Lavinia Wang
#1473704
CSC578-701

To begin with, first thing to figure out is the network structure. Node number of both input and output layers are already known, so I need to specify the node number of hidden layer(s). In Nielsen's book chapter 1, he wrote "simple learning algorithm + good training data ≥ sophisticated algorithm". So I decided to use only 1 layer of hidden nodes. I Googled about the rule of thumb of deciding the hidden node number, and I experimented with this formula.

$$N_h = \frac{N_s}{a * (N_i + N_o)}$$

$N_i = number\ of\ input\ neurons,\ N_o = number\ of\ output\ neurons,$
$N_s = number\ of\ samples\ in\ training\ dataset,$
$a = an\ arbitrary\ scaling\ factor\ usually\ 2 - 10$

I ended up with trying $a\ \in [2,10] \equiv\ N_h \in [2,11]$.

Then is the number of epoch, mini-batch and learning rate. Also from Google search, some articles pointed out that good epoch sizes are often between 10 and 200 examples but depends largely on the diversity of example set. Since we only have 150 samples in the dataset, my max epoch is tested with 150. With respect to learning rate, I have found that there is a big improvement if you do a little training with a very low learning rate (if it is done on a small epoch number). I started with learning rate η = 5, epoch = 150, mini-batch = 10. Then I got results like a roller coaster, which consists of continuous low accuracy at the beginning and a sudden 98% in the middle then a steep drop back to 66% as where it began.

I also tried a combination of epoch = 30, mini-batch = 1, η = 0.1, with incremented hidden nodes starting from 2 to 11. When hidden node > 6, I started to get accuracy of 98.7% in my first attempt. When it reached 11, I have my accuracy of 99.3% in the first attempt.

```
[149,
 0.9933333333333333,
 2.8819683428044286e-05,
 0.0009666258525111964,
 0.048799898412842185],
```

I also tried hidden node = 15 and 20. The results were no better than what I got from hidden node = 11. So I believe it's my best configuration: network [4,11,3] with 30 epochs, 1 mini-batch and η = 0.1.

After this experiment, I felt that the process of configuration is mysterious, or more like art than science. There are 4 undefined parameters and they are interacting with one another at each optimization. Also, manually changing those parameters means a lot of work which I believe can be done by computers using some automated procedure.