AI Homework No. 4 submission

KimJaeHwan

December 3, 2024

Problem 1e: Clustering 2D points [2 points]

You have now completed all parts needed to run main(). Now, let's verify your implementation by running the algorithm on 2D points. At the bottom of kmeans.py, there are codes for loading the 2D data points and the initial centroids we provided in from .csv files, and calling the main function. Run the program by typing:

```
python kmeans.py
```

in ther terminal window. If you are successful, you should see:

```
K-means converged after 7 steps.
```

as the output of the program. In addition, there should be 7 plots generated, in the `results/2D` folder. **Attach the 7 plot images to** `submission_studentid.pdf`

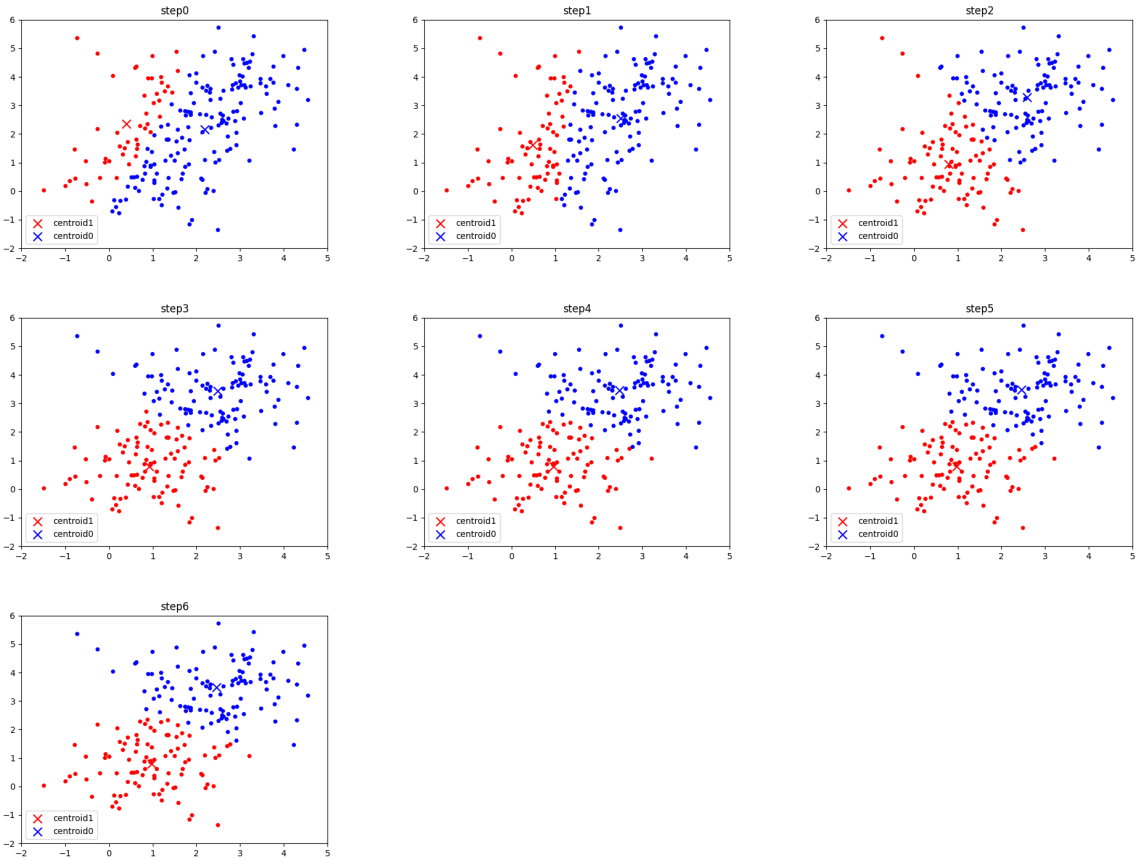

Figure 1: Result plots of K-mean algorithm

Problem 1f: Trying the Algorithm on MNIST [2 points]

... Description of MNIST is omitted. Note `ai_assn4.pdf` for more details

Why the centroid of each cluster looks like an actual digit? This question will be ungraded. But if you do not write any answers, you will get 0 points. We want you to at least think about what happened to the centroids. **Write your answers in `submission_studentid.pdf`.**

Answer

흰색인 값들의 (x, y) 좌표의 평균을 찾는 것이 아니라, 각 픽셀의 값들의 평균을 찾는 것이기 때문이다. 즉, 센트로이드 이미지의 각 픽셀은 클러스터에 포함된 이미지들의 각 픽셀의 평균값을 가지므로, 데이터포인터에서 흰색의 빈도가 높은 픽셀은 센트로이드 이미지에서도 흰색의 빈도가 높게 나타나게 된다. 따라서, 센트로이드가 실제 이미지와 유사한 형태를 띄게 된다.

Problem 2d: Clustering 2D points [2 points]

... Description of soft K-means is omitted. Note `ai_assn4.pdf` for more details

Let's compare the newly generated plots with the plots generated in Problem 1e. **Also, attach the new 7 plots to `submission_studentid.pdf`.**

As you may have already noticed, the hyper-parameter β is used when calculating responsibility. Change the value of β to 50 in main() function of `soft_kmeans.py` file. What changes have been made to the plots? Think about how β affects soft K-means clustering and **write your answer with a `step6` plot when setting β to 50** in `submission_studentid.pdf`

Answer

일반 k-means clustering과 다르게, soft k-means clustering에서는 β의 값에 따라서 각 클러스터의 책임도가 결정된다. β의 값을 가정하고 식을 분석해본다면,

- $\beta = 0$인 경우:
  responsibility of x_i to cluster k,

$$r_{ki} = \frac{\exp(-\beta dist_{ki})}{\sum_{l \in K} \exp(-\beta dist_{li})} = \frac{\exp(0)}{\sum_{j=1}^{K} \exp(0)} = \frac{1}{K}$$

$$centroid_k = \frac{\sum_{i \in N} r_{ki} x_i}{\sum_{i \in N} r_{ki}} = \frac{\sum_{i \in N} \frac{1}{K} x_i}{\sum_{i \in N} \frac{1}{K}} = \frac{\sum_{i \in N} x_i}{N}$$

  즉, 모든 데이터 포인트가 동등하게 기여하므로, 모든 센트로이드가 데이터 포인터들의 평균 한 점으로 수렴하게 된다.

- $\beta \to \infty$인 경우:
  극한을 취하게 되면,

$$\exp(-\beta dist_{ki}) \to \begin{cases} 1 & \text{if } k \text{ is the nearest centroid to } i \\ 0 & \text{otherwise} \end{cases}$$

$$r_{ki} = \begin{cases} 1 & \text{if } k \text{ is the nearest centroid to } i \\ 0 & \text{otherwise} \end{cases}$$

  로 수렴한다. 따라서, 각 데이터 포인트는 가장 가까운 센트로이드에만 책임도를 부여하게 되고, 이는 일반적인 k-means clustering과 같아진다.

따라서, β는 가까운 데이터 포인터에 영향을 많이 받는지에 대한 가중치를 조절하는 하이퍼파라미터이다. 아래 이미지를 보면, β가 3인 경우에 비해서 β가 50인 경우 클러스터로부터 비교적 먼 경계에 위치한 데이터 포인트 또한 책임도를 부여받아 색이 진하게 나타나는 것을 확인할 수 있다.

Note. The plots are attached in the next page.

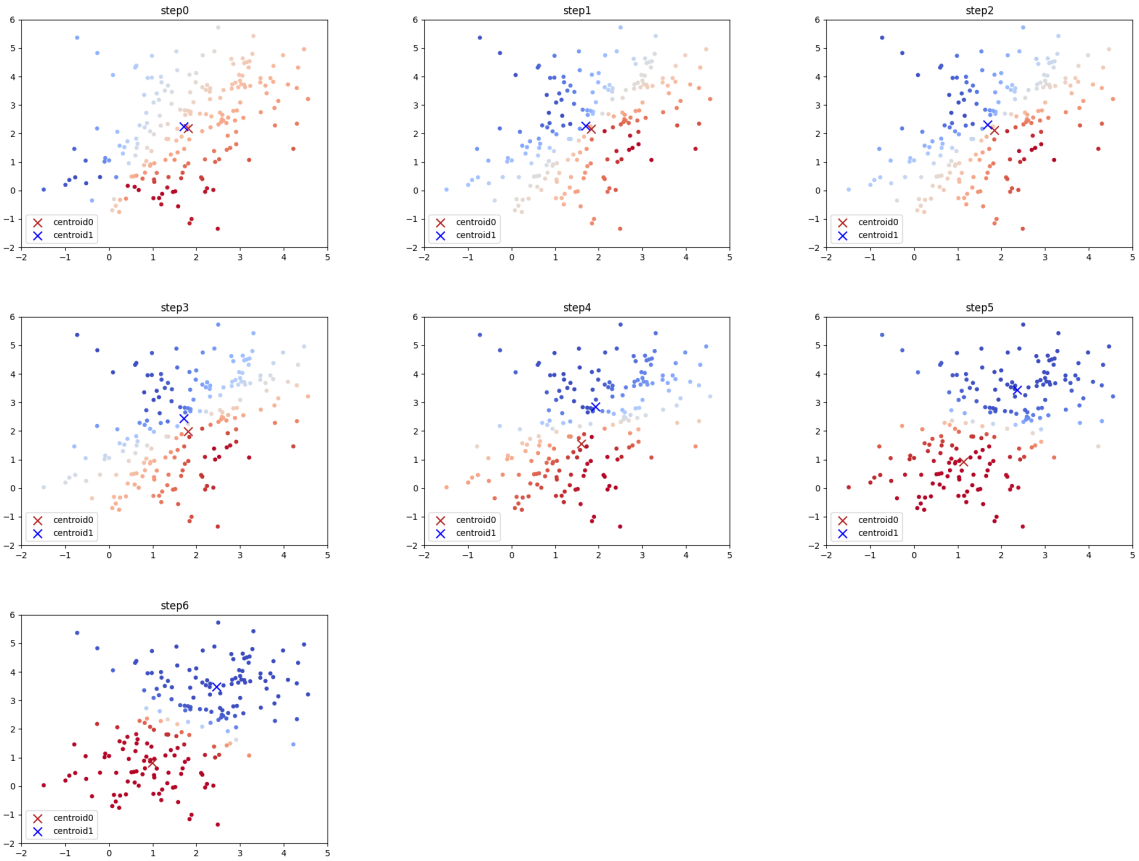

Figure 2: Result plots of soft-K-mean algorithm

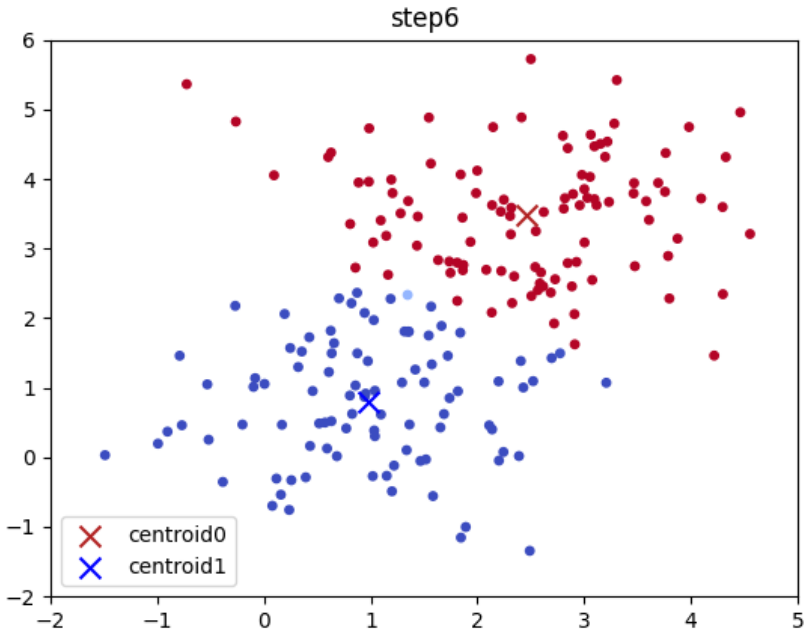

Figure 3: Result plots of soft-K-mean algorithm when β is set to 50