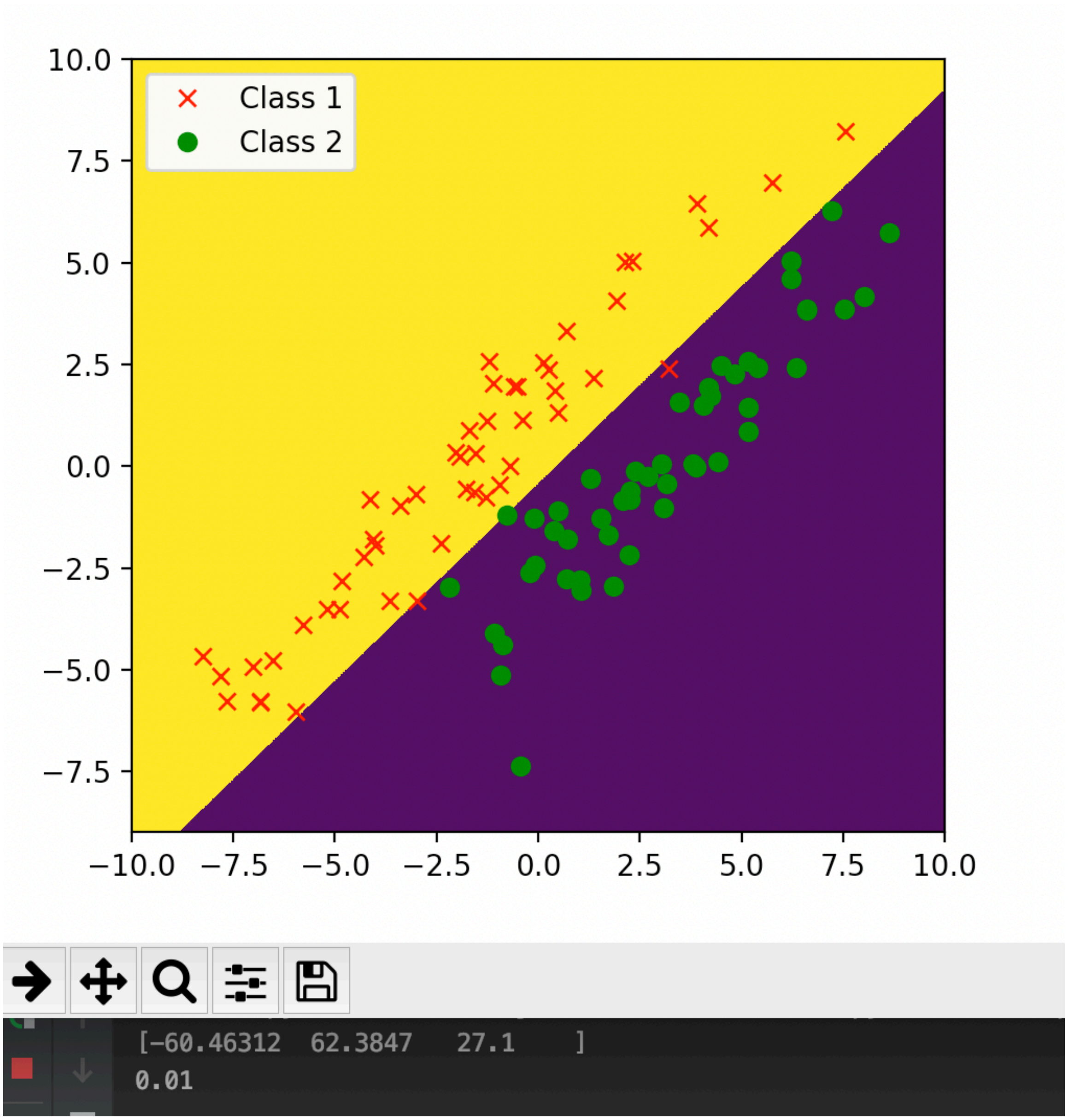
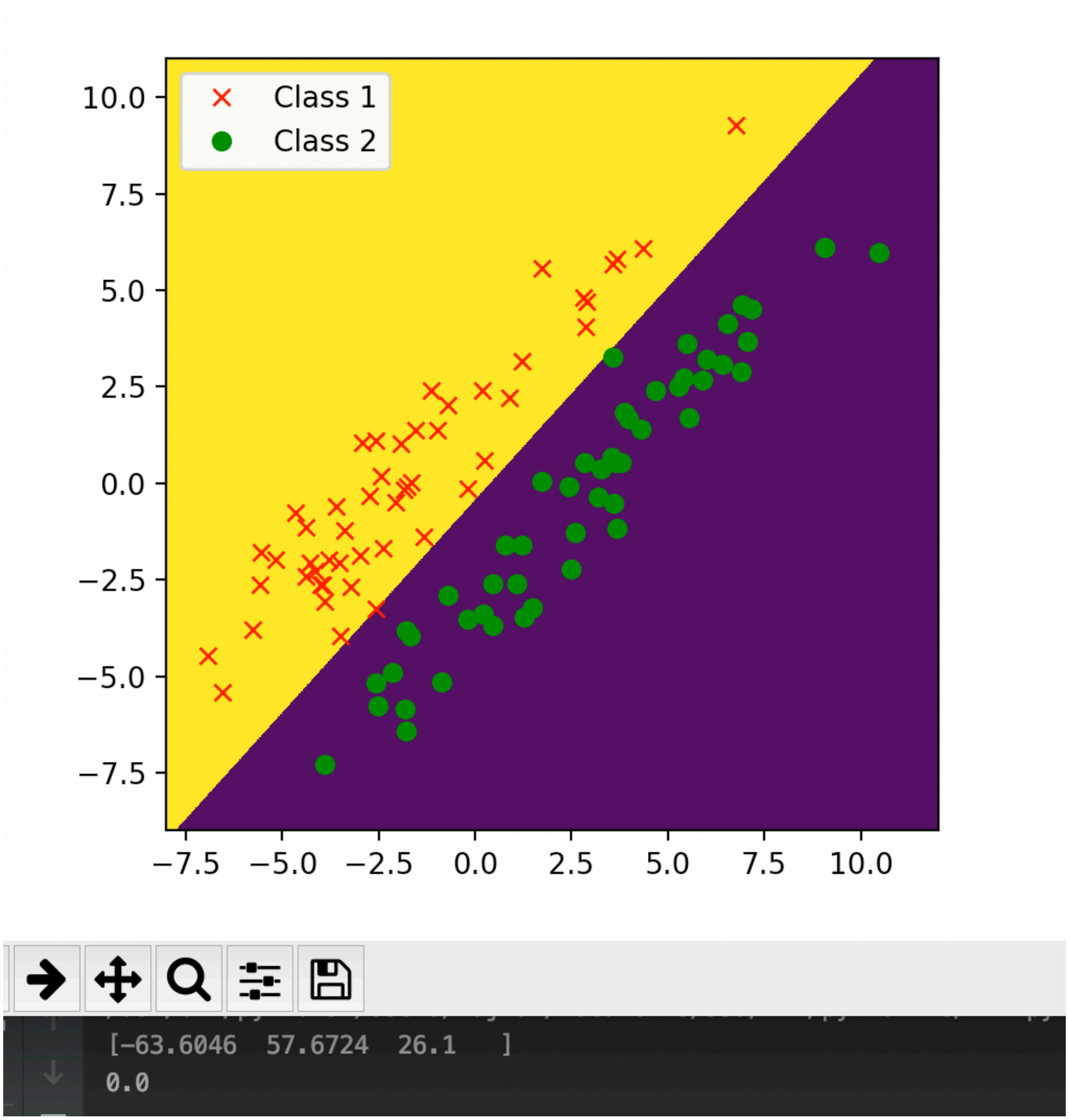
Jian Xu

jxu72364@usc.edu



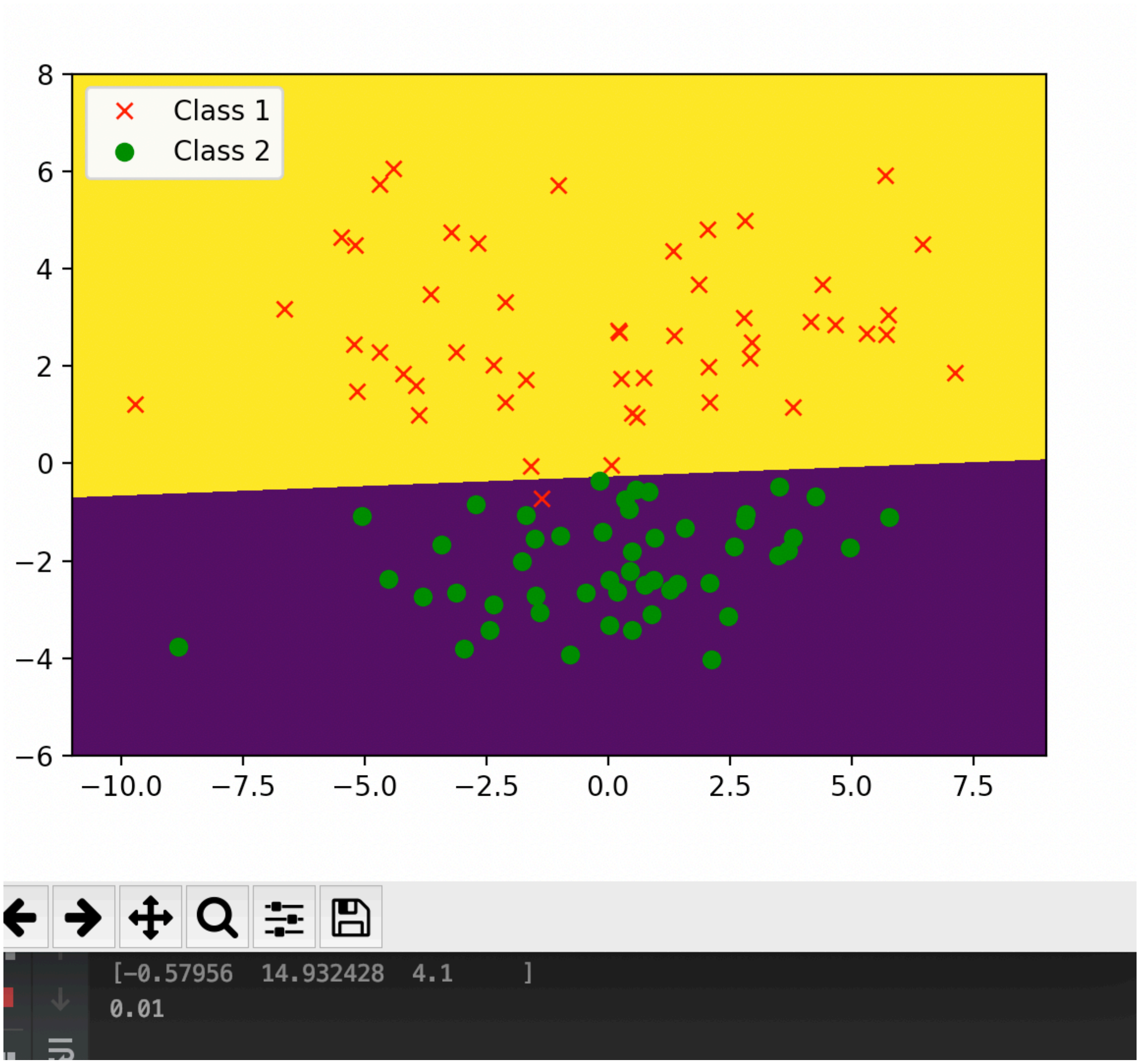
the classification error rate on the training set sythetic1 is 0.01

the weight vector is[-60.46312   62.3847   27.1]



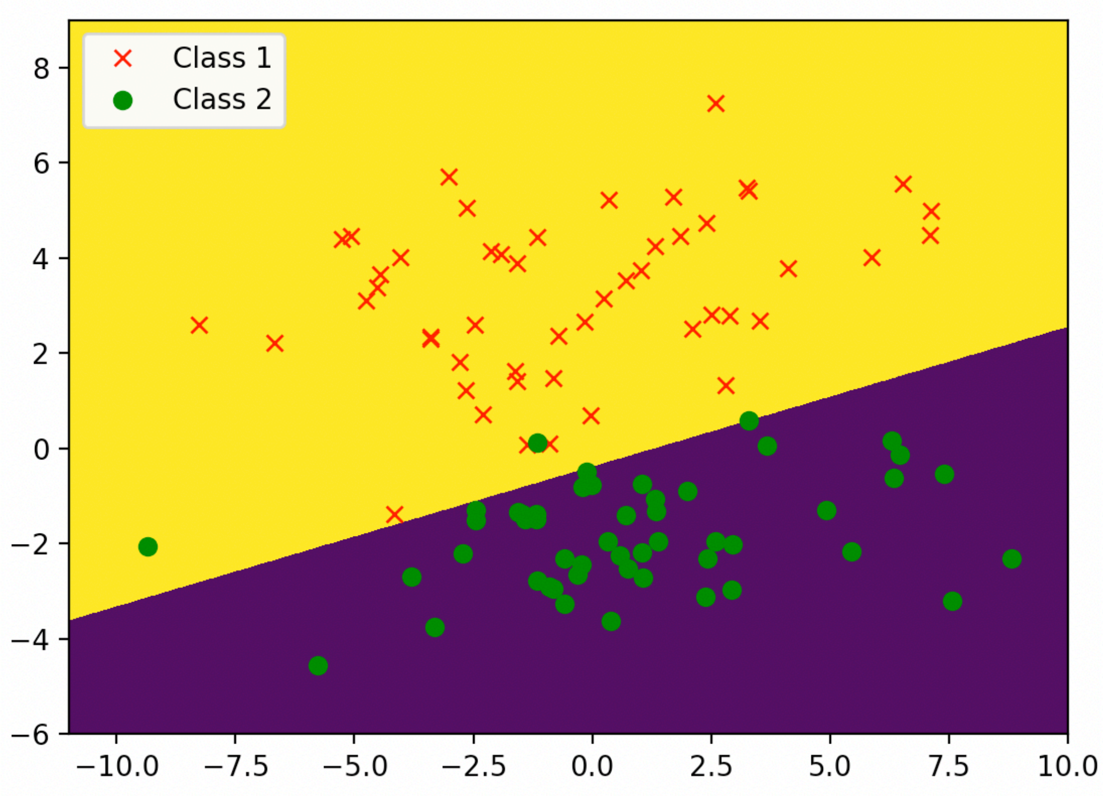
the classification error rate on the test set sythetic1 is 0.0

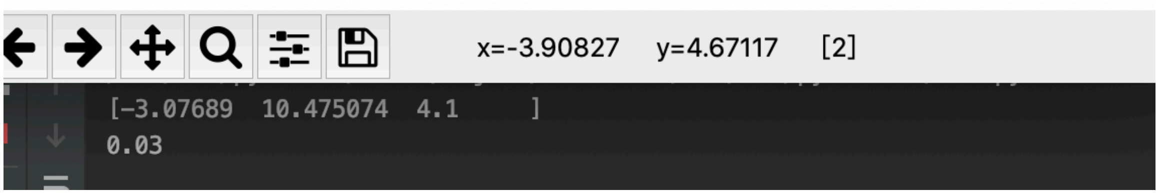
the weight vector is[-63.6046   57.6724   26.1]



the classification error rate on the training set sythetic2 is 0.01

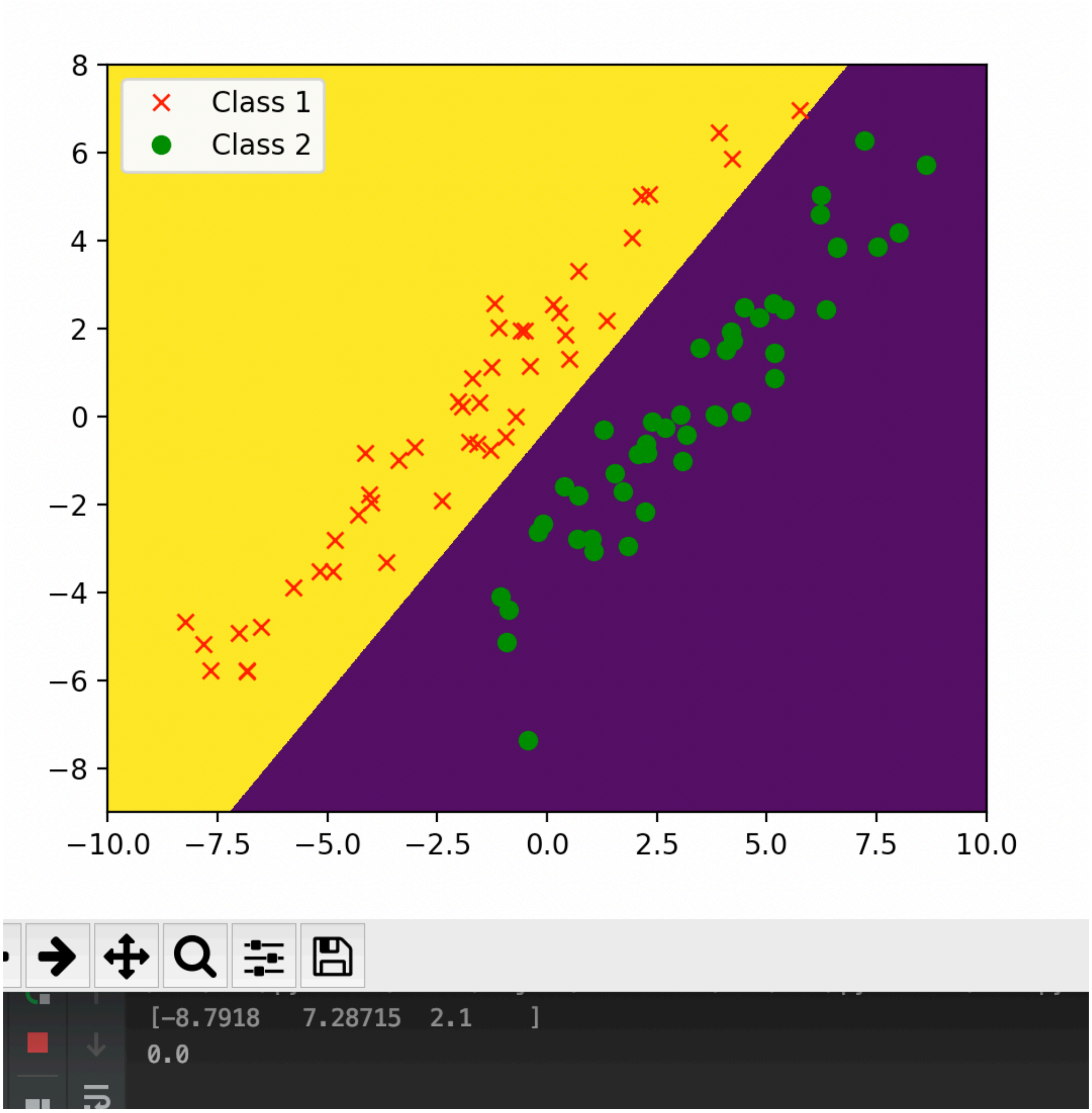
the weight vector is[-0.57956   14.932428   4.1]





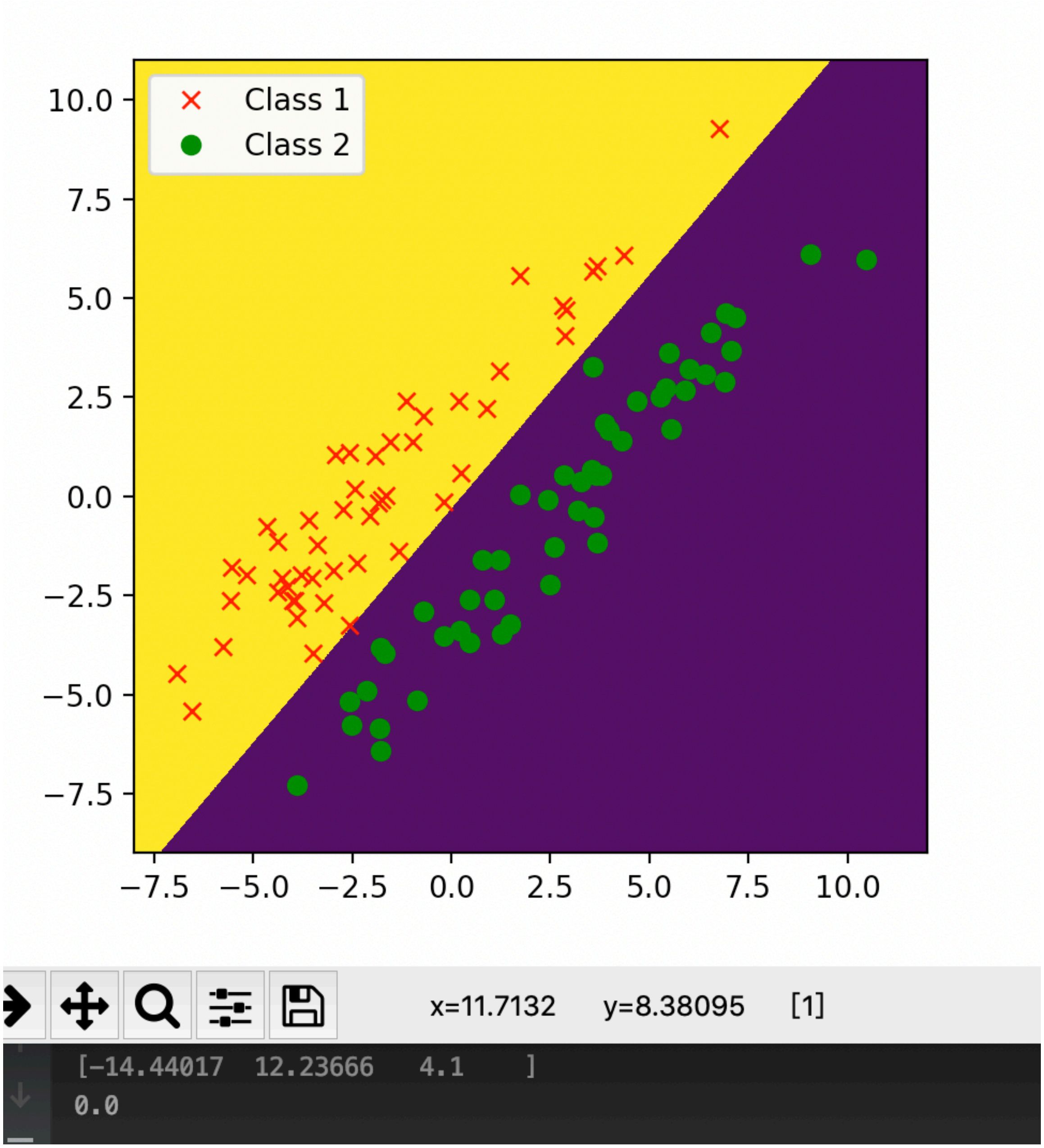
the classification error rate on the test set sythetic2 is 0.03

the weight vector is[-3.07689   10.475074   4.1]



the classification error rate on the training set sythetic3 is 0

the weight vector is[-8.7918   7.28715   2.1]



the classification error rate on the test set sythetic3 is 0

the weight vector is[-14.44017 12.23666 4.1]

(C)

Using nearest means, the error rates of training and test set for synthetic1 are 0.21 and 0.24, for sythetic2 are 0.03 and 0.04.

Using perceptron, the error rates of training and test set for synthetic1 are 0.01 and 0, for sythetic2 are 0.01 and 0.03.

We can find that because the distribution of the datasets is different, sometimes using nearest means can get the results as good as perceptron, but sometimes it cannot.

But in the data sets synthetic1 and synthetic2, perceptron always has a good result.

(a) $$E\{\Delta \underline{w}(i)\} = E\{\underline{w}(i+1)\} - E\{\underline{w}(i)\}$$

$$w(i+1) = w(i) - \eta(i) \nabla_{\underline{w}} J_{\underline{w}}(i)$$

so $$E\{\Delta \underline{w}(i)\} = -\eta(i)\, E\{\nabla_{\underline{w}} J_{\underline{w}}(i)\}$$

$$E\{\Delta \underline{w}(i)\} = -\eta \cdot \frac{1}{N} \sum_{i=1}^{N} \nabla_{\underline{w}} J_{\underline{w}}(w_i)$$

(b) for batch gradient descent

$$\Delta \underline{w}(i) = \underline{w}(i+1) - \underline{w}(i)$$

$$\Delta w(i) = -\eta(i) \sum_{i=1}^{N} \nabla_{\underline{w}} J_{\underline{w}}(w_i)$$

Because from (a) we have:

$$E\{\Delta \underline{w}(i)\} = -\eta \cdot \frac{1}{N} \sum_{i=1}^{N} \nabla_{\underline{w}} J_{\underline{w}}(w_i)$$

so, we have $\Delta w(i) = N\, E\{\Delta \underline{w}(i)\}$

so, we know the expected value of SGD equals to $\frac{1}{N}$ batch GD update

explaination: Because batch GD using the whole data set to calculate loss function, so $\frac{1}{N}$ of batch GD means average loss of the whole data set.

SGD means we choose date randomly from the dataset, and expected value of SGD also means "average" of the data we chosen.

so, we can know that the expected value of SGD equals to $\frac{1}{N}$ batch GD update