**Machine Learning**

**HW2**

**Yun Li 6035800416**

Problem 1 

Solution 1.1a



Solution 1.1b

Maximize over



Take derivative：



Solution 1.2a



Solution 1.2b



Solution 1.3a





Because we have , that is:



Suppose  is reversible, we get



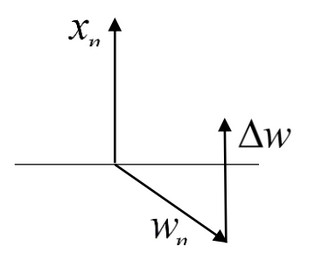
Thus, we have



Problem 2 

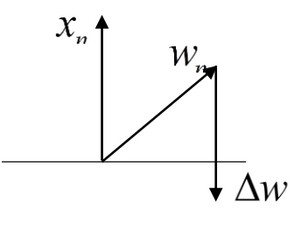
We need to update only when the previous model does not classify correctly on the current sample. That is  . Suppose the angle between and  is . We have the following claims:

Case1: 



In this case, we need to rotate ( suppose a bit more than ), so that , and , where, 

Case2: 



In this case, we need to rotate ( suppose a bit more than), so that , and , where, 

In both cases, we have this formula:



Thus, when , we set the updating rules to:



Problem 3 

Soluton 3.a

Since K1 and K2 are positive semi-definite, that is:



Because

Then we have:

.

Thus K3 is positive semi-definite.

Solution 3.b



is a symmetric matrix, which could be written in the following forms:





Thus K4 is positive semi-definite.

Solution 3.c

Let x1,…,xn be an orthonormal basis of eigenvectors for K2 with eigenvalues λ1,…,λn≥0. Then



Thus K5 is positive semi-definite.

Also ref: <https://en.wikipedia.org/wiki/Schur_product_theorem>

Problem 4 

Solution 4.a



Solution 4.b



Solution 4.c

Note that, not considering the biased term, we have 



Solution 4.d

When  is small (tending to zero), the bias term would tend to zero, while the variance term would tend to be constant, .

When  is large, the bias term would tend to be very larger, while the variance term would tend to be smaller.

Problem 5 

5.2 Feature Representation

Enron: 600 of occurrence, Will: 351 of occurrence, Please: 291 of occurrence

5.3 Batch Gradient Descent

(2) Updating Equations

a. For unregularized logistic regression:

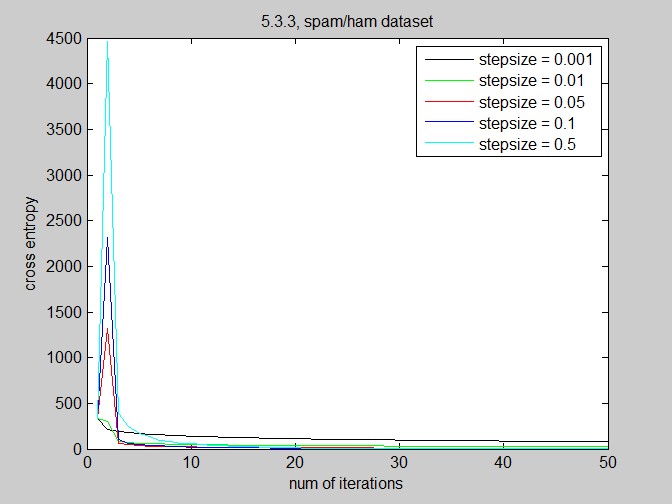
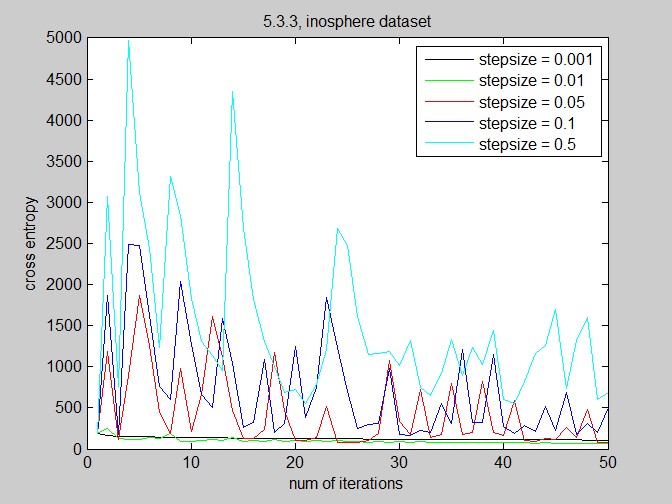


b. For regularized logistic regression:



(3)Results for different stepsizes without regularization

a. cross entropy value

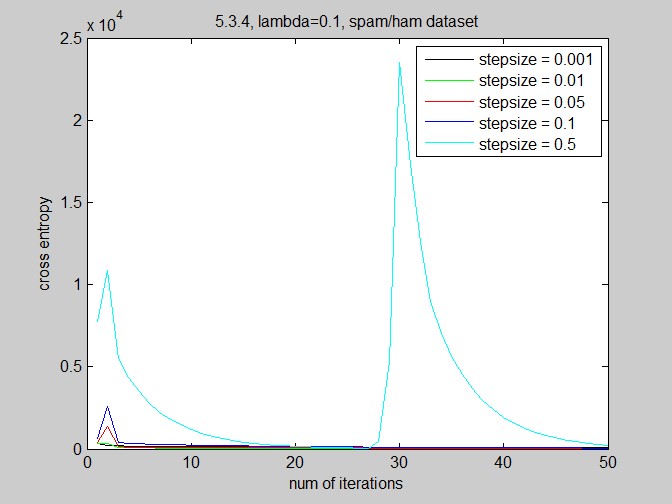
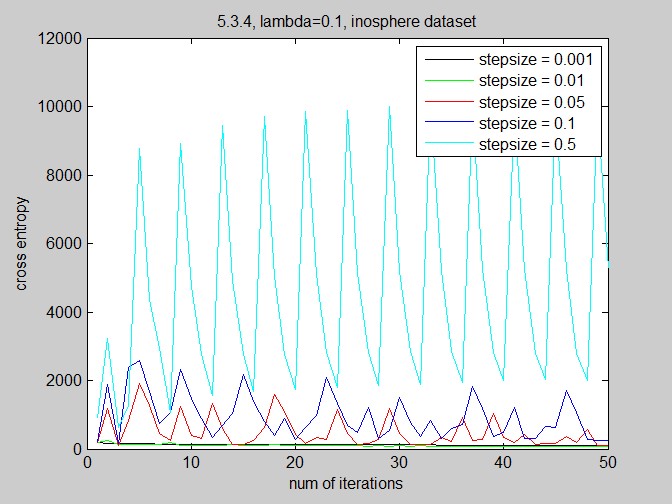


b. L2 norm

|  |  |  |
| --- | --- | --- |
| stepsize | L2 Norm of w | |
| Ionosphere | EmailSpam |
| 0.001 | 1.49462031809725 | 2.60178597372220 |
| 0.01 | 4.65527606736429 | 8.00474552175571 |
| 0.05 | 18.6195776224861 | 28.5167563308863 |
| 0.1 | 37.2185798331789 | 55.6326122952729 |
| 0.5 | 190.270625301818 | 275.759304647348 |

(4) Results for different stepsizes with regularization

a. lambda= 0.1, plot cross-entropy function value

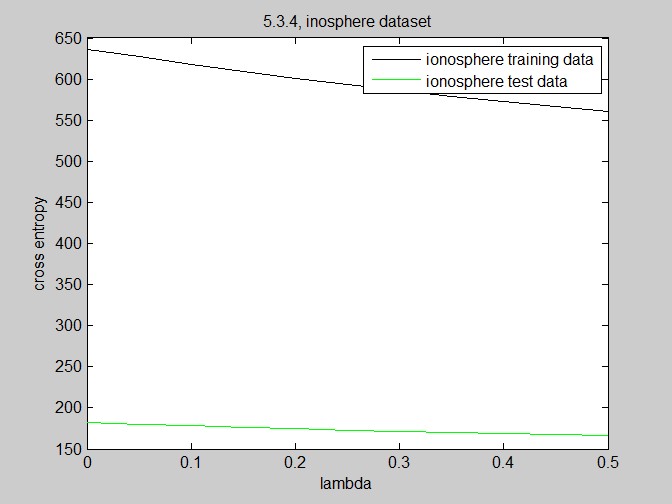
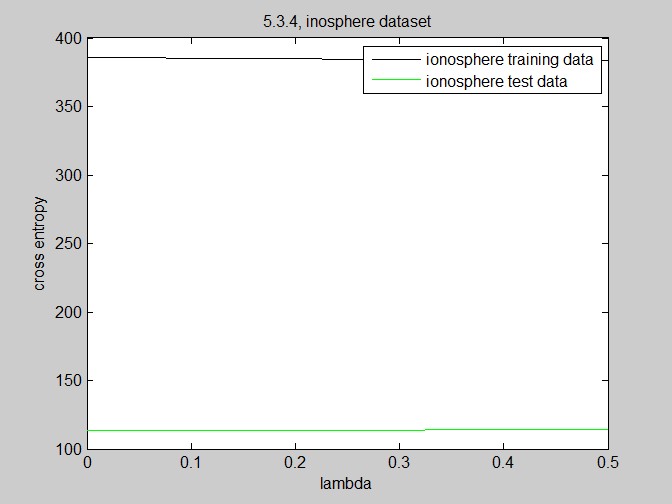


b. report the L2 norm of vector w after 50 iterations

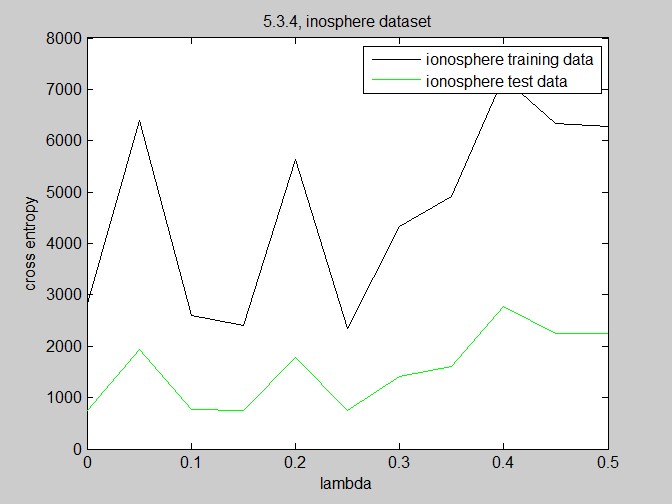
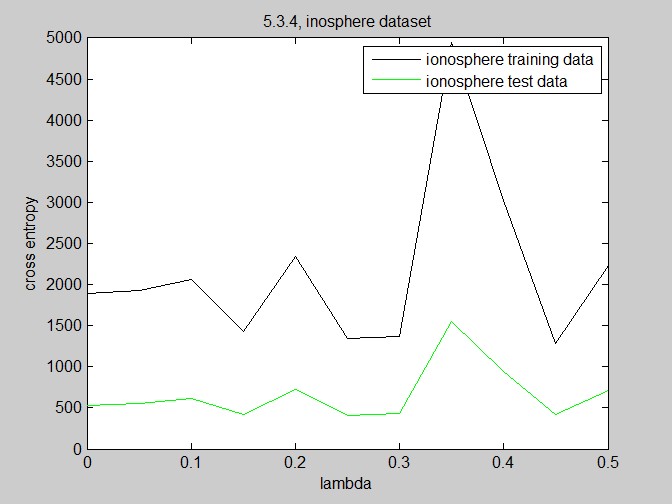
|  |  |  |
| --- | --- | --- |
| lambda | L2 Norm of w | |
| Ionosphere | EmailSpam |
| 0 | 4.65527606736429 | 8.00474552175571 |
| 0.05 | 4.57540848783625 | 7.75374929063780 |
| 0.1 | 4.49881052442936 | 7.51626905081810 |
| 0.15 | 4.42531694902240 | 7.29163819978431 |
| 0.20 | 4.35475034538736 | 7.07921400707528 |
| 0.25 | 4.28692034030427 | 6.87837691896061 |
| 0.30 | 4.22161799969738 | 6.68852989828170 |
| 0.35 | 4.15860954095679 | 6.50909780846072 |
| 0.40 | 4.09767287694895 | 6.33952685318458 |
| 0.45 | 4.03867909966482 | 6.17928408081901 |
| 0.5 | 3.98159242553231 | 6.02785695580808 |

c. cross entropy function value, for both training and test data

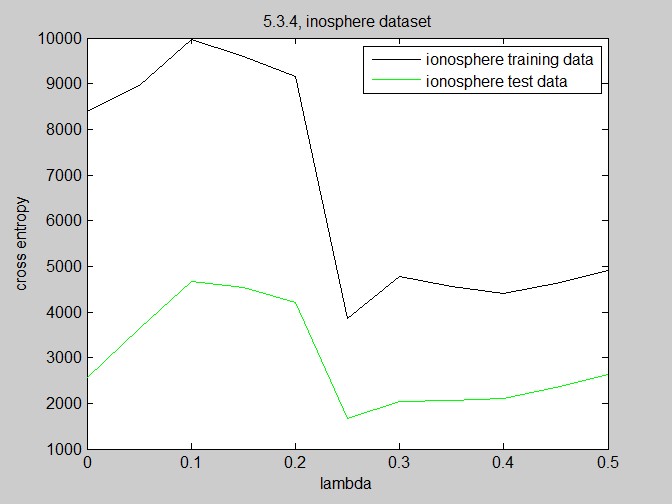
1. ionosphere dataset



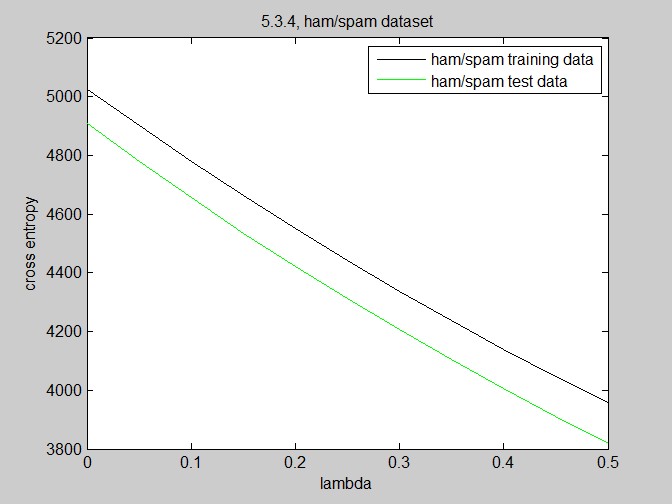
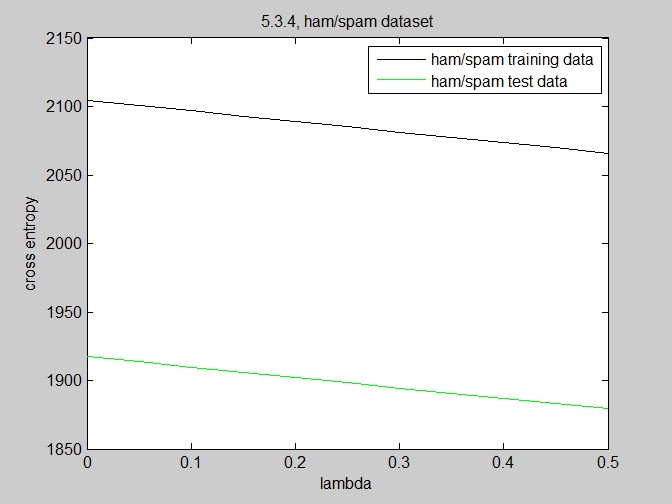


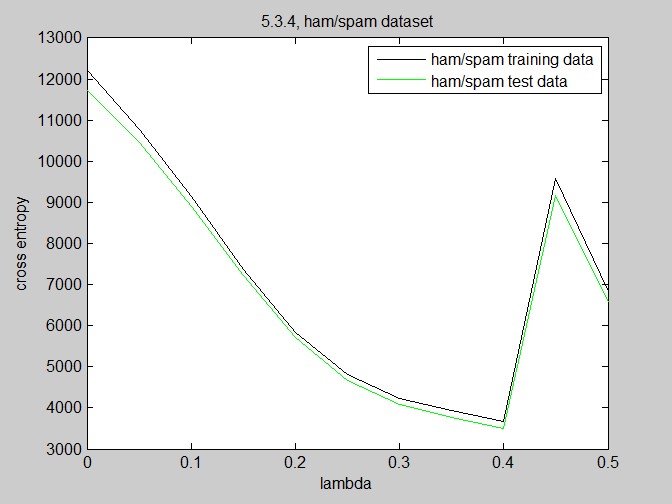
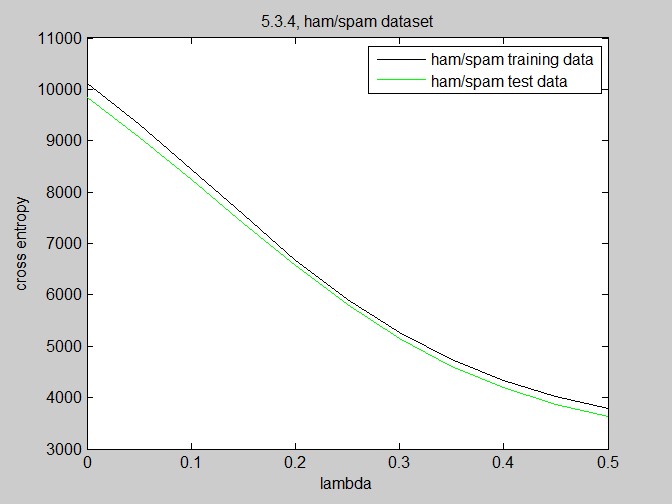




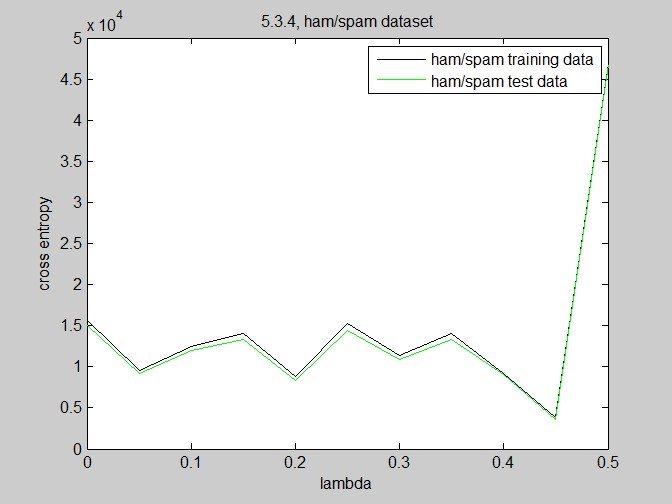
2. spam/ham dataset









Newton’s Method

(5) Updating Equations

a. For unregularized logistic regression



where 



b. regularized logistic regression



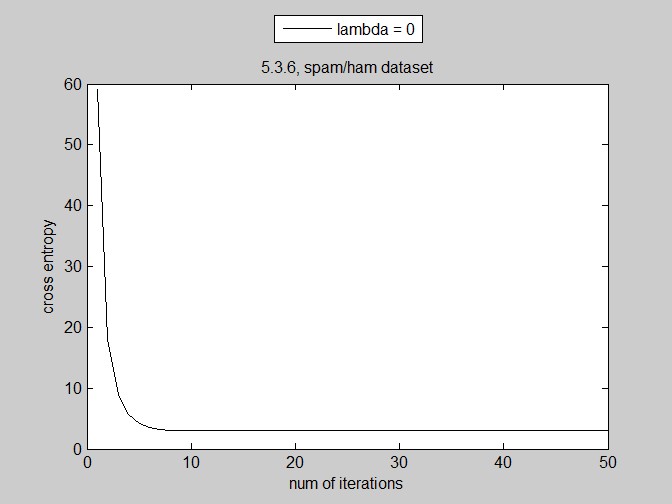
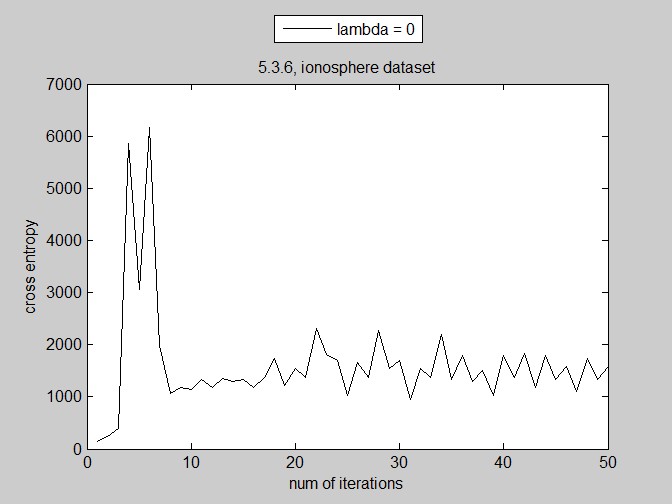
where 



Where 

(6) Run the Newton's method on both of datasets without regularization

a. Plot cross-entropy function value as the number of steps T = [1,...,50] (make two plots)



b. Report L2 norm of vector w after 50 iteration.

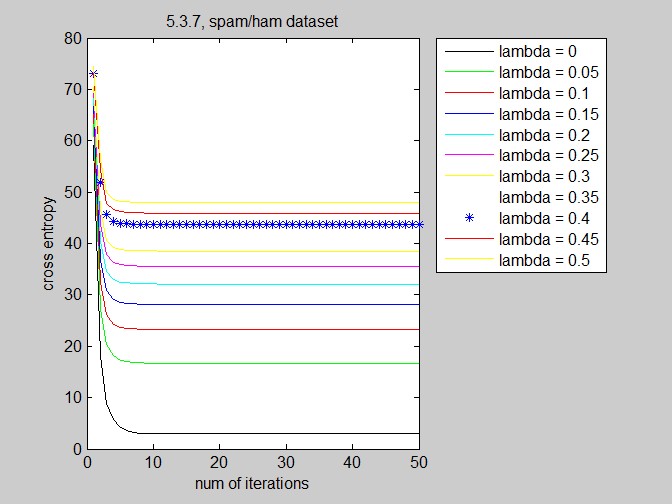
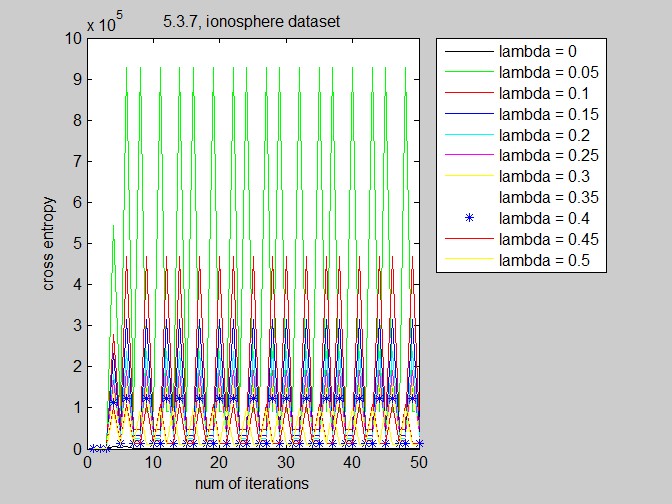
|  |  |  |
| --- | --- | --- |
| lambda | L2 Norm of w | |
| Ionosphere | EmailSpam |
| 0 | 1.06283926239763e+16 | 640.239642816382 |

c. Report cross-entropy function value for the test data.

|  |  |  |
| --- | --- | --- |
| lambda | Cross-entropy value | |
| Ionosphere | EmailSpam |
| 0 | 2540.17184613733 | 11681.0730646370 |

(7) Repeat the problem (6) with regularized case.

a. Plot cross-entropy function value as the number of steps T = [1,...,50] (make two plots)



b. Report L2 norm of vector w after 50 iterations

|  |  |  |
| --- | --- | --- |
| lambda | L2 Norm of w | |
| Ionosphere | EmailSpam |
| 0 | 1.06283926239763e+16 | 640.239642816382 |
| 0.05 | 1315.64547045816 | 12.6555893635315 |
| 0.1 | 657.822735226999 | 10.5102787234012 |
| 0.15 | 438.548490151793 | 9.34381711648633 |
| 0.20 | 328.911367614021 | 8.56110780185517 |
| 0.25 | 263.129094091300 | 7.98136860009019 |
| 0.30 | 219.274245076129 | 7.52617375861859 |
| 0.35 | 187.949352922425 | 7.15462907446191 |
| 0.40 | 164.455683807140 | 6.84281405927605 |
| 0.45 | 146.182830050907 | 6.57558260785219 |
| 0.5 | 131.564547045816 | 6.34276969604776 |

c. Report cross-entropy function value for the test data.

|  |  |  |
| --- | --- | --- |
| lambda | Cross Entropy value | |
| Ionosphere | EmailSpam |
| 0 | 478.937699342762 | 3147.28210815344 |
| 0.05 | 88419.7270259075 | 113.832371764247 |
| 0.1 | 44415.1573042781 | 109.166392708338 |
| 0.15 | 29990.7989382876 | 107.785652608200 |
| 0.20 | 22778.6197552701 | 107.461252728636 |
| 0.25 | 18451.3122454520 | 107.600698071689 |
| 0.30 | 15566.4405722369 | 107.975999932512 |
| 0.35 | 13505.8179485104 | 108.481444071480 |
| 0.40 | 11960.3509807147 | 109.061896901703 |
| 0.45 | 11489.8157398151 | 109.685952808035 |
| 0.5 | 10528.1918487389 | 110.334441303831 |

(8) Briefly explain the results you got from (3), (4)

After introducing the regularization, the convergence slows down and the cross entropy becomes more unstable.

Given the stepsize’s impact, the bigger the stepsize is, the quicker the convergence will achieve, and the bigger the magnitude of cross entropy the result would be.

But, when stepsize is too big, it may not converge.

(9)Comparison between Gradient Descent and Newton's method

Newton’s method converges more faster than gradient descent, but sometimes Newton’s method fails to converge (depends on the initialized w).