Machine Learning, 2024 Spring
Assignment 4

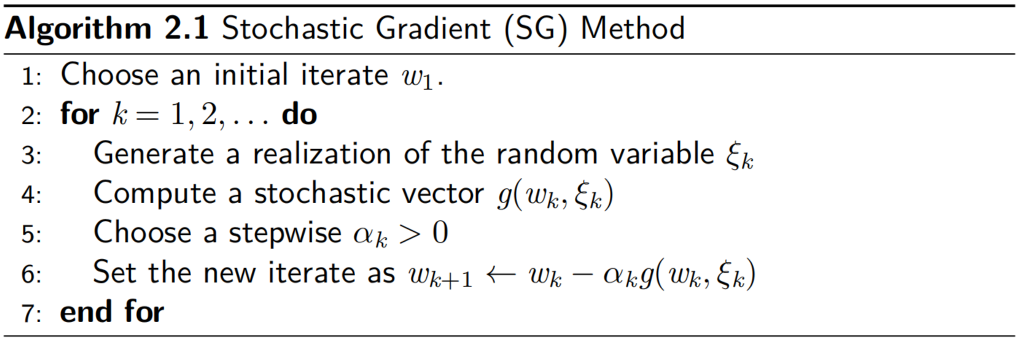
Notice

Plagiarizer will get 0 points.
LaTeXis highly recommended. Otherwise you should write as legibly as possible.

Problem 1 For problem 3 in assignment 3, change your GD code to SGD and complete the tasks below:

- Present your code.
- How to choose (mini -) batch size?
- How to choose learning rate?
- How to terminate?
- Demonstrate the impact of different learning rates on the accuracy of the solution. In other words, your program should output an image similar to the image on page 34 of the Lecture 6-SGD PPT.

Solution:



Algorithm 2.1 Stochastic Gradient (SG) Method

```
Choose an initial iterate $w_1$.
for $k = 1, 2, \ldots$ do
    Generate a realization of the random variable $\xi_k$
    Compute a stochastic vector $g(w_k, \xi_k)$
    Choose a stepwise $\alpha_k > 0$
    Set the new iterate as $w_{k+1} \leftarrow w_k - \alpha_k g(w_k, \xi_k)$
end for
```

Figure 1: Stochastic Gradient Descent

The SGD algorithm is implmented following the Pseudo code in Figure 1.

0. Data preparation: The data is treated in the same way as the data in assignment 3:
The 'User ID' column is dropped, the 'Age' and 'EstimatedSalary' are separated into different intervals. At last, a bias term '1' is add for each input fetures.
And the features are normalized to make the data have zero mean and unit variance.
And before seperating the data, we random shuffled the total 400 data with a fixed seed in order to

make sure the randomness, but could be stable reproduction.
The data is divided into training set, validation set, and testing set. The training set is used to train the model, and the validation set is used to evaluate the model for selecting the most suitable learning rate, and the testing set is used to evaluate the model for the final performance. The size of training, validation, and testing set is 240, 80, 80 respectively, which is 60%, 20%, 20% of the total data.

1. The code and the method to run the code are all in the folder ‘code’.

2. The mini-batch size is set to be 1 as we are applying the stochastic gradient descent.
And the diction is set to be

$$g(\mathbf{w}_k, \xi_k) = -\nabla e_k(\mathbf{w}_k) = \frac{-y_k \mathbf{x}_k}{1 + e^{y_k \mathbf{w}_k^T \mathbf{x}_k}}$$

3. The learning rate η is tried with the following values: $0.08, 0.04, 0.02, 0.01, 0.005, 0.0025$.
From the results show in Figure 2, we can see that the learning rate $\eta = 0.01$ has the best accuracy on the validation set. So we take the $\eta = 0.01$ as the final learning rate to test on the testing set.

4. The termination condition is chosen by reaching the number of iterations: 50000, which is large enough to make the loss converge.

5. The impact of different learning rates on the accuracy of the solution is shown in Figure 2. And we can see that in the suitable range, the loss is convergence to lower value as the learning rate get smaller, which suits the regular to the image on page 34 of the Lecture 6-SGD PPT. From the last 1000 iterations of training process, it could be seen clearly.
But we can also notice that if the learning rate is too small($\eta = 0.005, 0.0025$), the loss will converge to a non-better value than $\eta = 0.01$. Perhaps this is because the learning rate is too small and unsuitable for the model to learn the data. But from the tends of $\eta = 0.08, 0.04, 0.02, 0.01$, we would see that the lower learning rate could reach a better convergence loss, and have a better performance.

And the Figure 3 is the accuracy and loss on validation set, form the validation set, we choose the learning rate that has the best accuracy, so $\eta = 0.01$ is selected.

And the Figure 4 is the test accuracy and loss with the best learning rate $\eta = 0.01$.
The final prediction accuracy is $\dfrac{73}{80} = 0.9125$.
And the testing loss curve could be seen has the overfitted trend.

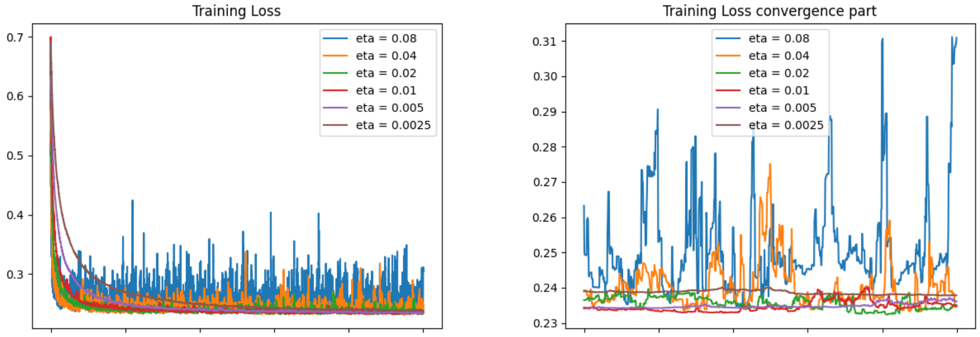


Figure 2: Impact of different learning rates on the training loss

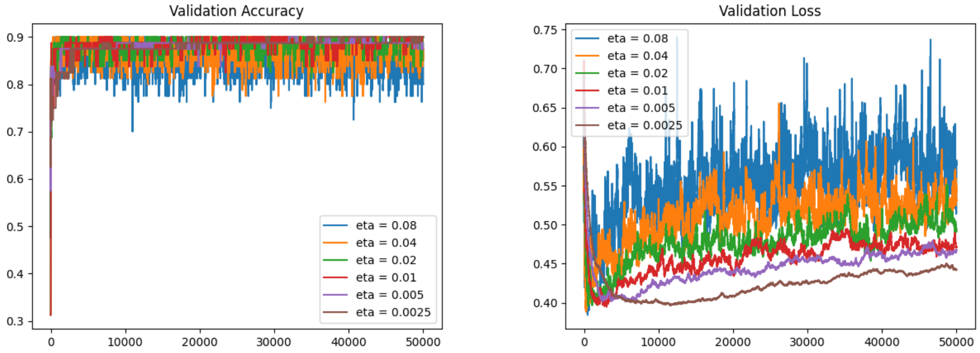


Figure 3: Impact of different learning rates on the validation accuracy and loss

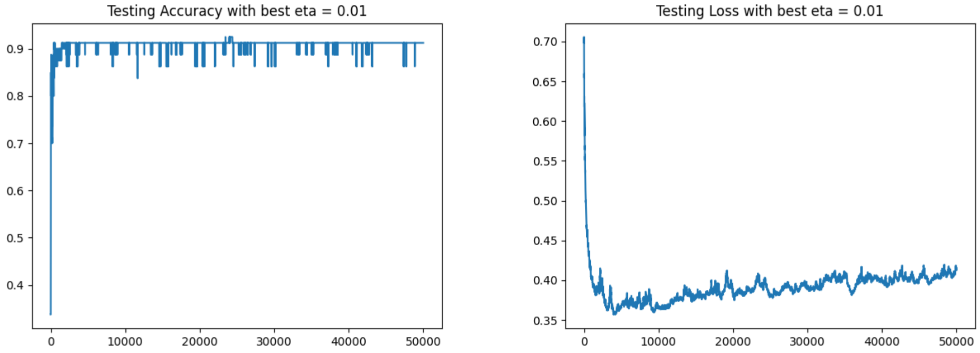


Figure 4: Accuracy and loss for best learning rate $\eta = 0.01$ on the testing set