



DEEP LEARNING IN DATA SCIENCE DD2424

SOLUTION TO ASSIGNMENT 2 (FOR BONUS)

YIMING FAN
yimingf@kth.se

1 Optimize the performance of the network

We tried the following methods of optimization:

1. Trained the network with more than 60 epochs until the network converges (i.e. when the gradient stops decreasing). We used the last 800 data for validation.
2. We did a finer grid search and found the ‘better’ combination of hyper-parameters: $\lambda = 0.001$, $\eta = 0.02$. The learning rate η decays by 0.95 after every 10 epochs.
3. We tried increasing the number of hidden nodes from 50 to 300 and also increase the regularization term λ respectively. The number of hidden nodes that provides best training results are 100. Basically increasing the number of hidden nodes will increase the complexity of the network, hence improve the accuracy. However too many hidden nodes may cause the network to overfit.

2 Train network using a different activation to ReLu

We tried `sigmoid` function for activation function. The training result was terrible:

1. With momentum the cross-entropy loss doesn’t decrease - the network is worse than random guesser.
2. Without momentum the loss converge to around 2.2 after 5 epochs - a little better than random guesser.
3. `sigmoid` is usually 2~3 times slower than `ReLU`.

Conclusion: `sigmoid` is slow and inefficient. We will discard using it in the future.
