



수치해석 Term Project

2018008613 안상욱



1. Overview

5개의 cluster 의 중심을 임의로 지정해 주었습니다.

Cluster의 중심에서 각각 normal distribution 을 이용해 중심마다 300개씩 총 1500개의 데이터를 만들어 주었습니다.

1500개의 데이터를 5개의 중심을 가지고 K-means clustering 을 10번 시행해 새로 만들어진 5개의 중점과 각각의 중점으로부터 그 중점이 포함된 cluster 에서의 최대 길이를 저장해 주었습니다. 이 최대 길이보다 거리가 짧은 것들만 cluster 에 포함하도록 설정했습니다.

테스트를 할 때 초기에 선정했던 5개의 중심에서 같은 방식의 normal distribution 을 이용해 중심마다 100개씩 총 500개의 데이터를 생성했고, 임의의 한 점을 지정해서 그곳에서 100개의 데이터를 생성해 총 600개의 데이터를 테스트해서 각각 100개의 데이터가 어느 cluster에 포함되는지 결과를 출력해 보았습니다.

2. 중심 설정

cluster0의 중심점은 0, 0, 0으로 설정했습니다. X, y, z 각각 $N(0,1)$, $N(0,1)$, $N(0,1)$ 을 이용해 300개의 데이터를 만들었습니다.

cluster1의 중심점은 4, 0, 0으로 설정했습니다. X, y, z 각각 $N(4,1)$, $N(0,4)$, $N(0,1)$ 을 이용해 300개의 데이터를 만들었습니다.

cluster2의 중심점은 4, 8, 0으로 설정했습니다. X, y, z 각각 $N(4,1)$, $N(8,4)$, $N(0,4)$ 을 이용해 300개의 데이터를 만들었습니다.

cluster3의 중심점은 0, 8, 0으로 설정했습니다. X, y, z 각각 $N(0,1)$, $N(8,4)$, $N(0,1)$ 을 이용해 300개의 데이터를 만들었습니다.

cluster4의 중심점은 2, 4, 3으로 설정했습니다. X, y, z 각각 $N(2,2.25)$, $N(4,2.25)$, $N(3,2.25)$ 을 이용해 300개의 데이터를 만들었습니다.

Test에서 different distribution에서 사용한 중심점은 2, 4, -5이고, x, y, z 각각 $N(2, 1)$, $N(4, 1)$, $N(-5,1)$ 을 이용해 100개의 테스트 케이스를 만들었습니다.

이와 같이 만들었을 때 cluster0은 cluster1, 3, 4와 겹치는 부분이 있고

Cluster1은 cluster0, 2, 4와 겹치고, cluster2는 cluster 1, 3, 4와 겹치고, cluster3은 cluster 0, 2, 4와 겹치고, cluster4는 cluster 0,1,2,3과 모두 겹치도록 설정해 주었습니다.

그리고 test를 위한 distribution 은 cluster 1,2,3,4와 겹치도록 설정해 주었습니다.

3. K-means clustering

K-means clustering 과정을 10번 수행해 주었습니다.

먼저 1500개의 점을 5개의 중심 중 가장 가까운 중심 쪽 cluster 에 각각 넣어주었습니다. 1500개의 데이터의 분리가 끝나면 cluster 별로 중심을 구해서 새로운 중심을 만들어 주었습니다.

10번의 clustering 과정이 끝나면 각각의 cluster에서 중심으로부터의 거리가 가장 먼 값을 저장해 주었습니다.

Test에서 이 값보다 거리가 먼 점은 cluster 에 포함시키지 않도록 기준을 설정해 주었습니다.

4. Testing

- 기존에 설정해 주었던 5개의 중점에서 기존의 Normal distribution을 이용해 각각 100개의 데이터를 생성했고, 앞에서 말한 test를 위해 생성한 (2,4,-5)에서도 100개의 데이터를 생성해 각각 test를 해 주었습니다.
- 100개의 데이터를 가장 가까운 cluster로 분류했는데, 중심과의 거리가 앞에서 설정한 값보다 작은 경우만 그 cluster로 분류해 주었습니다.

5. 출력 결과

출력 결과 오른쪽 그림과 같은 결과가 출력되었습니다.

Cluster 0, 1, 2, 3, 4의 중심은 각각 (0,0,0) (4,0,0) (4,8,0) (0,8,0) (2,4,3)이고, test를 위해 설정한 중심은 (2,4,-5)였습니다.

Cluster0은 cluster1, 3, 4와 겹치는 부분이 있어 cluster1로 1개 잘못 분류되어 99%의 정확도를 가짐을 확인해 볼 수 있었습니다.

Cluster1은 cluster0, 2, 4와 겹치는 부분이 있어 cluster0으로 1개, cluster4로 1개 잘못 분류되어 98%의 정확도를 가짐을 확인해 볼 수 있었습니다.

Cluster2은 cluster1, 3, 4와 겹치는 부분이 있어 cluster4로 7개 잘못 분류되어 93%의 정확도를 가짐을 확인해 볼 수 있었습니다.

Cluster3은 cluster0, 2, 4와 겹치는 부분이 있어 cluster2로 3개, cluster4로 3개, 거리가 기준보다 길어서 cluster에 포함되지 않은 데이터로 1개 잘못 분류되어 93%의 정확도를 가짐을 확인해 볼 수 있었습니다.

Cluster4은 cluster0, 1, 2, 3과 겹치는 부분이 있어 cluster0으로 7개, cluster1로 3개, cluster2로 3개, cluster3으로 5개 잘못 분류되어 82%의 정확도를 가짐을 확인해 볼 수 있었습니다.

새로 만들어준 점에서의 distribution은 cluster 0, 1, 2, 3과 겹치는 부분이 있어 cluster0으로 1개, cluster1로 5개, cluster2로 31개, cluster3으로 8개, 거리가 기준보다 길어서 cluster에 포함되지 않은 데이터로 55개 잘못 분류된 것을 확인해 볼 수 있었습니다.

```
cluster0
[99, 1, 0, 0, 0, 0]
cluster1
[1, 98, 0, 0, 1, 0]
cluster2
[0, 0, 93, 0, 7, 0]
cluster3
[0, 0, 3, 93, 3, 1]
cluster4
[7, 3, 3, 5, 82, 0]
another
[1, 5, 31, 8, 0, 55]
```

6. 결과 분석

cluster 0, 1, 2, 3은 서로 다른 2개의 cluster와 겹치는 부분이 있고 cluster4와 공통적으로 겹치지만, cluster4는 cluster 0, 1, 2, 3 모두와 겹치도록 설정해 주었습니다. 따라서 cluster4가 비교적 다른 cluster와의 충돌이 많으므로, test를 해 봤을 때 cluster4에서 정확도가 가장 낮은 것을 확인해 볼 수 있었습니다.

그리고 cluster 2에서만 z 값에 대해 분산을 4로 설정해 주어서 cluster4와 test를 위해 생성한 점 2, 4, -5로부터의 데이터들과 많이 충돌할 가능성이 크기 때문에 cluster2의 테스트 결과 cluster 0, 1, 3보다 cluster4로 많이 clustering 되었고, cluster4의 테스트 결과 cluster 0, 1, 3보다 cluster2로 잘못 clustering 된 데이터가 많은 것을 확인해 볼 수 있었습니다.

그리고 2, 4, -5에서의 테스트 결과에서도 cluster 0, 1, 3보다 cluster2로 분류된 것이 더 많은 것을 확인해 볼 수 있었습니다.