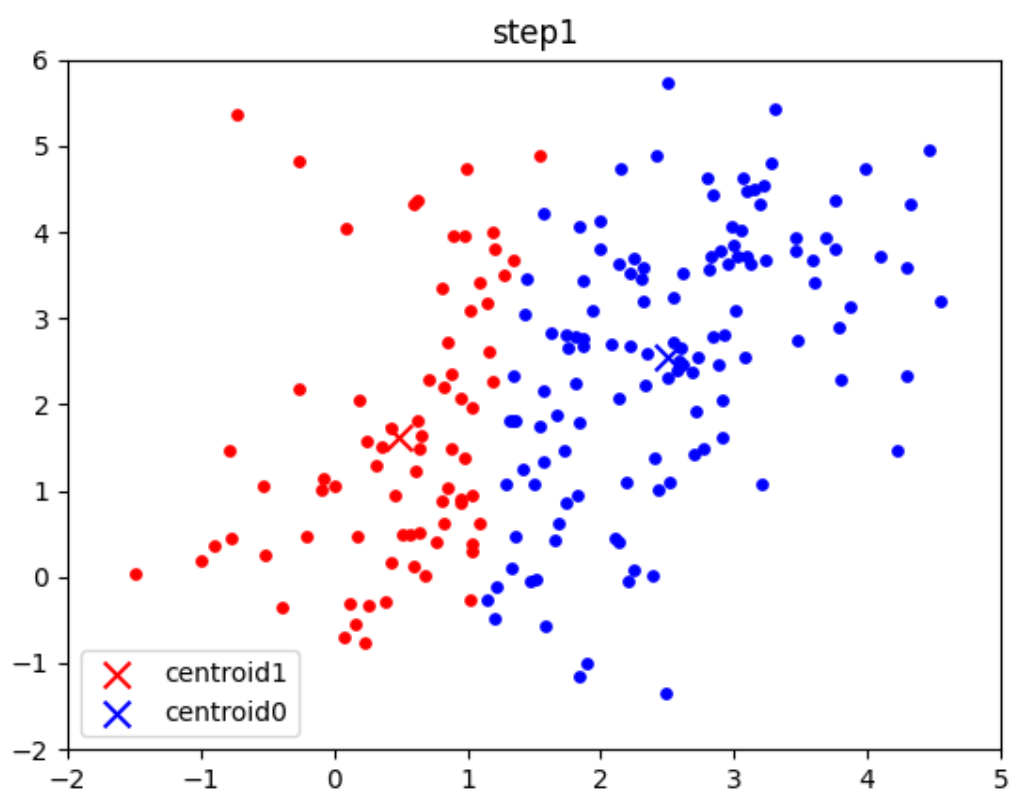
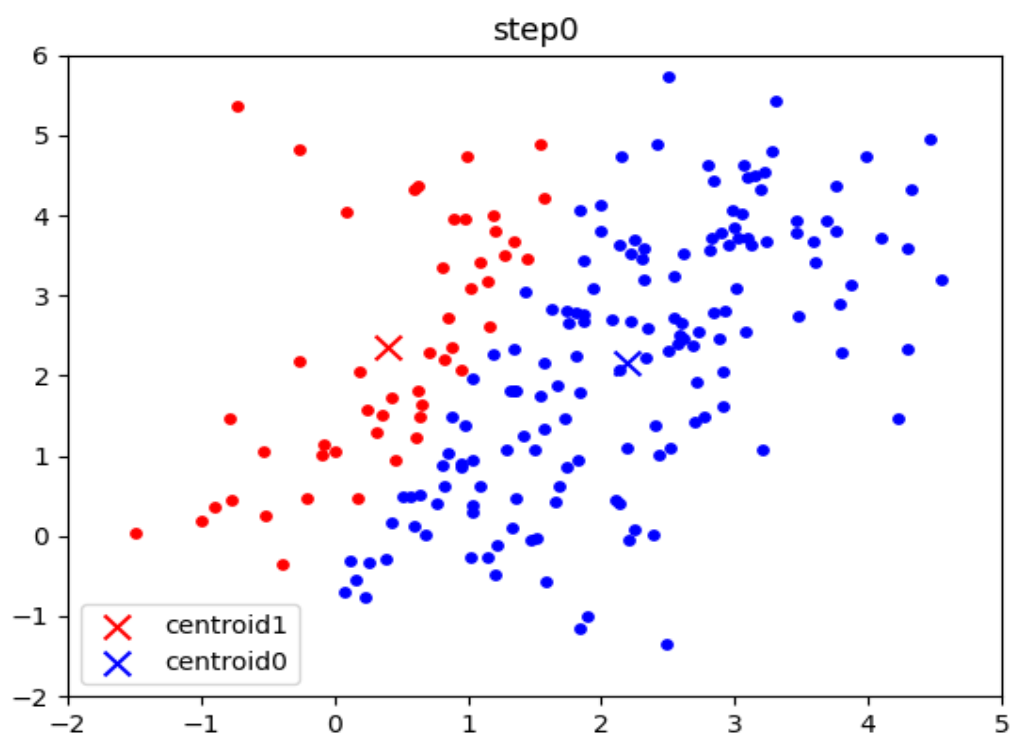
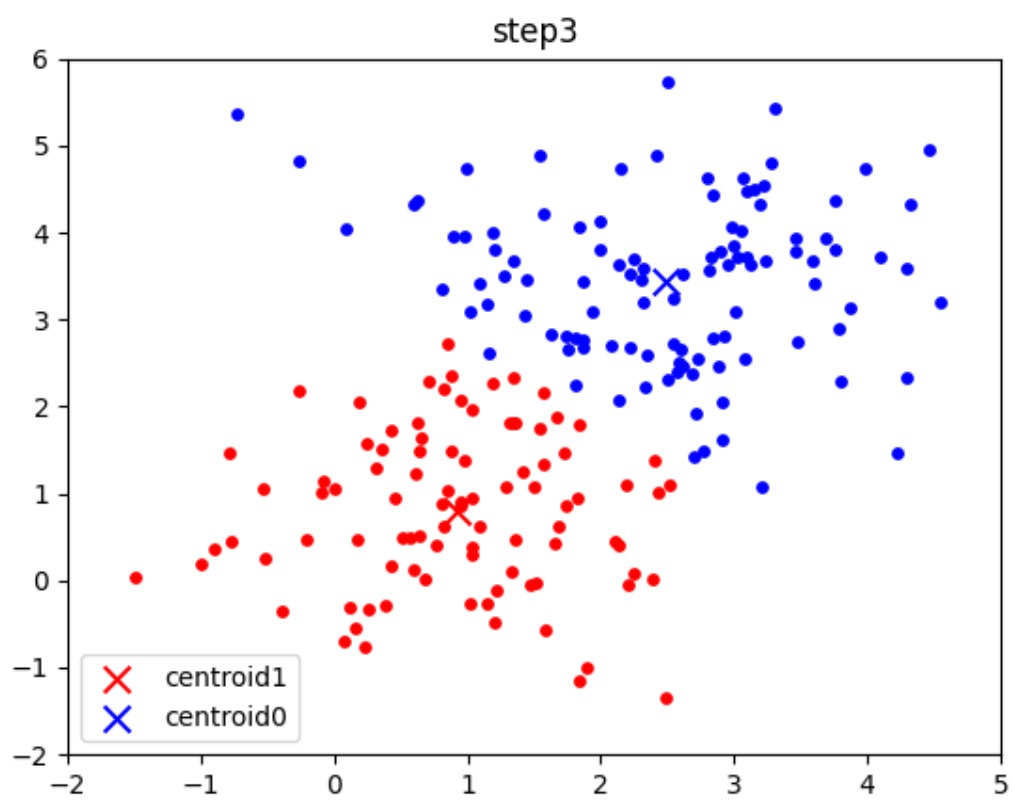
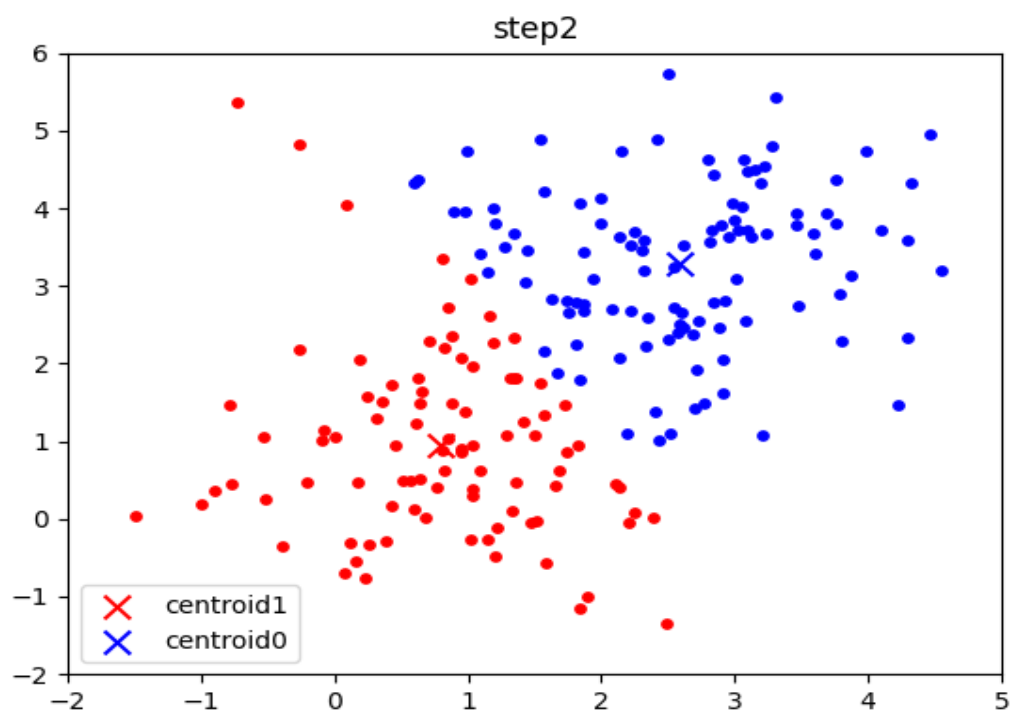


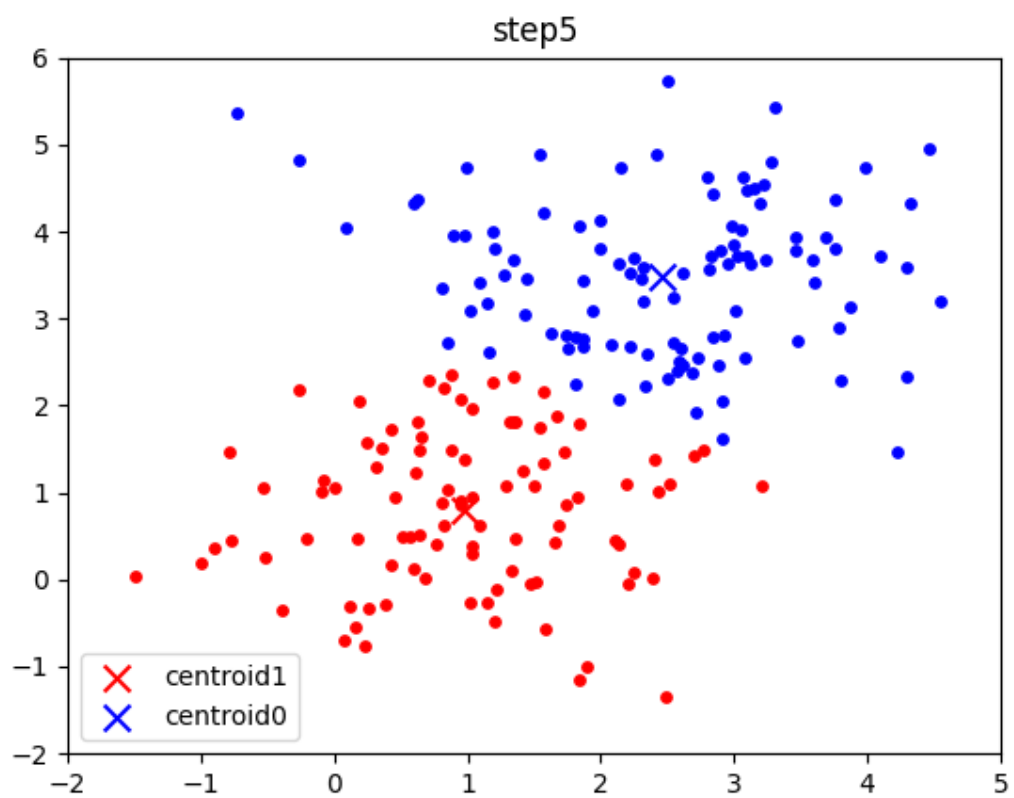
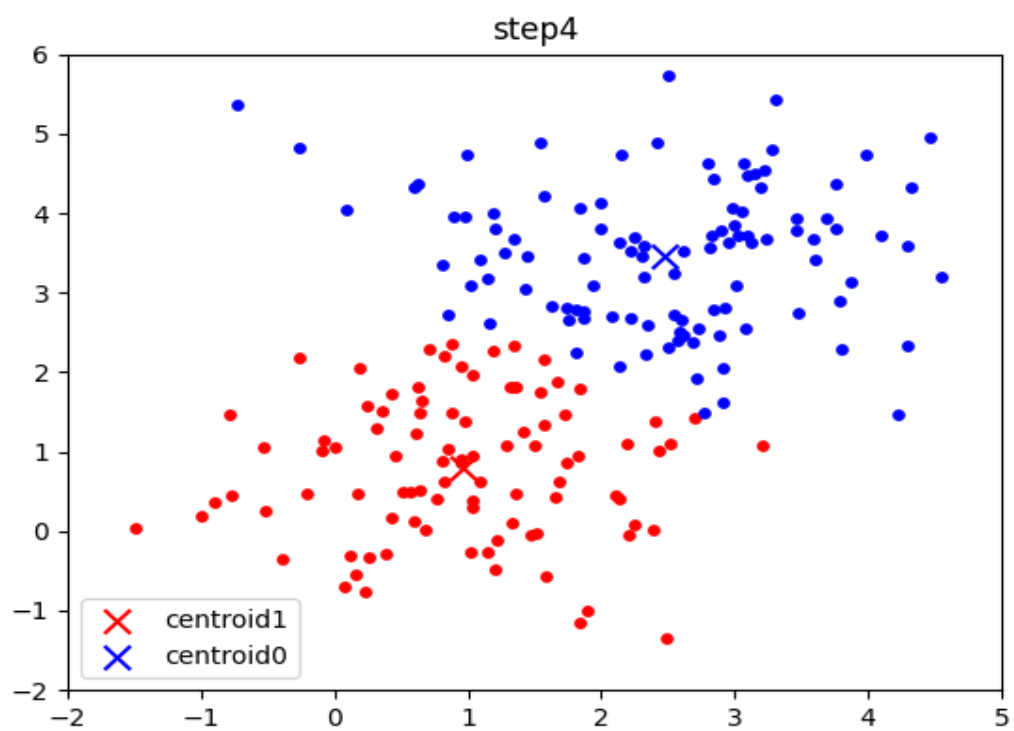
# **CSED342 – Artificial Intelligence**

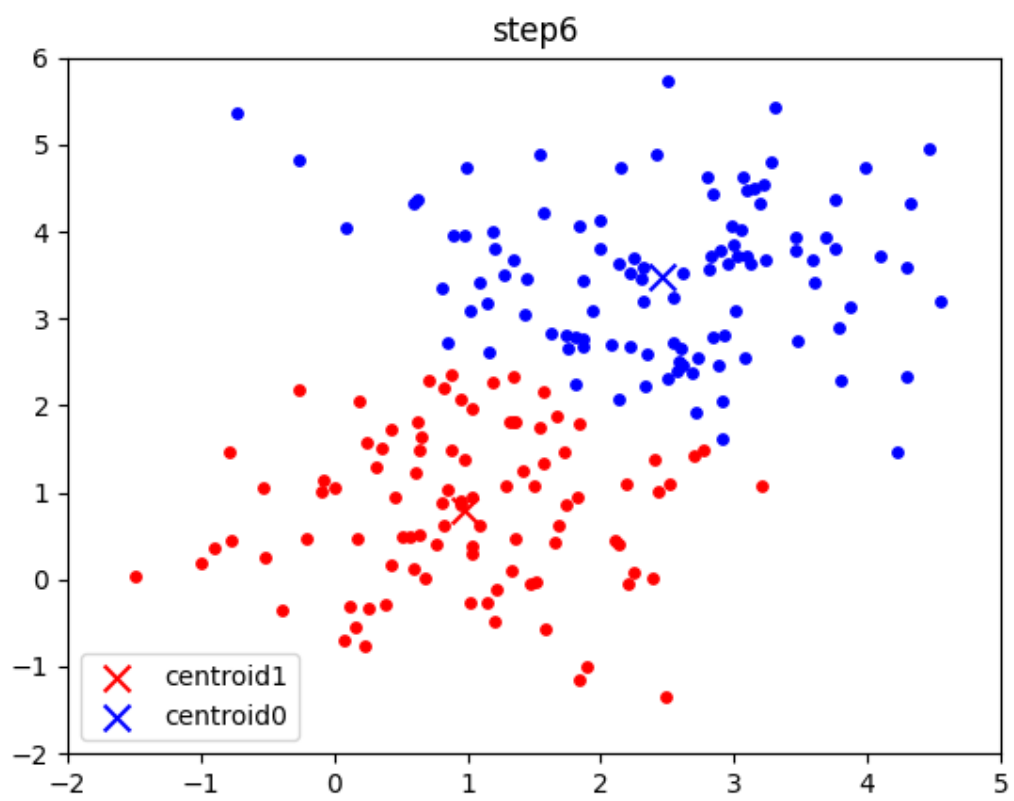
## **Assignment 4. K – Means Clustering**

### **Assignment Report**







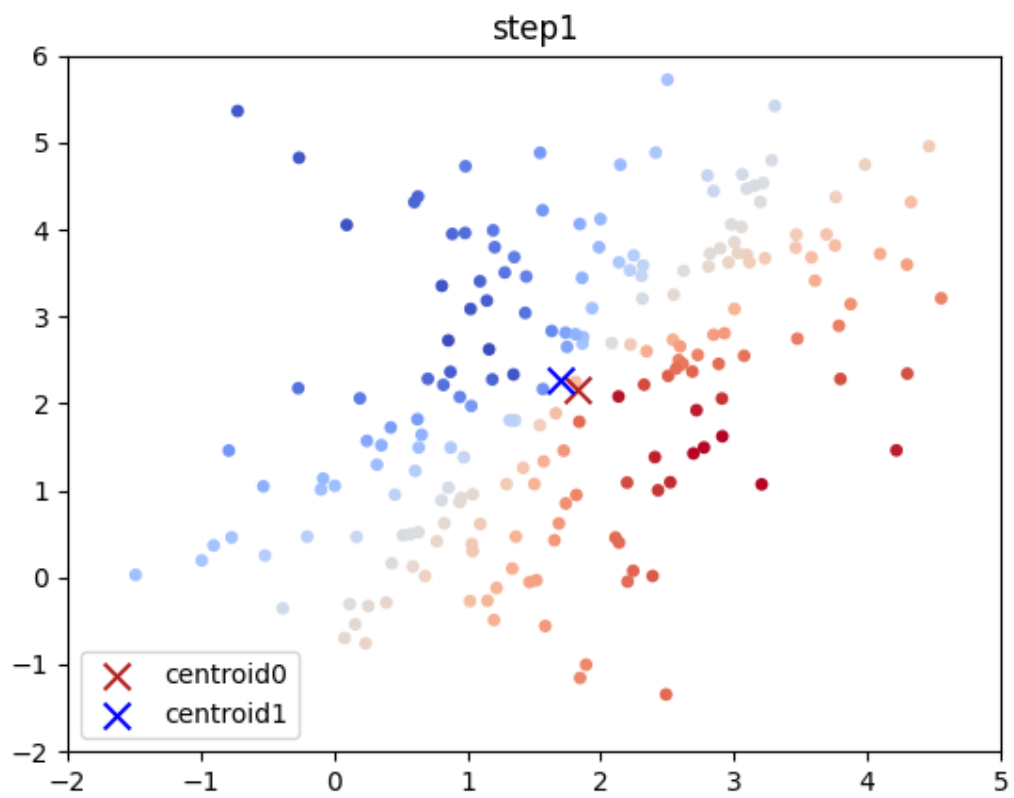
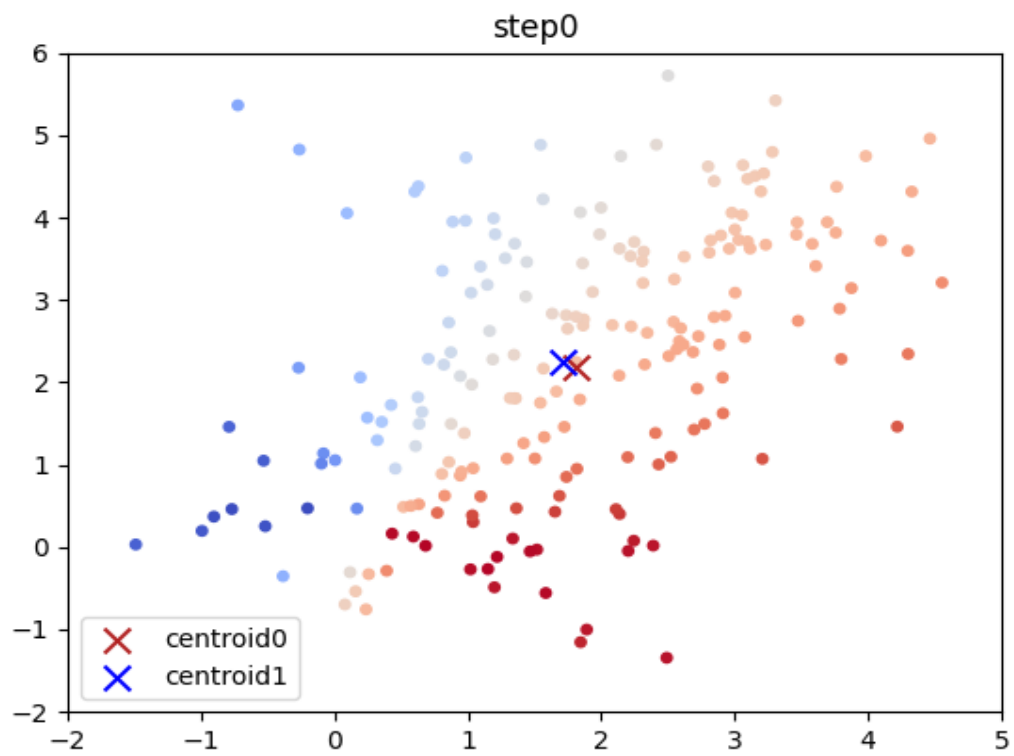


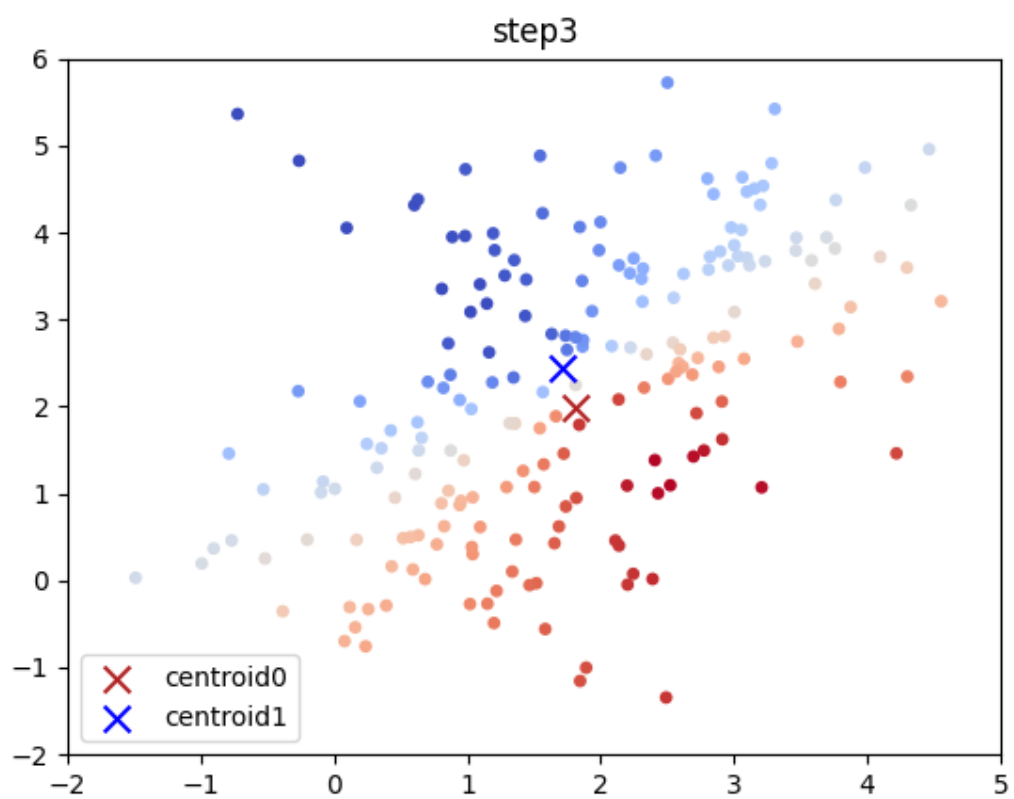
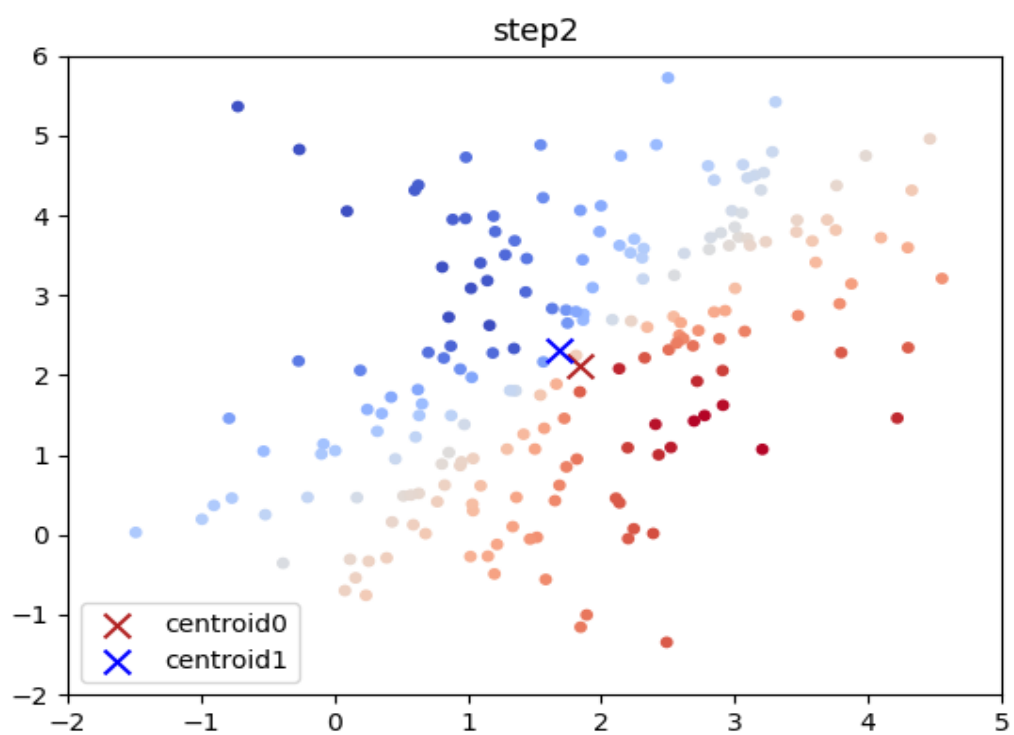
## Trying the Algorithm on MNIST

K-Clustering은 K개의 cluster로 data들이 군집하게 하는 기법을 말합니다. 이때, 각각 군집 된 data들의 중심을 centroid라고 명명합니다.

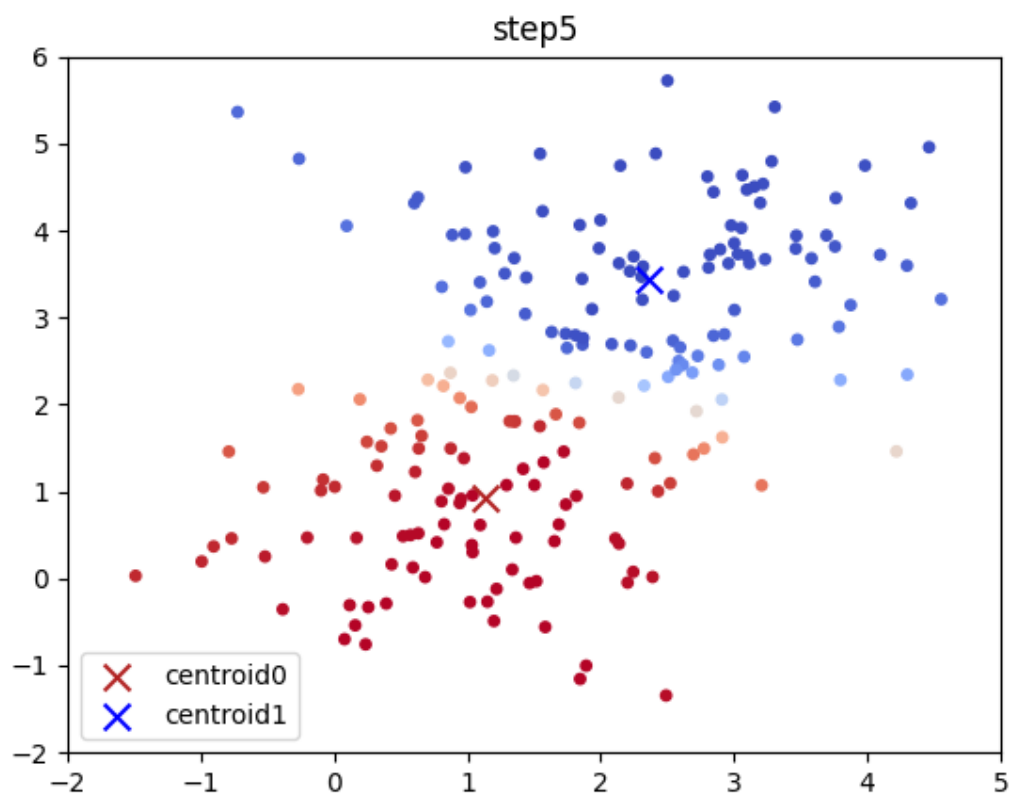
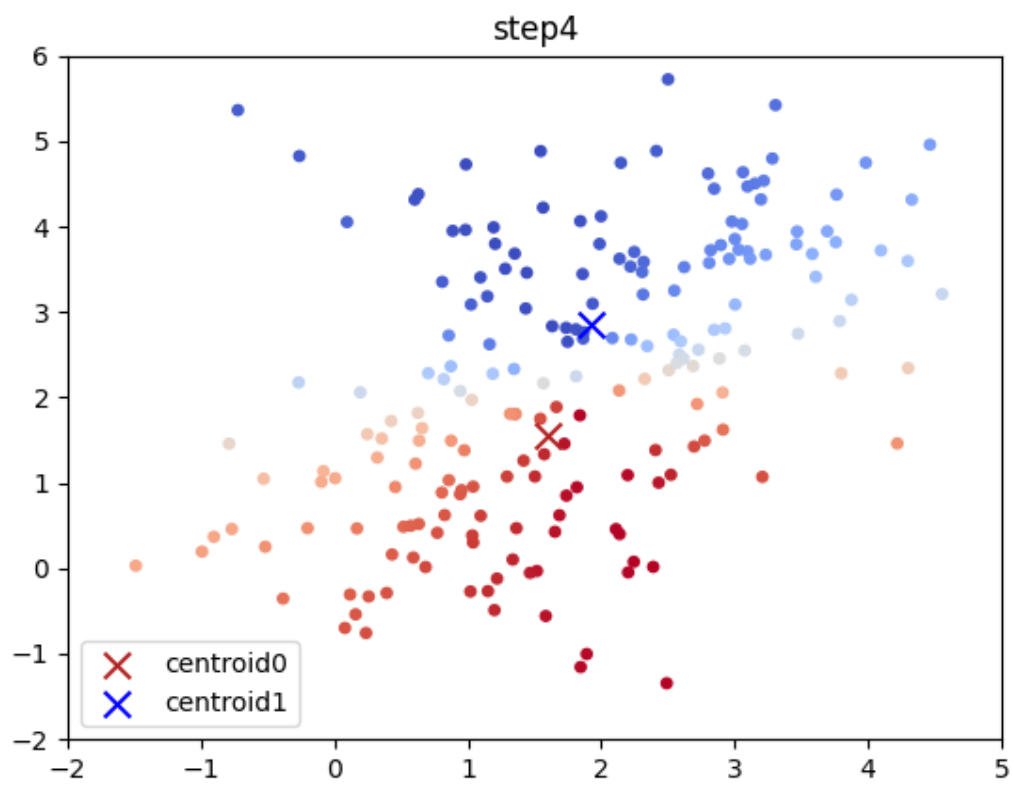
초기 상태에 Gaussian noise 상태의 이미지를 K-clustering을 통해서 data를 군집시켰고, 이 data point들이 centroid를 중심으로 모이면서 이 centroid들이 digit처럼 표현되는 것입니다.

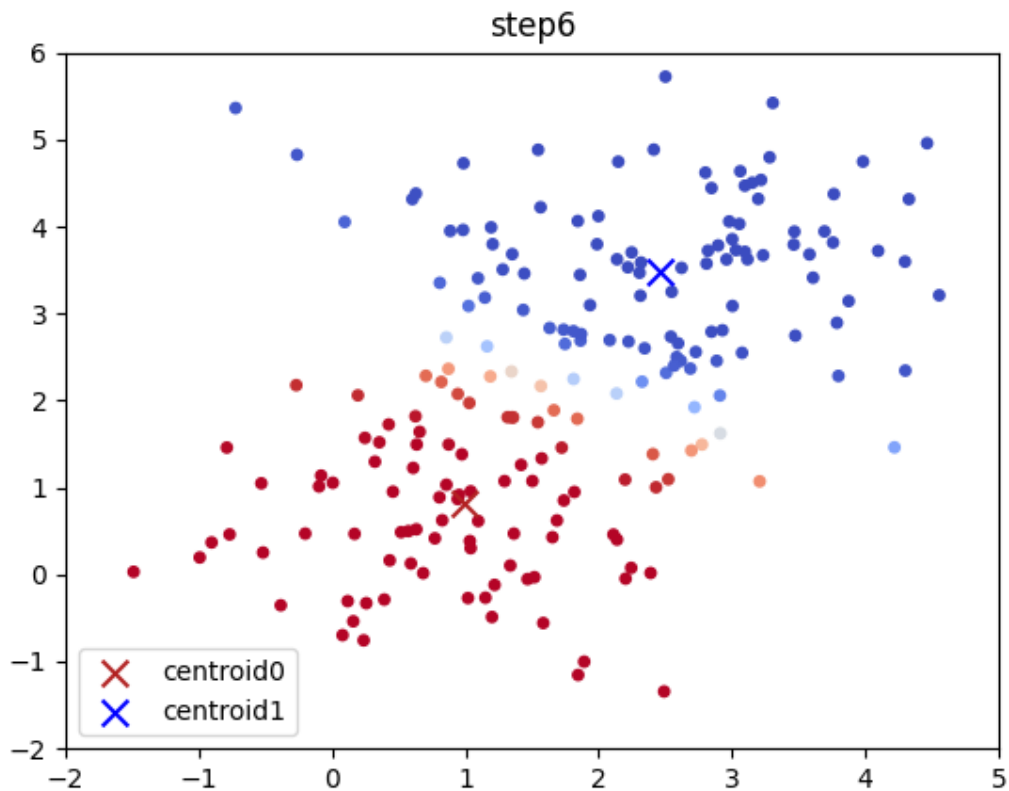
위 조건은  $K = 10$  상황을 부여하고 알고리즘을 실행했을 것입니다. 재미있는 점을 보면, /result/MNIST/final 폴더에 있는 결과를 보면 centroid 0과 centroid 8, 비슷한 결과를 보이고, centroid 6은 centroid 2와 centroid 5의 중간 값으로 보이는 결과를 보입니다. 이것은 각 digit의 dominant한 특징을 가진 cluster가 부족하고, 이는 K-clustering이 unsupervised learning이기 때문에 algorithm을 supervise할 mechanism이 부재하기 때문이라고 생각합니다.











## Answer for Problem 2d

$\beta$ 는 Soft K-means에서 'stiffness'를 조절합니다. 이때, 'stiffness' 라는 것은 cluster가 가질 수 있는 다양성을 조절하는 척도입니다.  $\beta$ 가 커질수록 Hard K-means에 도달합니다.

Hard K-means vs. Soft K-means

Hard K-means는 각 data point들이 한 cluster에 속해 있는 것을 말합니다. Soft K-means는 이와 다르게 인접 cluster에 포함될 가능성도 고려하여 확률로 나타냅니다. 이것이 과제 전반부에서와 달리 후반부에서는 확률을 계산 한 이유입니다. 다른 cluster에 속할 수 있다는 정보를 버리지 않고, 이를 soft label에 포함시켜 확률로 나타내는 것입니다.

우리 과제에서 보면, data point의 색깔의 진하기 정도가  $\beta$ 가 커질수록 진해지는 것을 볼 수 있습니다. 이는 위의 분석과 동일한 맥락의 결과를 낸다는 것을 의미합니다.