1. Comparison of errors 'in-sample' and 'out-sample'

The following tables are represented the in-sample errors of Stochastic and Batch Gradient Descent:

Learning Rate	Classification, Ein	Classification, Eout	Like hood, Ein	Like hood, Eout
0.1	3956	603	1.683869e+01	1.673531e+01
1.0	3961	604	3.558817e+01	3.545315e+01
10	3895	623	4.077497e+02	4.060328e+02
50	3600	800	4.760392e+02	4.698733e+02

Table 1: Stochastic Gradient Descent errors

The learning rate has directly impact on in & out sample errors. As we know, the linear learning algorithm has stability condition 0<learning rate<1. When algorithm satisfies the condition, it converges fast and produces less error rate. The discussed phenomenon we can observe from table 1 and 2. As the learning rate increased, the error rate becomes high,].

Learning Rate	Classification, Ein	Classification, Eout	Like hood, Ein	Like hood, Eout
0.1	3600	800	1.710687e+01	1.674029e+01
1.0	3544	811	2.896754e+01	2.876534e+01
10	3467	883	3.765439e+02	3.754379e+02
50	3600	960	3.960392e+02	3.968733e+02

Table 2: Batch Gradient Descent errors

2. Choice and justification of stopping criteria

There are two ways that we can set our stopping criteria. One is based on gradient. The idea is straightforward. When the slopes become zero, the algorithm will converge and stop. But it never happens in reality. The step size (learning rate) plays an important rules where larger step size may miss the convergent point and algorithm runs forever for convergence. The other option is based on squared error calculation. When the error has not changed in a significant rate, the algorithm stopped. Therefore, we need to set the stopping criteria. In our implementation, we have considered the above discussed criterion.

Stochastic Gradient Descent: The stopping criteria has been set based on gradient. When the gradient is not changing a significant rate, the algorithm stopped. This method helps to algorithm convergence at fast rate.

Batch Gradient Descent: The stopping criteria has been set based on squared error rate. Batch algorithm works slow for large dataset. Also, larger learning rate leads to instability of the algorithm. When the error is not changing a significant rate, the algorithm stopped. This method helps to algorithm convergence at fast rate.

3. Discussion and results of the effect of learning rate

The learning rate has affected the algorithm convergence time and error rate. Because learning rate or step size requires to determine slope of the function. If it is larger, the chance is higher to become overfitting and algorithm does not reach convergence point. Also, the stability condition: 0<learning rate<1 must satisfies for the fastest convergence and less error.
For the single point divergent algorithm, it takes larger time to converge when learning rate >1 so that the stopping criterion must be chosen in accordingly.

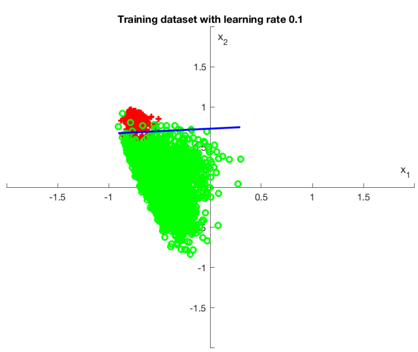

Fig. 1: Training dataset with learning rate 0.1

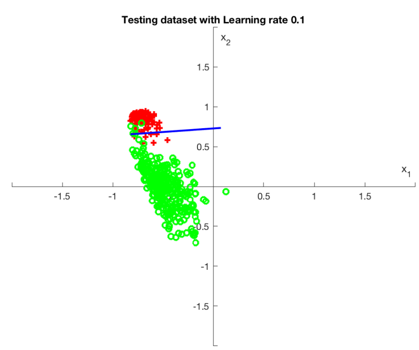

Fig. 2: Testing dataset with learning rate 0.1

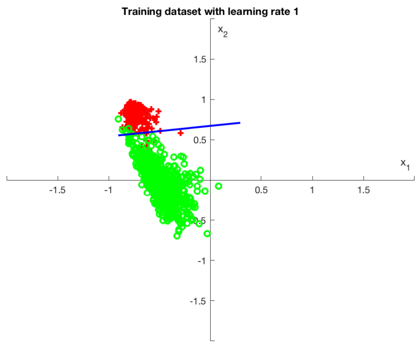

Fig. 3: Training dataset with learning rate 1

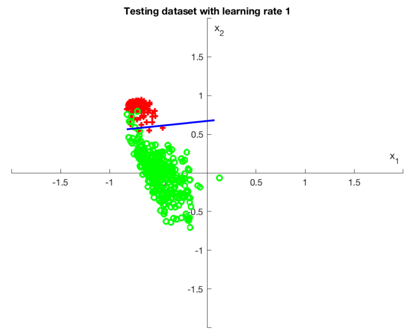

Fig. 4: Testing dataset with learning rate 1

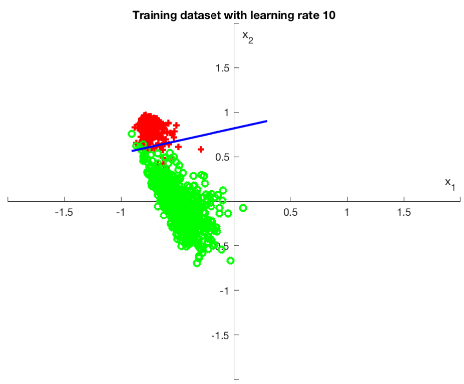

Fig. 5: Training dataset with learning rate 10

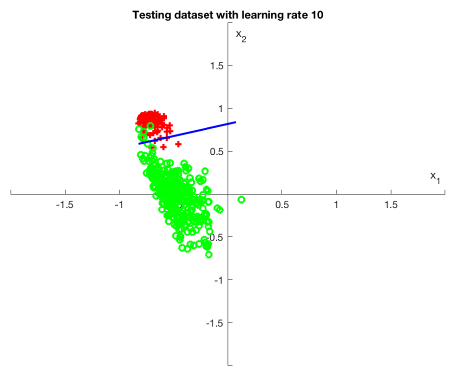

Fig. 6: Testing dataset with learning rate 10

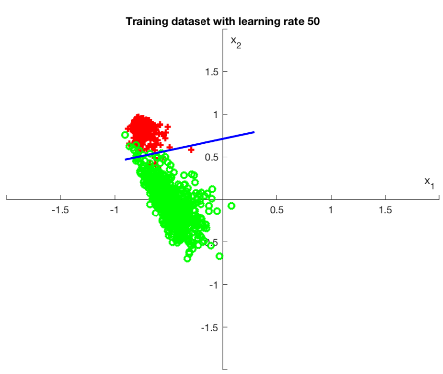

Fig. 5: Training dataset with learning rate 50

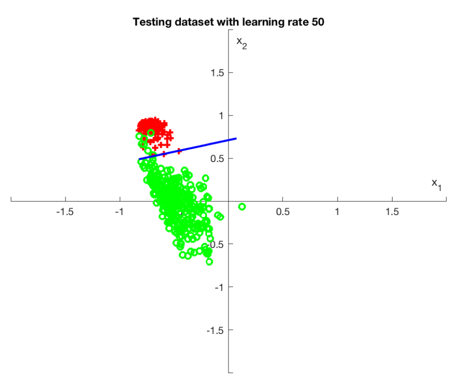

Fig. 5: Testing dataset with learning rate