HW2

George Zhou

September 2019

1 Batch GD

Derivative of gradient for one datapoint ($\mathbf{x}$, y):

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} y \log \sigma(\theta^T \mathbf{x}) + \frac{\partial}{\partial \theta_j}(1-y)\log[1-\sigma^T(\mathbf{x}\theta^T)] \tag{1}$$

$$= [\frac{y}{\sigma(\theta^T x)} - \frac{1-y}{1-\sigma(\theta^T x)}]\frac{\partial}{\partial \theta_j}\sigma(\theta^T x) \tag{2}$$

$$= [\frac{y}{\sigma(\theta^T x)} - \frac{1-y}{1-\sigma(\theta^T x)}]\sigma(\theta^T x)[1-\sigma(\theta^T x)]x_j \tag{3}$$

$$= [\frac{y-\sigma(\theta^T x)}{\sigma(\theta^T x)[1-\sigma(\theta^T x)]}]\sigma(\theta^T x)[1-\sigma(\theta^T x)]x_j \tag{4}$$

$$= [y-\sigma(\theta^T x)]x_j \tag{5}$$

Where theta equals w in this case, and taking the sum for all data points for the negative form:

$$\frac{\partial NLL(w)}{\partial w} = -\sum_{i=1}^{N}[y_i - \sigma(w^T x_i)]x_i \tag{6}$$

2 SGD

a)

$$[(1-y_t)\log(1-\sigma(w^T x_t)) + y_t \log \sigma(w^T x_t)] \tag{7}$$

b)

$$w_t = w_{t-1} + \eta[y_t - \sigma(w_{t-1}^T x_t)]x_t \tag{8}$$

c) If it is dense, the time complexity is equal to O(Nx) where N is number of samples and x is number of features or average number of features. If it is very sparse, the time complexity approaches O(N) where N is size of dataset.

d) A large value in learning rate leads to large changes of the weight of the feature along the direction of gradient which can lead to the algorithm to look over the global minimum for optimization and find a local minimum instead. Smaller learning rates usually will take much longer,however, at finding a converging minima.

e)

$$w_t = w_{t-1} + \eta[(y_t - \sigma(w_{t-1}^T x_t))x_t - 2\mu w_{t-1}^T] \tag{9}$$

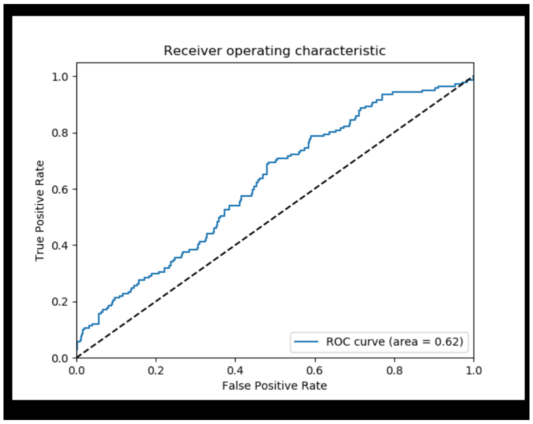
Time complexity is O(x) where x is number is features and O(Nx) where N is number of samples and in this case, it approaches O(x).

3 2.1 b) Descriptive Statistics

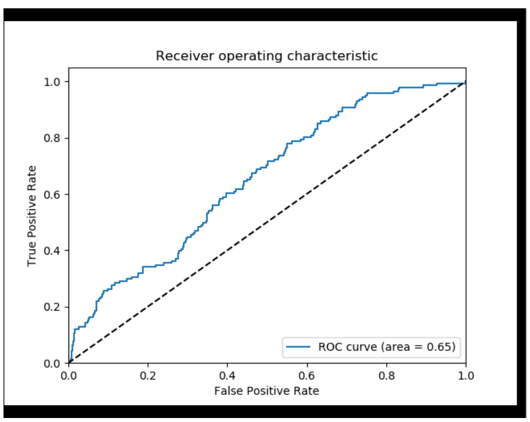
Metric	Deceased patients	Alive patients
Event Count		
1. Average Event Count	1027.74	683.2
2. Max Event Count	16829	12627
3. Min Event Count	2	1
Encounter Count		
1. Average Encounter Count	24.84	18.7
2. Median Encounter Count	NA	NA
3. Max Encounter Count	375	391
4. Min Encounter Count Record Length	1	1
Record Length		
1. Average Record Length	157.04	194.7
2. Median Record Length	25	16
3. Max Record Length	5364	3103
4. Min Record Length	0	0
Common Diagnosis	DIAG320128 DIAG319835 DIAG313217 DIAG197320 DIAG132797	DIAG320128 DIAG319835 DIAG317576 DIAG42872402 DIAG313217
Common Laboratory Test	LAB3009542 LAB3023103 LAB3000963 LAB3018572 LAB3016723	LAB3009542 LAB3000963 LAB3023103 LAB3018572 LAB3007461
Common Medication	DRUG19095164 DRUG43012825 DRUG19049105 DRUG956874 DRUG19122121	DRUG19095164 DRUG43012825 DRUG19049105 DRUG19122121 DRUG956874

4 2.3 b) SGD LR Single Model Approach

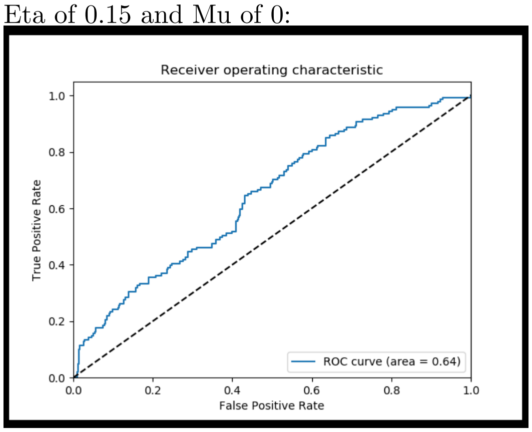
Eta of 0.01 and Mu of 0:



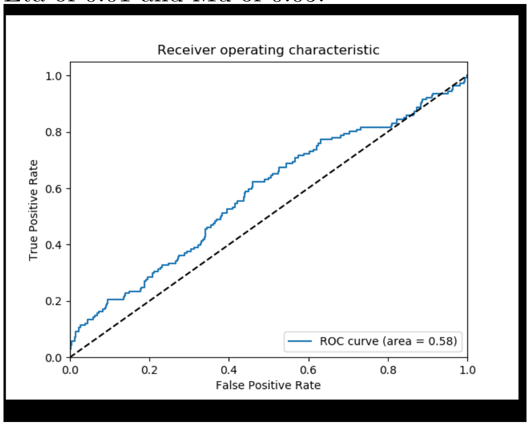
Eta of 0.07 and Mu of 0:



Eta of 0.15 and Mu of 0:



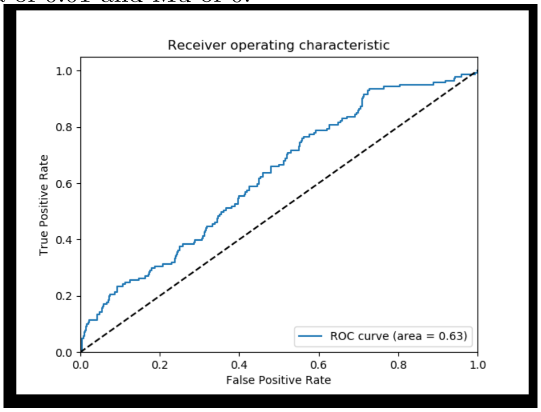
Eta of 0.01 and Mu of 0.05:



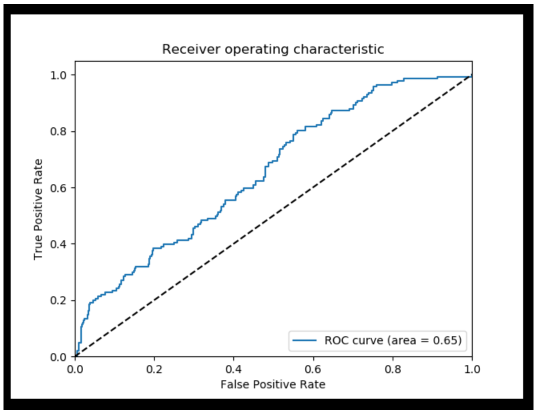
It seems that as compared to the default hyper-parameters, a higher mu constant of 0.05 made the model perform even worse (makes sense as this type of model and data may not need much regularization). Furthermore, after testing higher learning rates with constant regularization constant, the model performed better with eta of 0.07 but then got worse as it went to 0.15. It seems that this particular model may have an optimal learning rate somewhere between 0.01 and 0.15. With more tuning the best performance can be achieved.

5 2.4c) SGD LR Ensemble Approach

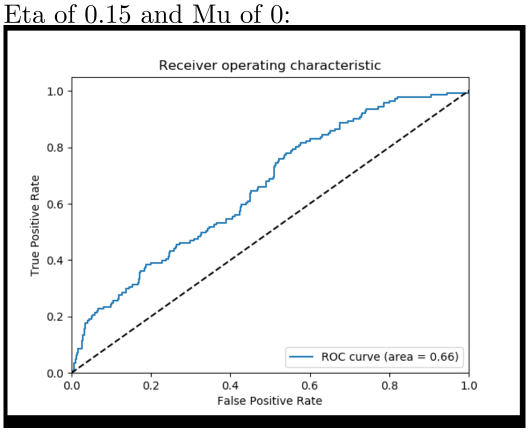
Eta of 0.01 and Mu of 0:



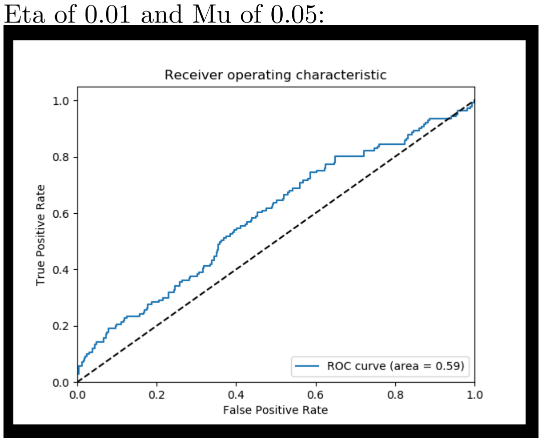
Eta of 0.07 and Mu of 0:



Eta of 0.15 and Mu of 0:



Eta of 0.01 and Mu of 0.05:



Looking at the ensemble results, there does not seem to be too much of a difference between the ensemble and single model approach. Although, for every hyperparameter, the AUC was definitely better higher. The small difference in performance may be important to some production level systems but, one reason that the difference is so small, may be attributed to the fact that it uses the same model. Ensemble modeling is typically more significant when different types or models are used.