 것이 pattern recognition이었는데, 여기서 하나의 문제가 있다.

학습 데이터 집단과 아주 멀리 떨어진 데이터의 경우 실제론 어느 cluster에도 속한다고 보기 어려우나 단순히 vector distance를 비교하는 것만으로는 이러한 경우의 처리가 불가능했다. 그래서 어떤 데이터가 어떤 cluster 에 속한다고 할 수 있는 vector distance 의 상한이 필요했다. 이것이 maximum distance 이다.

필자의 경우 K-Means Clustering 을 사용한 직후에 maximum distance 를 구했는데, 그 방식은 다음과 같다. 우선 구해진 각 cluster 에 대해, cluster 에 속한 데이터와 mean vector 의 vector distance의 최댓값을 구한다. 이렇게 하면 각 cluster 의 maximum distance를 구할 수 있다. 이후 이 distance를 평균을 내서, 그 값에 0.8을 곱하여 전체 cluster 에 대한 maximum distance 를 구했다.

모델링 – maximum distance의 설정

단순하게 말하자면, 각 cluster 의 maximum distance 의 평균을 약간 scaling 한 것이라 볼 수 있다. 이런 방식을 선택한 이유는 다음과 같다.

1. Cluster에 속하기 위한 상한은 결국 각 cluster 의 maximum distance와 관련이 있다.
   만약 각 cluster 를 생성하는 분포의 분산이 크다면 당연히 maximum distance 의 크기도 클 것이다. 반대로 cluster 를 생성하는 분포들의 분산이 작다면 maximum distance 도 작아져야 한다. 따라서 각 cluster의 maximum distance 값을 모두 반영할 수 있는 평균을 선택했다.

2. 평균만으로는 정보가 부족하다.
   단순히 cluster의 maximum distance의 평균만을 사용하게 되면, 특정 경우에서 적절하지 못한 maximum distance 를 구할 수도 있다. 예를 들어, 데이터 셋의 어떤 cluster가 특정 방향으로만 분산이 크고 나머지 방향으로는 분산이 작을 경우, 혹은 다른 cluster 들은 모두 maximum distance가 작은데 하나의 cluster만 maximum distance가 클 경우, 평균만으로는 왜곡이 생길 수 있다. 따라서 이 값을 어느정도 scaling하는 것이 필요하다고 판단했고, 여러 테스트의 결과 0.8이 제일 정확도가 높은 것을 확인했다.
   (단 이는 데이터의 분포에 따라 차이가 있을 수 있다.)

테스트

모델링한 코드가 제대로 동작하는지 확인할 때 학습 데이터를 생성하는 분포에 약간의 변경을 더해가면서 분포의 형태에 따라 정확도가 얼마나 변화하는지를 분석했다.

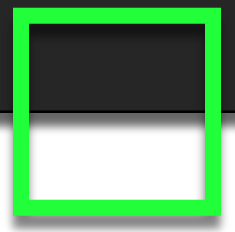
매 테스트 실행마다 학습 데이터와 테스트 데이터는 새로 생성했으며, 이로 인해 같은 분포로도 다른 결과가 나올 수 있었다. 따라서 매 분포마다 5번 실행을 해서 그 대략적인 경향성을 분석하려 한다.

테스트하는 데이터 셋은 크게 2가지로 나뉜다.

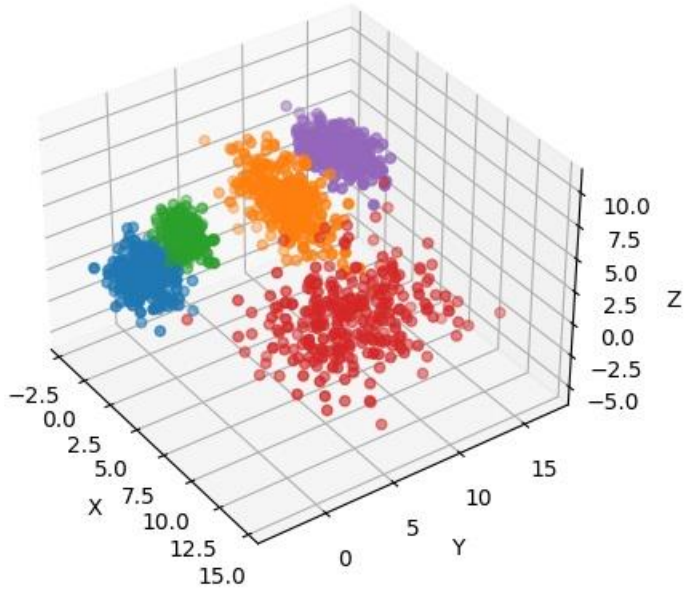
1. 각 클러스터 간의 mean vector 의 거리에 따른 변화
    1. 기존 분포
    2. 기존 분포보다 거리가 2배, 4배 멀 때
2. 각 클러스터의 분포 정도(분산)에 따른 변화
    1. 기존 분포
    2. 기존 분포보다 분산이 0.5, 1 클 때

#2

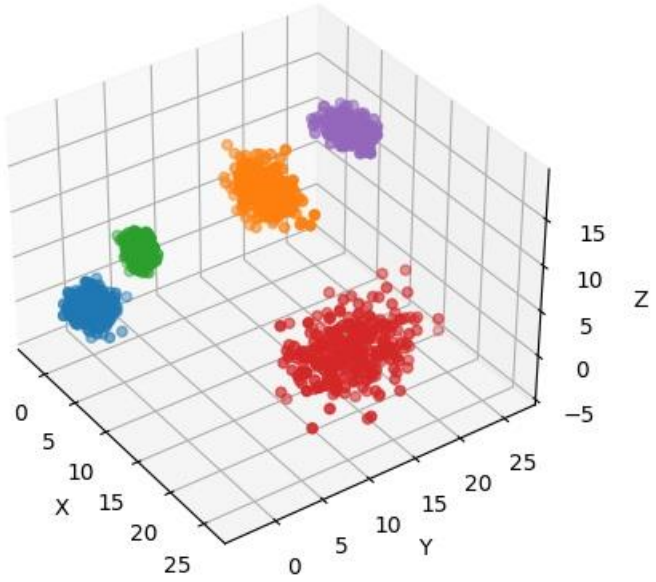
결과와 비교



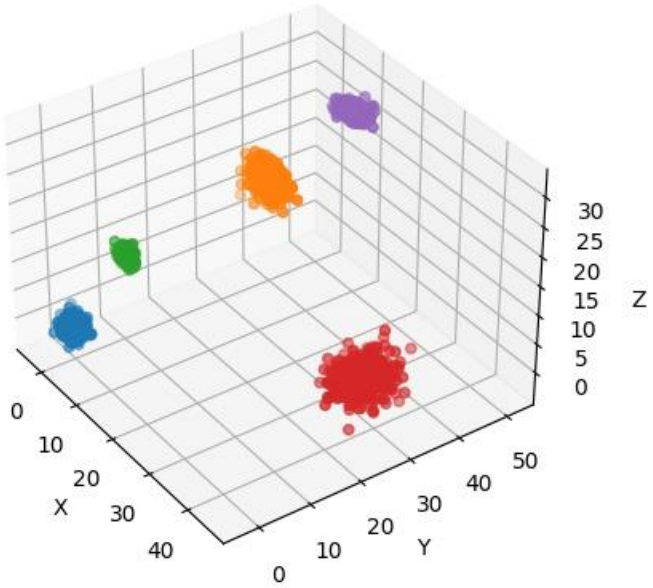
Mean Vector 간의 거리에 따른 데이터 분포



기본 데이터 분포



Mean Vector 간 거리 2배



Mean Vector 간 거리 4배

Mean Vector 간의 거리에 따른 결과

```
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 85.333333%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 87.333333%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 89.500000%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 84.166667%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 88.833333%
```

기본 데이터 분포

```
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 92.166667%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 95.500000%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 96.833333%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 95.333333%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 96.166667%
```

Mean Vector 간 거리 2배

```
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 95.833333%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 91.833333%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 92.166667%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 94.000000%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 95.166667%
```

Mean Vector 간 거리 4배

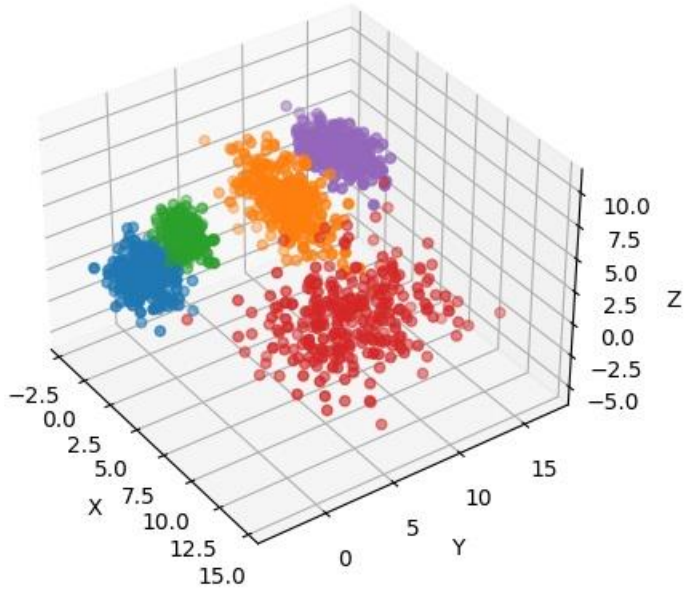
Mean Vector 간의 거리에 따른 분석

기본으로 설정한 데이터 분포가 어느 정도의 중첩이 있어서인지, 85% 내외의 정확도를 보이고 있었다. 한편, 거리를 2배로 한 것과 거리를 4배로 한 것은 93% 내외의 정확도를 보여주고 있었다. 이는 각 cluster 간의 분산은 유지되고 있는데 반해, 거리가 멀어지면서 cluster 간의 중첩이 줄어들었기 때문일 것이다.

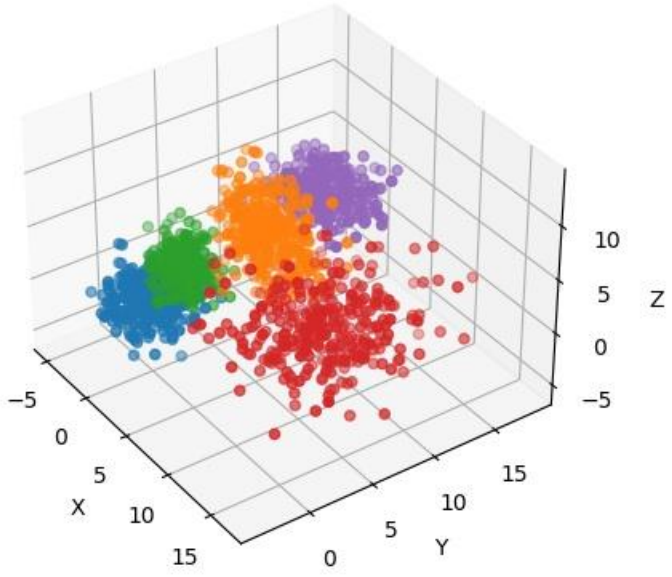
다만, cluster간의 경계가 명확해졌음에도 불구하고 정확도가 100%가 되지는 않았는데, 이는 maximum distance를 평균을 통해 구하면서 상대적으로 분산이 큰 cluster 가 false negative를 야기하는 일이 많아졌기 때문이다.

또한 거리를 2배로 한 결과와 4배로 한 결과 간의 정확도 차이가 거의 없는 것을 확인할 수 있는데, 이는 Mean Vector 간의 거리는 결국 각 cluster 간의 중첩되는 정도에만 관여하기 때문이다. 따라서 중첩되는 것이 없어지고 cluster 간의 경계가 명확해지는 거리에 도달한 경우, 그 이상으로 거리가 멀어져도 정확도가 높아질 수 없는 것이다. 즉, cluster 간의 거리가 모델링의 정확도에 완전히 비례한다고 볼 수는 없고, 어느 임계지점이 있다는 것을 확인할 수 있었다.

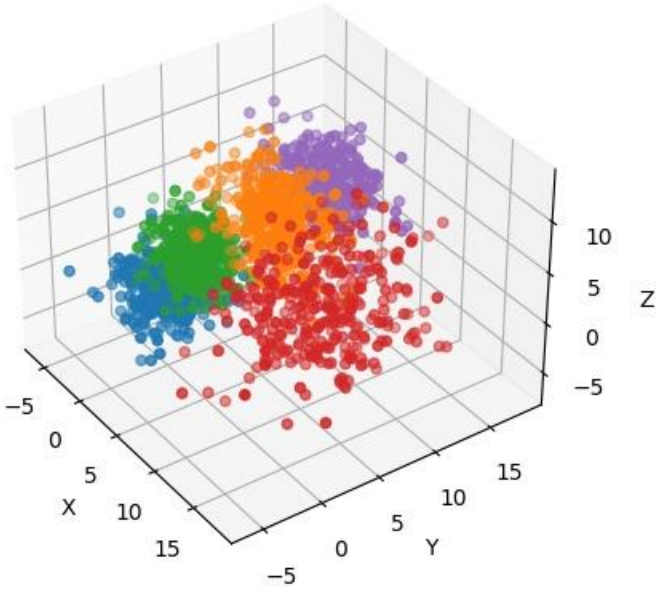
분산의 크기에 따른 데이터 분포



기본 데이터 분포



분산 + 0.5



분산 + 1

분산의 크기에 따른 결과

```
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 85.333333%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 87.333333%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 89.500000%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 84.166667%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 88.833333%
```

기본 데이터 분포

```
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 76.000000%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 77.000000%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 75.833333%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 76.833333%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 76.166667%
```

분산 + 0.5

```
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 58.666667%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 68.833333%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 67.666667%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 70.500000%
PS C:\Users\pch68\HYU_MAT3008\Project 2> python
Making Random Clustered Data... (for Learning)
Evaluate Clustering...
Making Random Clustered Data... (for Testing)
Testing...
Clustering Accuracy : 67.666667%
```

분산 + 1

분산의 크기에 따른 분석

분산이 0.5 늘어날 때마다 거의 10% 정도의 정확도 저하가 일어나는 것을 볼 수 있었다. 이는 분산이 늘어나면 곧 각 cluster 의 분포 자체가 퍼지게 되고, 이로 인해 cluster 간의 중첩이 더 많아지기 때문이다. cluster 간의 중첩이 늘어난다는 것은 데이터 분포 그래프만 보아도 알 수 있다.

또한 분산이 커질 수록 5번의 시도 간의 정확도 차이도 커지는 것을 볼 수 있는데, 이는 분산이 커지면서 데이터의 분포가 더욱 무작위에 의존하게 되기 때문이다. 분산이 크더라도 어떤 데이터는 평균점에 더 모이게 되고, 어떤 데이터는 좀더 바깥쪽에 퍼지게 되는데, 이 차이는 분산의 크기에 따라 달라지기 때문이다. 즉, 분산이 커지면 커질 수록 학습하기가 더 어렵다는 것을 알 수 있다.

다만 분산이 커질 수록 각 cluster의 크기 차이가 작아지는 것을 확인할 수 있었는데, 이로 인해 maximum distance 로 인한 정확도 저하는 적어졌으리라 예상한다. 즉, 지금 결과에서 확인할 수 있는 정확도 저하는, maximum distance 로 인한 정확도 저하가 적어졌음에도 발생한 만큼, 실제론 10%보다 클 것이라는 것을 알 수 있다.