CS 260 : Machine Learning Algorithm
Homework #3

Ronak Sumbaly
UID - 604591897

October 29, 2015

Question 1

The 3 most frequent words in the training dataset of EmailSpam

$$
\begin{array}{ll}
\textbf{enron} & [600] \\
\textbf{will} & [351] \\
\textbf{please} & [291]
\end{array}
$$

Question 2

Updating equation for **w** and **b** in Unregularized Gradient Descent

$$w_{t+1} = w_t - \eta \sum_1^n (\sigma(w^T x_i + b) - y_i) x_i$$

$$b_{t+1} = b_t - \eta \sum_1^n (\sigma(w^T x_i + b) - y_i)$$

Updating equation for **w** and **b** in Regularized gradient descent

$$w_{t+1} = w_t - \eta \left[\sum_{i=1}^n \{\sigma(b + w^T x_i) - y_i\} x_i + \sum_d 2.\lambda.w_d \right]$$

$$b_{t+1} = b_t - \eta \left[\sum_{i=1}^n \{\sigma(b + w^T x_i) - y_i\} \right]$$

Question 3

Question 3.a

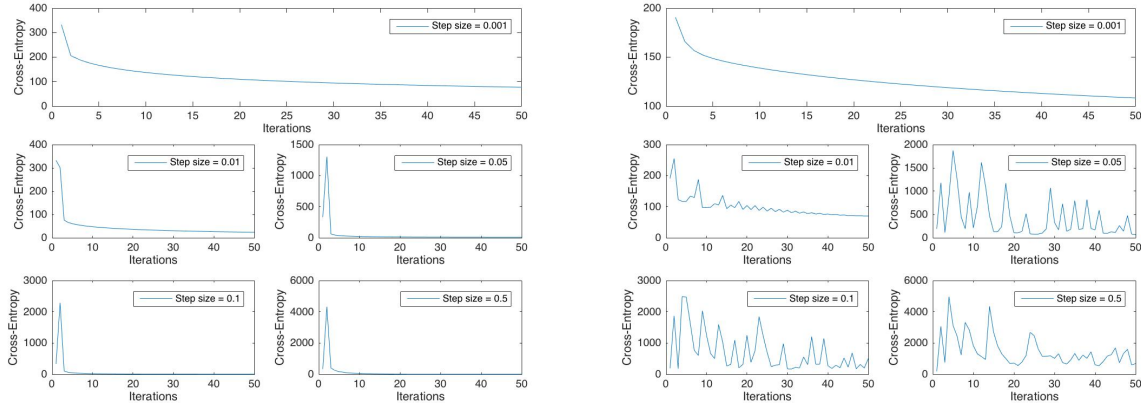Gradient Descent Without Regularization

Figure 1: Cross Entropy : Ionosphere(Left), EmailSpam(Right)

Question 3.b

L_2 Norm of vector w

$L_2\mathrm{Norm}(\lambda = 0)$	0.001	0.01	0.05	0.1	0.5
Ionosphere	1.4946	4.6553	18.6196	37.2186	190.2706
EmailSpam	2.5933	7.9696	28.4815	55.6097	275.7817

Question 4

Question 4.a

Gradient Descent With Regularization

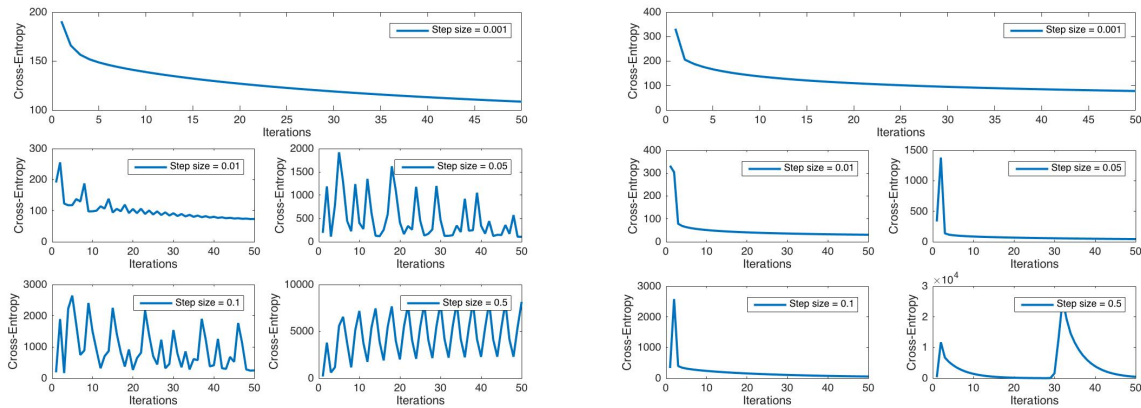
Figure 2: Cross Entropy : Ionosphere(Left), EmailSpam(Right)

Question 4.b

L_2 norm of vector $\mathbf{w}$ for each regularization coefficient.

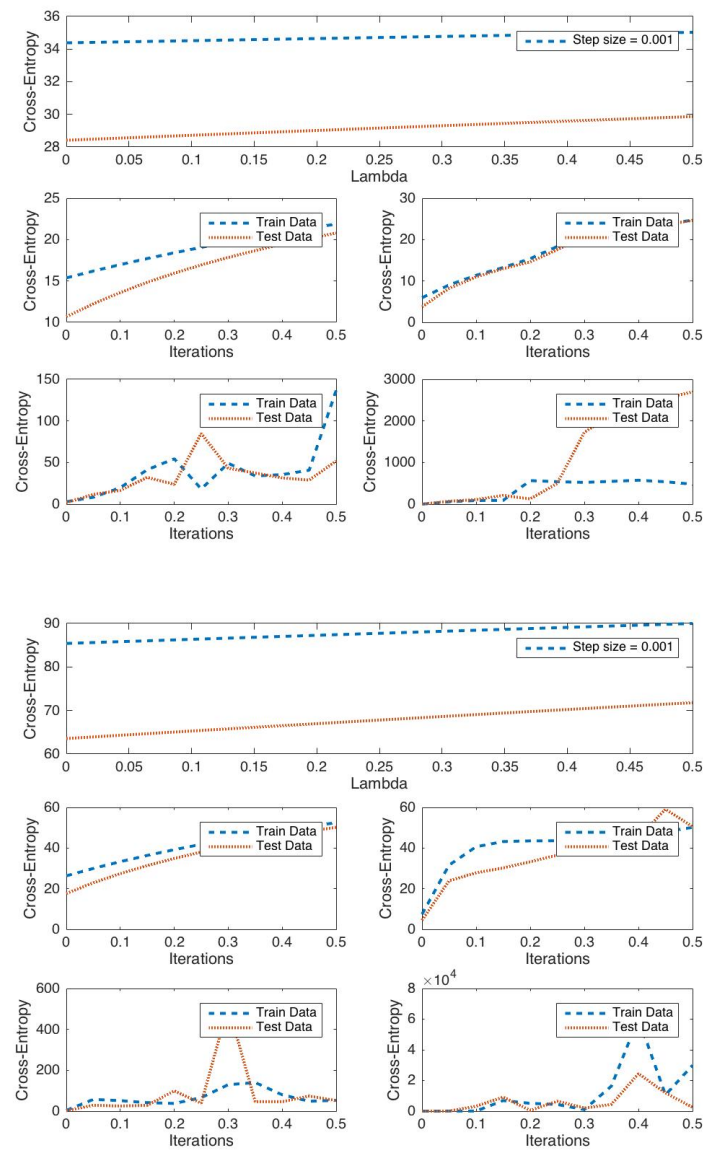| L_2 Norm	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
Ionosphere	4.6553	4.5754	4.4988	4.4253	4.3548	4.2869	4.2216	4.1586	4.0977	4.0387	3.9816
EmailSpam	7.9696	7.7183	7.4806	7.2558	7.0432	6.8422	6.6522	6.4726	6.3029	6.1426	5.9911

Question 4.c

Gradient Descent - Training Testing Data

Figure 3: Cross Entropy : Ionosphere(Top), EmailSpam(Bottom)

Question 5

Updating equation for **w** and **b** in unregularized Newton Method

$$w_{t+1} = w_t - \{(H)^{-1} \bigtriangledown \varepsilon(w^t)\}$$

where

$$H = \frac{\partial \varepsilon}{\partial w w^T} = \sum_{i=1}^{n} \sigma(b + w^T x_i).(1 - \sigma(b + w^T x_i)).x_i.x_i^T$$

$$\bigtriangledown \varepsilon(w^t) = \sum_{i=1}^{n} \{\sigma(b + w^T x_i) - y_i\} x_i$$

and

$$b_{t+1} = b_t - \{(H)^{-1} \bigtriangledown \varepsilon(b_t)\}$$

where

$$H = \frac{\partial \varepsilon}{\partial^2 b} = \sum_{i=1}^{n} \sigma(b + w^T x_i).(1 - \sigma(b + w^T x_i))$$

$$\bigtriangledown \varepsilon(b_t) = \sum_{i=1}^{n} \{\sigma(b + w^T x_i) - y_i\}$$

Updating equation for **w** and **b** in regularized Newton Method

$$w_{t+1} = w_t - \{(H)^{-1} \bigtriangledown \varepsilon(w^t)\}$$

where

$$H = \frac{\partial \varepsilon}{\partial w w^T} = \sum_{i=1}^{n} \sigma(b + w^T x_i).(1 - \sigma(b + w^T x_i)).x_i.x_i^T + 2.\lambda.I_d$$

$$\bigtriangledown \varepsilon(w^t) = \sum_{i=1}^{n} \{\sigma(b + w^T x_i) - y_i\} x_i + \sum_{d} 2.\lambda.w_d$$

and

$$b_{t+1} = b_t - \{(H)^{-1} \bigtriangledown \varepsilon(b_t)\}$$

where

$$H = \frac{\partial \varepsilon}{\partial^2 b} = \sum_{i=1}^{n} \sigma(b + w^T x_i).(1 - \sigma(b + w^T x_i))$$

$$\bigtriangledown \varepsilon(b_t) = \sum_{i=1}^{n} \{\sigma(b + w^T x_i) - y_i\}$$

Question 6

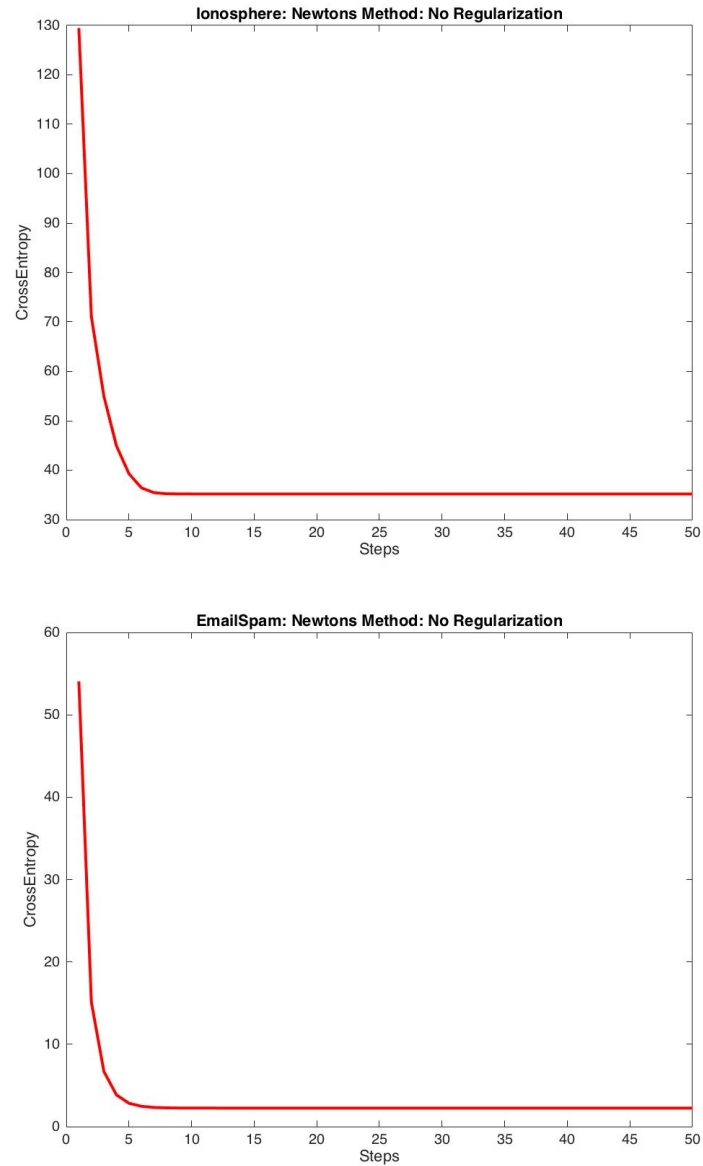## Question 6.a

Newton Method - No Regularization

Figure 4: Cross Entropy : Ionosphere(Top), EmailSpam(Bottom)

Question 6.b

Newton Method - L2 Norm Value

$$Ionosphere = 94.3533$$
$$EmailSpam = 296.5734$$

Question 6.c

Newton Method - Cross Entropy

$$Ionosphere = 9.1038e - 15$$
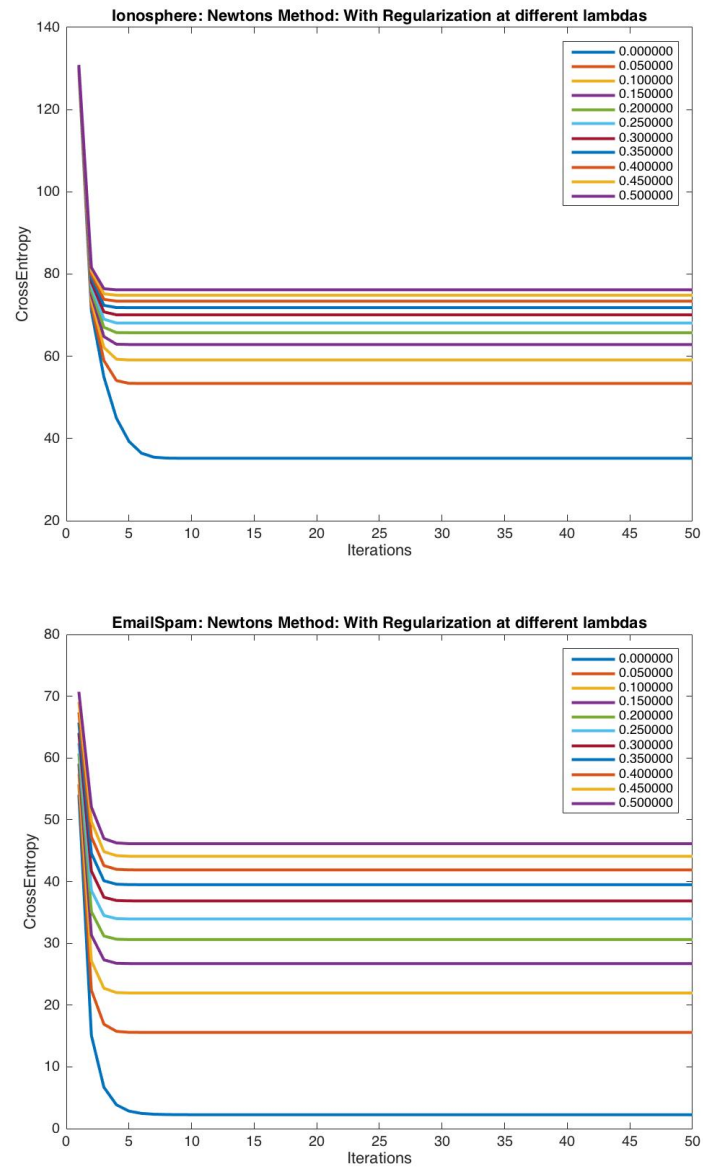$$EmailSpam = 5.3069e - 14$$

Question 7

Question 7.a

Figure 5: Cross Entropy : Ionosphere(Top), EmailSpam(Bottom)

Question 7.b

L_2 Norm of $\mathbf{w}$ - Regularization

L_2 Norm	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
Ionosphere	94.353	16.006	12.228	10.454	9.3622	8.6009	8.0294	7.5786	7.2101	6.9008	6.6257
EmailSpam	296.57	12.507	10.399	9.2522	8.4821	7.9112	7.4625	7.0958	6.7879	6.5238	6.2936

Question 7.c

Cross Entropy Newton Method - Regularization

L_2 CrossEntropy	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
Ionosphere	35.186	53.396	59.103	62.862	65.73	68.078	70.083	71.842	73.416	74.845	76.157
EmailSpam	2.2493	15.58	21.961	26.717	30.606	33.935	36.865	39.495	41.888	44.089	46.13

Question 8

For Batch Gradient Descent it can be seen from the graphs that at the same level of regularization, convergence is negatively related to step size η, while at the same level of step size, convergence is negatively proportional to the level of regularization, fluctuating alot with change in value.

Question 9

In generality Gradient Descent is much better than Newton's Method but if the initial weights of the latter are chosen proper then we can see that it rapidly converges. It was also observed that Newton's Method takes a much longer time to converge than Gradient Descent for the same level of regularization.