黃偉祥 X1136010

1. Sigmoid is used for binary classification, cause sigmoid function can map input into a 0 to 1 range of output, represent the possibility of output = 1, but we can use 1-P to get the possibility of output = 0. While SoftMax will give different possibilities of each class.
2. Learning rate can make different step scale when updating weight, while learning rate is not small enough it will change too much on the weight. Mini-batch give random mini batch of training set and when training different mini set will have different losses. Bias vector could affect much on the weight also, at the early training phase it would not be accurate. This may make loss oscillating during model training. Loss function is not linear, and also our activation is not linear too, so in the training process we will have oscillating losses.
3. A big learning rate will increase the time to make the model converge, because each update step will be large with a big learning rate so the model needs more training epoch to reach the converge point. A small batch size will make the model converge faster, because each batch will update the weight once so if we got 100 batches in a training set, it will update the weight 100 times with a training set epoch.
4.