

CSE6250: Big Data Analytics in Healthcare

Homework 2 Answer

February 11, 2018

1 Logistic Regression [25 points]

1.1 Batch Gradient Descent

a. Derive the gradient of the negative log-likelihood in terms of \mathbf{w} for this setting. [5 points]

$$\begin{aligned} NLL(D, \mathbf{w}) &= -\sum_{i=1}^N [(1 - y_i) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}_i)) + y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i)] \\ &= -\sum_{i=1}^N [(1 - y_i) (-\mathbf{w}^T \mathbf{x}_i + \log(\sigma(\mathbf{w}^T \mathbf{x}_i))) + y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i)] \\ &= -\sum_{i=1}^N [-\mathbf{w}^T \mathbf{x}_i + \log \sigma(\mathbf{w}^T \mathbf{x}_i) + y_i (\mathbf{w}^T \mathbf{x}_i)] \\ &= -\sum_{i=1}^N \left[-\mathbf{w}^T \mathbf{x}_i - \log(1 + e^{-\mathbf{w}^T \mathbf{x}_i}) + y_i (\mathbf{w}^T \mathbf{x}_i) \right] \\ &= -\sum_{i=1}^N \left[-\log(1 + e^{\mathbf{w}^T \mathbf{x}_i}) + y_i (\mathbf{w}^T \mathbf{x}_i) \right] \\ \frac{\delta NLL(\mathbf{w})}{\delta \mathbf{w}} &= -\sum_{i=1}^N \mathbf{x}_i (-\sigma(\mathbf{w}^T \mathbf{x}_i) + y_i) \end{aligned}$$

1.2 Stochastic Gradient Descent

a. Show the log likelihood, l , of a single (\mathbf{x}_t, y_t) pair. [5 points]

$$l(\mathbf{w}) = (1 - y_t) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}_t)) + y_t \log \sigma(\mathbf{w}^T \mathbf{x}_t)$$

b. Show how to update the coefficient vector \mathbf{w}_t when you get a patient feature vector \mathbf{x}_t and physician feedback label y_t at time t using \mathbf{w}_{t-1} (assume learning rate η is given). [5 points]

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta(\mathbf{x}_t (-\sigma(\mathbf{w}_{t-1}^T \mathbf{x}_t) + y_t))$$

c. What is the time complexity of the update rule from **b** if \mathbf{x}_t is very sparse? [2 points]

If we are using (feature-id \rightarrow value) mapping, suppose there are k such pairs, we only need to update k of d dimensions of \mathbf{w}_t . The time complexity is just $\mathcal{O}(k)$, note that k is very small since \mathbf{x}_t is very sparse, in the best case the time complexity is close to $\mathcal{O}(1)$.

d. Briefly explain the consequence of using a very large η and very small η . [3 points]

η tells how big each gradient descent “step” is. If η is small, it may take much more iterations to reach convergence, but small η also has larger probability to eventually reach a convergence point. If η is very large, it may only take a few iterations to reach convergence, but large η also has a larger probability of missing the convergence point and in the worst case the iterations may never stop.

e. Show how to update \mathbf{w}_t under the penalty of L2 norm regularization. In other words, update \mathbf{w}_t according to $l - \mu\|\mathbf{w}\|_2^2$, where μ is a constant. What’s the time complexity? [5 points]

$$\frac{\delta(l - \mu\|\mathbf{w}\|_2^2)}{\delta\mathbf{w}} = \mathbf{x}_t \left(-\sigma(\mathbf{w}^T \mathbf{x}_t) + y_t \right) - 2\mu\mathbf{w}$$

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta \left(\mathbf{x}_t \left(-\sigma(\mathbf{w}_{t-1}^T \mathbf{x}_t) + y_t \right) - 2\mu\mathbf{w}_{t-1} \right)$$

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta\mathbf{x}_t \left(-\sigma(\mathbf{w}_{t-1}^T \mathbf{x}_t) + y_t \right) - 2\eta\mu\mathbf{w}_{t-1}$$

Suppose \mathbf{w}_t has d dimensions, since we have to update every dimension each time, the time complexity is $\mathcal{O}(d)$.

f. When you use L2 norm, you will find each time you get a new (\mathbf{x}_t, y_t) you need to update every element of vector \mathbf{w}_t even if \mathbf{x}_t has very few non-zero elements. Write the pseudo-code on how to update \mathbf{w}_t lazily. [Extra 5 points] **(no partial credit!)**

Rewrite the update equation as $\mathbf{w}_t = \mathbf{w}_{t-1} (1 - 2\eta\mu) + \eta\mathbf{x}_t (-\sigma(\mathbf{w}_{t-1}^T \mathbf{x}_t) + y_t)$. We only need to update \mathbf{w}^j where $x^j \neq 0$. Suppose there are T pair of (\mathbf{x}_t, y_t) . $I[j]$ records the value of the last time k $W[j]$ was updated.

```

input :  $(\mathbf{x}_t, y)$ 
output:  $\mathbf{W}$ 
1  $k = 0$ 
2  $I = [0.0] * \langle \text{number of features} \rangle$ 
3  $W = [0.0] * \langle \text{number of features} \rangle$ 
4 for  $t \leftarrow 1$  to  $M$  do
5    $k = k + 1$ 
6    $\text{sum} = 0$ 
7   for  $j, \text{value in } x_t$  do
8      $\text{sum} = \text{sum} + W[j] * \text{value}$ 
9      $W[j] = W[j] * (1 - 2\eta\mu)^{k-I[j]}$ 
10     $I[j] = k$ 
11  end
12  for  $j, \text{value in } x_t$  do
13     $W[j] = W[j] + \eta * \text{value} * (y - \sigma(\text{sum}))$ 
14  end
15 end

```

Algorithm 1: pseudo-code on updating \mathbf{w}_t lazily

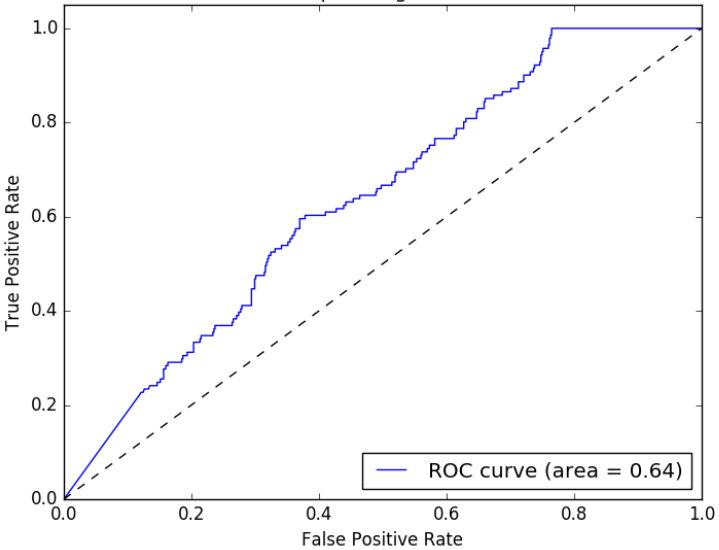
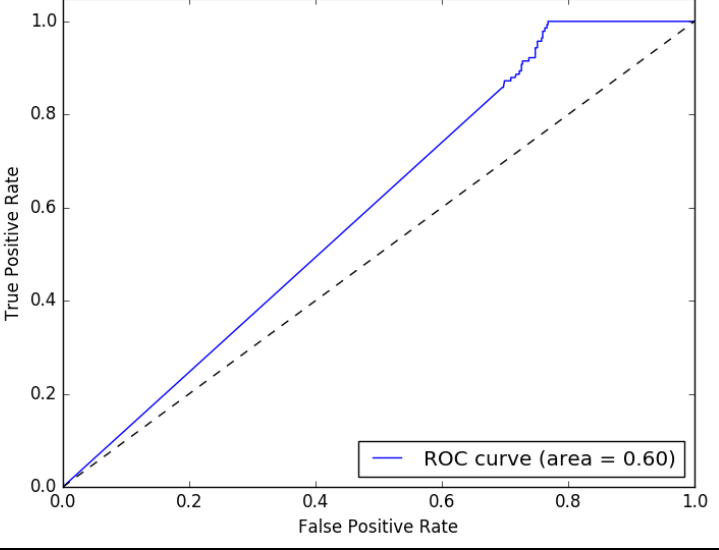
References

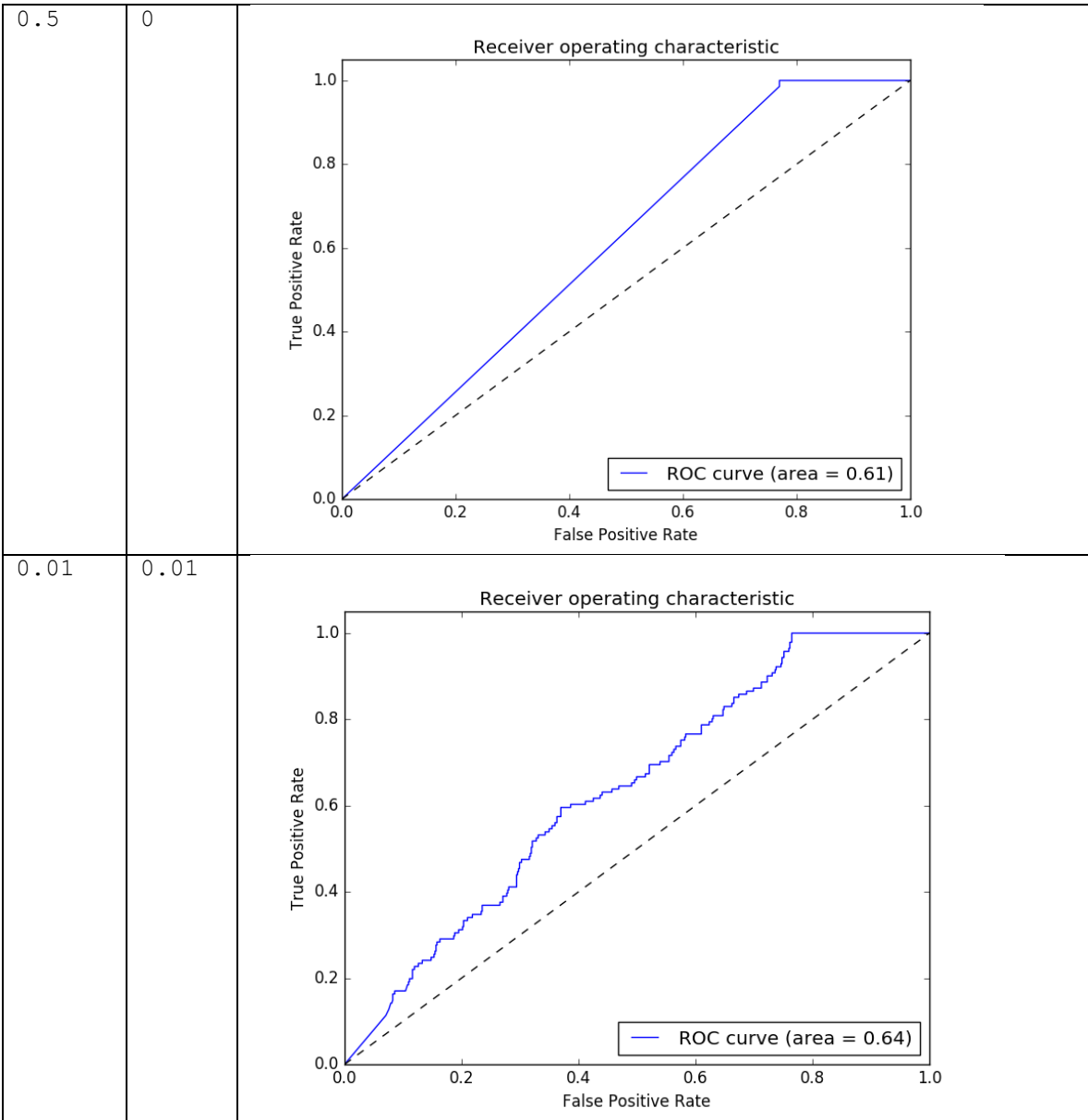
- [1] Bob Carpenter: Lazy Sparse Stochastic Gradient Descent for Regularized Multinomial Logistic Regression.
<https://lingpipe.files.wordpress.com/2008/04/lazysgdregression.pdf>

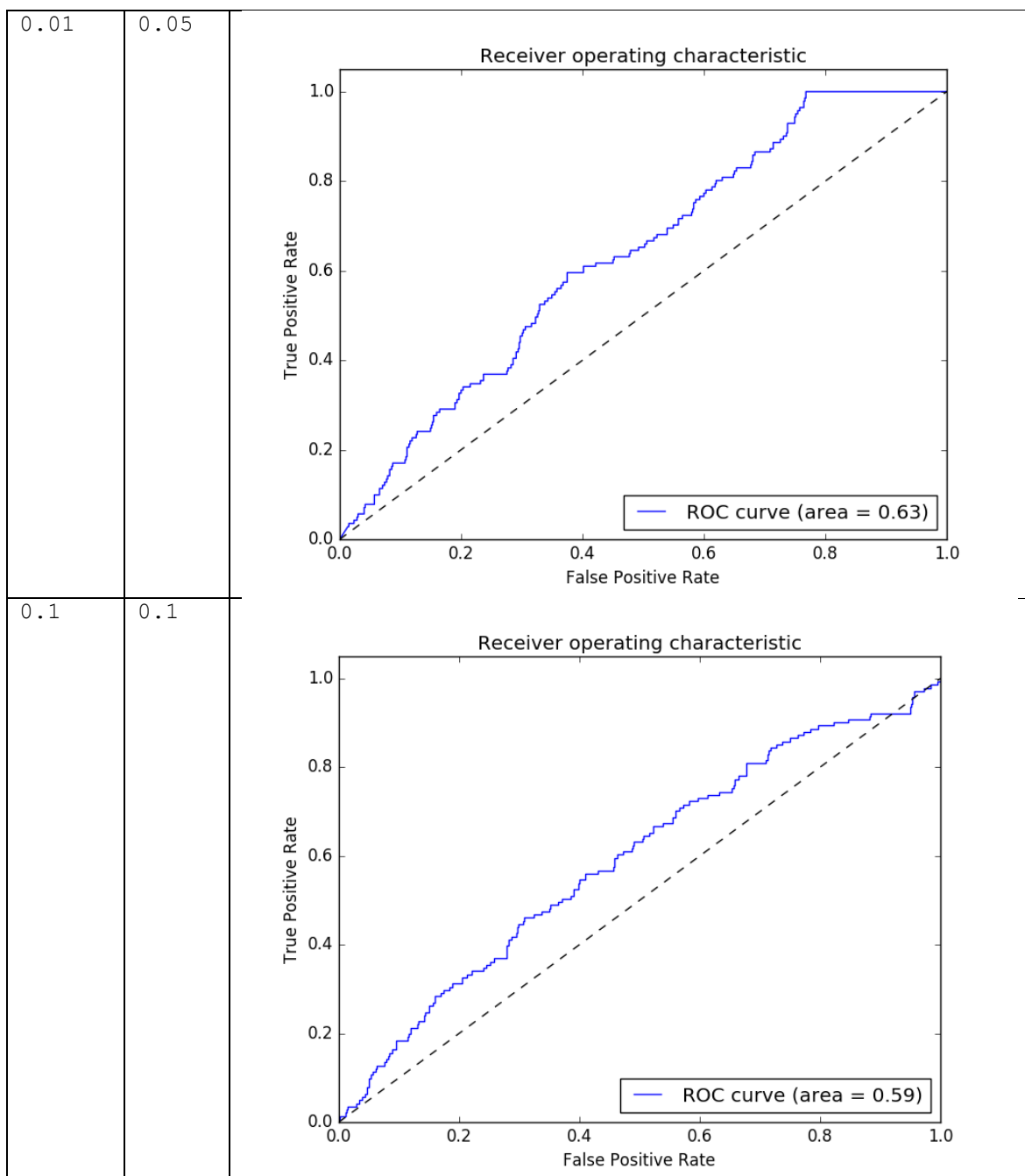
2.1 Descriptive Statistics

Metric	Deceased patients	Alive patients
Event Count		
1. Average Event Count	1027.7385229540919	683.1552587646077
2. Max Event Count	16829	12627
3. Min Event Count	2	1
Encounter Count		
1. Average Encounter Count	24.839321357285428	18.695492487479132
2. Max Encounter Count	375	391
3. Min Encounter Count	1	1
Record Length		
1. Average Record Length	157.04191616766468	194.70283806343906
2. Median Record Length	25	16
3. Max Record Length	5364	3103
4. Min Record Length	0	0
Common Diagnosis	DIAG320128 416 DIAG319835 413 DIAG313217 377 DIAG197320 346 DIAG132797 297	DIAG320128 1018 DIAG319835 721 DIAG317576 719 DIAG42872402 674 DIAG313217 641
Common Laboratory Test	LAB3009542 32765 LAB3023103 28395 LAB3000963 28308 LAB3018572 27383 LAB3016723 27060	LAB3009542 66937 LAB3000963 57751 LAB3023103 57022 LAB3018572 54721 LAB3007461 53560
Common Medication	DRUG19095164 6396 DRUG43012825 5451 DRUG19049105 4326 DRUG956874 3962 DRUG19122121 3910	DRUG19095164 12468 DRUG43012825 10389 DRUG19049105 9351 DRUG19122121 7586 DRUG956874 7301

2.3 SGD Logistic Regression

η	μ	
0.01	0	<div><p>Receiver operating characteristic</p><p>ROC curve (area = 0.64)</p></div>
0.1	0	<div><p>Receiver operating characteristic</p><p>ROC curve (area = 0.60)</p></div>

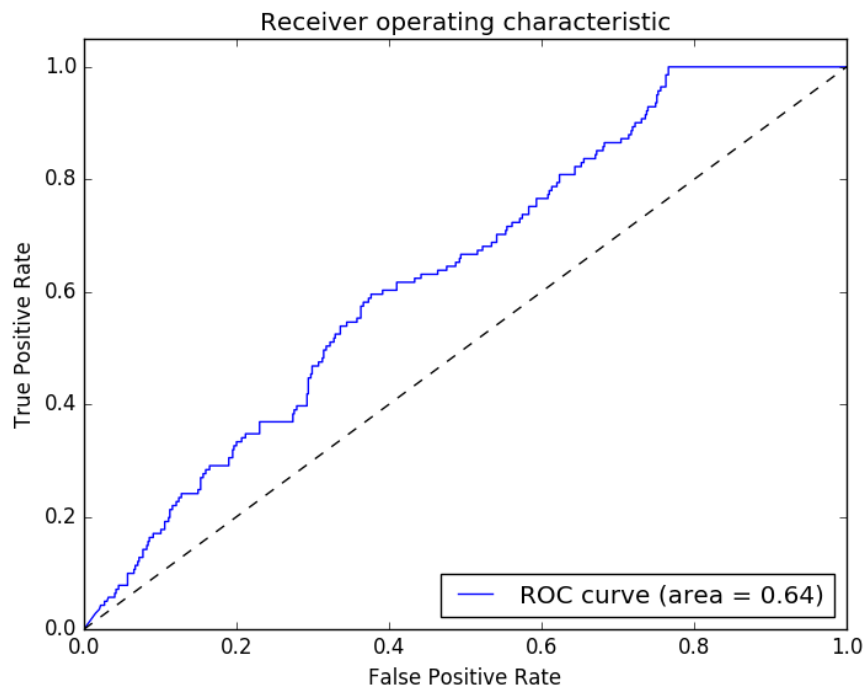




When $\mu = 0$, as η increases, the ROC curve becomes smoother, which means that when the 'learning rate η ' increases, there are less data points making the curve.

To prevent overfitting in logistic regression, we provide the L2 norm regularization, which reverses each "searching step" for the same η , when μ is larger, the AUC gets smaller.

2.4 Hadoop



This is the ROC curve of training 5 ensembles and averaging the test results. Ensemble method is effective in reducing bias and variance. However in this case, the AUC didn't change much, which means the dataset probably has low bias or variance.