Logistic Regression and Stochastic Gradient Ascent Analysis

Paul Laliberte' | CSCI-5622

1. How did the learning rate affect the convergence of your SGA implementation?

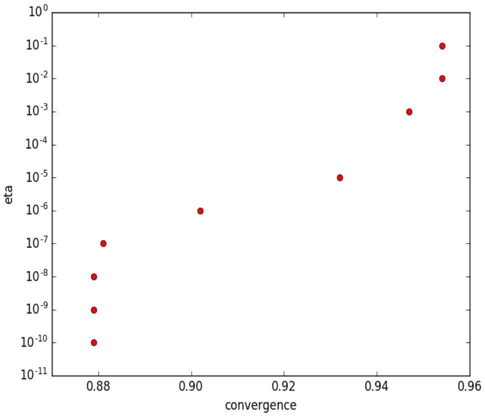
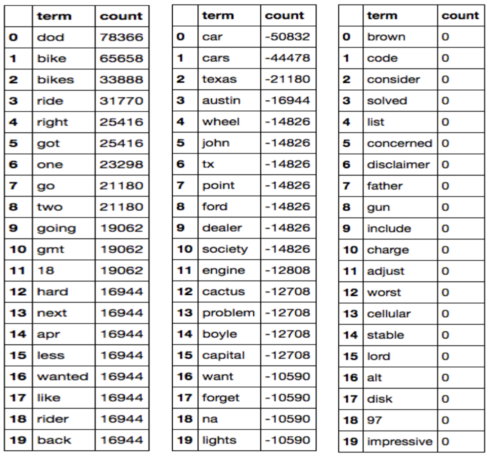


Figure 1: Learn Rate and Convergence



	term	count
0	dod	78366
1	bike	65658
2	bikes	33888
3	ride	31770
4	right	25416
5	got	25416
6	one	23298
7	go	21180
8	two	21180
9	going	19062
10	gmt	19062
11	18	19062
12	hard	16944
13	next	16944
14	apr	16944
15	less	16944
16	wanted	16944
17	like	16944
18	rider	16944
19	back	16944

	term	count
0	car	-50832
1	cars	-44478
2	texas	-21180
3	austin	-16944
4	wheel	-14826
5	john	-14826
6	tx	-14826
7	point	-14826
8	ford	-14826
9	dealer	-14826
10	society	-14826
11	engine	-12808
12	cactus	-12708
13	problem	-12708
14	boyle	-12708
15	capital	-12708
16	want	-10590
17	forget	-10590
18	na	-10590
19	lights	-10590

	term	count
0	brown	0
1	code	0
2	consider	0
3	solved	0
4	list	0
5	concerned	0
6	disclaimer	0
7	father	0
8	gun	0
9	include	0
10	charge	0
11	adjust	0
12	worst	0
13	cellular	0
14	stable	0
15	lord	0
16	alt	0
17	disk	0
18	97	0
19	impressive	0

Figure 2: Predictors: Best (+, −), Worst (0)

The initial learning rate was $\eta = 0.1$, with constant $\lambda = .25$ and passes = 3, for all runs. We then decreased the rate by magnitudes of .1, i.e. .01, .001, .0001, ect. From Figure 1, we see that a lower learning rate equates to a longer time period to find convergence versus a higher learning rate. It is important to note that a higher convergence is not always favorable. In the case where $\eta = .1$, we are severly over-fitting the data, and actually converged to an accuracy (test-HA accuracy) of .954 relatively early in the passes (roughly one pass through the data), whereas a $\eta = .1 \times 10^{-8}$ took longer to converge (all three passes).

2. What was your stopping criterion and how many passes over the data did you need before stopping?

The stopping criterion used had two factors to consider: 1) Was convergence a viable possibility? 2) What was the rate of convergence?. The first analyzed the issue of whether there was even a chance to converge, or were we trapped in a local minima. To give an example, running the program with a $\eta = .1$ and $\lambda = .25$ produced a divergent solution, no matter the number of passes through the data we assigned. The testing accuracy (HA) would vary from around .62 to .86. The variance over the data was consistent for both training data (TP, TA) and testing data (HP, HA), where TP and HP are the log probabilities. If there seemed to be a definitive convergence we then consider the second, what was the rate of convergence. To be more specific, running the program with a $\eta = .001$, $\lambda = .25$, and passes = 10 we would eventually find a solution that was at, or really close, to convergence at about 6 to 7 passes through the data. At this point we were getting minimal, if any, improvement in the convergence rate of TP, HP, TA, and HA. Hence, we could stop at this point and be confident in the results. The number of passes through the data varied between what the initial learning rate, η, was set at and what the regularization term, λ, was initially set to. A low η and λ would require several passes through the data (upper bound around 10) to confidently say that convergence was close to, if not met. A high η (.1 would be considered high) and a respectable λ (we say respectable is .25) usually resulted in a divergent solution, and a high η and low λ (assume low is .001 or unregularized) would converge to a solution in less than, or equal to, 1 pass through the data.

3. What words are the best predictors of each class? How (mathematically) did you find them?

To find the words that were the best predictors of each class we assigned a weight of (+1) if a word was in a document about motorcycles, and (-1) if a word was in a document about automobiles. Running the program through several passes of the data, we amounted to 20 of the best words for each class, which are present in Figure 2.

4. What words are the poorest predictors of classes? How (mathematically) did you find them?

Similarly to question 3, words that are poor predictors are those with an overall score of 0 (Figure 2). These words had we in both motorcycle articles and automobile articles evenly. Note that we cleaned the data of of instances where words in the vocab did not appear at all. Any words that have a score of 0 did appear in the articles at least twice.

Extra Credit 1.

Extra Credit 2.