Homework 2)

1.1

NLL(w) $= -\sum_{i=1}^{n}(1 - y_i) * \log(1 - \sigma(\mathbf{w^T x})) + y_i * \log(\sigma(\mathbf{w^T x}))$

-NLL(w) = $\sum_{i=1}^{n}(1 - y_i) * \log(1 - \sigma(\mathbf{w^T x})) + y_i * \log(\sigma(\mathbf{w^T x}))$

Taking the derivative of each piece of the summation separately (with respect to **w**), we get:

d/dw [$(1 - y_i) * \log(1 - \sigma(\mathbf{w^T x}))$] = $[(1 - y_i) * 1 / (1 - \sigma(\mathbf{w^T x}))] * \sigma(\mathbf{w^T x}) * (1 - \sigma(\mathbf{w^T x})) * x * -1$

= $(1 - y_i) * \sigma(\mathbf{w^T x}) * x * -1$

= $x * (y_i - 1) * \sigma(\mathbf{w^T x})$

= $xy_i\ \sigma(\mathbf{w^T x}) - x\ \sigma(\mathbf{w^T x})$

Now, taking the derivative of the other piece:

d/dw [$y_i * \log(\sigma(\mathbf{w^T x}))$] = $y_i * (1 / \sigma(\mathbf{w^T x})) * \sigma(\mathbf{w^T x}) * (1 - \sigma(\mathbf{w^T x})) * x$

= $y_i * (1 - \sigma(\mathbf{w^T x})) * x$

= $xy_i - xy_i\ \sigma(\mathbf{w^T x})$

Now, adding this derivative with the above result yields:

$xy_i\ \sigma(\mathbf{w^T x}) - x\ \sigma(\mathbf{w^T x}) + xy_i - xy_i\ \sigma(\mathbf{w^T x}) = xy_i - x\sigma(\mathbf{w^T x})$

= $x(y_i - \sigma(\mathbf{w^T x}))$

So the final result is:

d/dw[-NLL(w)] = $= \sum_{i=1}^{n} x(y\_i - \sigma(\mathbf{w^T x}))$

Therefore, d/dw[NLL(w)] $= -\sum_{i=1}^{n} x(y\_i - \sigma(\mathbf{w^T x}))$

1.2

a.

$$\sum_{i=1}^{n}(1 - y_i) * \log(1 - \sigma(\mathbf{w^T x_t})) + y_i * \log(\sigma(\mathbf{w^T x_t}))$$

b. Update $\mathbf{w_{t-1}}$ by:

$$\mathbf{w_t} = \mathbf{w_{t-1}} + \eta * (\mathbf{y_t} - \sigma(\mathbf{w_t^T x_t})) * \mathbf{x_t}$$

c. O(k*log(d)), where k is the number of nonzero elements of x and d is the total number of features in x

d.

Using a large value of η (learning rate) may cause the optimal solution to be skipped over as the iterations occur. A small value of η will increase the time of convergence, and thus, cause too many iterations to result in the optimal solution.

e.

Update each w_j in $\mathbf{w_t}$ by the following rule:

$\mathbf{w_{updated\_j}} = w_j + \eta * ((y_t - \sigma(\mathbf{w^T x_t})) * x_j - (2 * \mu * w_j))$ where $\mathbf{x_t}$ is the corresponding vector of feature values for the t^{th} patient, y_t is the label for the t^{th} patient, x_j is the feature value associated with weight w_j, and η is the learning rate

The time complexity is O(n) where n is the number of features.

f.

For x_{ti} in $\mathbf{x_t}$:

If x_{ti} not zero:

Update $\mathbf{w_{ti}}$ by:

$$\mathbf{w_{updated\_ti}} = \mathbf{w_{ti}} + \eta * ((y_t - \sigma(\mathbf{w^T x_t})) * x_{ti} - (2 * \mu * \mathbf{w_{ti}}))$$

Else

Do Nothing

2.2

Metric	Alive Patients	Deceased Patients
Event Count		
Average	682.6741	1029.059
Max	12627	16829
Min	1	2
Encounter Count		
Average	18.6694	24.861
Max	391	375
Min	1	1
Record Length		
Average	194.6541	151.397
Max	3103	2601
Min	0	0

Common Diagnosis

Alive DIAG

DIAG320128	1019
DIAG319835	721
DIAG317576	719
DIAG42872402	674
DIAG313217	641

Dead DIAG

DIAG320128	415
DIAG319835	413
DIAG313217	374
DIAG197320	346
DIAG132797	297

Common Laboratory Test

Alive LAB

LAB3009542	66910
LAB3000963	57733
LAB3023103	56967
LAB3018572	54667
LAB3007461	53548

Dead LAB

LAB3009542	32747
LAB3023103	28376
LAB3000963	28288
LAB3018572	27364
LAB3016723	27041

Common Medication

Alive DRUG

DRUG19095164	12452
DRUG43012825	10388
DRUG19049105	9329
DRUG19122121	7586
DRUG956874	7294

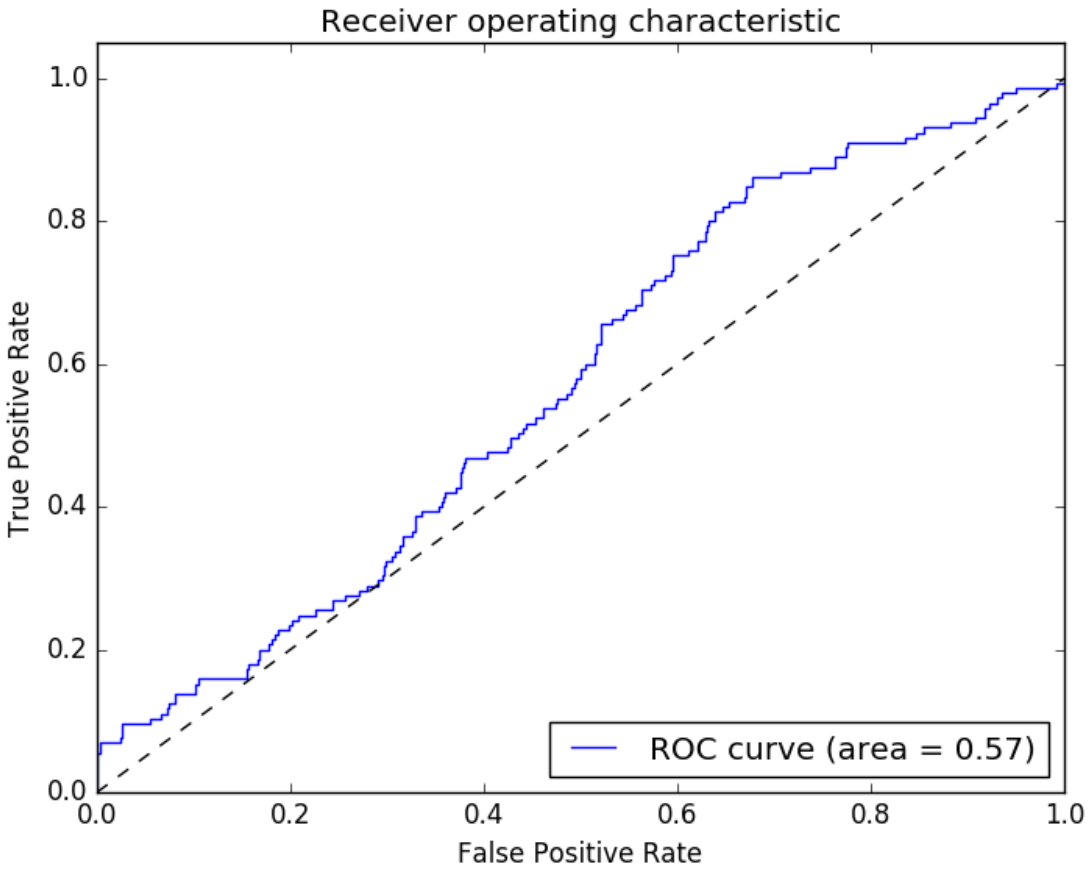
Dead DRUG

DRUG19095164	6394
DRUG43012825	5446
DRUG19049105	4323
DRUG956874	3962
DRUG19122121	3908

2.3

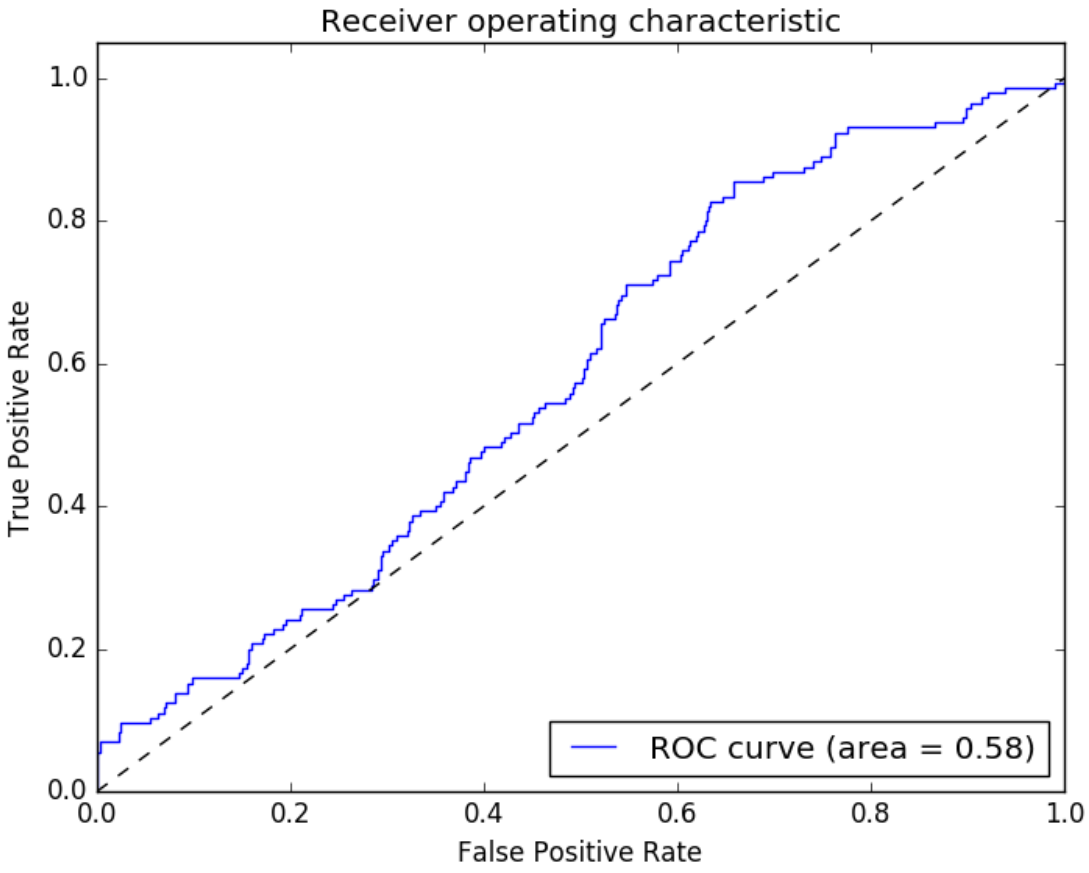
Learning Rate: .01

μ = .01



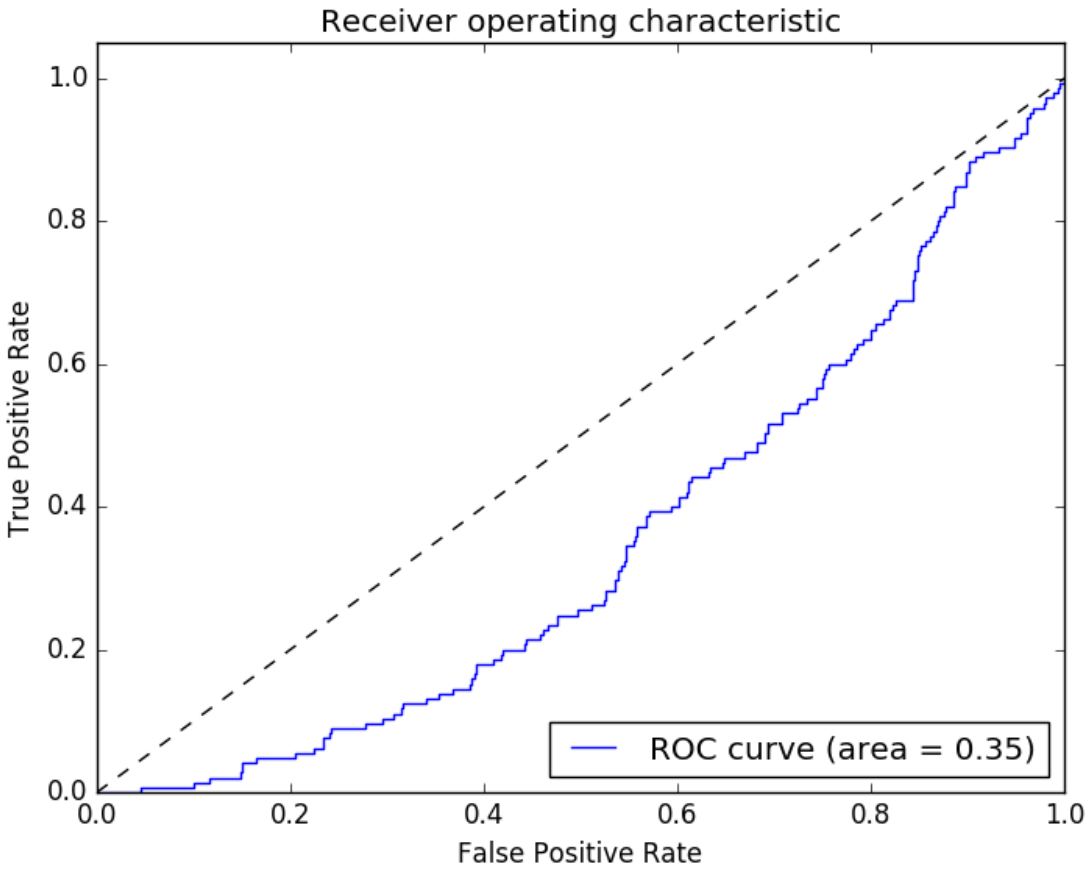
Learning Rate: .01

μ = .0001



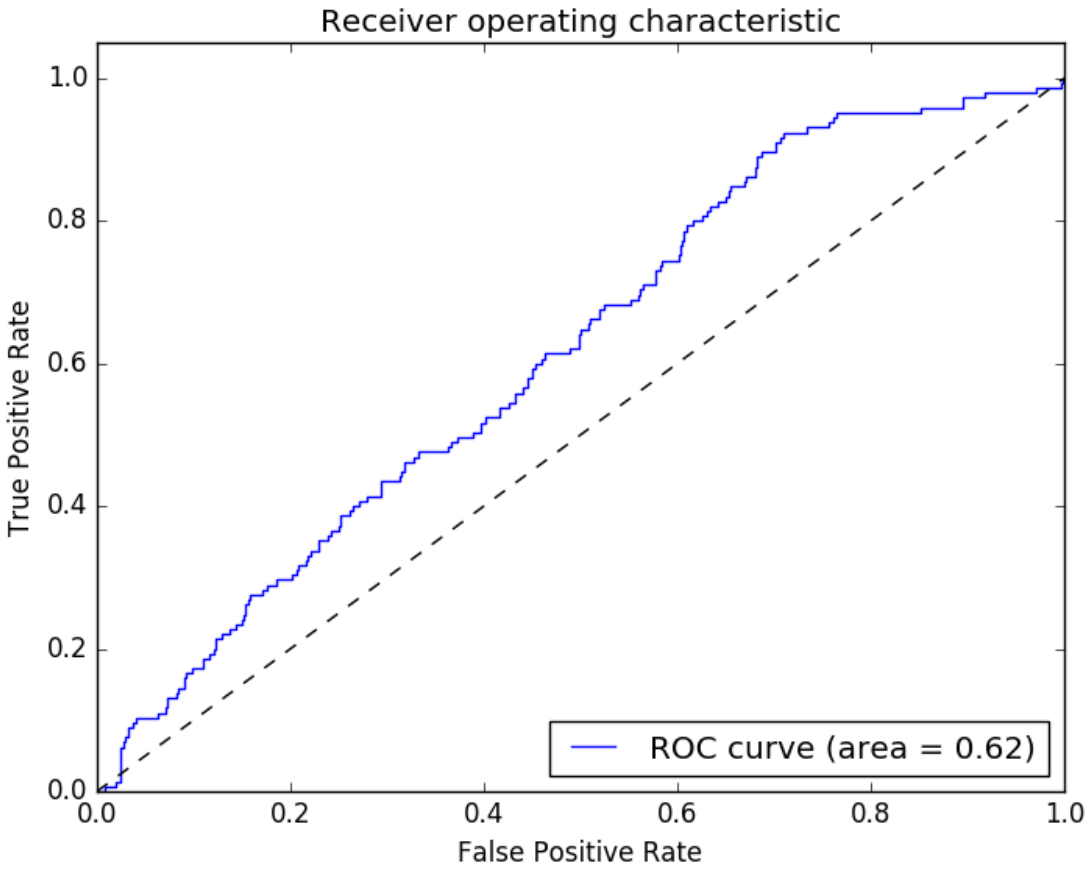
Learning Rate: .50

μ = .10



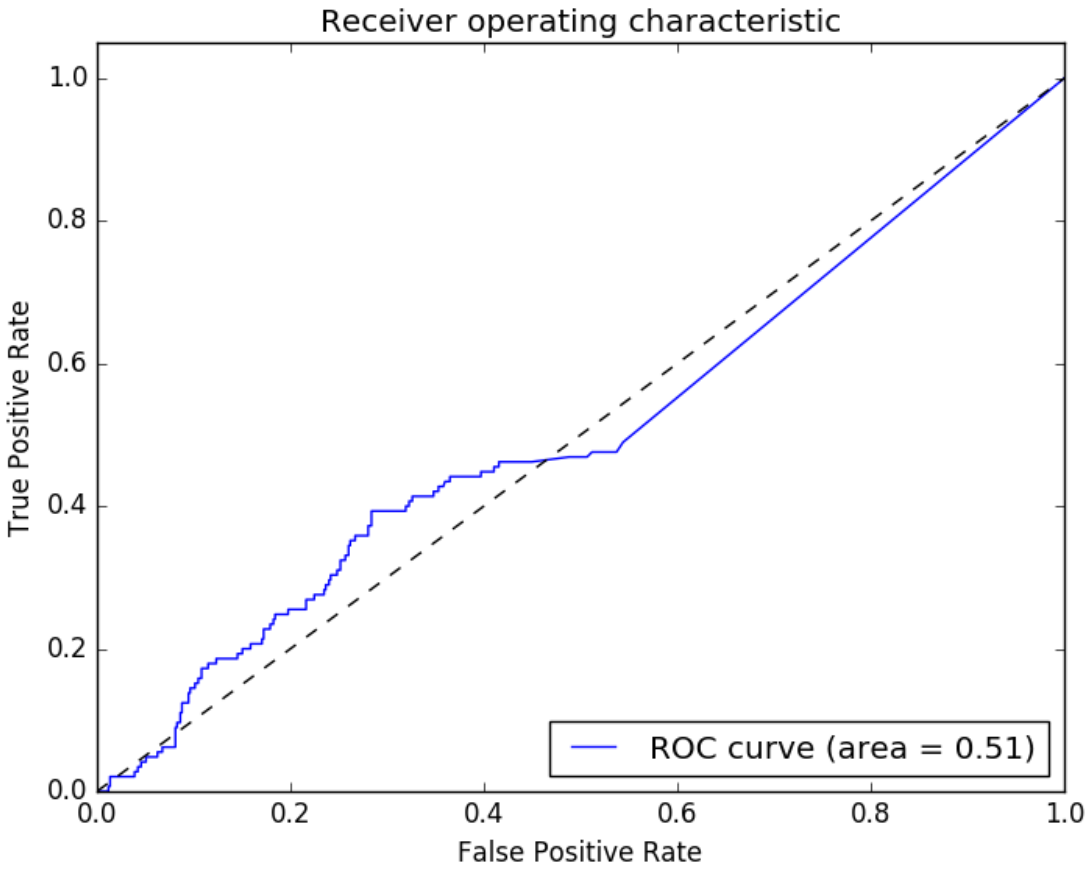
Learning Rate: 0.25

μ = .0001



Learning Rate: 5.0

$\mu = 10^{-6}$

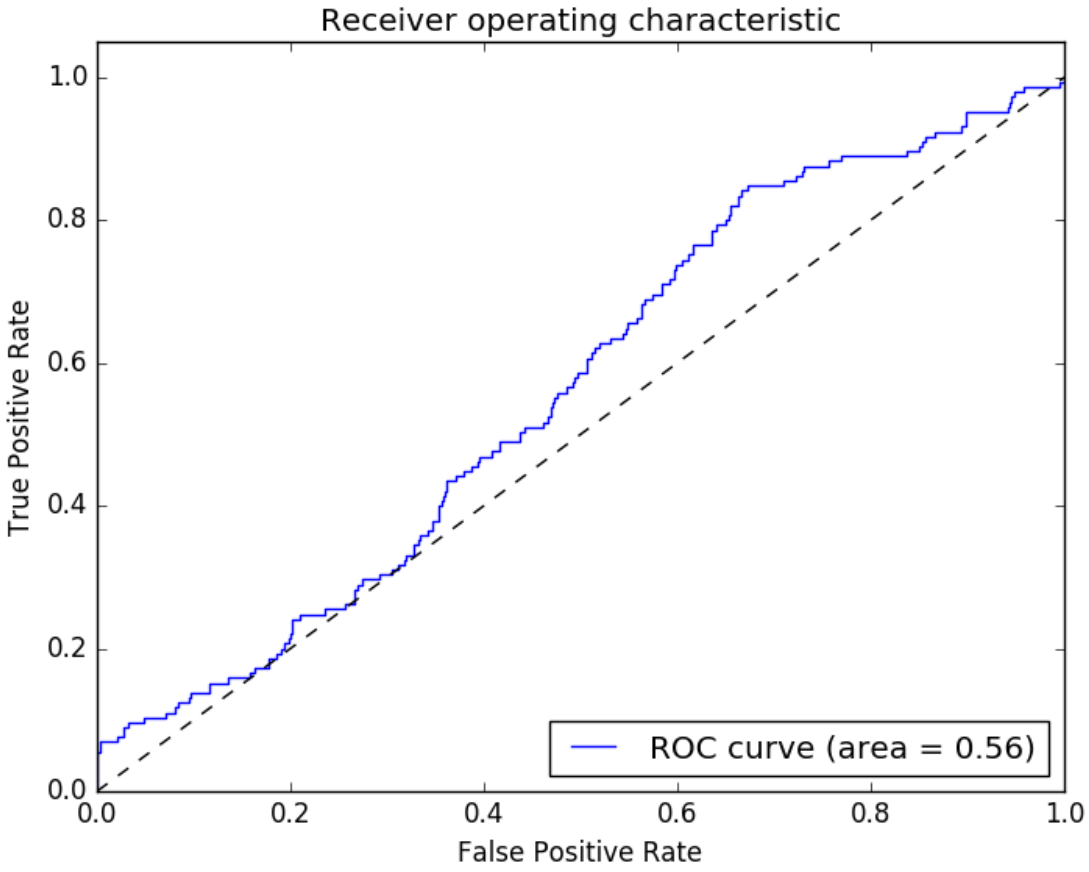


The results in the ROC charts above show the need to tune the learning parameter, η. A more aggressive (greater) value for the learning parameter is leading to worse results, which is most likely caused by the fact that the optimal solution for the logistic regression problem is being passed over by choosing too high of a value. In general, lower values of μ result in better AUC values. This is due to the fact that lower μ values result in less of an L2 norm penalty.

2.4

b.

ROC curve using five models.



c.

The ensemble methods vary depending upon the number of models used. A larger number of models results in smaller subsets of the data being trained, which results in lower AUC performance than the single logistic regression model used initially. The AUC from the ROC curve chart above results in roughly the same range as the initial runs of the logistic regression model.