Michael Lee

CSE 446

HW2

1 a) False, is not typically sparse. This is because by taking the 2-norm of the vector we don’t select sparse solutions optimally. For example:

If we calculate the L2 norm of the vectors:

Versus if we calculate the L1 norm:

In the L2 case we consider where as in the L1 case we consider

, hence we don’t optimize for solutions of sparsity.

b) True, because as the weights increase with lambda = 0 our algorithm will optimize for the reduction in training error. And be increasing the values of the weights the training will subsequently decrease.

c) False, the increase in lambda will cause the coefficients of to decrease thereby increasing the error on the training set. Hence, this will cause the likelihood to decrease.

d) At lambda=0 there will be over fitting and high error as a result. Then as lambda increases the test error will decrease to due to an extent, but the error will then increase eventually with high lambda due to under fitting.

e)

i)

ii) The relationship between is that are accounting for the same feature. Since are dealing with the same feature the algorithm will optimize such that these two weights are reduced equally. In comparison to the first data set would have a value larger value as it wouldn’t have its data point duplicated.

iii) My answer would not change because the effect would be the same, when we optimize for the L2 norm we still have the same effect with the coefficients.

2) Boosting

a)

b) show:

We start by rewriting the form of

-- identifying the pattern

--using the pattern

\*Note: If we multiply by and we arrive at the definition of .

Hence,

c)

i)

= 0

Now we show

ii)

Note:

Since were given

<=

<=

<=

<=

<=

iii)

Since we’re told that

Further then:

Therefore:

Finally, we conclude:

3.

Root

a)

Age?

<=32

>3232

Income?

>=65,00032

Income?

<65,000

>25,00032

<=25,00032

Yes-2

No-3

Yes-4

No- 1

Depth = 2 and error = 0

B) We choose , this will provide a perfect classification of the data, depth = 1.

C)

Disadvantages:

1) The model complexity increases with the number of parameters that are being chosen.

2) Choosing the right values can be computationally expensive since there exists an infinite amount of possible solutions.

Advantages:

1) More specificity at each split of the data, in our example we see that the tree was able to better classified at each step.

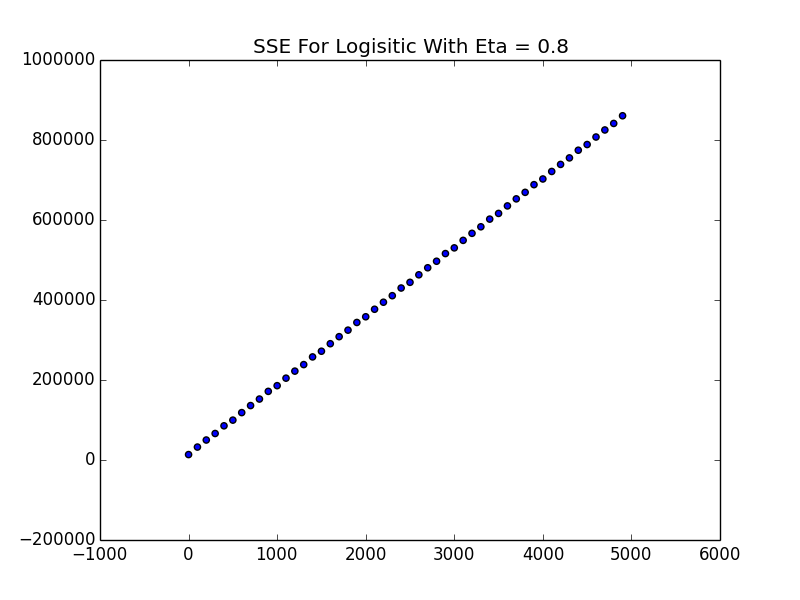
2) It can lead to shorter classification trees and reduce classification time as a result.

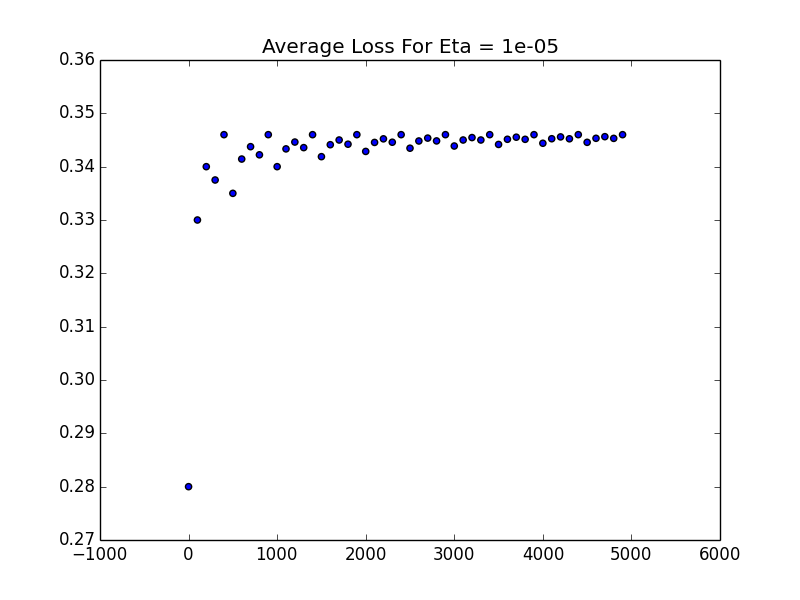
4.2

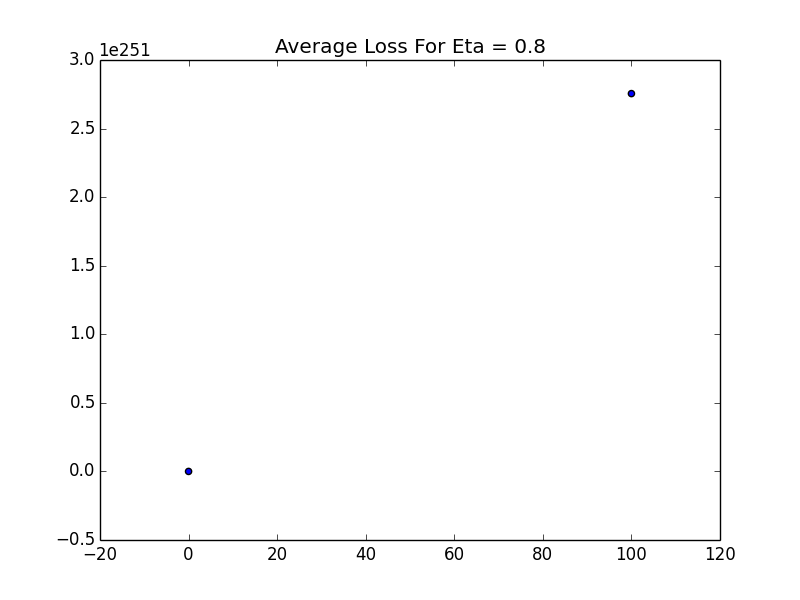
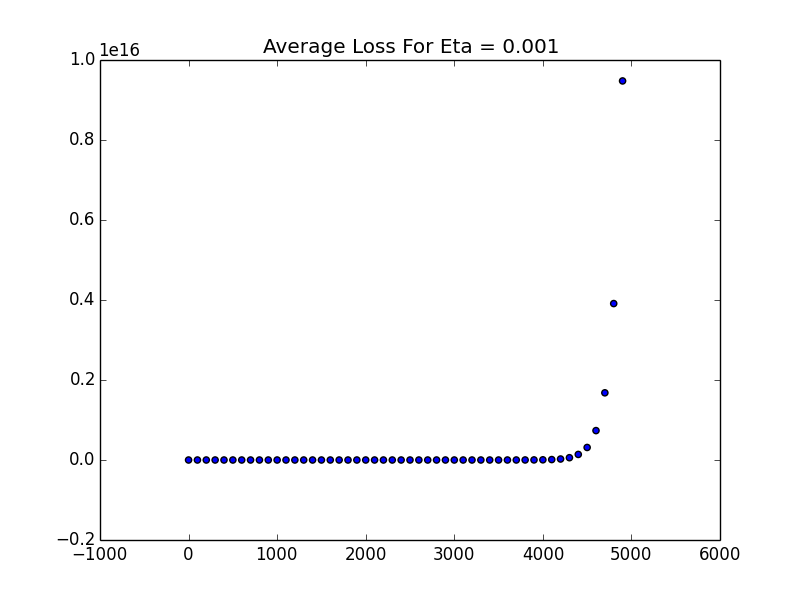
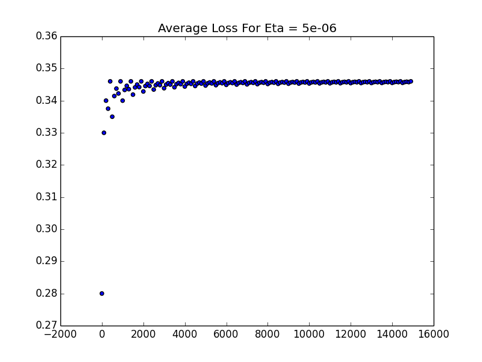
a) – logistic regression.

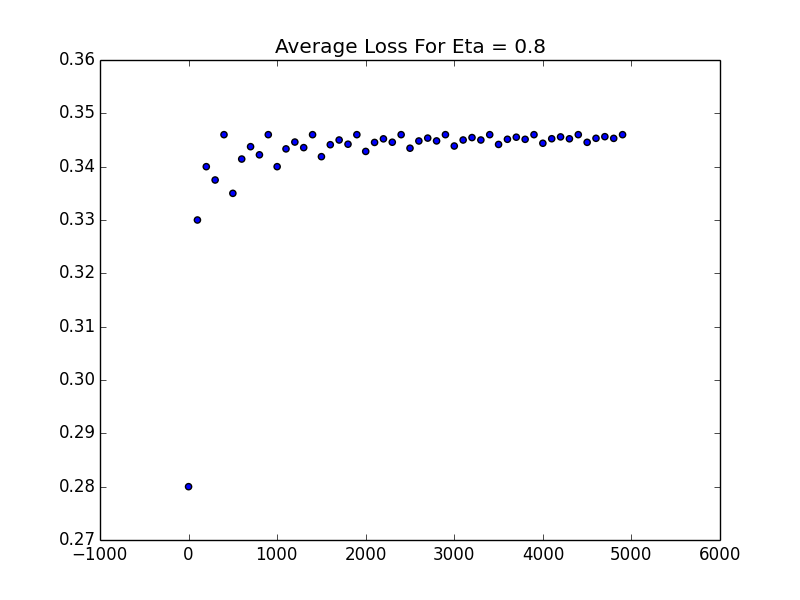
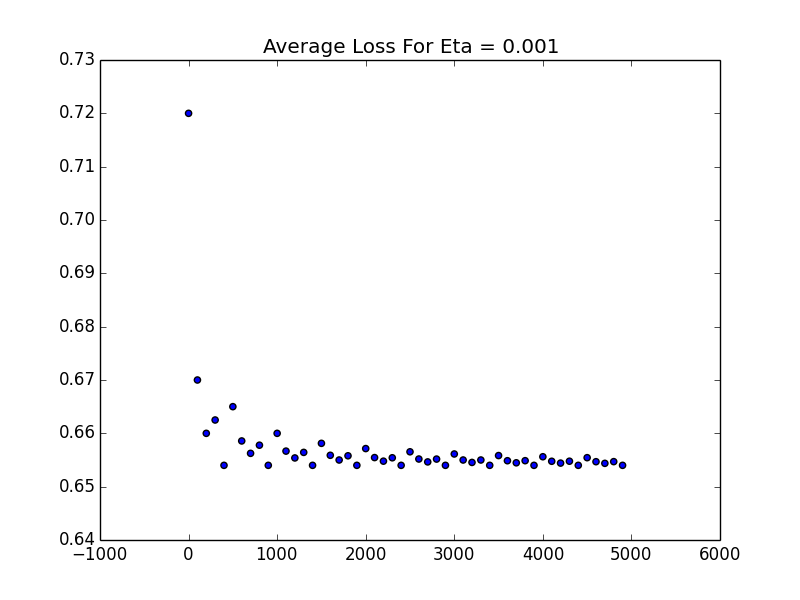
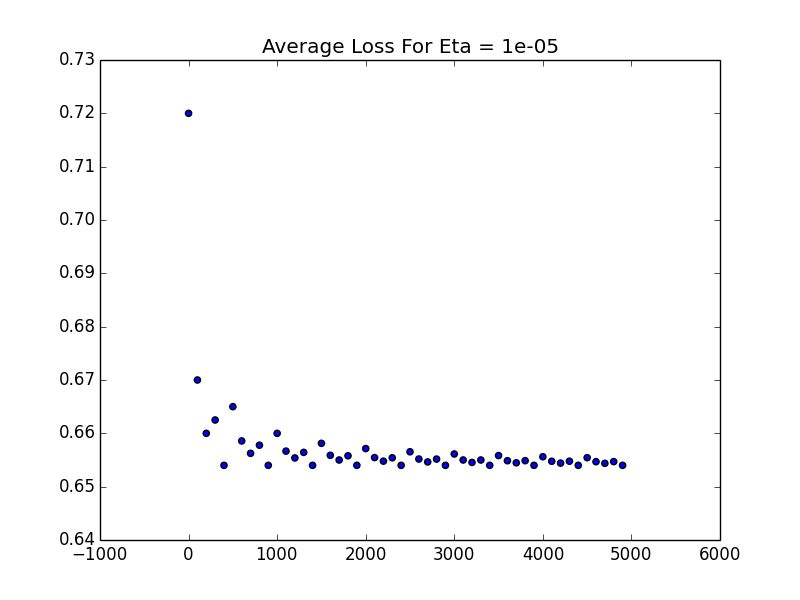
– linear regression

b) I will start by a couple of notes on the calculations and graphs below. I used the original formulas given in the homework prior to the email of the revised homework. So I did not use any permutations of my data or averages of w in calculating my guess for the outcome. One thing to note, when finding average loss I used the threshold heuristic that was presented to us.

i) Note: Below I recorded a version of the average loss for a total of 30 passes with linear model. Also, “eta” means step size.

****

**Linear Average Loss Plots**

**Logistic** Average Loss Plots

ii)

**Linear** **L2 Norm** After Each Pass For Increasing Step Size

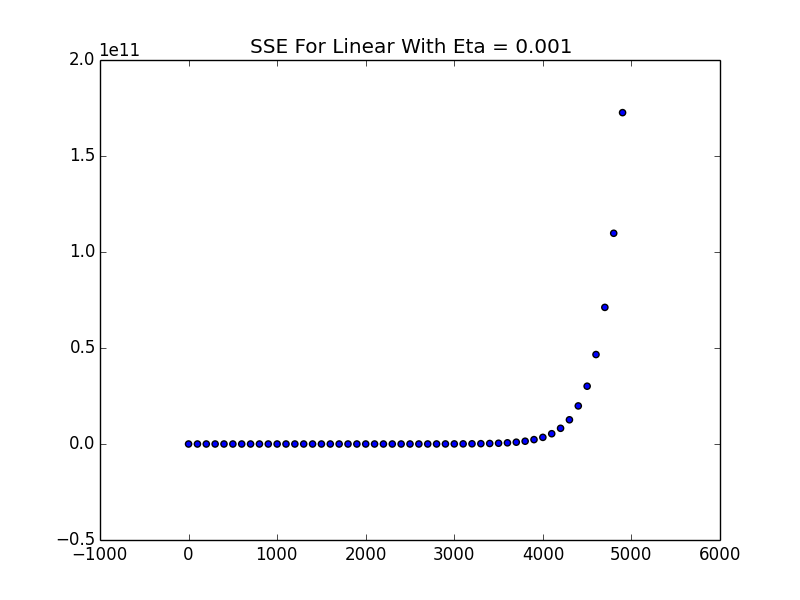
|  |  |  |  |
| --- | --- | --- | --- |
| Iterations/Step Size | .8 | .001 | .00001 |
| 1 | nan | 1.27374355206 | 0.00542761491925 |
| 2 | nan | 12.9517303027 | 0.0108552298385 |
| 3 | nan | 114.461674407 | 0.0162828447577 |
| 4 | nan | 998.072683797 | 0.021710459677 |
| 5 | nan | 8687.78609077 | 0.0271380745962 |
| 6 | nan | 75608.2521198 | 0.0325656895155 |
| 7 | nan | 657990.336701 | 0.0379933044347 |
| 8 | nan | 5726229.57008 | 0.043420919354 |
| 9 | nan | 49833096.0552 | 0.0488485342732 |
| 10 | nan | 433677579.771 | 0.0542761491925 |

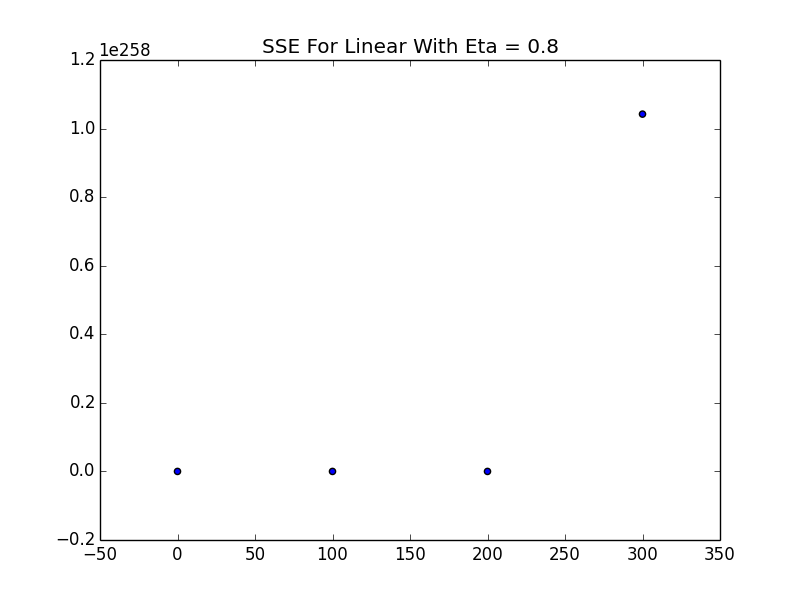
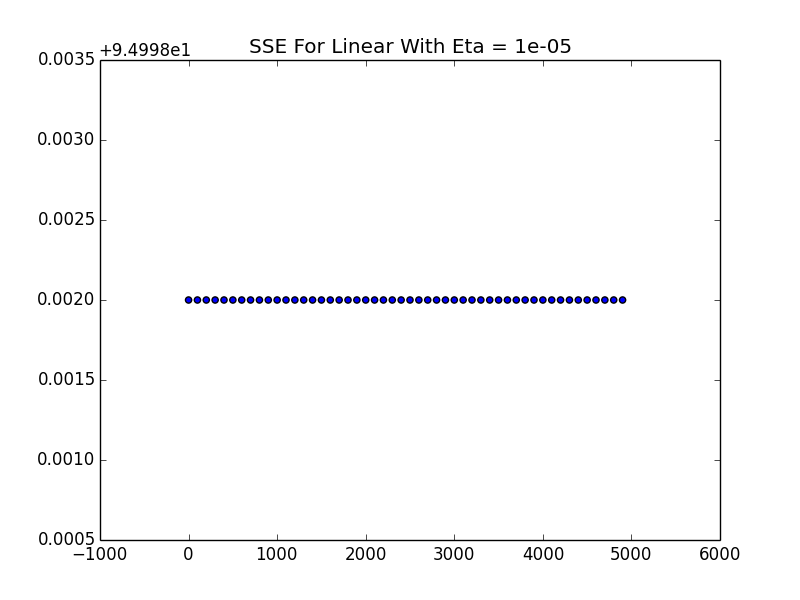
**Logistic** **L2 Norm** After Each Pass For Increasing Step Size

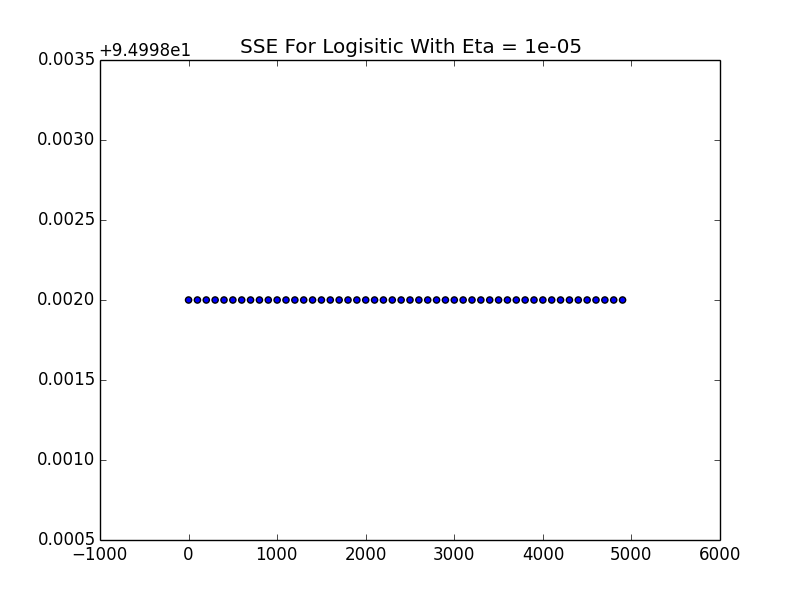
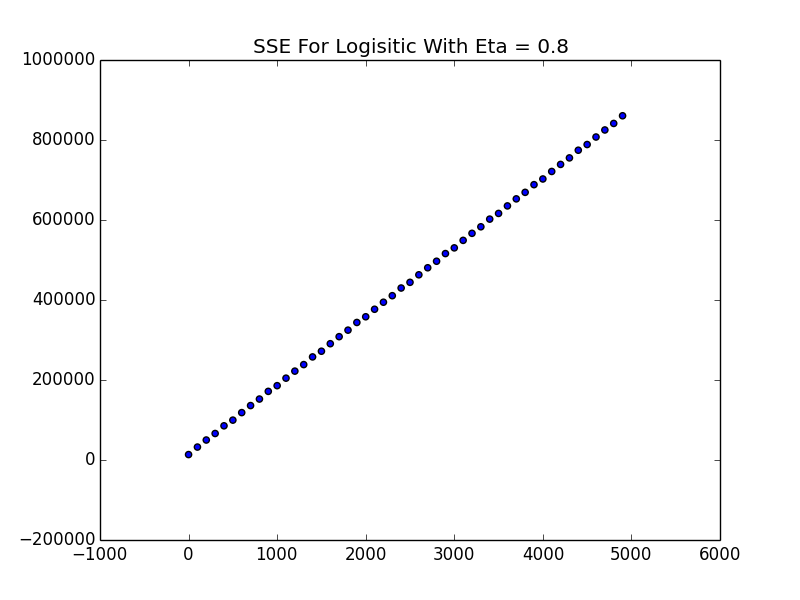
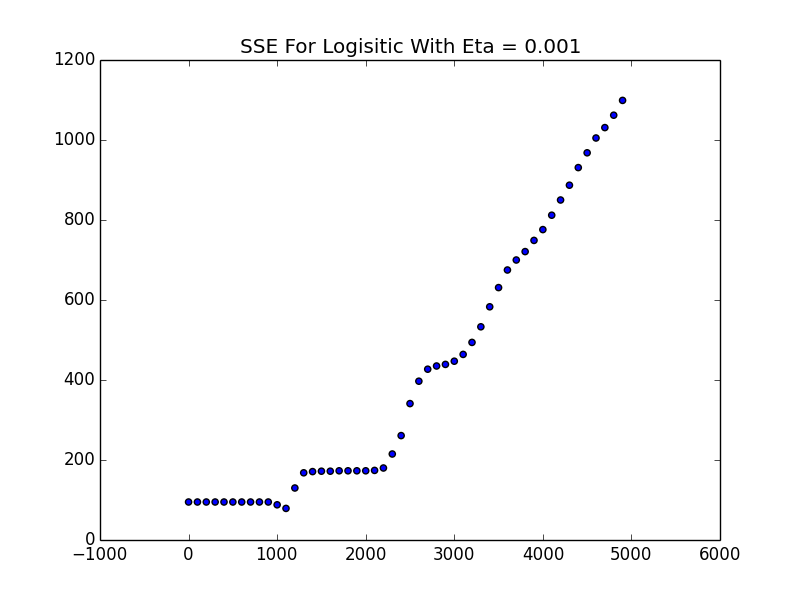
|  |  |  |  |
| --- | --- | --- | --- |
| Iterations/Step Size | .8 | .001 | .00001 |
| 1 | 215.186779099 | 0.112949614948 | 0.000980499183092 |
| 2 | 432.291276989 | 0.259291425644 | 0.00196351035673 |
| 3 | 649.395840994 | 0.447987750878 | 0.00294904061318 |
| 4 | 866.500421418 | 0.687328389345 | 0.0039370970612 |
| 5 | 1083.60500839 | 0.981674200981 | 0.00492768682601 |
| 6 | 1300.70959864 | 1.32827585624 | 0.00592081704931 |
| 7 | 1517.81419075 | 1.71739973604 | 0.00691649488926 |
| 8 | 1734.91878402 | 2.13639258252 | 0.00791472752052 |
| 9 | 1952.02337808 | 2.57407411706 | 0.0089155221342 |
| 10 | 2169.12797268 | 3.02258004237 | 0.00991888593789 |

iii)

**Linear** SSE Plots For Different Values of Eta

****

****

******Logistic** SSE After Each Pass For Increasing Step Size

C) **After 100,000 Steps:**

Note- I chose lambda = .00001 for the best logistic regression model. I chose this because it had lower SSE error than with .001 and it had a slightly better average loss.

**PGC** 0.079537

**insulin** 0.015663

**BMI** 0.077129

Looking at the values it appears that PGC and BMI both have a high impact on whether one has diabetes or not. They also share roughly the same weight in determining a person’s diagnosis. Insulin has a much lower impact than the other two variables, this is somewhat surprising given the disease.