1. What is the example space, X, for this problem?
    a. It would be the set of the three different possible characters to the 784th power
2. Which model do you expect to be more expressive: a binary classifier perceptron or a binary classifier multi-layer perceptron with one hidden layer? Why?
    a. Multilayer perceptron is more expressive as it is able to show any linear or nonlinear functions. While a binary classifier can only show linear separator functions
3. Which one of the models described in question 2 do you think is more likely to overfit, and why?
    a. The Multilayer is also the more likely to overfit as it more complex, a perceptron can not overfit non-linear data
4. Do you expect your perceptron to converge (assuming a constant learning rate)? Why or why not?
    a. It will not converge, if the data is not linearly separable and the learning rate is constant it will not converge.

1) Try adjusting each setting listed above, and report which combination (learning rate, training example ordering, and weight initialization) worked best. If you used random initial weights, specify what distribution the weights came from (uniform, Gaussian, or another distribution).

    alpha = 1 #decaying
    n_epoch = 20 # number of epochs
    training_order = 0 #0 for random, 1 for fixed
    weights = 1 #0 for zero weights or 1 for random weights

2) How many epochs were necessary to train your perceptron?

    20

3) What accuracy did you achieve on the training set?

    89%

4) What accuracy did you achieve on the test set?

    81.7%

1. Which combinations of hyperparameters did you try, and which combination worked best?
   activation = 'relu',
   max_iter = 500,
   hidden_layer_sizes = (200,200,200),
   alpha = .01,
   momentum = .9

2. How many epochs were necessary to train your multilayer perceptron?
   1000
3. What accuracy did you achieve on the training set?
   93.8%

4. What accuracy did you achieve on the test set?
   90.8%

5. How did the results obtained using the multilayer perceptron compare to those obtained using the single-layer perceptron in part 2? Explain the differences.
   a. Much higher, the network is more perceptive and expressive with the hidden layers so it is much more accurate

Citations:
http://scikit-learn.org/stable/modules/neural_networks_supervised.html
http://www.kdnuggets.com/2016/10/beginners-guide-neural-networks-python-scikit-learn.html/