Deep Learning in Data Science
DD2424

Report to Assignment 1

Yiming Fan
yimingf@kth.se

1 Introduction

This assignment aims at training a multiple linear one-layer network using *gradient descent* method. The dataset used in this assignment is CIFAR-10.

2 Methods & Mechanisms

The network consists of a linear classifier $\mathbf{W}$ and a bias vector $\mathbf{b}$. In the beginning linear scoring function

$$\mathbf{s} = \mathbf{W}\mathbf{x} + \mathbf{b} \tag{1}$$

and *softmax* function

$$\mathbf{p} = \text{softmax}(\mathbf{s}) = \frac{\exp(\mathbf{s})}{\mathbf{1}^T \exp(\mathbf{s})} \tag{2}$$

were used as classifier. The cross-entropy loss plus a regularization term was to be minimized. We calculate the gradient with regularization terms at each mini-batch by:

$$\frac{\partial J}{\partial \mathbf{W}} = \frac{1}{|\mathcal{D}|} \sum \mathbf{g}^T \mathbf{x}^T + 2\lambda \mathbf{W} \tag{3}$$

$$\frac{\partial J}{\partial \mathbf{b}} = \frac{1}{|\mathcal{D}|} \sum \mathbf{g} \tag{4}$$

where

$$\mathbf{g} = -\frac{\mathbf{y}^T}{\mathbf{y}^T \mathbf{p}} \left(\text{diag}(\mathbf{P}) - \mathbf{P}\mathbf{P}^T \right) \tag{5}$$

Then add them with the original terms:

$$\mathbf{W} = \mathbf{W} - \eta \frac{\partial J}{\partial \mathbf{W}} \tag{6}$$

$$\mathbf{b} = \mathbf{b} - \eta \frac{\partial J}{\partial \mathbf{b}} \tag{7}$$

Both train and validation loss decrease at each epoch.

3 Results

We used 90% of the data from a batch as the training dataset, and leave the rest as the validation dataset.

Some conclusions:

1. The learning rate should not be set too high, otherwise the learning process will be unstable. Too high learning rate may let the gradient ‘swing’ around the minima point.

2. The regularization terms could effectively avoid overfitting. However too high λ may decrease the size of the weight $\mathbf{W}$, which may increase the bias.

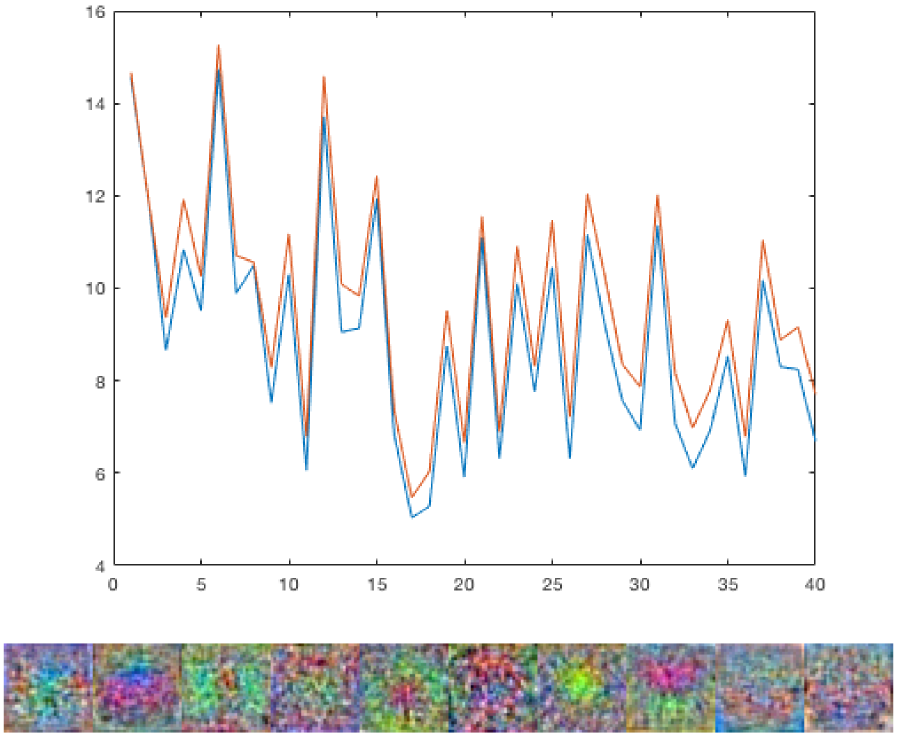

Figure 1: Result figure and the learnt matrix on: `lambda=0, n_epochs=40, n_batch=100, eta=.1`. The test accuracy is 26.63%.

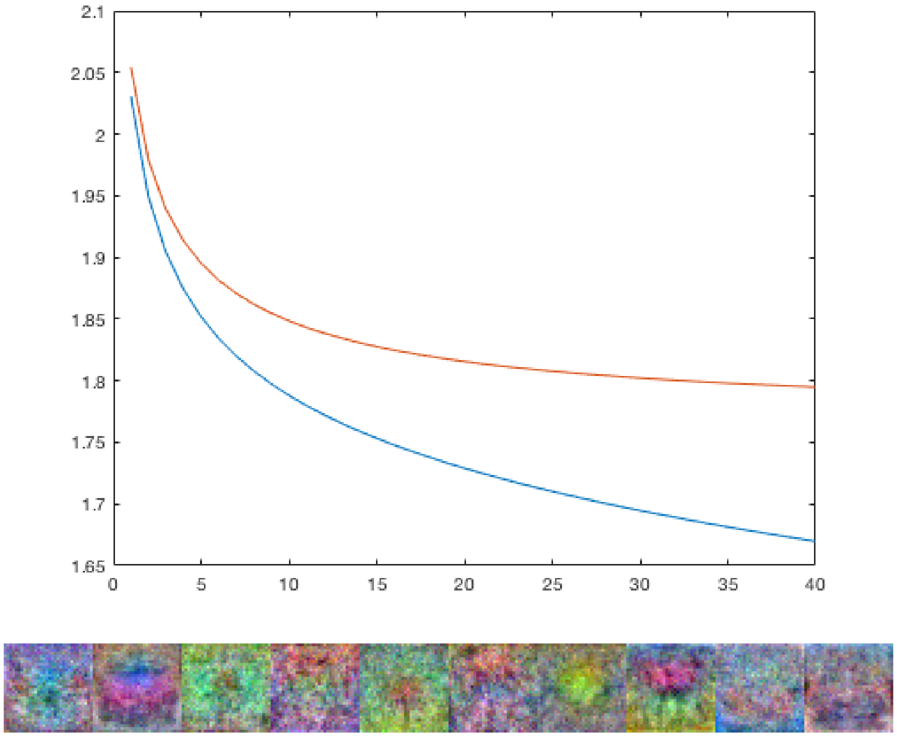

Figure 2: Result figure and the learnt matrix on: `lambda=0, n_epochs=40, n_batch=100, eta=.01`. The test accuracy is 42.24%.

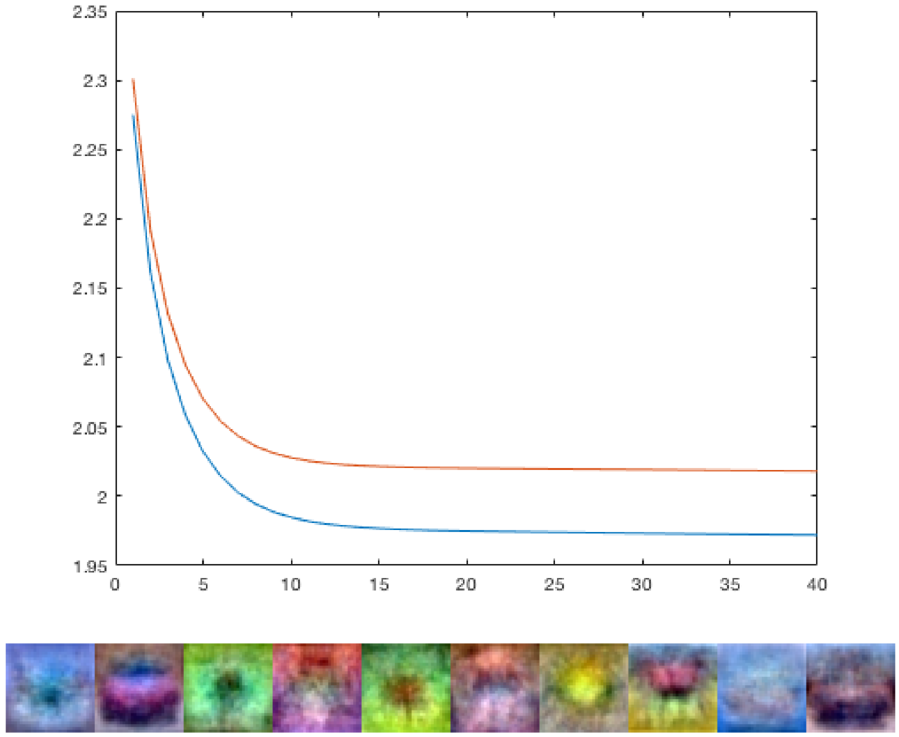

Figure 3: Result figure and the learnt matrix on: `lambda=0.1, n_epochs=40, n_batch=100, eta=.01`. The test accuracy is 36.03%.

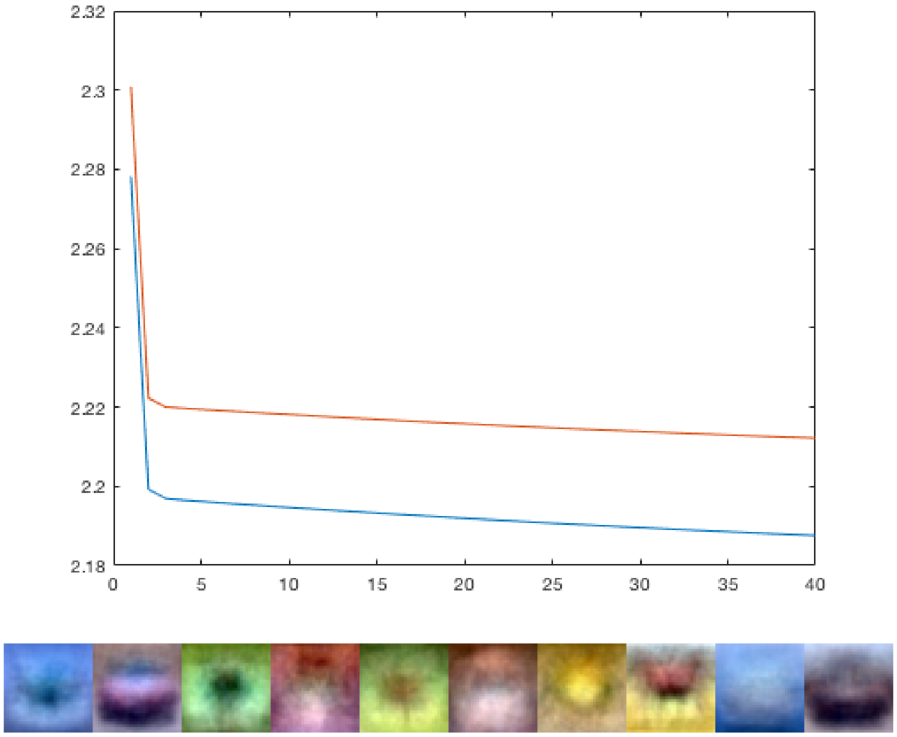

Figure 4: Result figure and the learnt matrix on: `lambda=1, n_epochs=40, n_batch=100, eta=.01`. The test accuracy is 24.56%.