[Takeaway messages]

For 3.4, I still have a little trouble following the process of Stochastic Gradient Descent. If the step makes the loss function result worse, it should be discarded instead of implemented, but I'm not sure where that is done.

For 4.1, I think it is interesting how the output layer can be molded by the Softmax function to become a probability distribution for whether the sample belongs to each class. The overall process isn't too difficult to understand, although I do not understand the function for cross entropy loss.

It seems like as the number of epochs increase, the model will converge better to the training data, but this could cause overfitting which makes the model have poor performance for real world data.

For 5.1, I think I understand that activation functions are important for introducing non-linearity to the output function, which allows for modeling nonlinear problems. However, it is still difficult to visualize how increasing or decreasing hidden layers will affect the performance of the model.