Machine Learning: Assignment #3

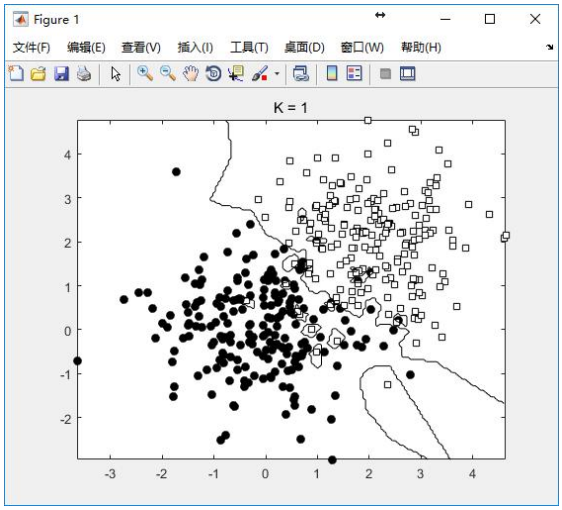
1. Neural Networks

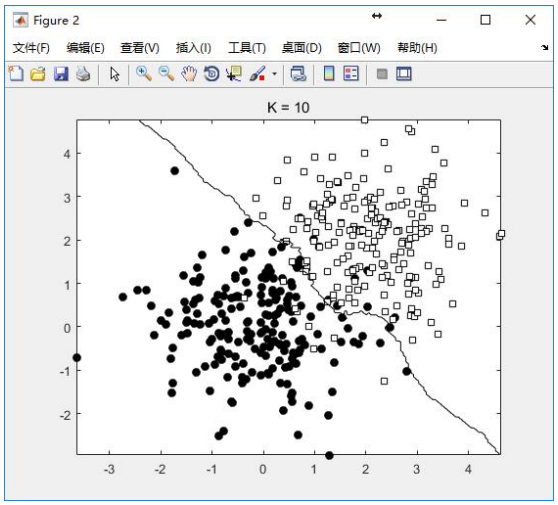
The code is in zip, we can train three layer neural networks by my code. And my test accuracy is 92.700%

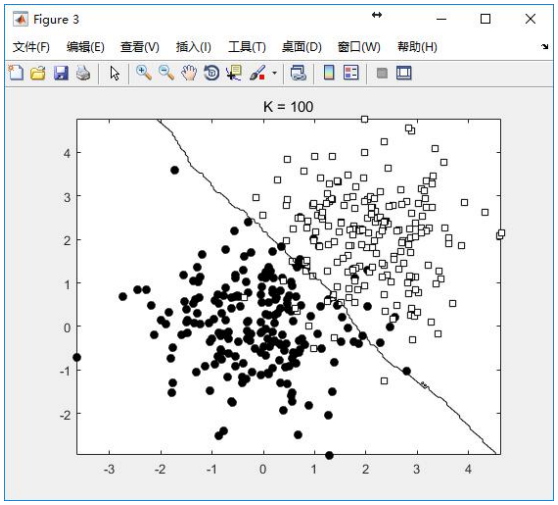
2. K-Nearest Neighbor

(a)

I have completed the knn.m, and the decision boundary of K = 1, 10, and 100 is as follow:







(b)

The parameter K is difficult to choose, on real-world data, we can use Cross-Validation to choose proper K.

(c)

Firstly, we should get the training set. I have write a script to get the CAPTCHA images. I have put the script to github, so everybody can use my code to get the CAPTCHA images.

The address of code is https://github.com/yysys/getDataFromURL

Then we tag the label of training set, and complete hack.m to recognize the CAPTCHA image using KNN algorithm.

3. Decision Tree and ID3

There are 540 samples, including 200 positive samples and 250 negative samples. So the entropy(S) is as follow:

Entropy(S) = $-\frac{200}{450}\log_2\frac{200}{450}-\frac{250}{450}\log_2\frac{250}{450}$ = 0.9911

记属性 gender 为 T1，属性 GPA 为 T2

$$\text{Entropy}(S\mid T1)=-\frac{215}{450}*\left(\frac{200}{215}\log_2\frac{200}{215}+\frac{15}{215}\log_2\frac{15}{215}\right)-\frac{235}{450}*\left(\frac{50}{235}\log_2\frac{50}{235}+\frac{185}{235}\log_2\frac{185}{235}\right)=0.5644$$

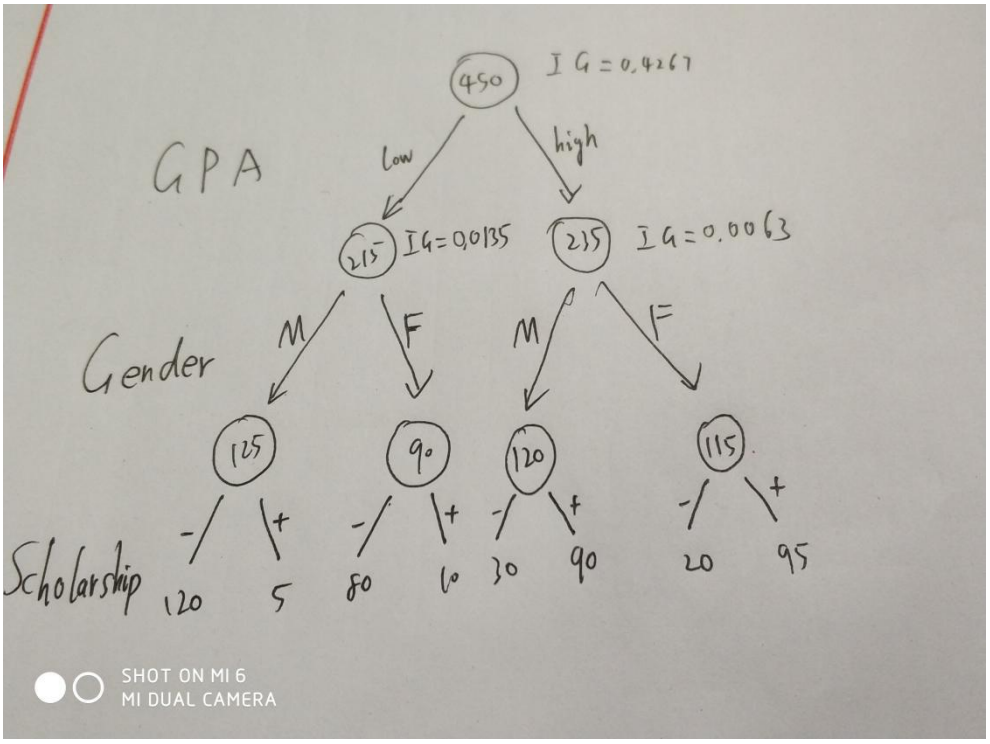
$$\text{Entropy}(S\mid T2)=-\frac{245}{450}*\left(\frac{150}{245}\log_2\frac{150}{245}+\frac{95}{245}\log_2\frac{95}{245}\right)-\frac{205}{450}*\left(\frac{100}{205}\log_2\frac{100}{205}+\frac{105}{205}\log_2\frac{105}{205}\right)=0.9798$$

$$\text{IG}(S\mid T1)=Entropy(S)-Entropy(S\mid T1)=0.4267$$

$$\text{IG}(S\mid T2)=Entropy(S)-Entropy(S\mid T2)=0.0113$$

So we choose the GPA to divide set, after that we choose the gender to divide the set.

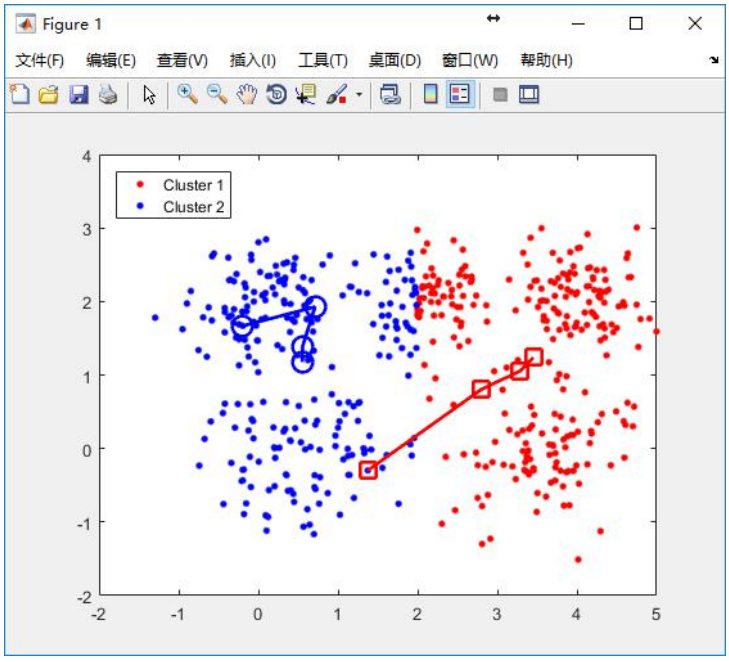
The decision tree is drew as follow:



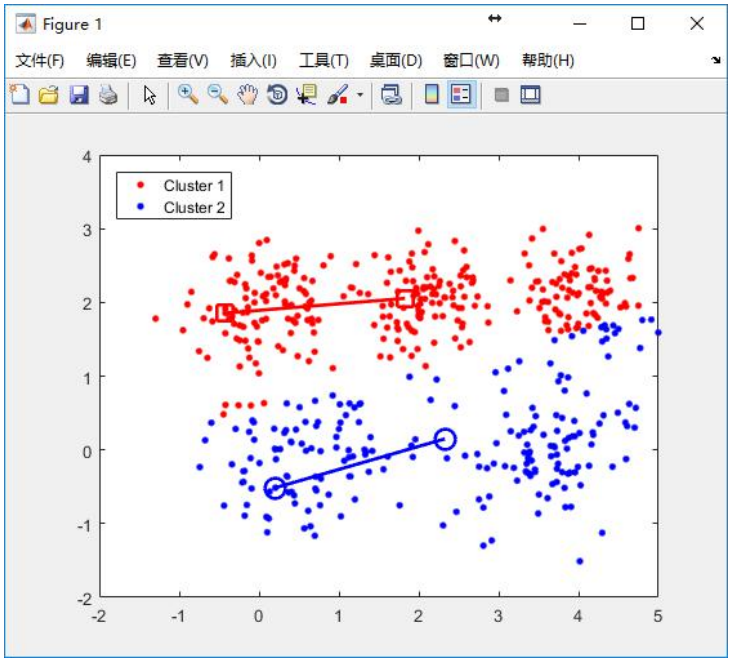
4. K-Means Clustering

(a)

The process of k-means algorithm for the two trials with smallest SD is as follow:



The process of k-means algorithm for the two trials with largest SD is as follow:



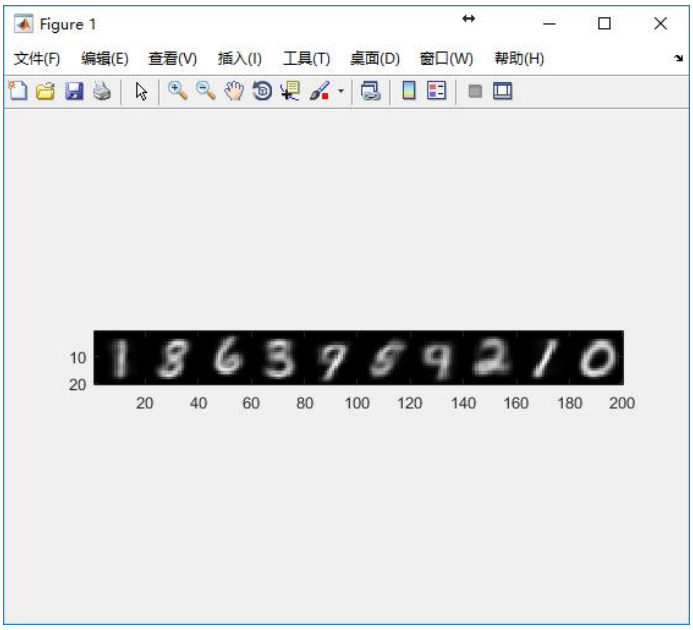
(b)

We can use k-means++ instead of k-means to reduce the influence of cluster centroids initialization.

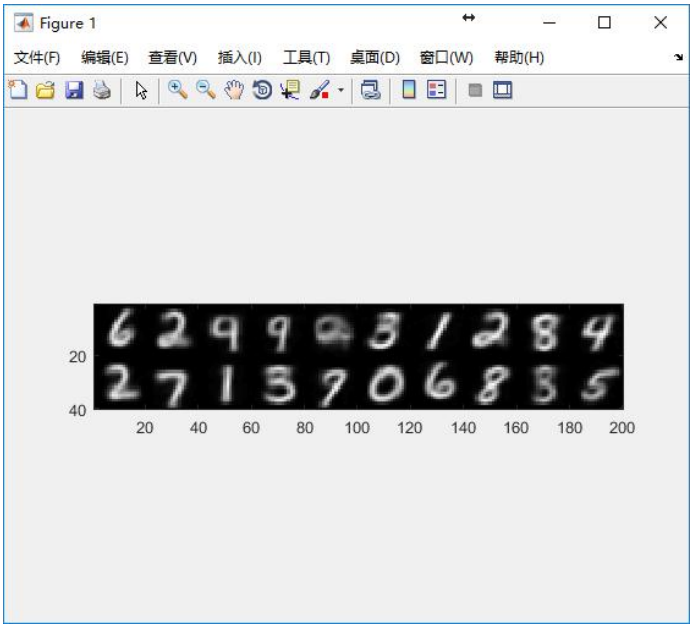
(c)

K-means algorithm can discover the patterns in dataset without any label information.

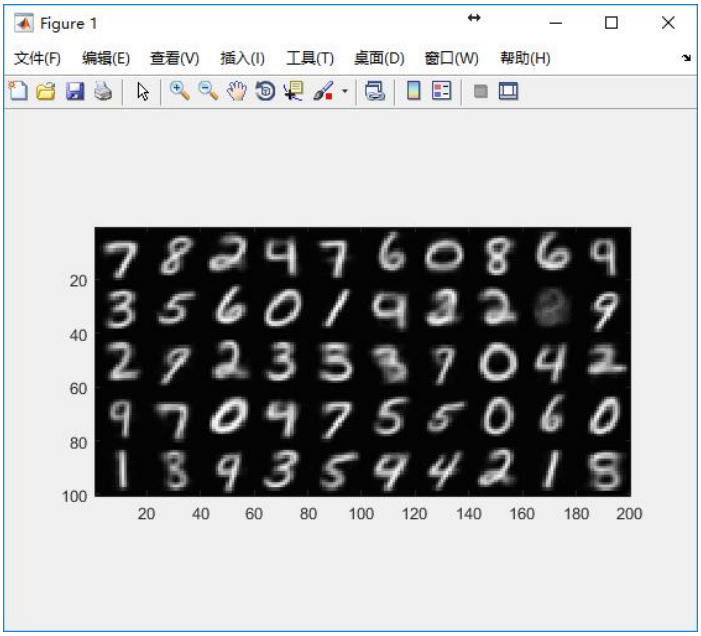
When K = 10, the picture is as follow:



When K = 20, the picture is as follow:



When K = 50, the picture is as follow:



(d)

We use k-means to vector quantization the three picture which is gave by TA.

In “sample0.jpg”, the file size before vector quantization is 80619. After vector quantization , the file size is 26565. The compress ratio of “sample1.jpg” if we set K to 64 is 32.95%

In “sample1.jpg”, the file size before vector quantization is 481536. After vector quantization , the file size is 92627. The compress ratio of “sample1.jpg” if we set K to 64 is 19.23%

In “sample2.jpg”, the file size before vector quantization is 2760774. After vector quantization , the file size is 361478. The compress ratio of “sample1.jpg” if we set K to 64 is 13.09%

The picture when K is 8, 16, 32, 64 is as follow:

K = 8:

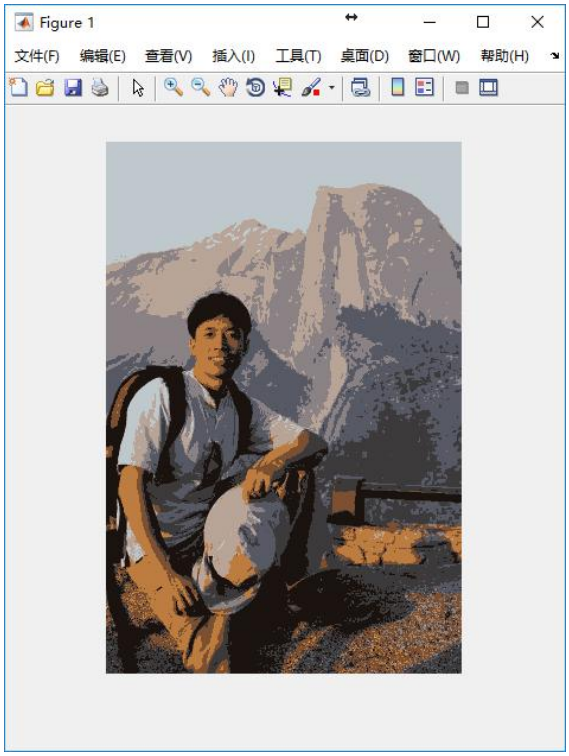


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)

K = 16

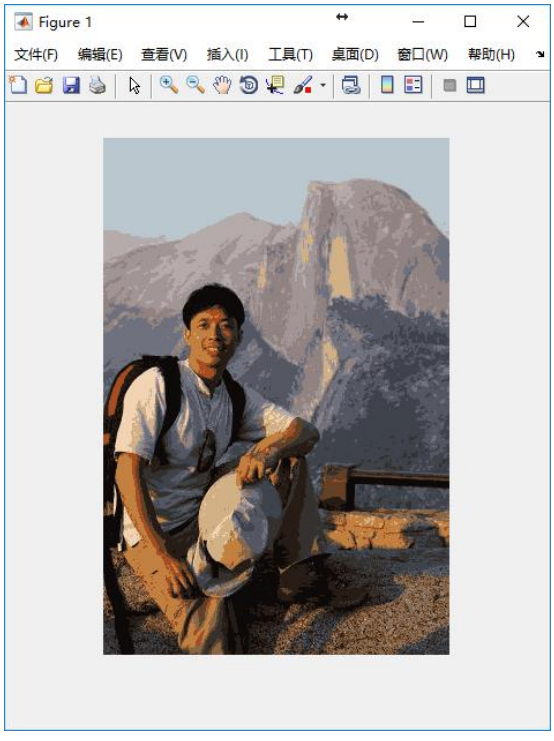


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)

K = 32

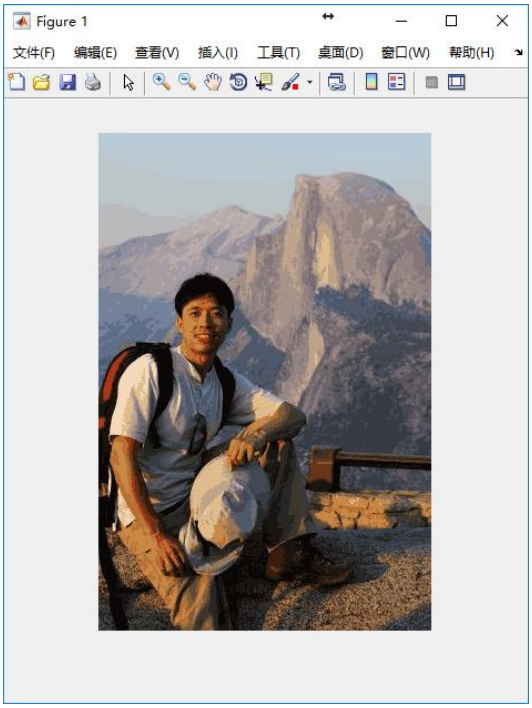


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)

K = 64

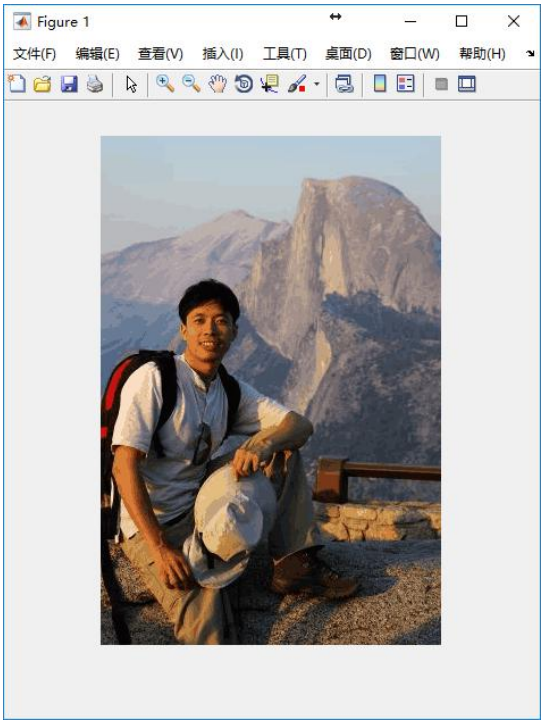


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)

K = 8

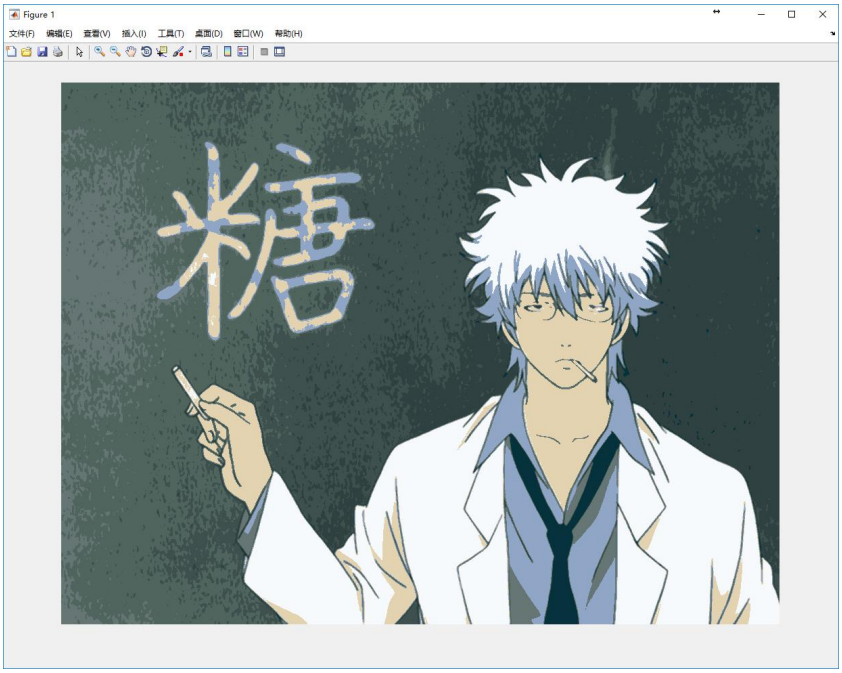


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
糖

K = 16

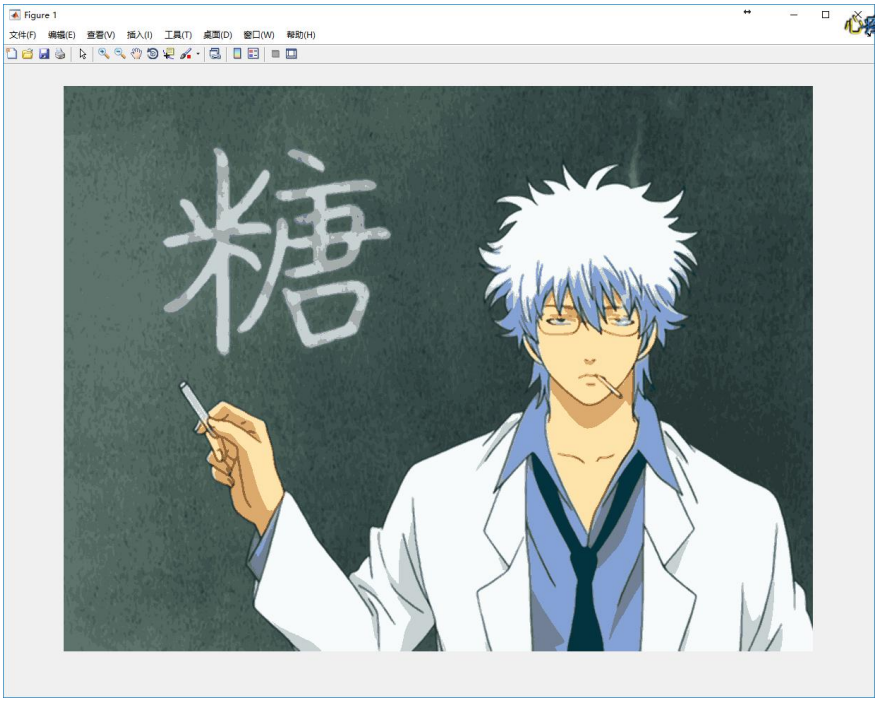


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
糖

K = 32

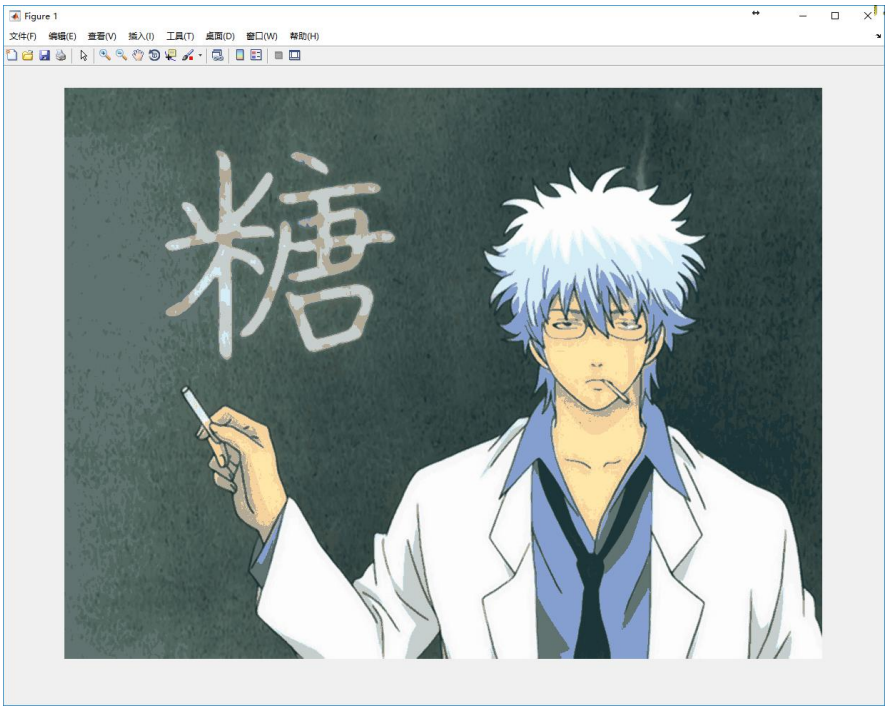


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
糖

K = 64

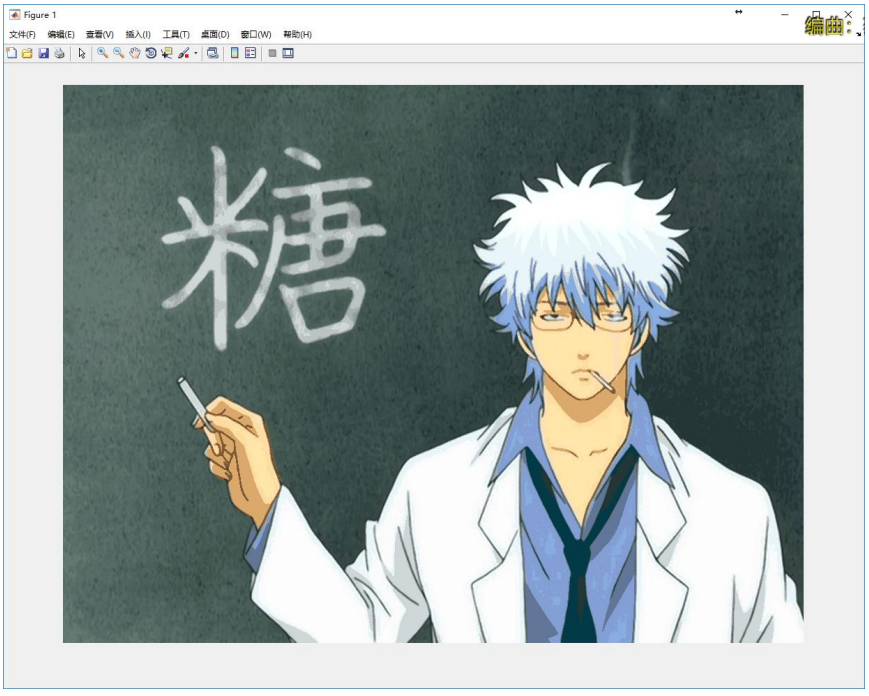


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
糖

K = 8

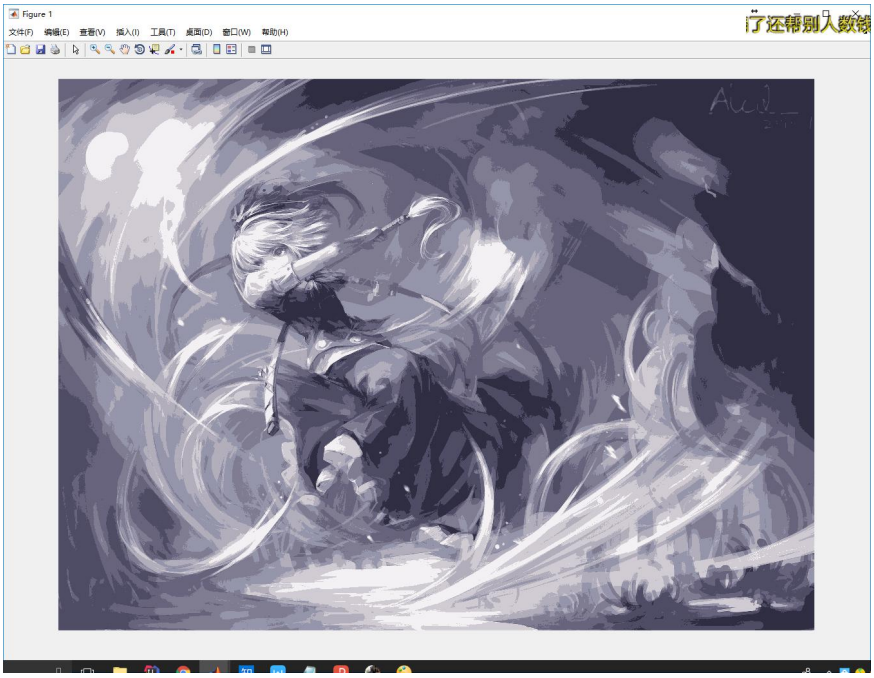


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)

K = 16

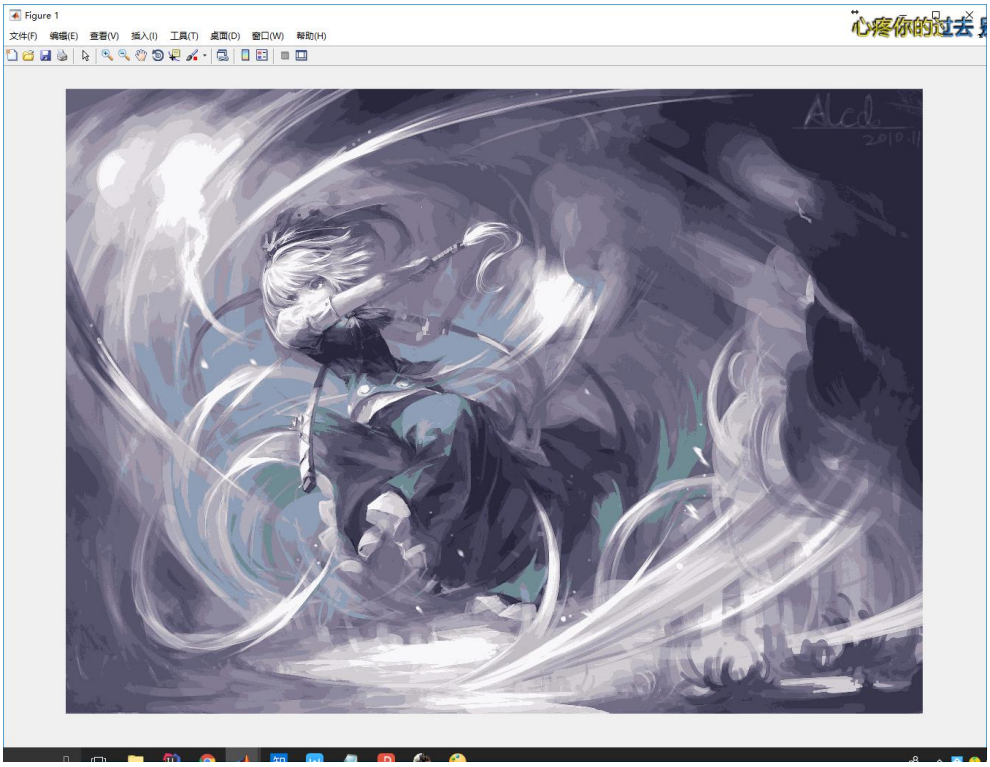


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)

K = 32

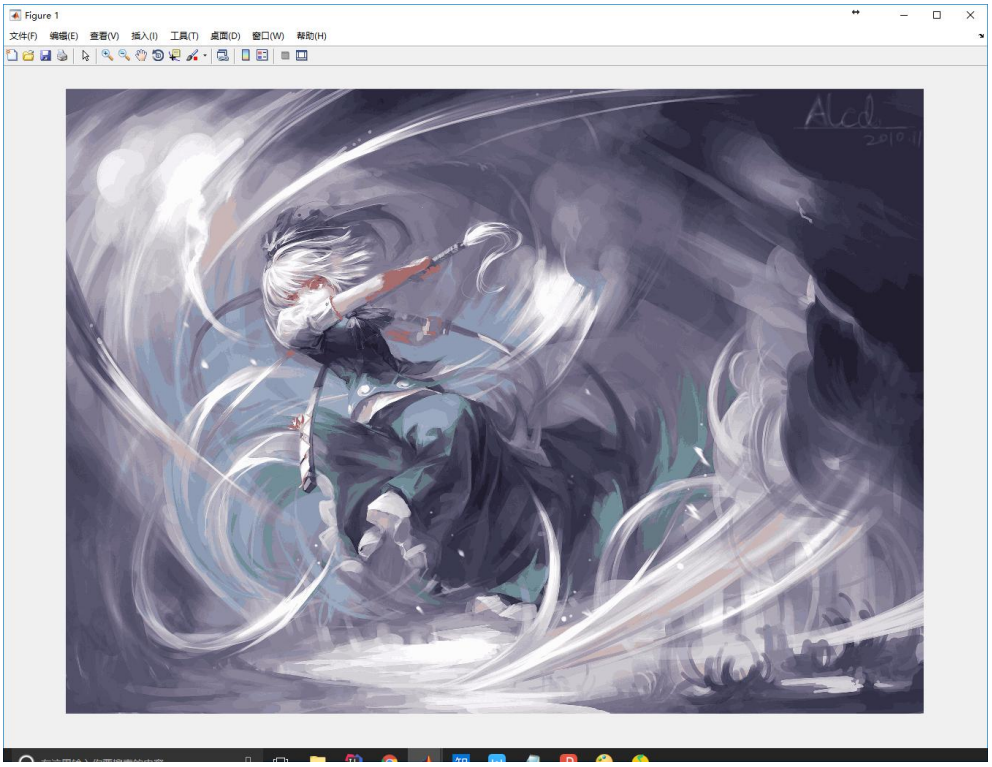


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)

K = 64

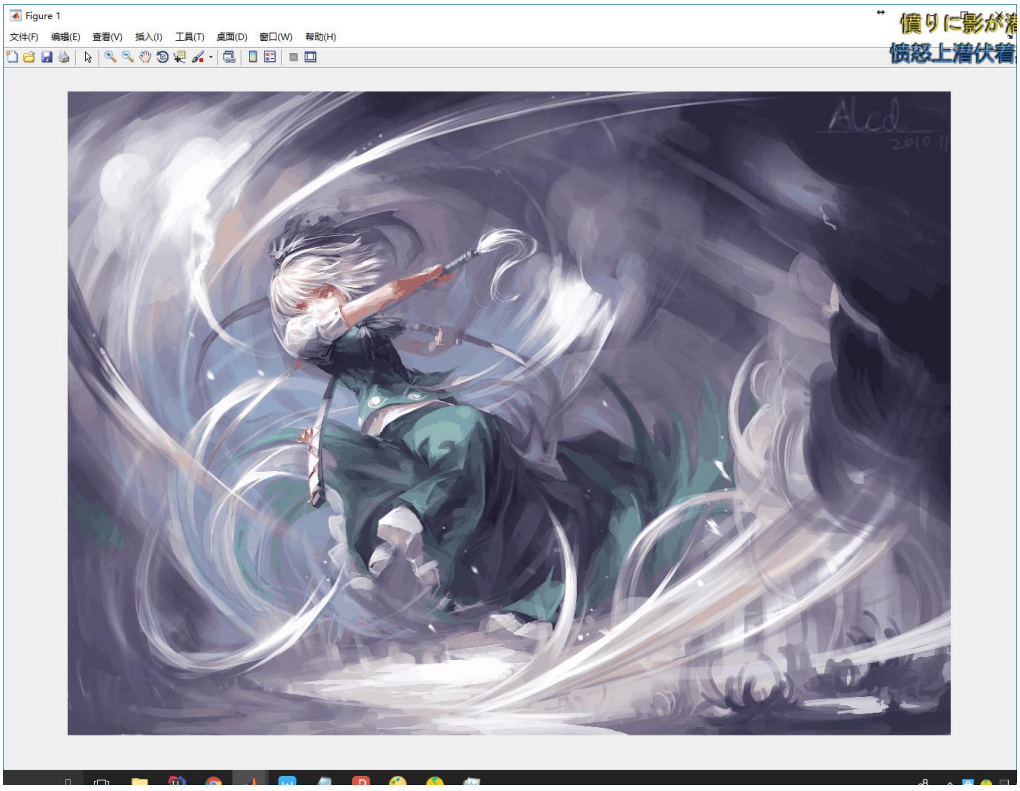


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)