

Homework 3

Collaborators:

Name: Tian Zijun

Student ID: 3160104043

Problem 3-1. Neural Networks

In this problem, we will implement the feedforward and backpropagation process of the neural networks.

(a) **Answer:** loss = 0.268 accuracy = 0.92

Problem 3-2. K-Nearest Neighbor

In this problem, we will play with K-Nearest Neighbor (KNN) algorithm and try it on real-world data. Implement KNN algorithm (in *knn.m/knn.py*), then answer the following questions.

(a) Try KNN with different K and plot the decision boundary.

Answer:

```
[0, 0, 0, ..., 1, 1, 1, ...]
```

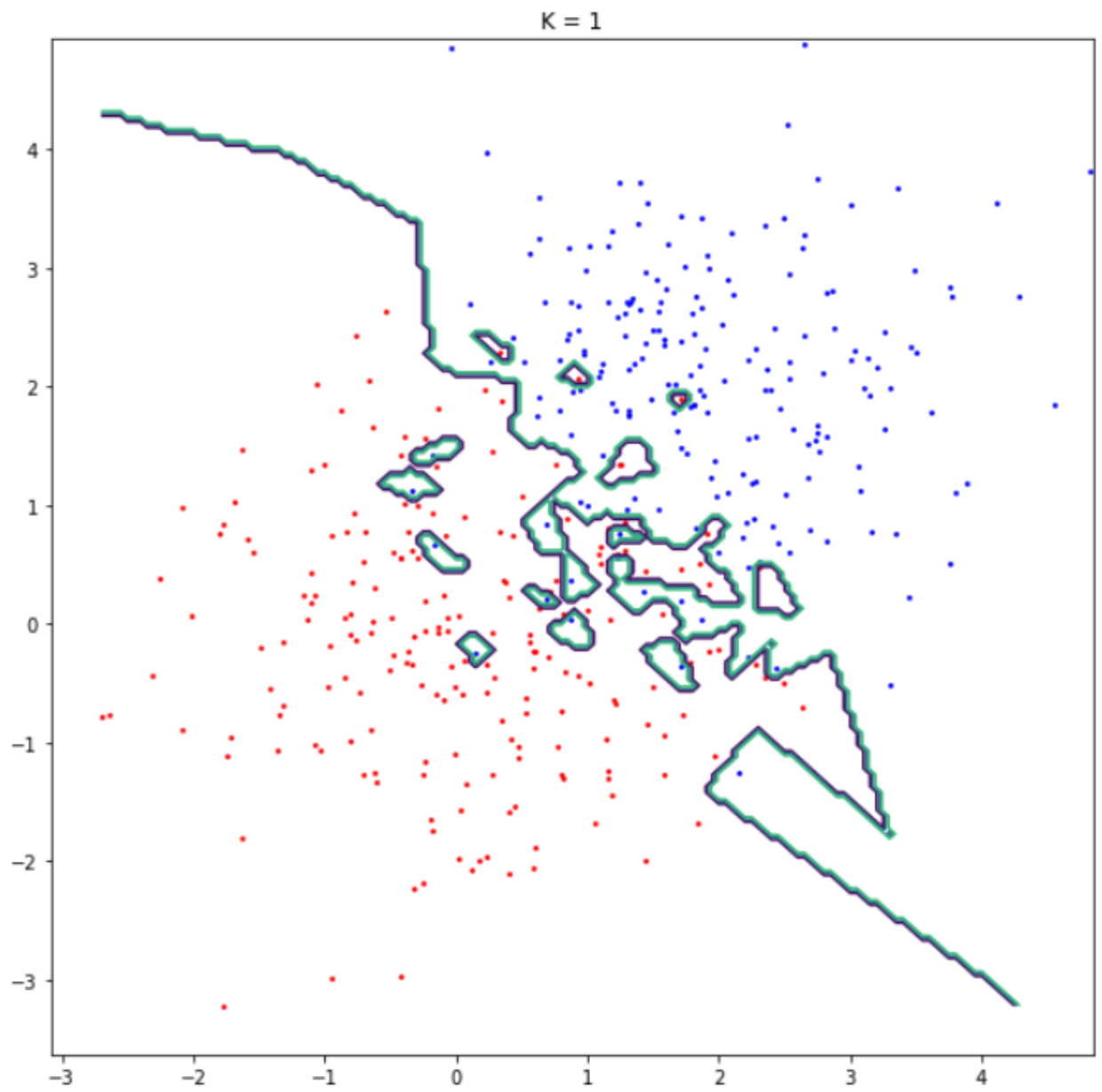
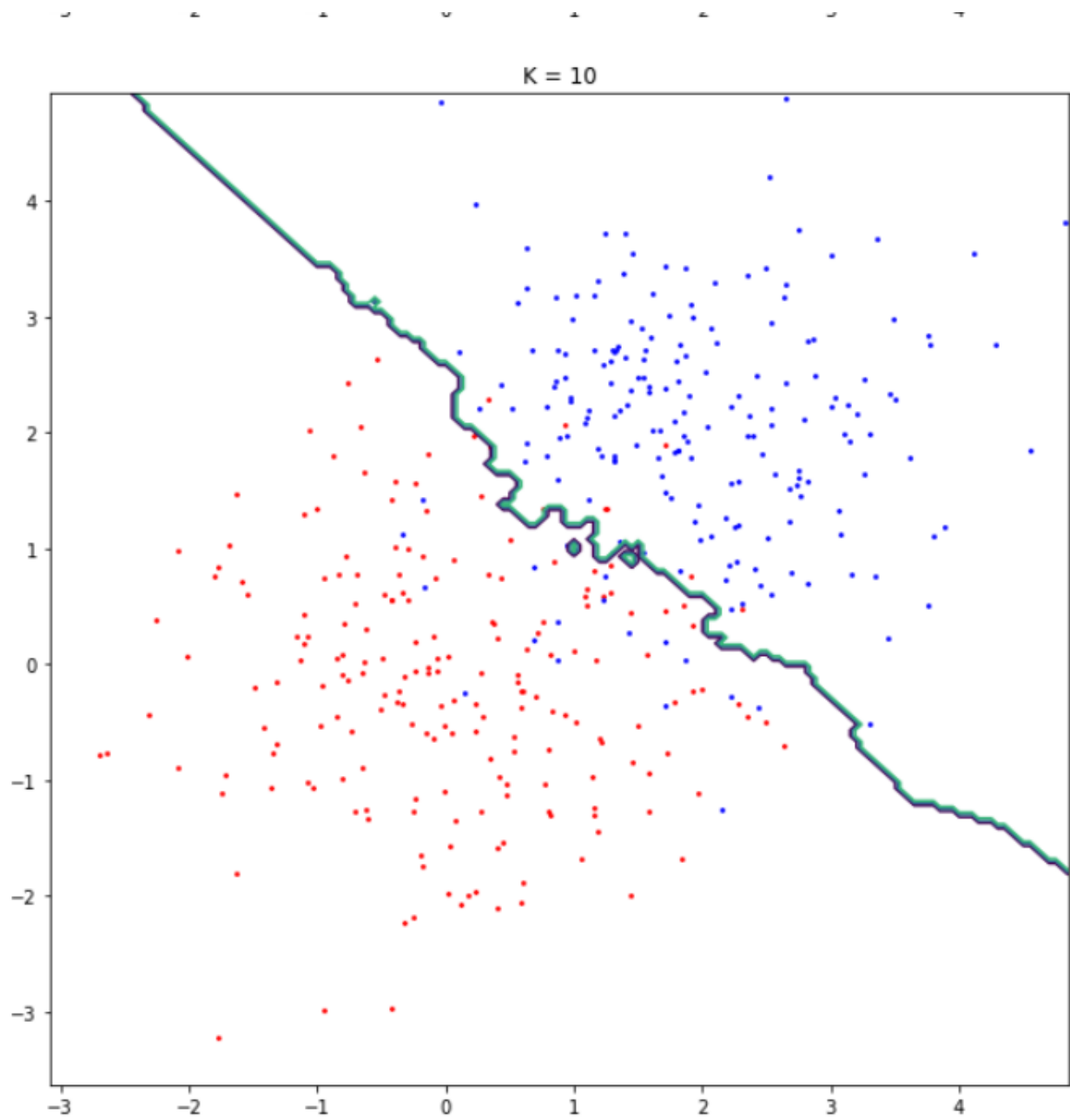


Figure 1: $k=1$

**Figure 2:** $k=10$

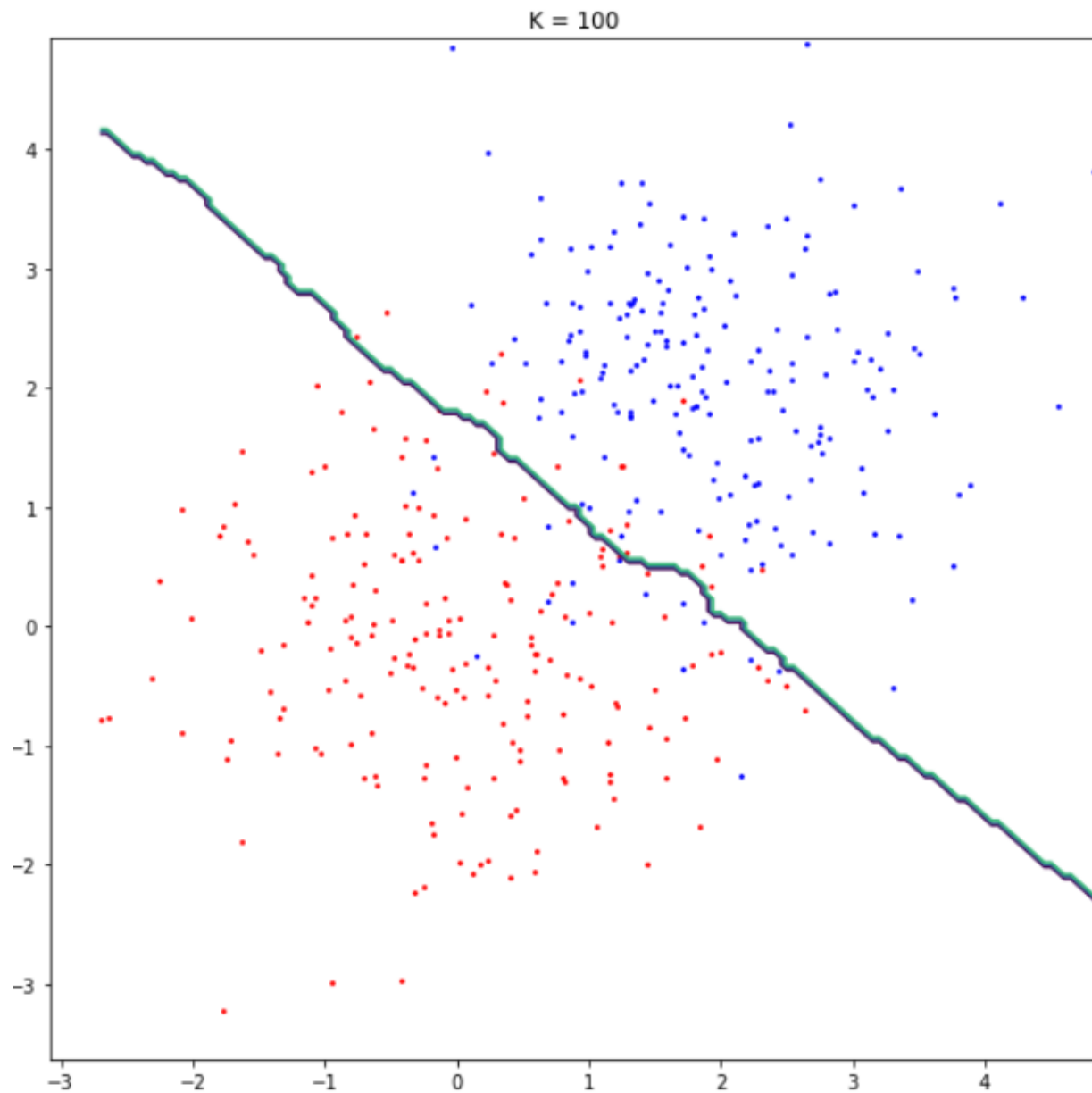


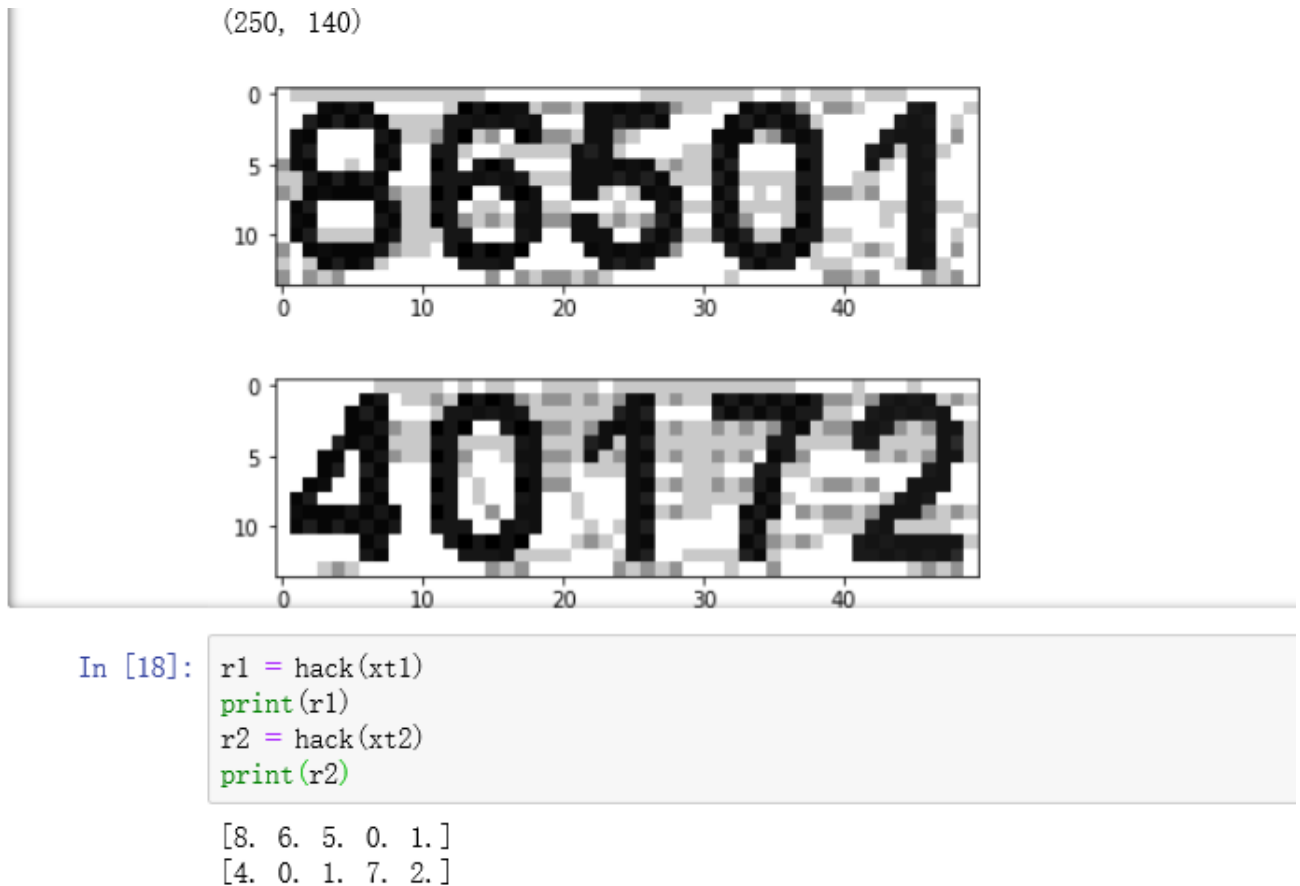
Figure 3: k=100

- (b) We have seen the effects of different choices of K. How can you choose a proper K when dealing with real-world data ?

Answer: By using cross validation, we can find and choose the k which has the best accuracy.

- (c) Finish *hack.ml/hack.py* to recognize the CAPTCHA image using KNN algorithm.

Answer:



Through the result above, we can find that under the condition of 50 training data the accuracy is 100

Problem 3-3. Decision Tree and ID3

Consider the scholarship evaluation problem: selecting scholarship recipients based on gender and GPA. Given the following training data:

Answer:

$E(S) =$

(T) Total = $10 + 95 + 5 + 90 + 80 + 20 + 120 + 30 = 450$

(TS) Total with scholarship = $10 + 95 + 15 + 90 = 210$

(TWS) Total without scholarship = $80 + 20 + 120 + 30 = 250$

Entropy (S) = $-\frac{4}{9} \log_2 \frac{4}{9} - \frac{5}{9} \log_2 \frac{5}{9} = 0.971$

① Entropy_{gender}(S) = $\sum_{i=1}^n \text{Entropy}_{\text{gender}}(S_i) \cdot \frac{S_i}{S}$

Attribute gender divides S into S_1 (gender = F), S_2 (gender = M)

Entropy_{gender}(S_1) = $-\frac{21}{41} \log_2 \frac{21}{41} - \frac{20}{41} \log_2 \frac{20}{41} = 0.79957$

Entropy_{gender}(S_2) = $-\frac{19}{49} \log_2 \frac{19}{49} - \frac{30}{49} \log_2 \frac{30}{49} = 0.96333$

Entropy_{gender}(S) = $\frac{S_1}{S} E_{\text{gender}}(S_1) + \frac{S_2}{S} E_{\text{gender}}(S_2) = \frac{21}{90} \times 0.79957 + \frac{49}{90} \times 0.96333 = 0.97989$

Gain (gender) = $E(S) - E_{\text{gender}}(S) = 0.0116$

② Entropy_{GPA}(S) = $\sum_{i=1}^n \frac{S_i}{S} E_{\text{GPA}}(S_i)$

Attribute GPA divides S into S_1 (GPA = High) and S_2 (GPA = Low)

$E_{\text{GPA}}(S_1) = -\frac{37}{47} \log_2 \frac{37}{47} - \frac{10}{47} \log_2 \frac{10}{47} = 0.74674$

$E_{\text{GPA}}(S_2) = -\frac{3}{43} \log_2 \frac{3}{43} - \frac{40}{43} \log_2 \frac{40}{43} = 0.56506$

$E_{\text{GPA}}(S) = \frac{47}{90} \times 0.74674 + \frac{43}{90} \times 0.56506 = 0.64338$

Gain (GPA) = $E(S) - E_{\text{GPA}}(S) = 0.42662$

③ $G(\text{GPA}) > G(\text{gender})$

```

graph TD
    GPA --> High
    GPA --> Low
    High --> Gender1[Gender]
    Low --> Gender2[Gender]
    Gender1 --> M1[M]
    Gender1 --> F1[F]
    M1 --> T1[+]
    F1 --> T1
    Gender2 --> M2[M]
    Gender2 --> F2[F]
    M2 --> T2[-]
    F2 --> T2
  
```

Problem 3-4. K-Means Clustering

Finally, we will run our first unsupervised algorithm k-means clustering.

(a) Visualize the process of k-means algorithm for the two trials.

Answer:

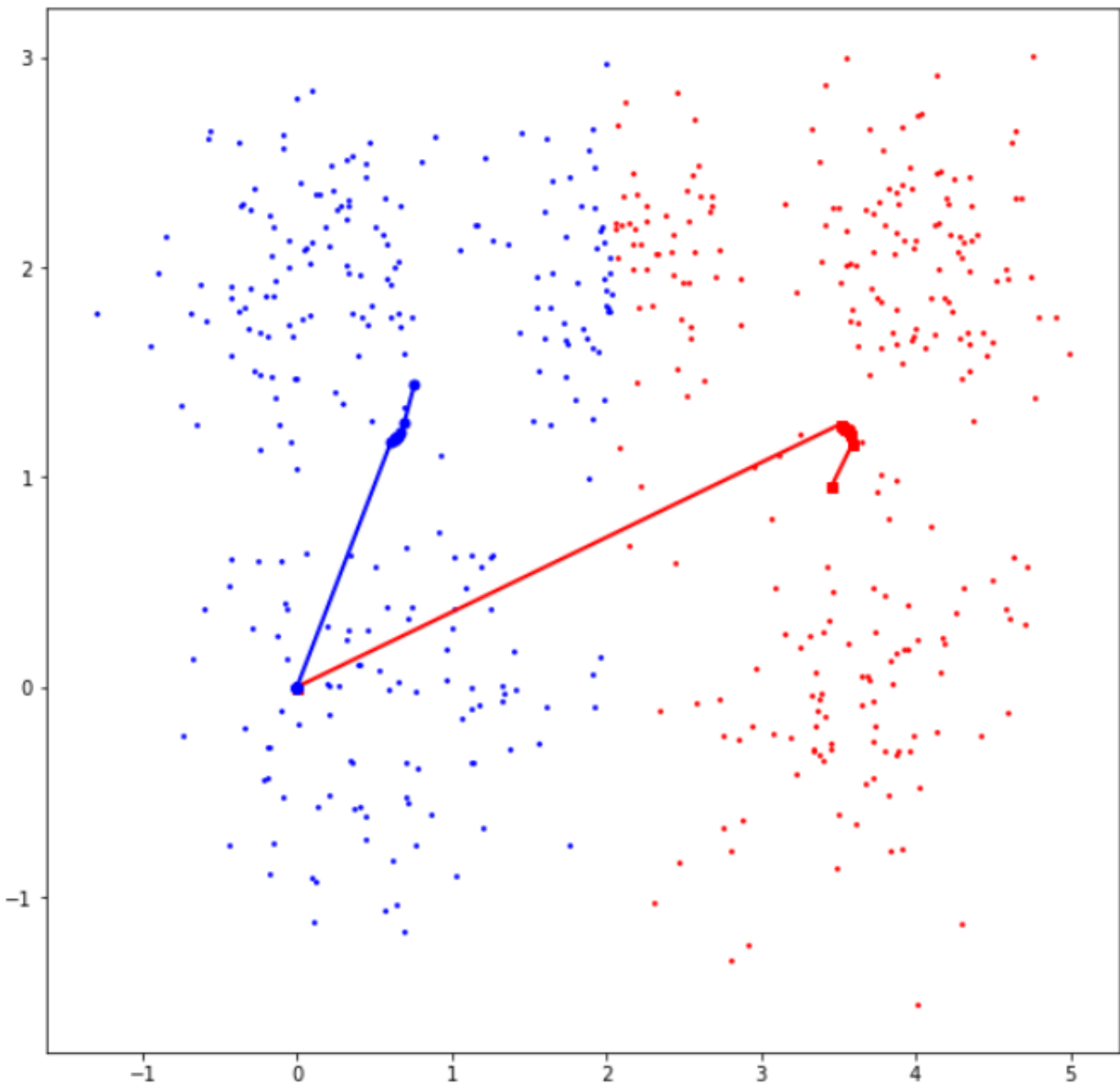


Figure 4: min SD

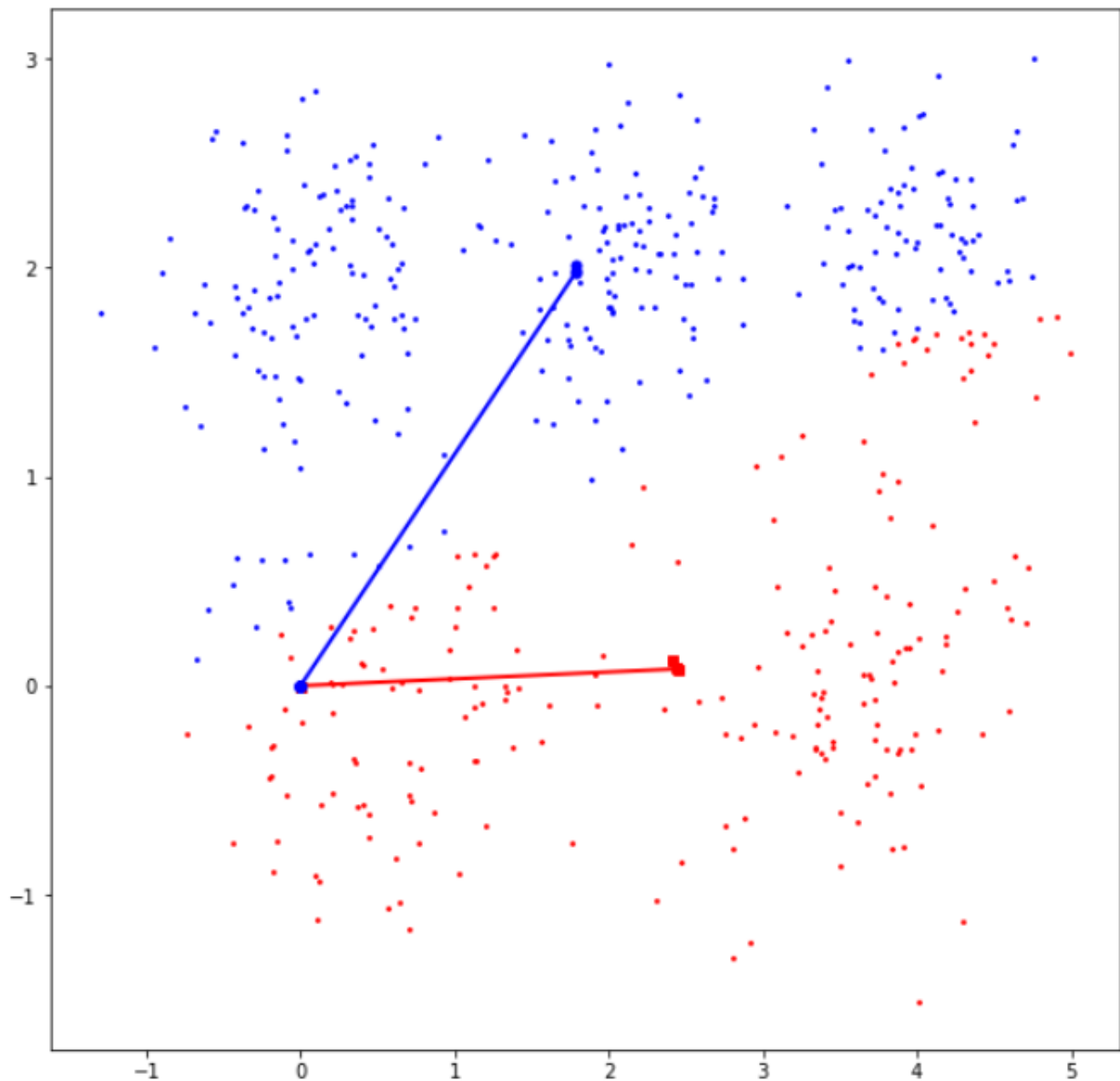


Figure 5: max SD

(b) How can we get a stable result using k-means?

Answer: Since the result of k-means differs due to different initial values of centers, we should run k-means algorithm for several times to get stable average result.

(c) Visualize the centroids.

Answer:

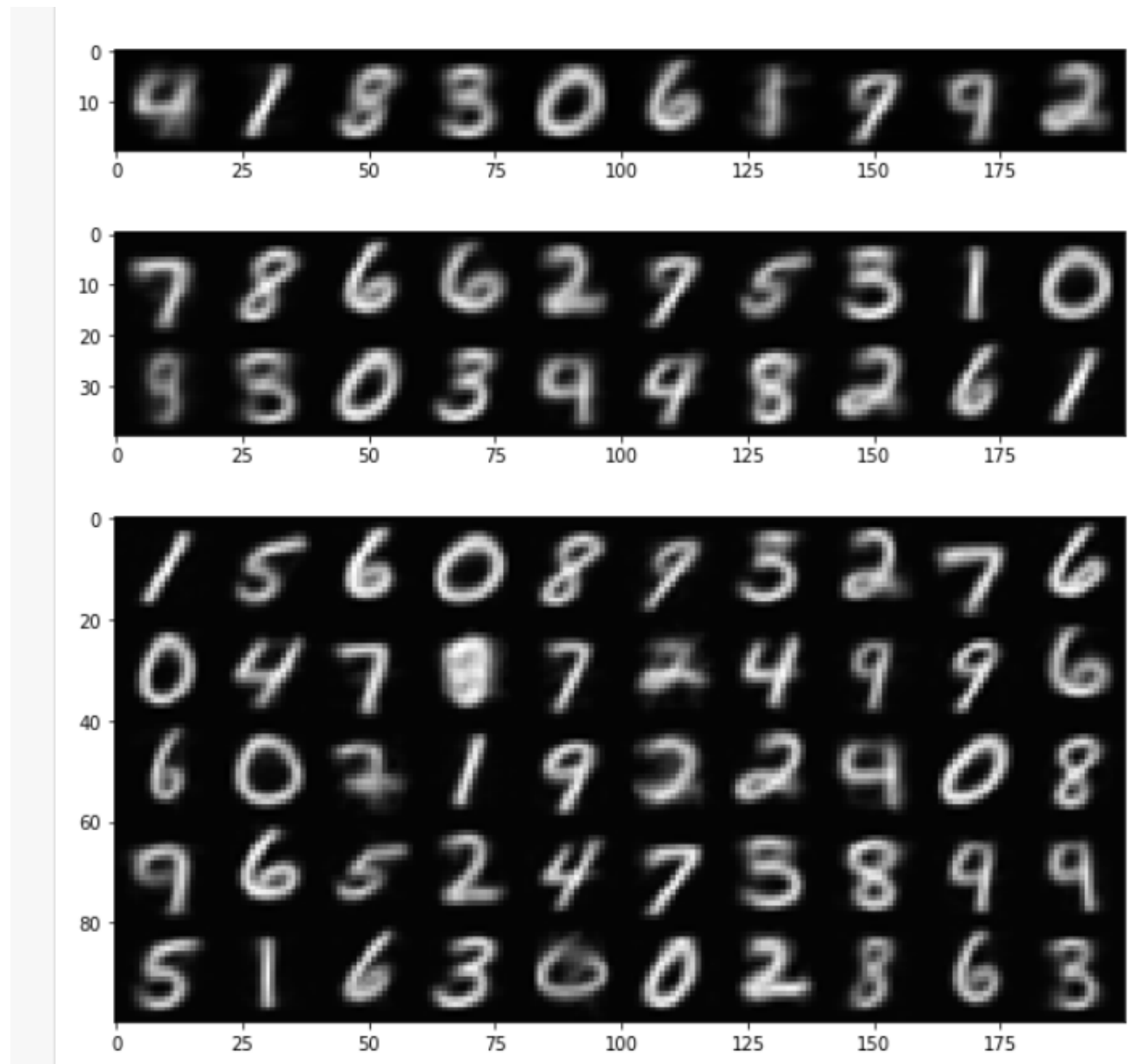


Figure 6: k-means

(d) Vector quantization.

Answer:

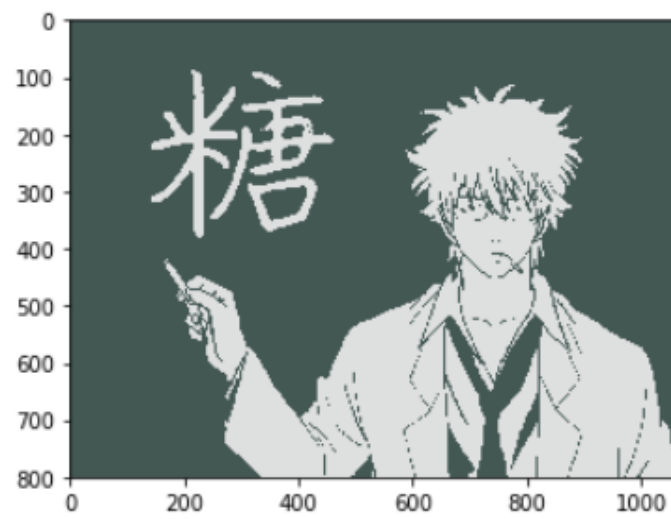


Figure 7: $k=8$

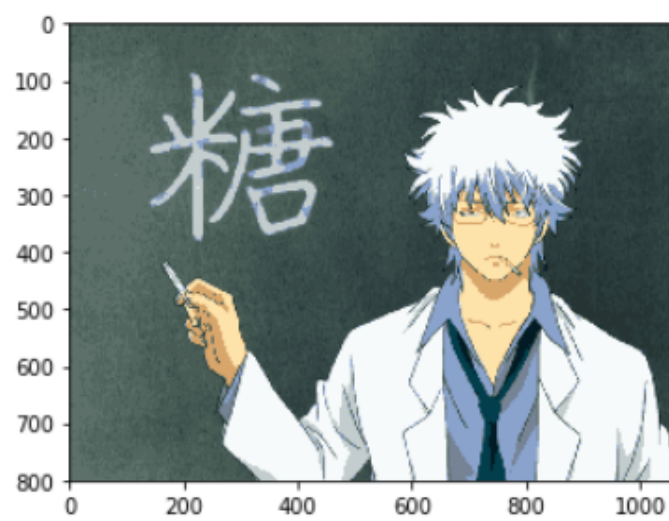
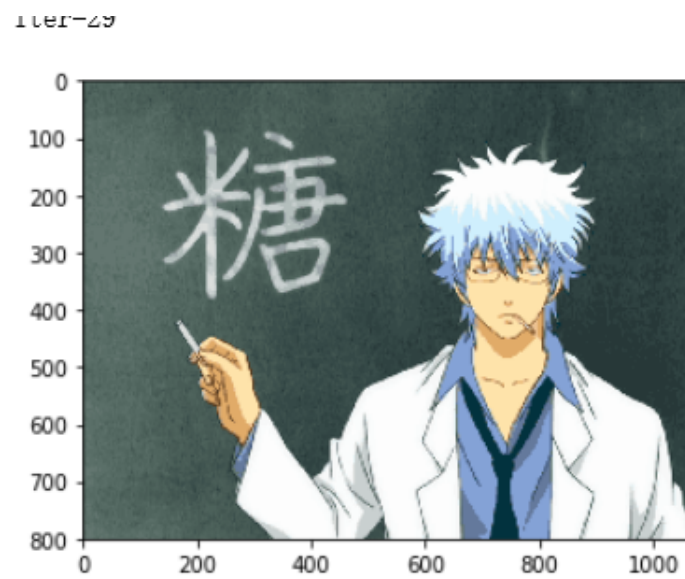
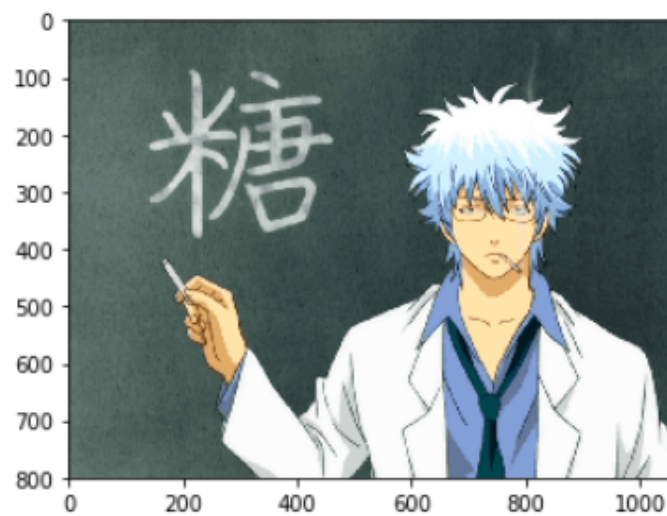


Figure 8: $k=16$

**Figure 9:** $k=32$ **Figure 10:** $k=64$