Part 1:

The early stopping method of stochastic gradient descent seems to be relatively effective at predicting the final loss of the fully trained model. Typically, the early stop's loss is within 0.01 of the final loss. When stopping the model at the stop point, the accuracy remains very good, if not slightly worse than the fully trained model.

Part 2:

The batch gradient descent method converges more consistently than SGD. They find convergence around the same time, but the batch gradient descent seems to be more consistent, where there are rare cases that the stochastic model will fail to converge.

Part 3:

The active learning method using stochastic gradient descent seems to be very effective at finding the minimal number of samples needed to train an accurate model. However, the compute time seems to be much longer than a normal stochastic or batch GD model. A strange behaviour I observed is that the hinge loss typically increases as the amount of samples increases, but will of course level out and converge afterwards.