**ML HW7**

1. **Advantage and disadvantage of stochastic gradient descent**

* Advantage: Using stochastic gradient descent is way computationally faster than gradient descent. By selecting only one observation per iteration instead of computing the sum of all observations, it fastens the process of optimization and could be especially efficient when some of the observations are redundant. Besides, when we get a new data point, we can just plug in the new data without redoing the gradient descent from scratch. Additionally, it is more likely for stochastic gradient descent to jump out of a local minimum so that it can reach a global minimum of the loss function.
* Disadvantage: The random choice of one observation every time will lead to a longer time for gradient descent to converge to a minimum. Meanwhile, it is possible that gradient descent may go in the wrong direction because of the random choice.

1. **Test the OOS accuracy for test data[6601:10000]**

The average OOS accuracy for test data[6601:10000] is 0.94463074

1. **Test the OOS accuracy for test data[6601:10000] after removing the dropout**

The OOS accuracy after removing dropouts is 0.9449838. The accuracy actually didn’t change a lot, which means this model is not overfitting, hence the dropout influence is not very significant. The intuition behind is that dropout cut out some of the neurons randomly according to the probability, this prevents the output layer’s over-reliance on some of the input that test dataset doesn’t have.

1. **What other conventional algorithm(s) could you have used for this task?  What was the advantage of using a neural network for this exercise, instead of any other potential conventional algorithms?**

* Decision tree: Represent the digit pictures as numeric vectors or matrices, then use the random forest method to categorize the data. However, the decision tree can be really biased while handling high dimensional data in this case.
* K-means: Represent the digit pictures as numeric vectors or matrices, and use K-means to cluster the digit pictures with K = 10 initially and look at different error rates and assign the final number of centroids. However, the K-means is of lower speed and lower accuracy compared to Neural Network.
* Multinomial Logistic Regression: Since y variable is categorical range from 0 to 9, we can use multinomial logistic regression to identify the number. However, as 28 \* 28 pixels, x variable is highly dimensional, with which MNL is hard to handle.

The advantage of the Neural network to be applied in this exercise is that three layers are utilized to figure out a better combination of weights that in a way reduces dimension.

1. **Use a multinomial logistic regression model to achieve the same handwriting identification task and compare the out-of-sample accuracy against the results of the neural network.**

The out-of-sample accuracy from multinomial logistic regression is also about 90%, close to the accuracy from NN, though MNL will take much longer time to run.