AI Assignment 5

aasmunhb

April 2017

1 Task 1

By plotting the values of the loss function, with a step of 0.1, we get the following plots. The 3D graph can be further explored by running the included source code. The minimum when using a step size of 0.1 is 0.0049, for the weights $w_1 = 5.9, w_2 = -2.9$. When using our derivative, we get an error of 0.0045.

For the gradient descent plot, fixed starting weights were selected, so that the resulting weights could be easily reproduced with different learning rates. The number of iterations used were 1000, and the learning rates were $10^i, i \in [-4, -3..., 3]$. We observe that both the low and high learning rates does not converge towards a good estimate of weights. For the low learning rates, we are not able to 'walk' far enough, since the number of iterations are low, so we end up in a bad spot. When the learning rate is too high, the algorithm jumps around a lot, and is not able to pinpoint a good combinations of weights. The change in weights is too large. Note that the algorithm may also not converge if it hits a 'ledge' and wants to go outside the given weight.

2 Task 2

The final formula for the derivative is $\frac{[\delta L_n(w,x_n,y_n)]}{\delta w_i} = (y - \sigma(w, x_n)) * x * \frac{e^{(-w^T x)}}{(1+e^{(-w^T x)})^2}$

By running the SGD and BGD we notice that BGD uses significantly more time, but performs a little better with few iterations. For both algorithms, the non-separable set is the hardest to train, and both end up with error rates bigger than zero. With enough iterations, both algorithms manage to separate the separable dataset perfectly. Figure 5 and 6 shows both datasets and the partitions using SGD. Figure 7 shows execution time and error rate as a function of number of iterations. We observe a linear increase in execution time, and an drastic decrease in error rate. The linear increase in time makes sense, since the algorithm is $O(T * i)$ where T is the number of iterations and i is the number of features.

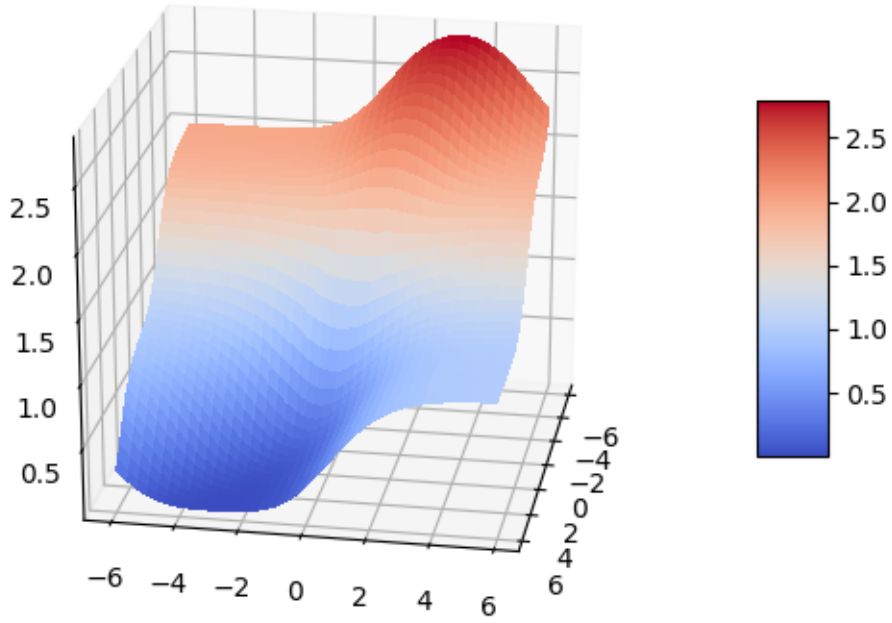


Figure 1: 3D plot of loss function

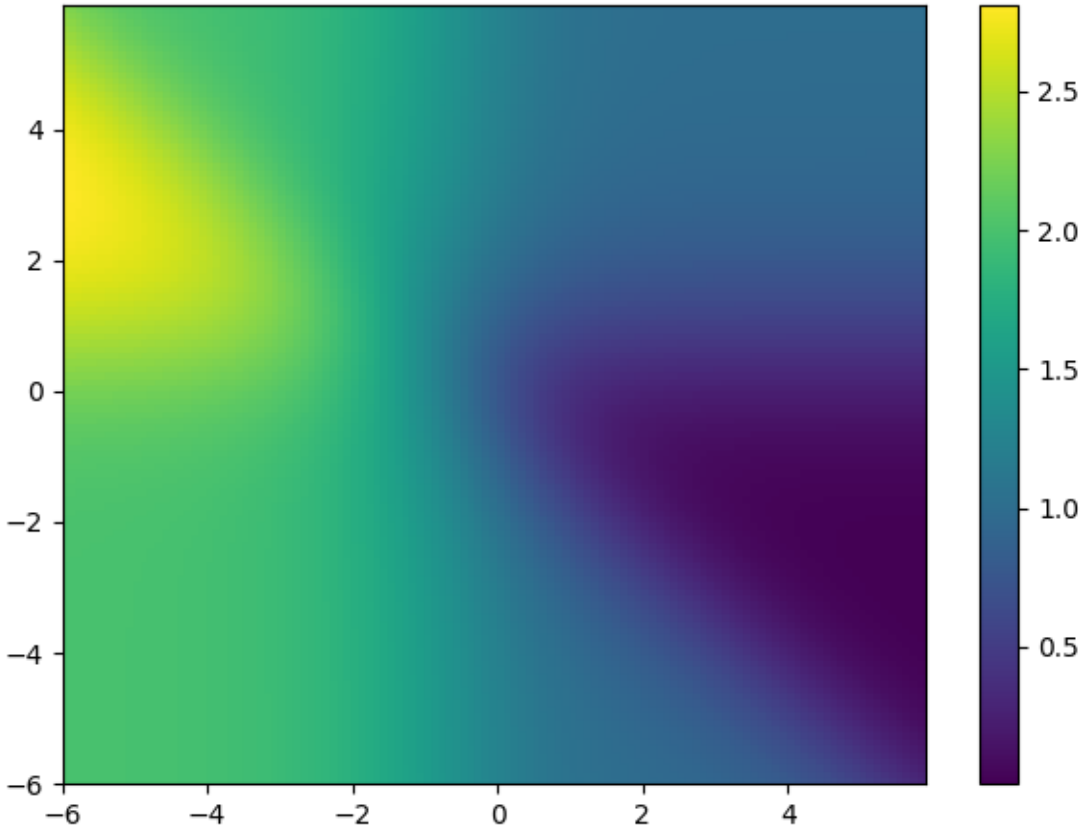
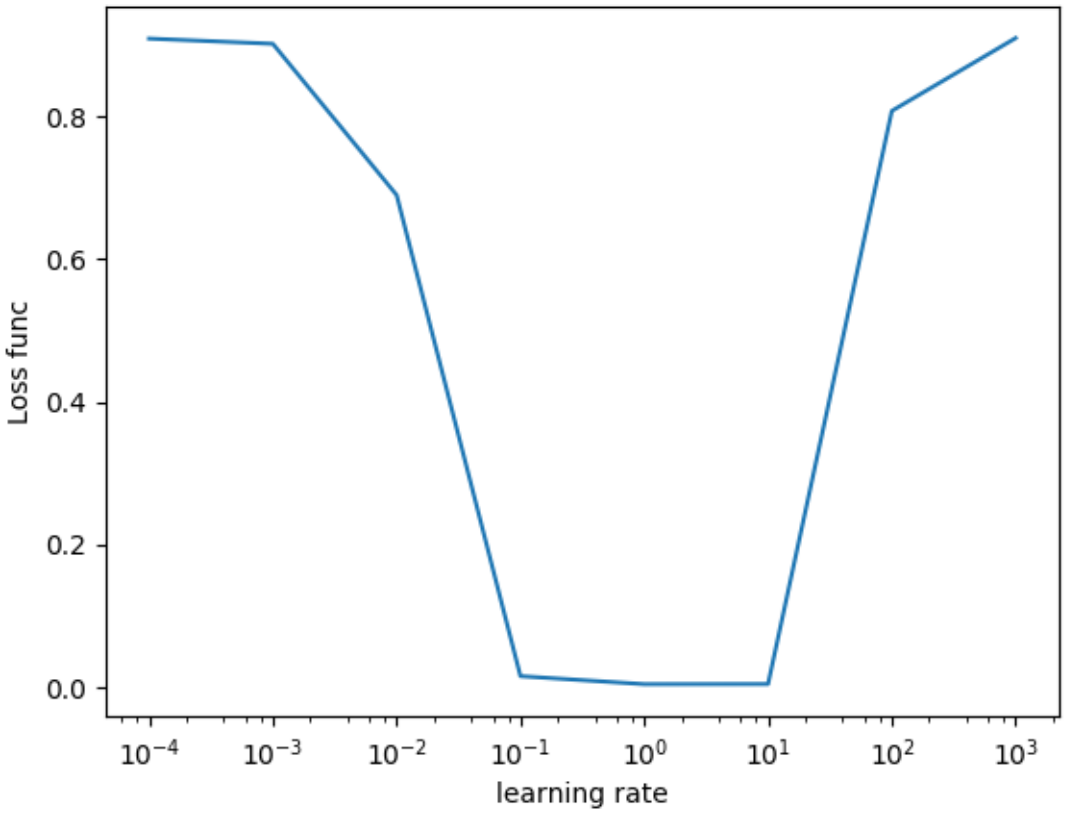


Figure 2: Heatmap of loss function


Loss func
0.0
0.2
0.4
0.6
0.8
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
learning rate

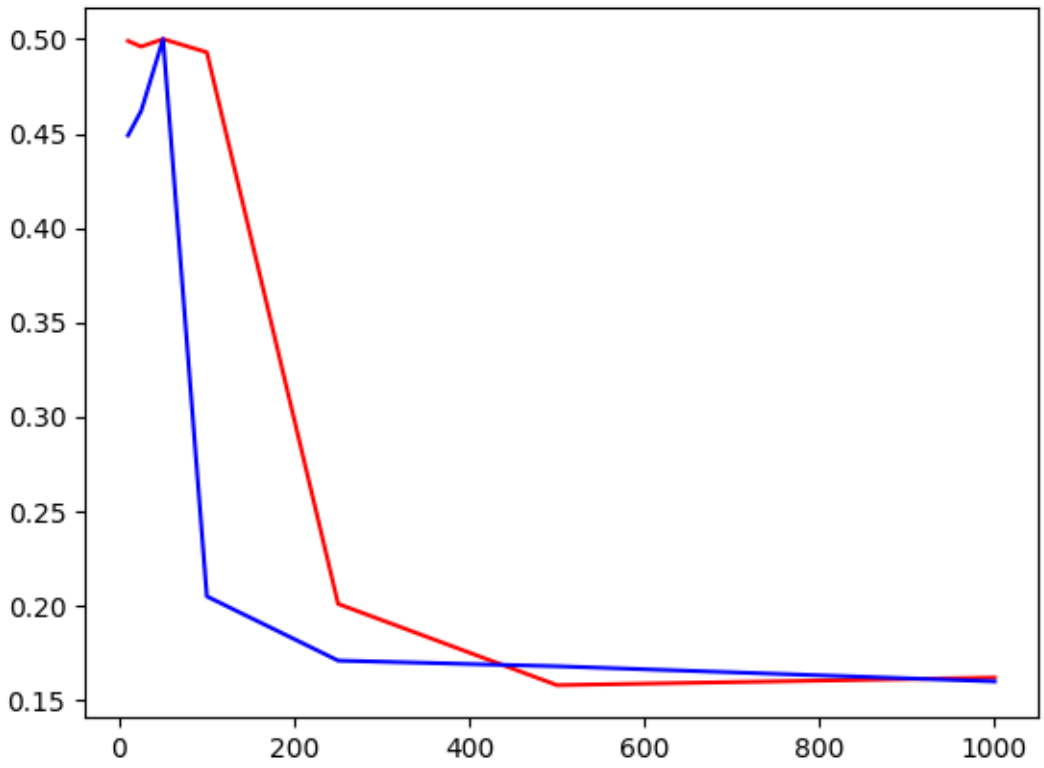


Figure 3: Nonseparable dataset. Red is SGD, Blue BGD. The x-axis is no. iterations, and y is error.

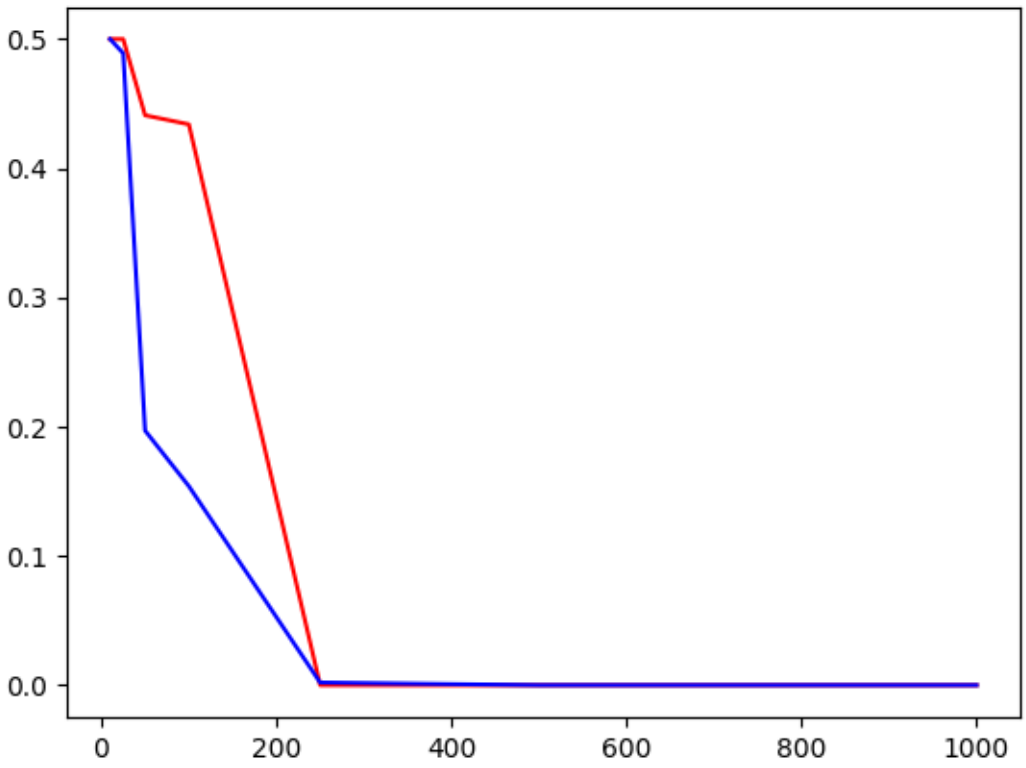


Figure 4: Separable dataset. Red is SGD, Blue BGD. The x-axis is no. iterations, and y is error.

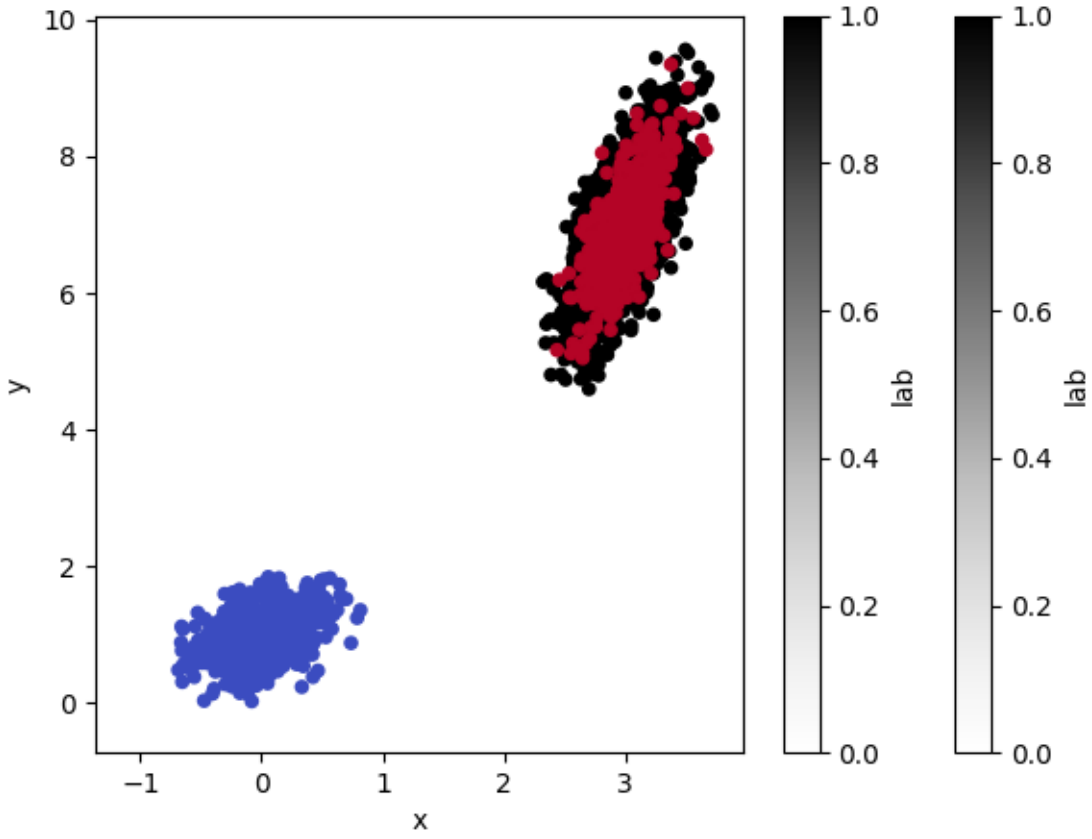


Figure 5: Separable dataset. Trained with SGD, error = 0

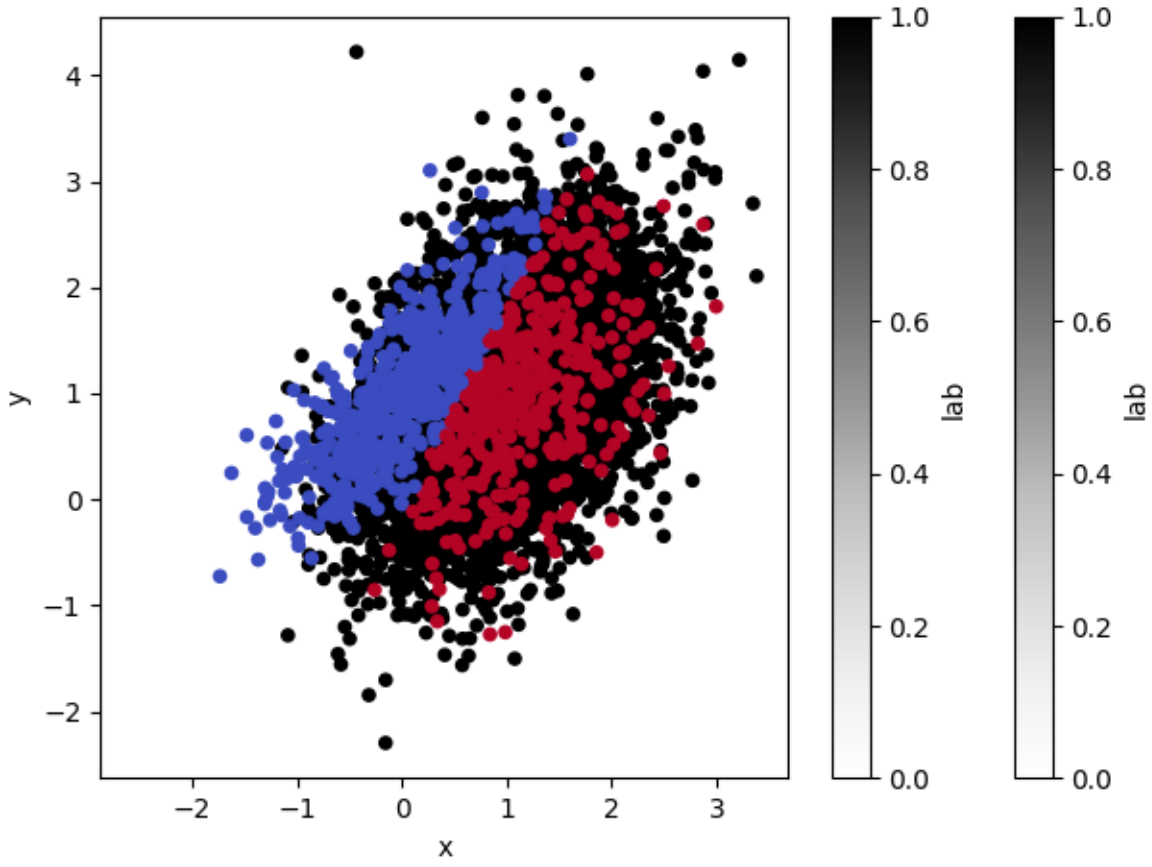


Figure 6: Non-separable dataset.Trained with SGD, error = 0.155

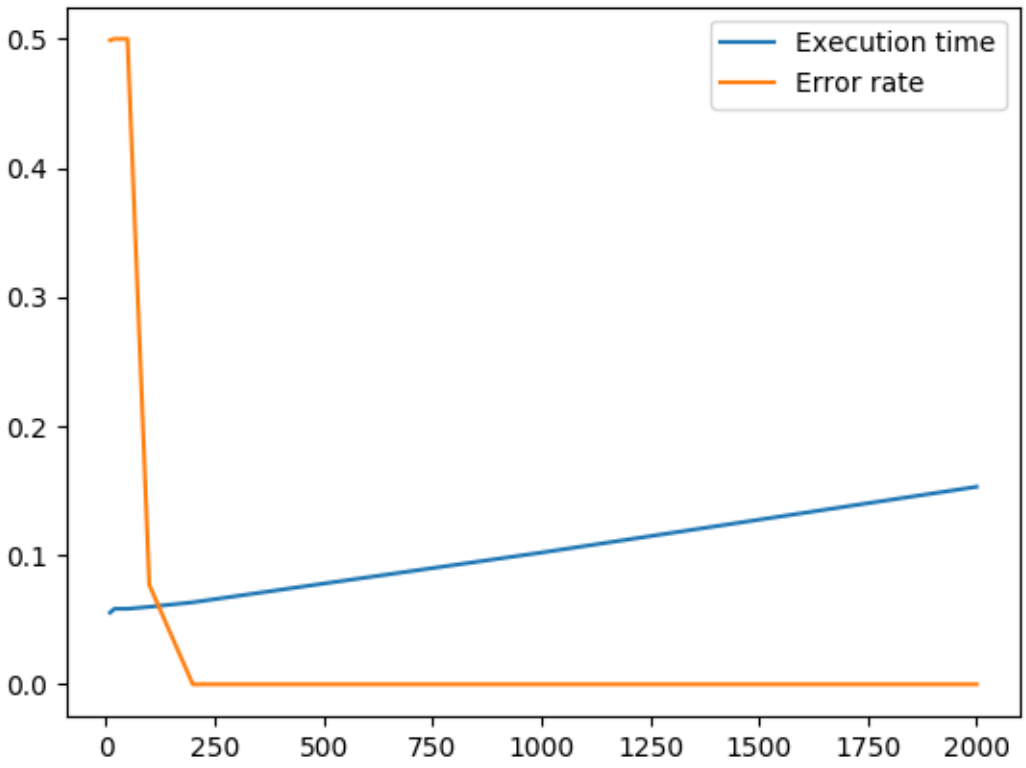


Figure 7: The execution time and error rate for SGD with number of iterations (x), and time (seconds) and error rate on y-axis