Question - 1

1. $E\left[\,|g(X)-Y|\,\right]$

$$= \int_Y \int_X |g(x)-y|\, f(x,y)\, dx\, dy$$

'y' represents ~~class~~ classes

$$= \sum_Y \int_X |g(x)-y|\, f(x,y)\, dx$$

for correct classification error (difference) is zero

$$= \int_X \sum_Y |g(x)-y|\, f(x,y)\, dx$$

'y' is one-hot ~~Loss~~ error/difference will be same for any mis classification

$$= \int_X \sum_Y |g(x)-y|\, f(x,y)\, dx$$

for every 'x' we try to minimze

$$\sum_{\forall Y \neq g(x)} |g(x)-y|\, f(x,y)$$

joint probability function.

$$= \lambda \sum_{y \neq g(x)} f(x,y)$$

$$= \lambda \sum_{\substack{\forall i \neq K \\ (1,2,\cdots n)}} P(x, C_i)$$

$$= \lambda \left(P(x) - P(x, C_K) \right)$$

minimizing this equation is
nothing but
maximizing $P(x, C_K)$ current label for that x.

maximize $P(x, y)$

maximize $P(y/x)\ P(x)$

Optimal bayes also select that g^* which maximizes above expression.

2.

$E[(g(x)-Y)^2]$

$$= \sum_{Y} \int_{x} (g(x)-Y)^2 f(x,y)\, dx$$

we want to minimize error for each point

$$= \int_{x} \sum_{Y} (g(x)-y)^2 f(x,y)\, dx$$

$f(x,y) \rightarrow P(y/x)\, P(x)$

$$= \sum (g(x)-y)^2 P(y/x)$$

for 'y' one-hot any mis-classification will incur same ~~loss~~ error

$$= \sum_{g(x) \neq y} (g(X) - Y)^2 P(Y/X)$$

$$= \lambda \sum_{g(x) \neq y} P(Y/X)$$

$$= \lambda \left(1 - P(y/x) \right)$$

$\Rightarrow$ minimozing above

$\Rightarrow$ maximize $P(y/x)$

$\Rightarrow$ we have predict based on the posterior $P(Y/X)$ to minimze $E[(g(X)-Y)^2]$.

3.

$g(x) \sim \eta(x)$

Let us assume this randomized rule $g(x)$ leads to a lower error than the determinstic optimal Bayes Rule.

$\Rightarrow$ The Optimal Bayes Rule with which we are Comparing doesn't maximize $P(y/x)$

The new Rule maximizes $P(y/x)$, Which is a contradiction to the definition of optimal Bayes Rule.

2.

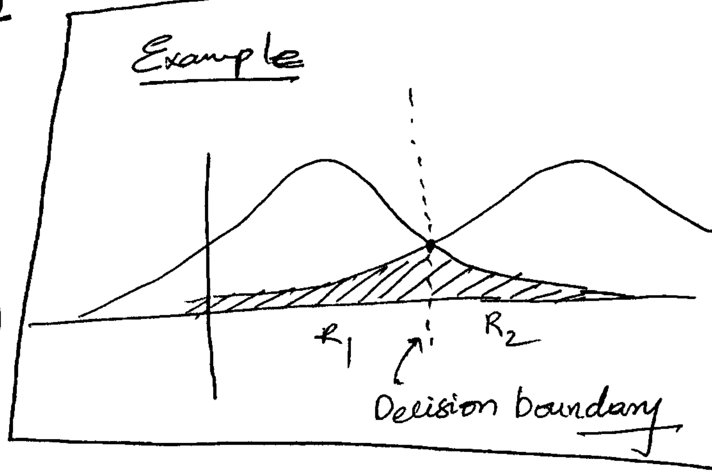
fix $P(error/\omega_1) = \epsilon$

minimize the error of the other class,

$\epsilon_2 = P(error/\omega_2)$

$$\epsilon = P(error/\omega_1) = \int_{R_1} P(x/\omega_2)\, P(\omega_2)\, dx$$

$$P(error/\omega_2) = \int_{R_2} P(x/\omega_1)\, P(\omega_1)\, dx$$



we use Lagrange multiplier

$$q = \theta\left(-\epsilon + \int_{R_1} P(x/\omega_2) P(\omega_2) dx\right) + \int_{R_2} P(x/\omega_1) P(\omega_1) dx$$

$$q = \theta\left(\int_{\alpha_0}^{+\infty} P(x/\omega_2) P(\omega_2) dx - \epsilon\right) + P(\omega_1)\int_{-\infty}^{\alpha_0} P(x/\omega_1) dx$$

$\frac{\partial q}{\partial \alpha_0} = 0$; $\frac{\partial q}{\partial \theta} = 0$ (this will give the constraint)

$$\rightarrow \frac{\partial q}{\partial \alpha_0} = -\theta P(\omega_2) P(\alpha_0/\omega_2) + P(\omega_1) P(\alpha_0/\omega_1)$$

$$\Rightarrow \theta = \frac{P(\omega_1) P(\alpha_0/\omega_1)}{P(\omega_2) P(\alpha_0/\omega_2)} = \frac{P(\omega_1, \alpha_0)}{P(\omega_2, \alpha_0)} = \frac{P(\omega_1/\alpha_0)}{P(\omega_2/\alpha_0)}$$

'α_0' is the decision boundary

for $P(\omega_1/\alpha_0) > \theta P(\omega_2/\alpha_0)$ we will classify α_0 as belonging to ω_1

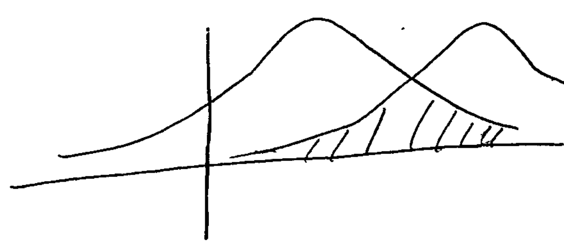
$$\Rightarrow \quad \omega_1 \quad \text{if} \quad \frac{P(\omega_1/x)}{P(\omega_2/x)} > \theta$$

3.

$$pdf = \frac{1}{(\sqrt{2\pi})^d |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$$\Sigma = \sigma^2 I$$
$$\Sigma^{-1} = \frac{1}{\sigma^2} I \quad \text{(diagnol matrix)}$$

$$P(\omega_1) = P(\omega_2)$$



Boundary where

$$P(x/\omega_1)\, \cancel{P(\omega_1)} = P(x/\omega_2)\, \cancel{P(\omega_2)}$$

$$\Rightarrow \quad P(x/\omega_1) = P(x/\omega_2)$$

$$\frac{1}{\cancel{(\sqrt{2\pi})^d |\Sigma|^{1/2}}} e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\mu_i^1)^2} = \frac{1}{\cancel{(\sqrt{2\pi})^d |\Sigma|^{1/2}}} e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\mu_i^2)^2}$$

just an index

$$\Rightarrow \quad \Sigma(x_i-\mu_i^1)^2 = \Sigma(x_i-\mu_i^2)^2$$

$$\cancel{\Rightarrow \quad x_0 = \frac{\mu_1+\mu_2}{2}}$$ (~~Square~~ terms cancel)

$$x^T x - 2\mu_1^T x + \mu_1^T \mu = x^T x - 2\mu_2^T x + \mu_2^T \mu$$

$$\Rightarrow \quad 2(\mu_2 - \mu_1)^T x = \frac{\mu_2^T \mu - \mu_1^T \mu_1}{2}$$

decision boundary is a hyperplane passing through $\frac{\mu_1 + \mu_2}{2}$ and orthogonal to $\frac{\mu_2 - \mu_1}{2}$.

$$p(\text{error}) = \int_{R_1} pdf_2 \, dx_1 \ldots dx_d + \int_{R_2} pdf_1 \, dx_1 \ldots dx_d$$

error distribution is symmetric over $\frac{\mu_1 + \mu_2}{2}$ and pdf_1 is symmetric over μ_1.

$$= \int_{R_2} pdf_1 \, dx_1 \ldots dx_d + \int_{R_2} pdf_1 \, dx_1 \ldots dx_d$$

$$= \int_{R^*} pdf_1 \, dx_1 \ldots dx_d + \int_{R_2} pdf_1 \, dx_1 \ldots dx_d$$

region, which is reflection of 'R_2' w.r.t to a plane passing through 'μ_1' and orthogonal to $(\mu_2 - \mu_1)$

R_2^*, R_2 are all set of points when projected on $(\mu_2-\mu_1)$ are farther than $\left(\frac{\mu_2+\mu_1}{2}\right)$ from 'μ_1'

Projecting the ~~distributions~~ points ~~into~~ along $(\mu_2-\mu_1)$, the distribution will remain gaussian (Spherical Covariance) with mean at the same and Variance 'σ'.

$p(w_1) = p(w_2) = 1/2$

$$= 2 \times 1/2 \times \int_{\frac{\mu_2+\mu_1}{2}}^{\infty} \frac{1}{(\sqrt{2\pi}\,\sigma)^d} \exp\left\{-\frac{(x-\mu_1)^T(x-\mu_2)}{2\sigma^2}\right\} dx_1 \dots dx_n$$

By above ~~Geave~~ Geomentrical interpretation

$$= \int_{\left(\frac{\cancel{\mu_2+\mu_1}}{2}\right)}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{z^2}{2\sigma^2}\right\} dz$$

$(\mu_2-\mu_1)^T\left(\frac{\mu_1+\mu_2}{2}-\mu_1\right)$

$$P(\text{error}) = \int_{\frac{\|\mu_2-\mu_1\|}{2}}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{z^2}{2\sigma^2}\right\} dz$$

$z = \sigma u$

$$\Rightarrow P(\text{error}) = \int_{\frac{\|\mu_2-\mu_1\|}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}\,\cancel{\sigma}} \exp\left\{-\frac{\sigma^2 u^2}{2\sigma^2}\right\} \cancel{\sigma}\, du$$

$$\therefore\ P(\text{error}) = \int_{\frac{\|\mu_2-\mu_1\|}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma^2}} du.$$

3) 2.)

$$\int_a^\infty e^{-u^2/2} du \le \frac{1}{a} e^{-a^2/2}$$

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-u^2/2} du \le \frac{1}{\sqrt{2\pi}\, a} e^{-a^2/2}$$

$$= \frac{1}{\sqrt{2\pi}} \left(\frac{e^{-a^2/2}}{a} \right)$$

$$a = \frac{\|\mu_2-\mu_1\|}{2\sigma}$$

as $d \to \infty$

$a \to \infty$

$$\Rightarrow \underset{d\to}{Lt}\ P_e = \underset{a\to\infty}{Lt} \frac{1}{\sqrt{2\pi}} \left(\frac{e^{-a^2/2}}{a}\right)$$

$$= 0$$

4.)

$$Y = WX \quad ; \quad X \sim N(0,1)$$

1.)

$$E[Y] = E[WX]$$
$$= E[W]\,E[X]$$
$$= 0 \times 0 = 0$$

$$E[(Y-0)^2] = E[Y^2]$$
$$= E[W^2 X^2]$$
$$= E[W^2]\,E[X^2]$$
$$= 1 \times 1 = 1$$

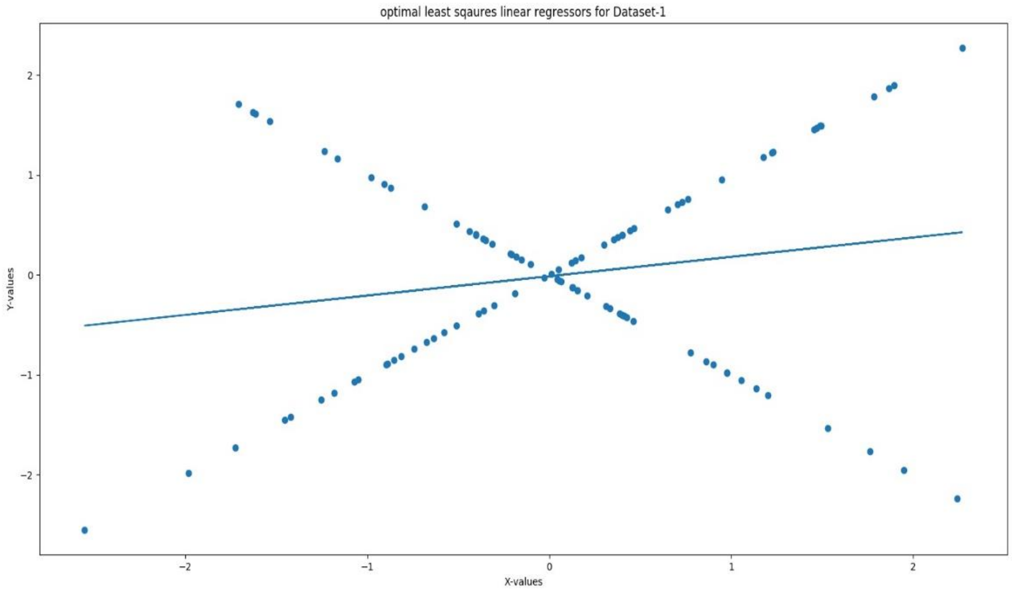
W^2 always 1

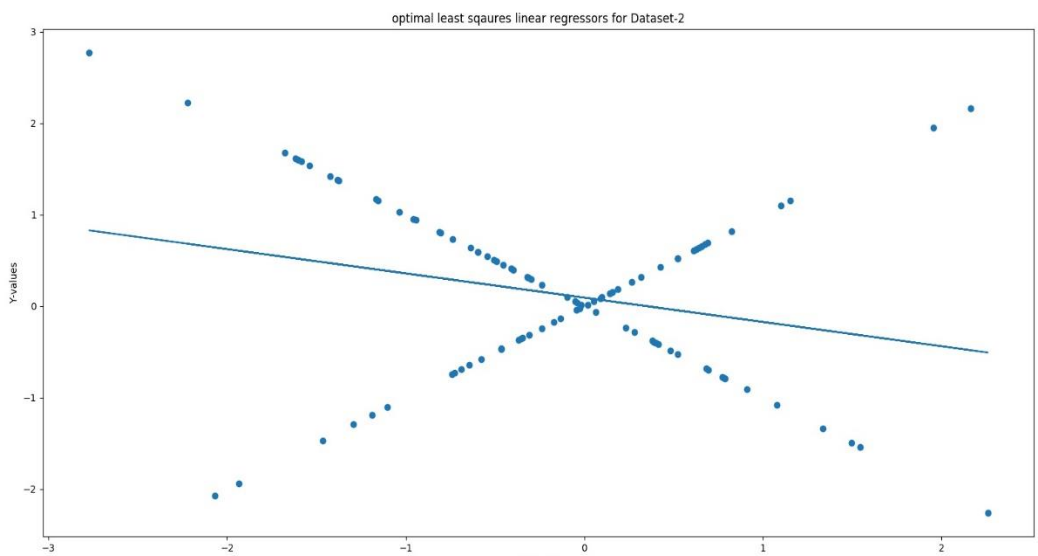
$$\Rightarrow Y \sim N(0,1)$$

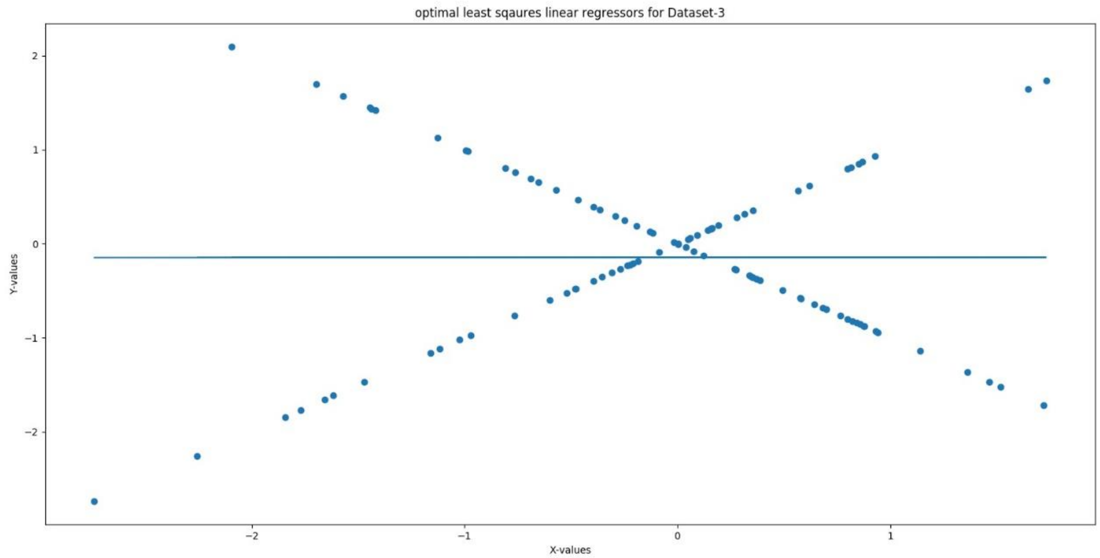
2.)

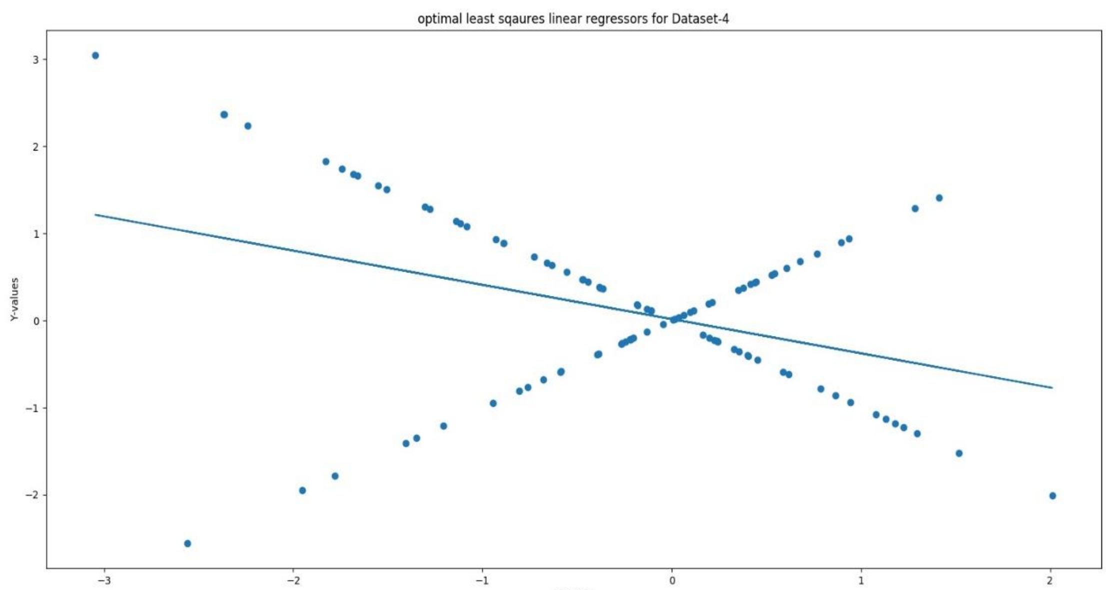
$$Cov[X,Y] = E[(X-0)(Y-0)]$$
$$= E[XY]$$
$$= E\left[E\left[\frac{XY}{W}\right]\right]$$
$$= \left[\tfrac{1}{2}E[X^2] + \tfrac{1}{2}E[-X^2]\right]$$
$$= 0.$$

Question-4 Continued:









From the plots we can observe that in some cases the fit is very close to origin, in some cases it is not. Majority of the points lie in the region where X belongs [-1,1].

For cost function Mean Square error, farther points have a significant effect in pulling the decision boundary towards them.

For **dataset-1**, we have positive slope for the fit, because upper right corner and lower left corner points are far away making the fit tilt towards these points more. (As our cost function is Mean Square error)

For **dataset-2**, we have negative slope for the fit, because probability density in upper left side is high and there some points in upper left, lower right corners which far away, making the fit tilt towards these points more. (As our cost function is Mean Square error)

For **dataset-3**, we have almost zero slope for the fit, even though there are farther points in upper right corner they are balanced out by points probability density in the lower right corner. Similar thing is observed in the left part. Both these things lead to a fit which is parallel to x-axis.

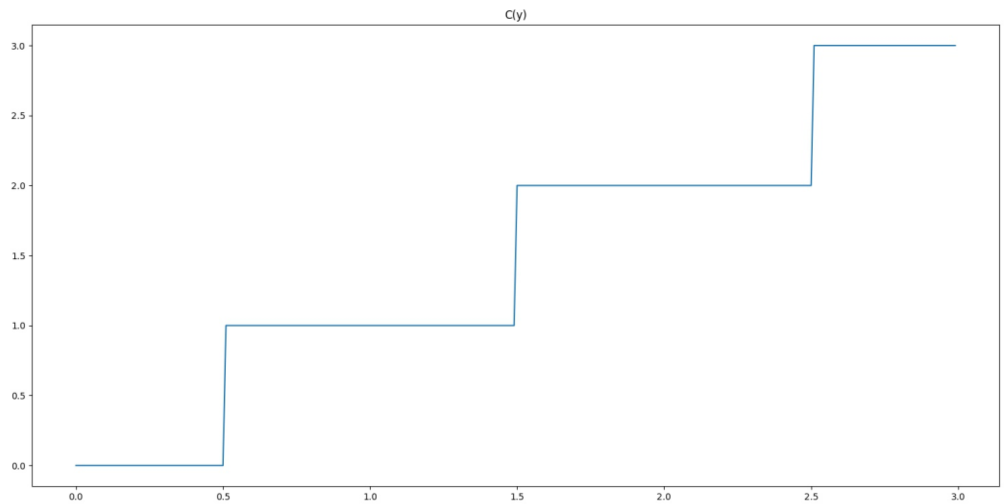
For **dataset-4**, we have negative slope for the fit, because upper left part has farther points and high probability distribution that left lower part. Though right lower contributes in pulling the fit towards it, that is not a major contribution when compared to the contribution of left upper. (effects of right lower and right upper are almost same)

Problem-5:

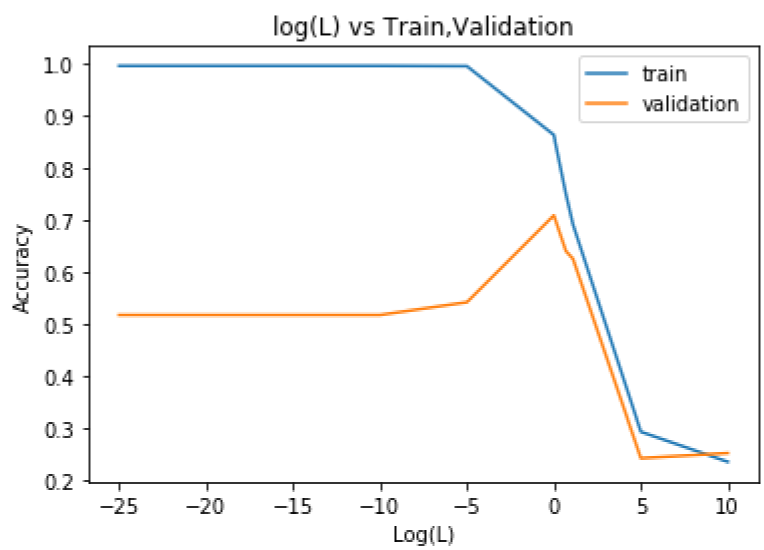
- Dimension of X is 1830 X 1500
- Dimension of Y is 1830 X 1
- Dimension of W is 1 X 1500

5.1.2 **Choice of C(y):** We used Step function.

Why: In Classification task we need to predict the class which is an integer. So, we need to have a function which maps any value to integer space. For this reason, we choose a Step function. (any values less than zero are also given 0 and any value greater the 3 are also given 3)



7. Plot of accuracies vs Log(L) values:



Based on the plot we choose L = 1, because at L we got maximum validation accuracy, which means our model will be able to generalize well on unseen data.

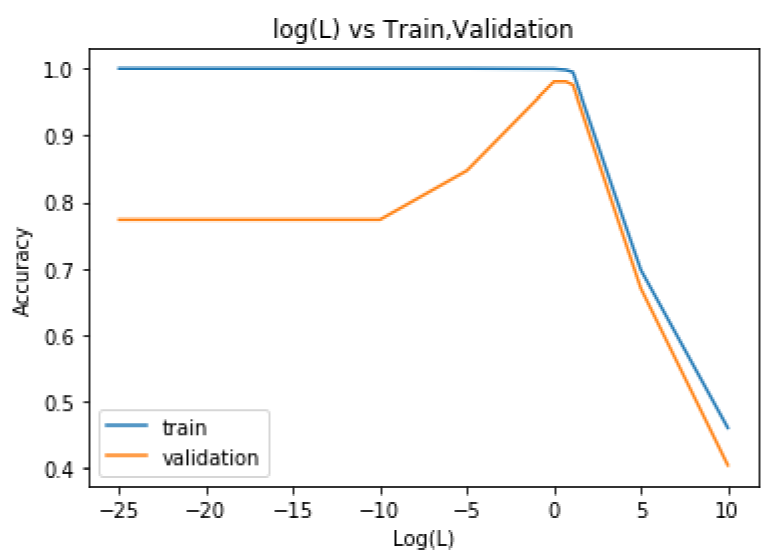
At L = 1, Test accuracy is `0.6060606060606061`

5.1.3.

- Dimension of X is 1830 X 1500
- Dimension of Y is 1830 X 4
- Dimension of W is 4 X 1500

Choice of C(y): SoftMax function. (this makes the summation of probabilities to 1)

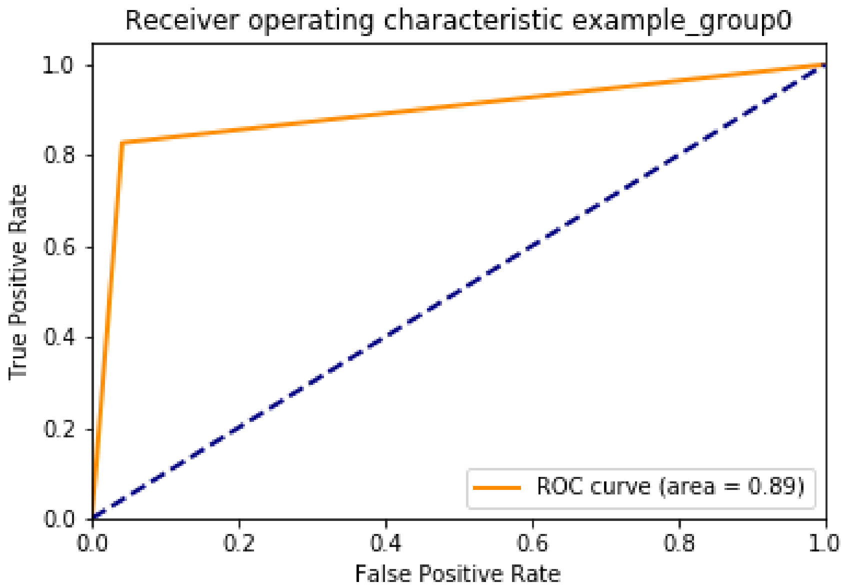
Plot of accuracies vs Log(L) values:

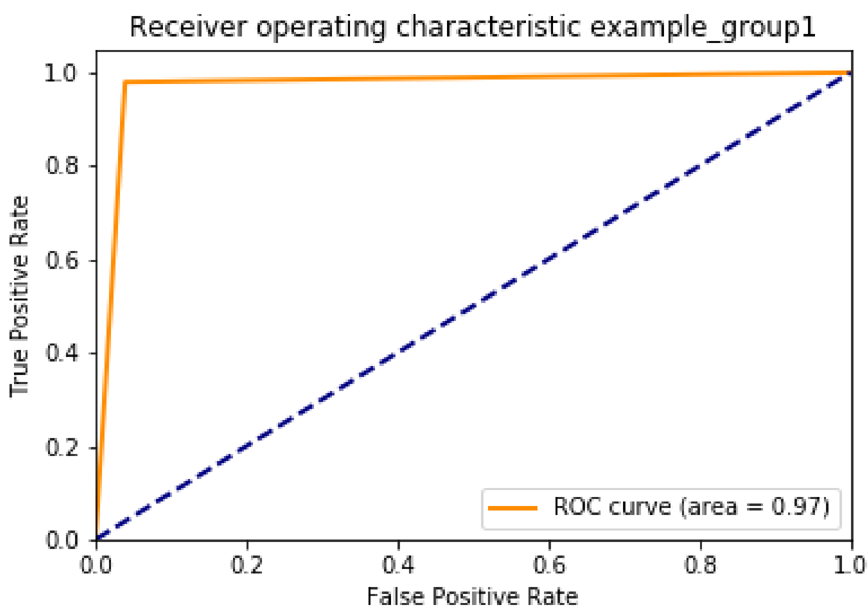


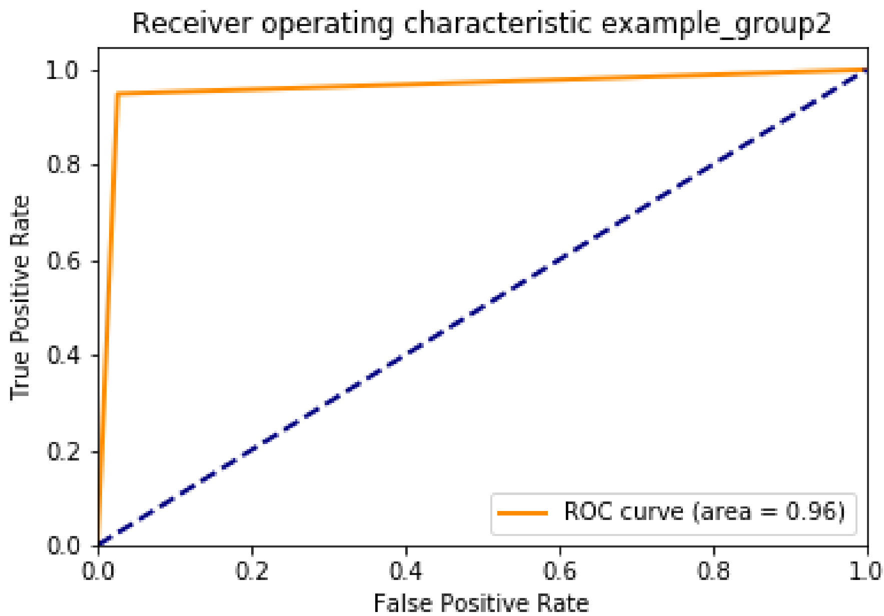
Choice of L: 1

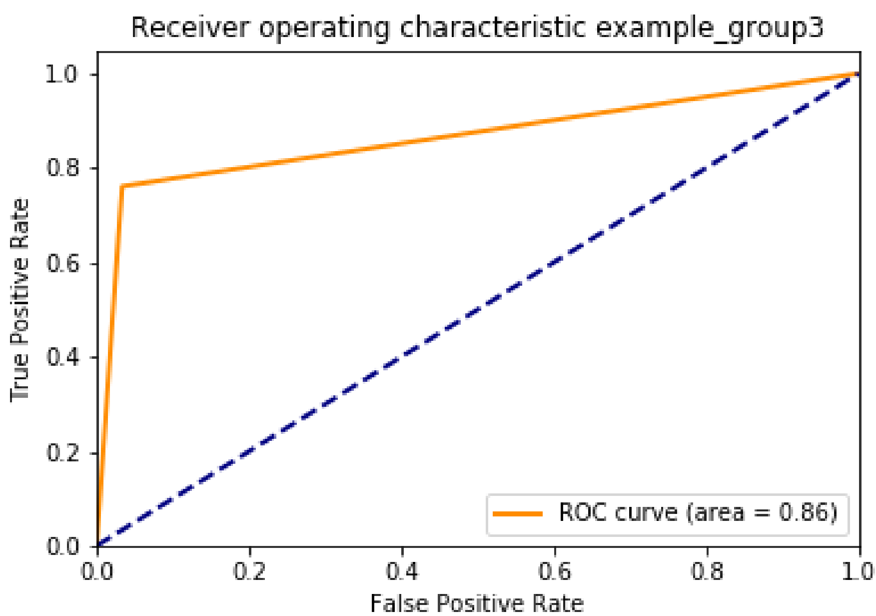
TEST Accuracy: 90.02 %

Roc plots:





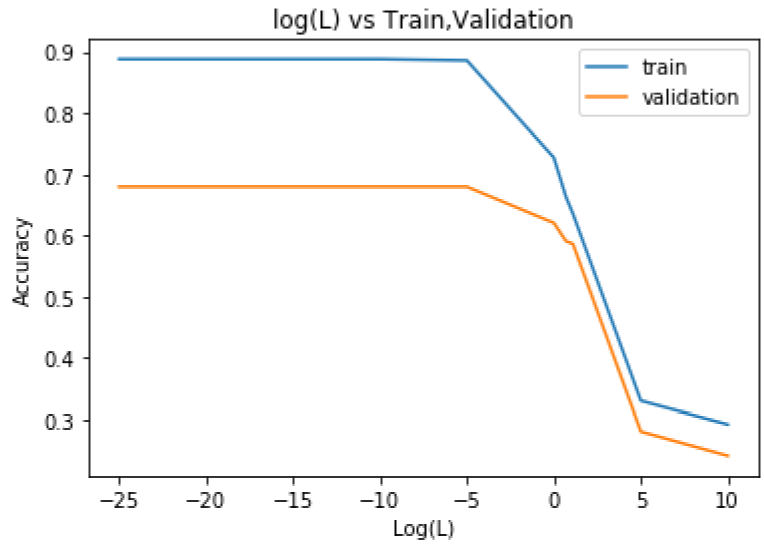

Receiver operating characteristic example_group2
True Positive Rate
False Positive Rate
ROC curve (area = 0.96)
0.0
0.2
0.4
0.6
0.8
1.0


Receiver operating characteristic example_group3
True Positive Rate
False Positive Rate
ROC curve (area = 0.86)
0.0
0.2
0.4
0.6
0.8
1.0

5.1.4

Scalar:

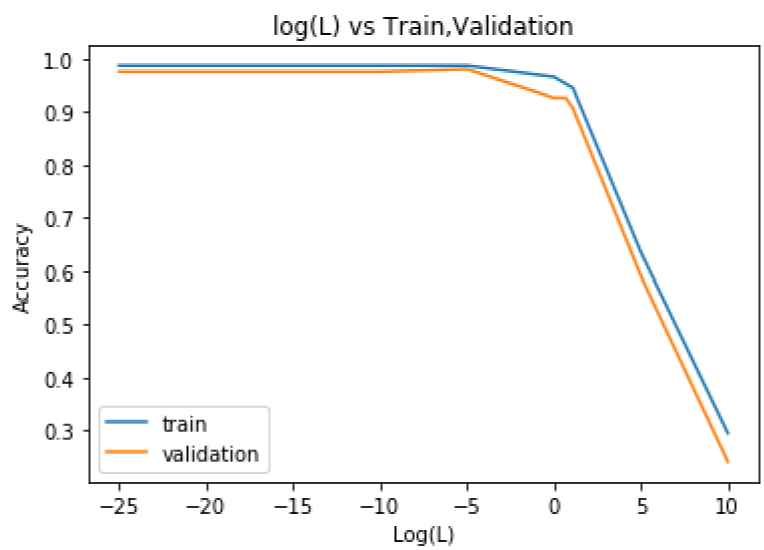
Curve:



At l = 0, we get a test accuracy of **58.98.**

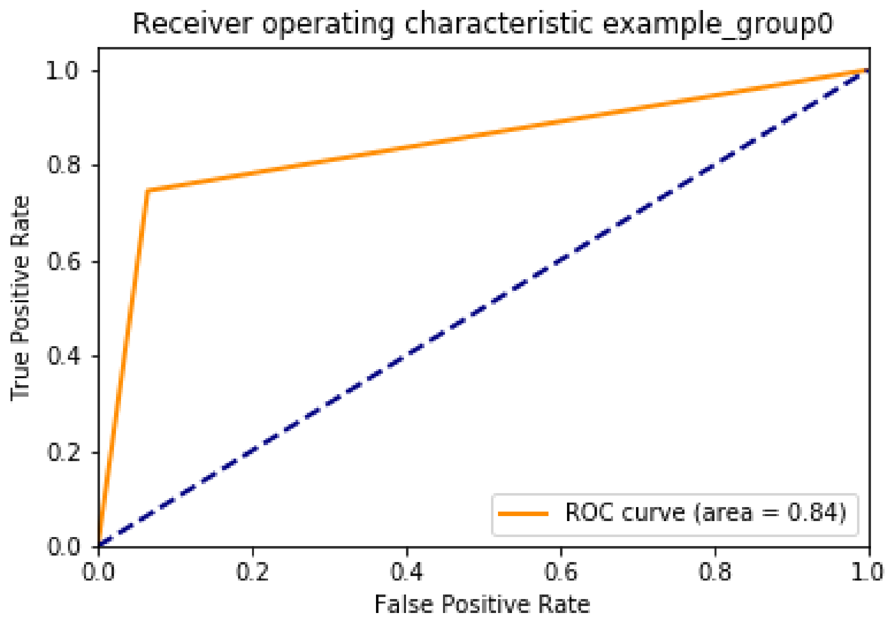
One-hot:

Curve:

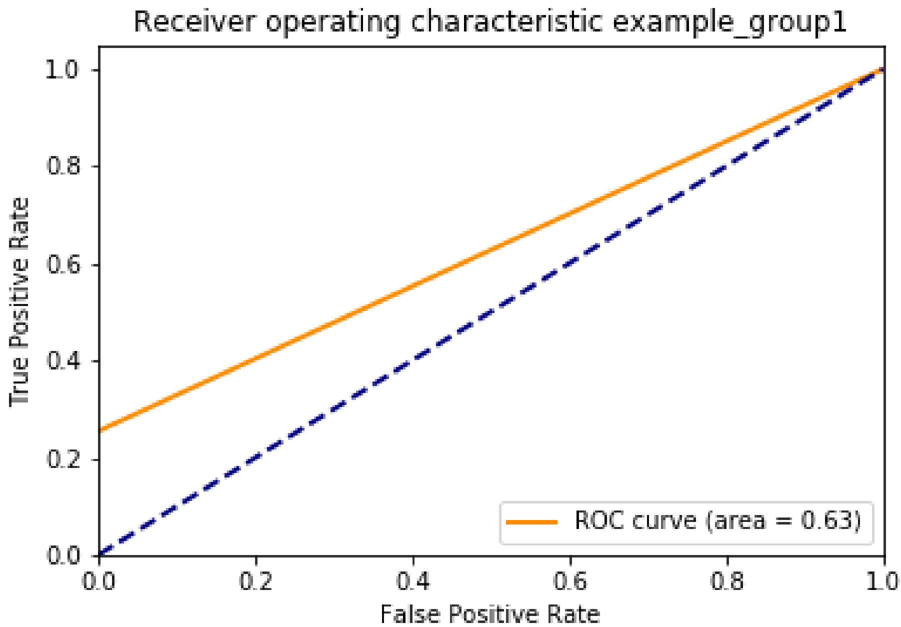


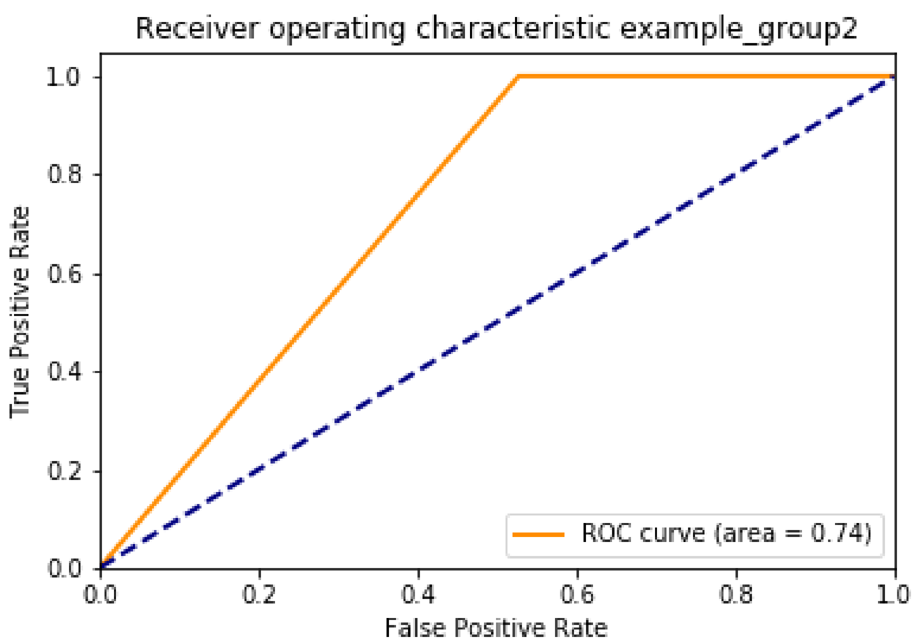
At l = 0, we get a test accuracy of **69.62**.

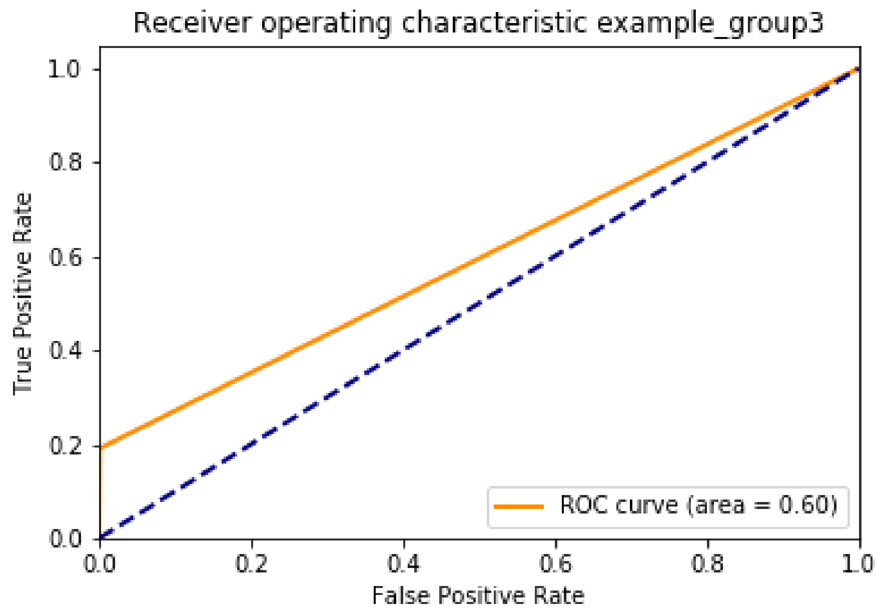
ROC Plots:



Receiver operating characteristic example_group0
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.84)


Receiver operating characteristic example_group1
True Positive Rate
False Positive Rate
ROC curve (area = 0.63)
0.0
0.2
0.4
0.6
0.8
1.0


Receiver operating characteristic example_group2
True Positive Rate
False Positive Rate
ROC curve (area = 0.74)
0.0
0.2
0.4
0.6
0.8
1.0



5.2. Analysis:

1. For Mean Square error our assumption is that the noise follows Normal Distribution, this may not be the case always. It is not always a good approach. For classification tasks it is better to use cross entropy loss.

2. Test accuracy for one-hot encoding is high. The one-hot encoding vectors of labels are uncorrelated with each other, while for scalar encoding they are correlated to some extent. But according to us the labels shouldn't be correlated, it will introduce unnecessary errors.

In one-hot, we have separate value which is predicted implies we have separate weights for each label, that means the algorithm can assign different weights to different features depending on the class. This doesn't happen in scalar encoding.

3.

RBF doesn't improve the encoding scheme, as we have principal components of dimension 1500 and reducing them 1000, we might incur in loss of information which is crucial in prediction of the right category because we are not considering the class labels of the data during the reduction process.

Basis Function	Encoding	
	Discrete Scalar	**One-Hot Vector**
$\phi(x) = x$	L=1, Test Accuracy= 60.60	L=1, Test Accuracy=90.02
RBF basisi Function	L= 0, Test Accuracy= 58.98	L= 0, Test Accuracy=69.62

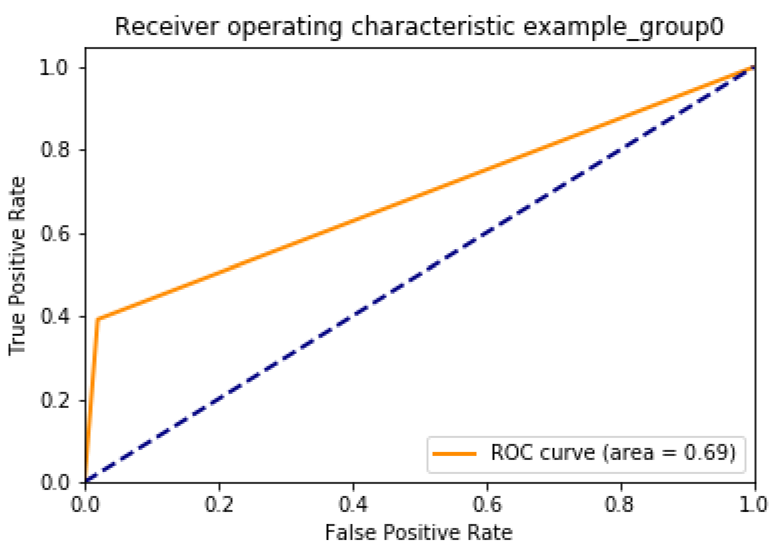
6.

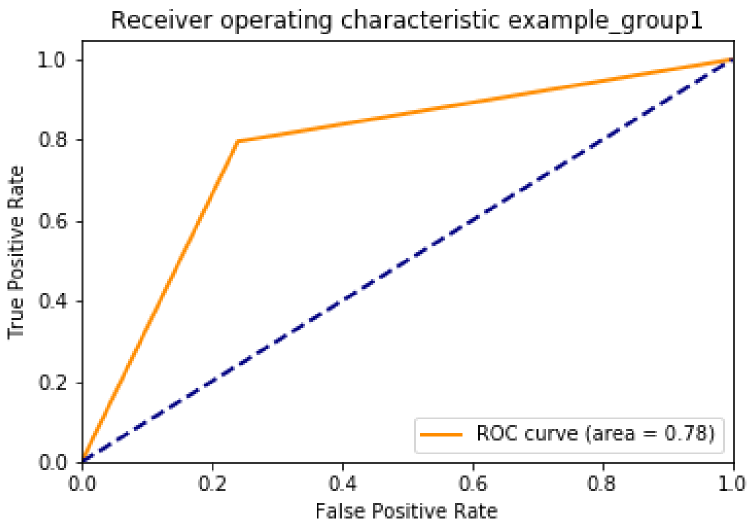
Case 0:

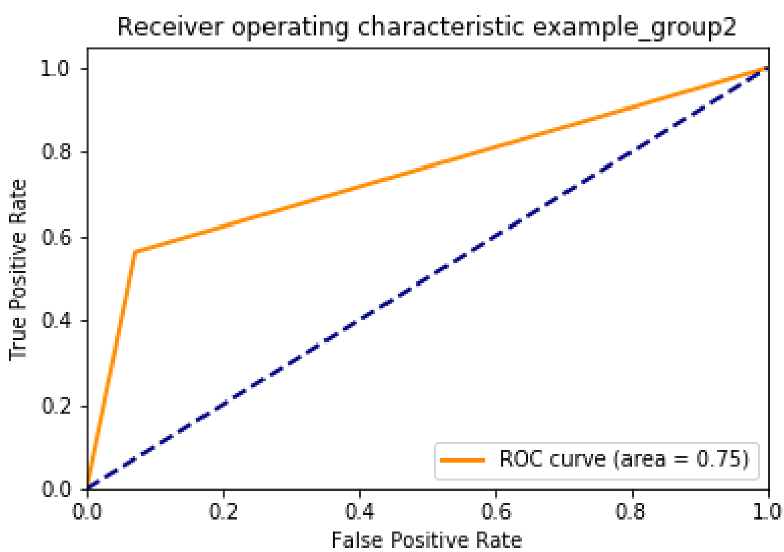
Confusion Matrix:

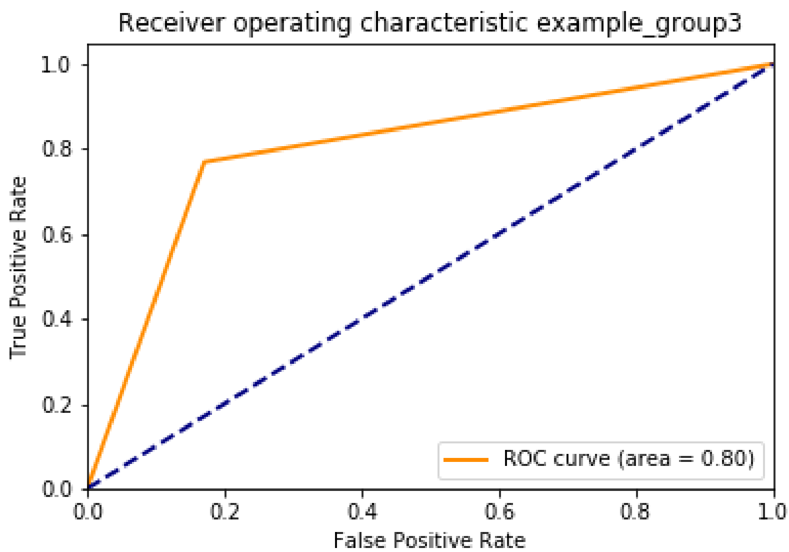
```
[[ 20, 3, 0, 28],
 [ 0, 39, 10, 0],
 [ 0, 28, 36, 0],
 [ 3, 6, 0, 30]]
```

Roc Curves:(group means class)

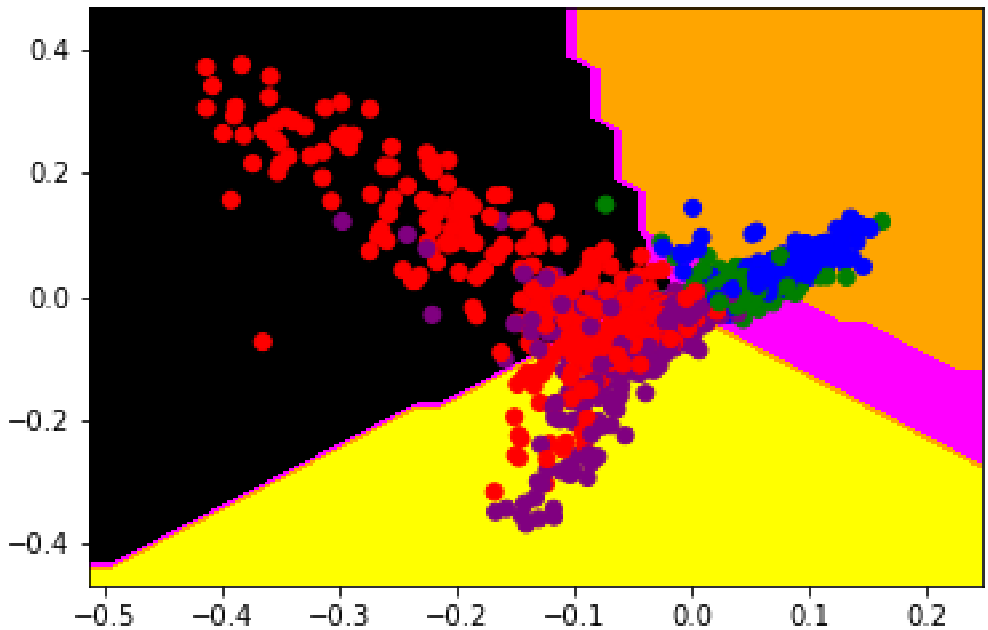



Receiver operating characteristic example_group1
True Positive Rate
False Positive Rate
ROC curve (area = 0.78)
0.0
0.2
0.4
0.6
0.8
1.0


Receiver operating characteristic example_group2
True Positive Rate
False Positive Rate
ROC curve (area = 0.75)
0.0
0.2
0.4
0.6
0.8
1.0



Decision Boundary and data points Scatter plot:



5.2. **Analysis**:

1. One disadvantage of Mean Square Error is that it is heavily effected by outliers.

Case-1:

Confusion Matrix:

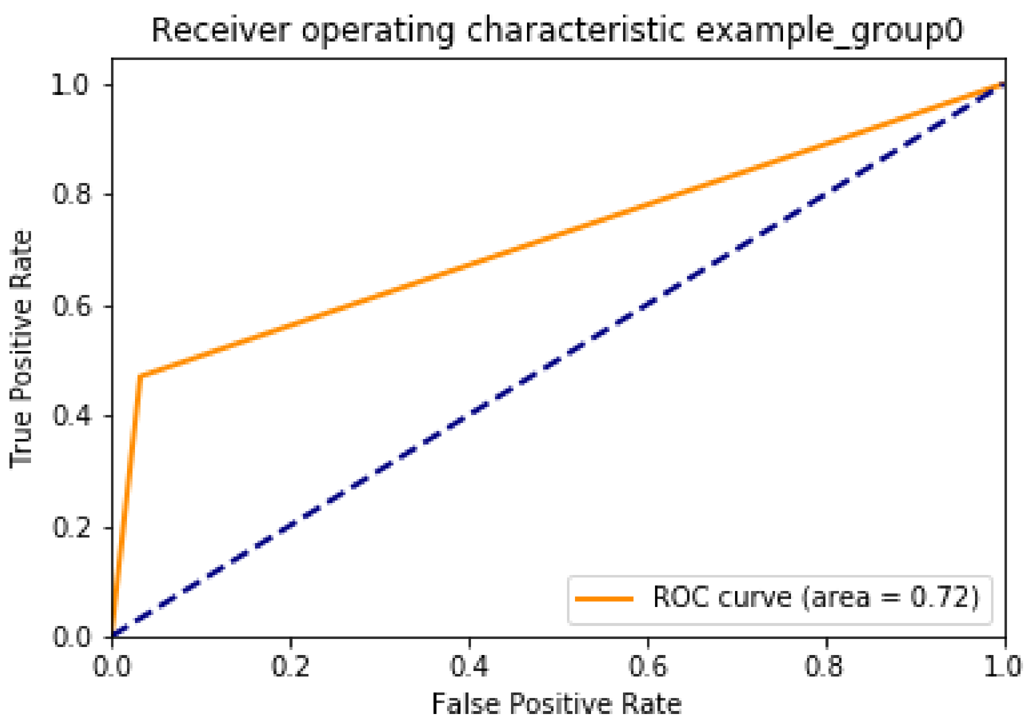
[[24 6 0 21]

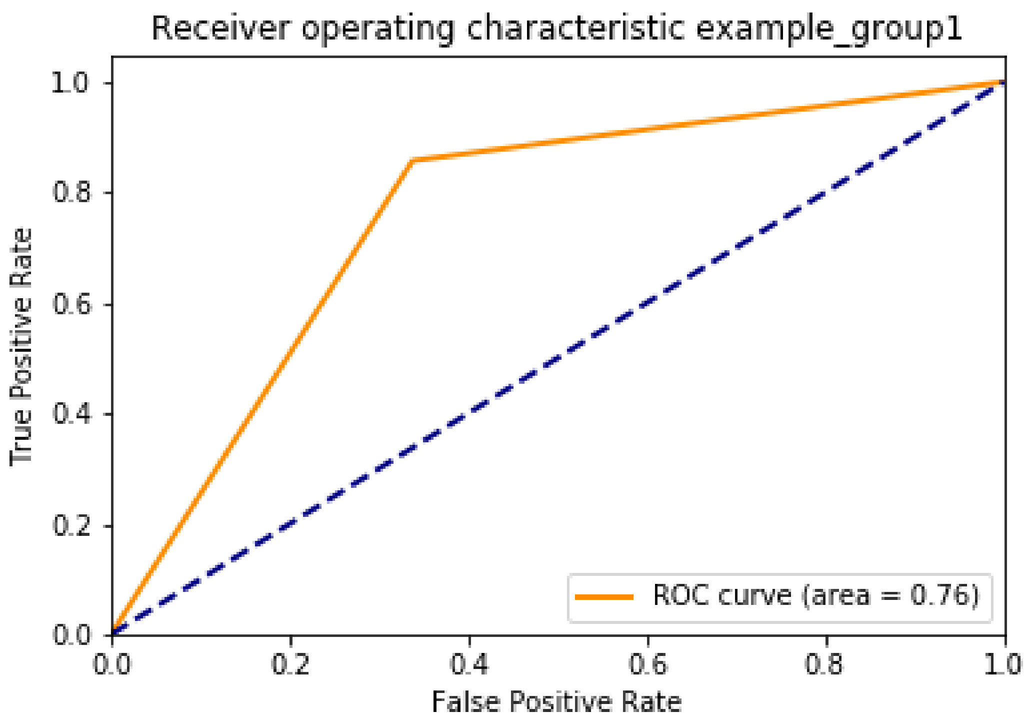
[0 42 7 0]

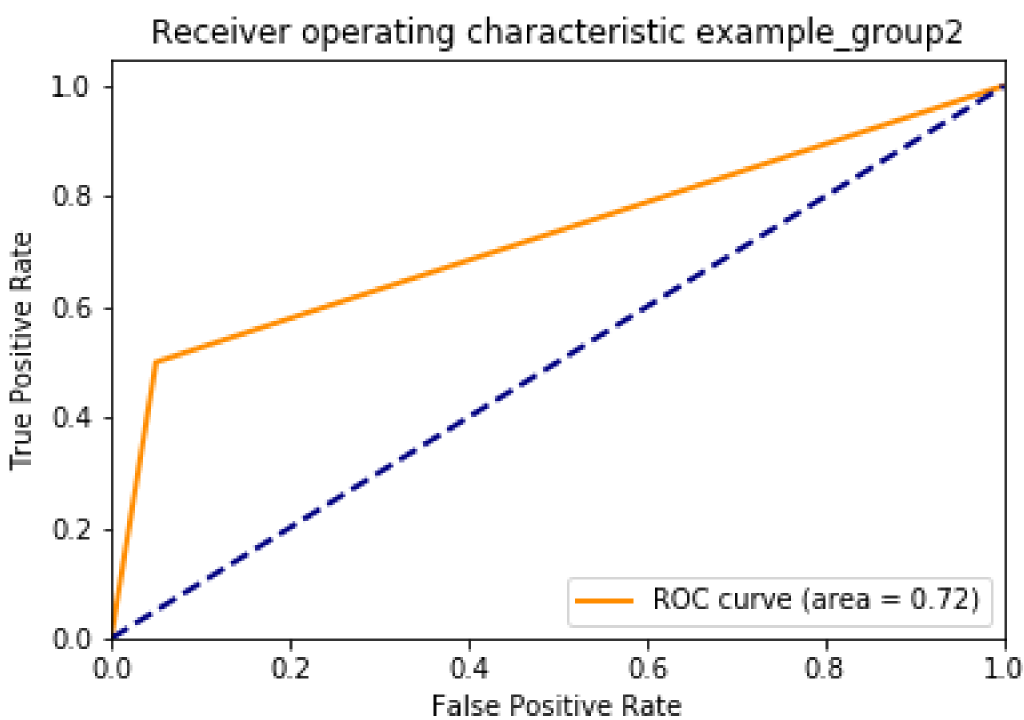
[0 32 32 0]

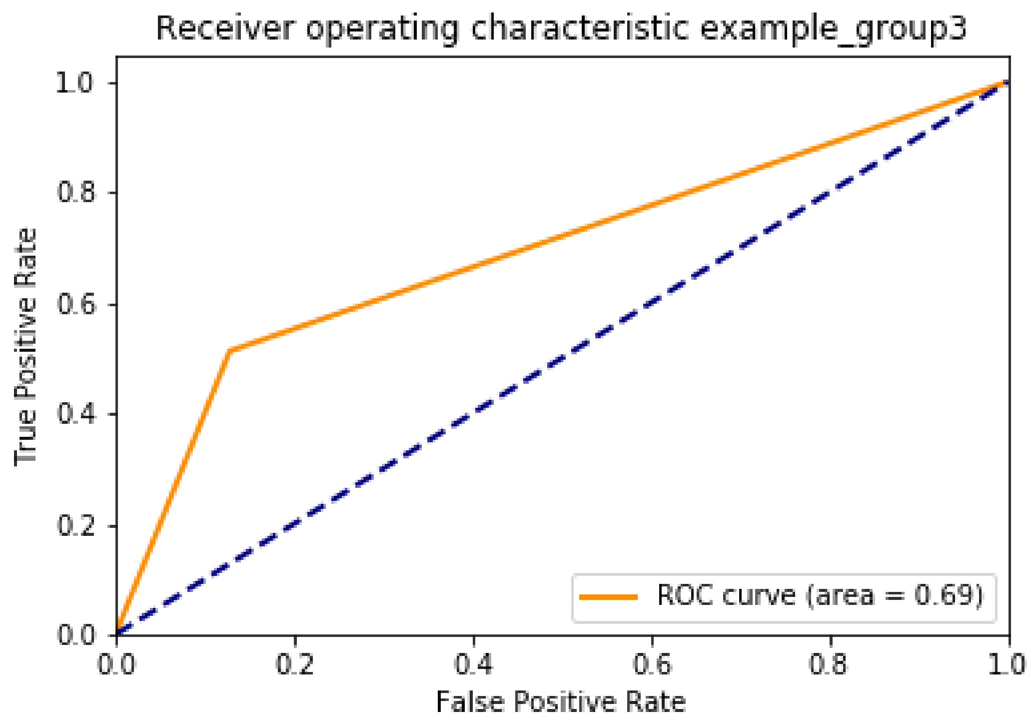
[5 14 0 20]]

ROC Curves:

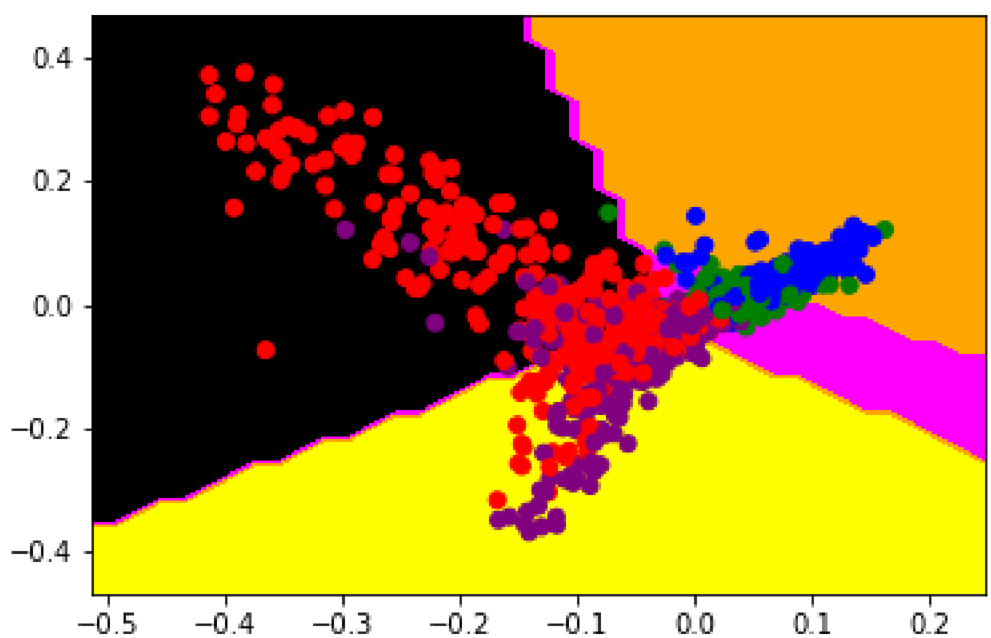



Receiver operating characteristic example_group1
True Positive Rate
False Positive Rate
ROC curve (area = 0.76)
0.0
0.2
0.4
0.6
0.8
1.0


Receiver operating characteristic example_group2
True Positive Rate
False Positive Rate
ROC curve (area = 0.72)
0.0
0.2
0.4
0.6
0.8
1.0



Decision Boundary and data points Scatter plot:



Case-2:

Confusion Matrix:

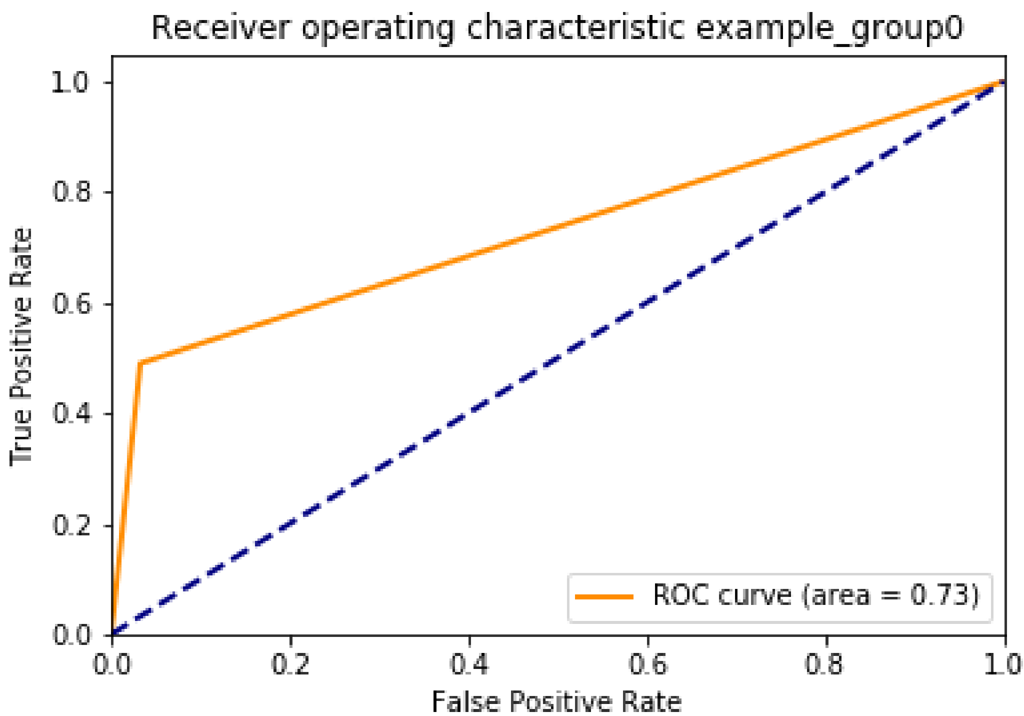
[[25 5 0 21]

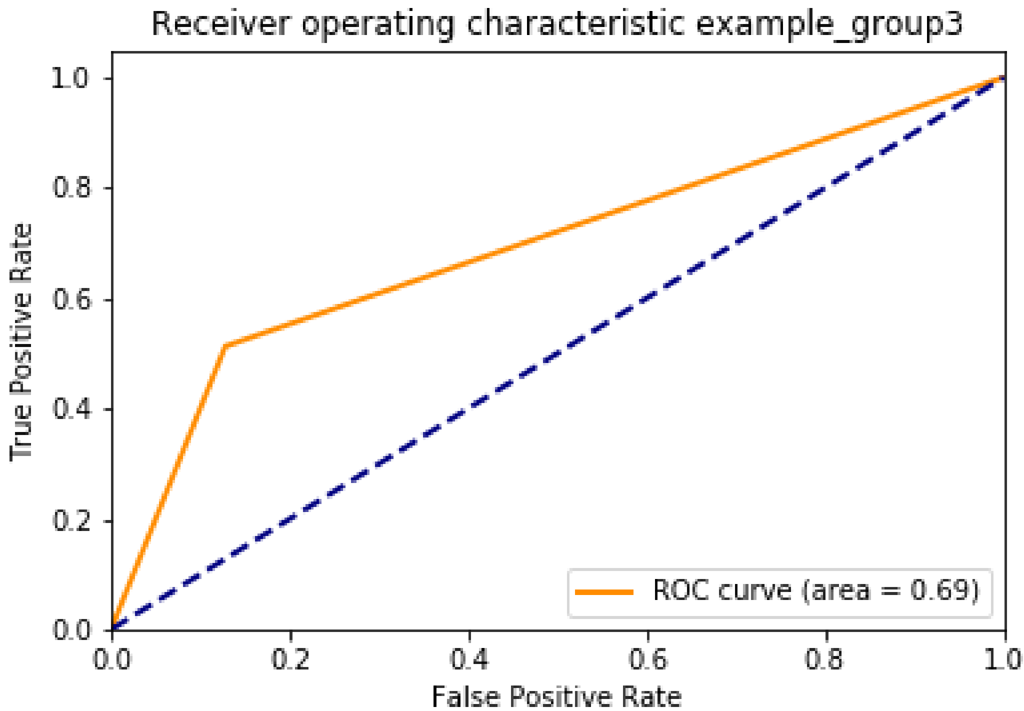
[0 42 7   0]

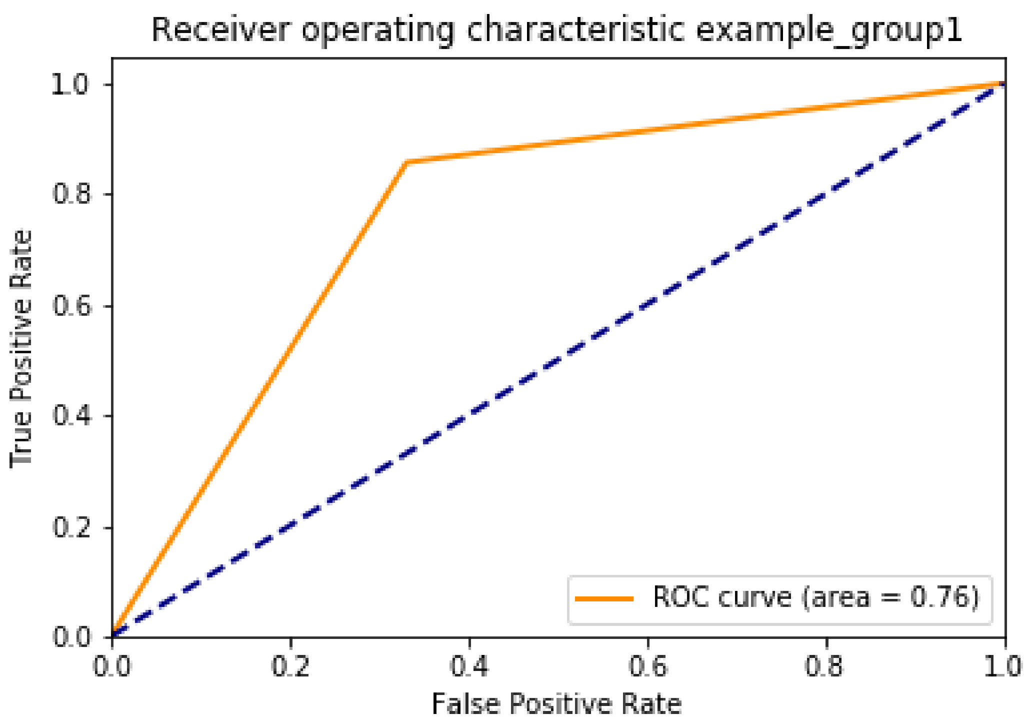
 [0 32 32 0]

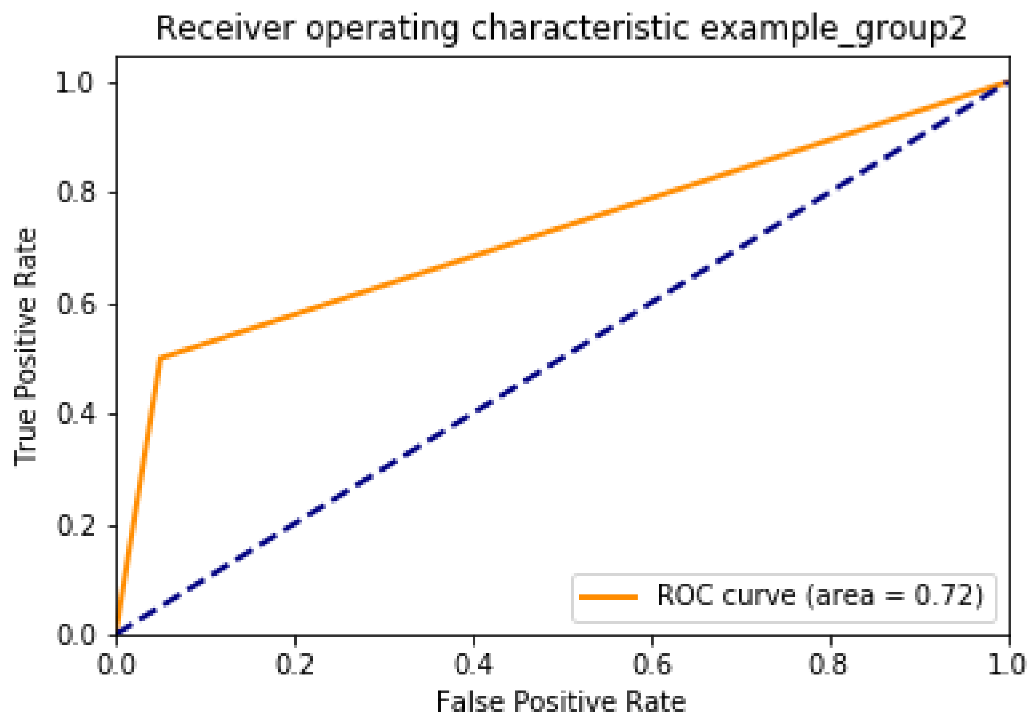
 [5 14 0 20]]

ROC Curves:

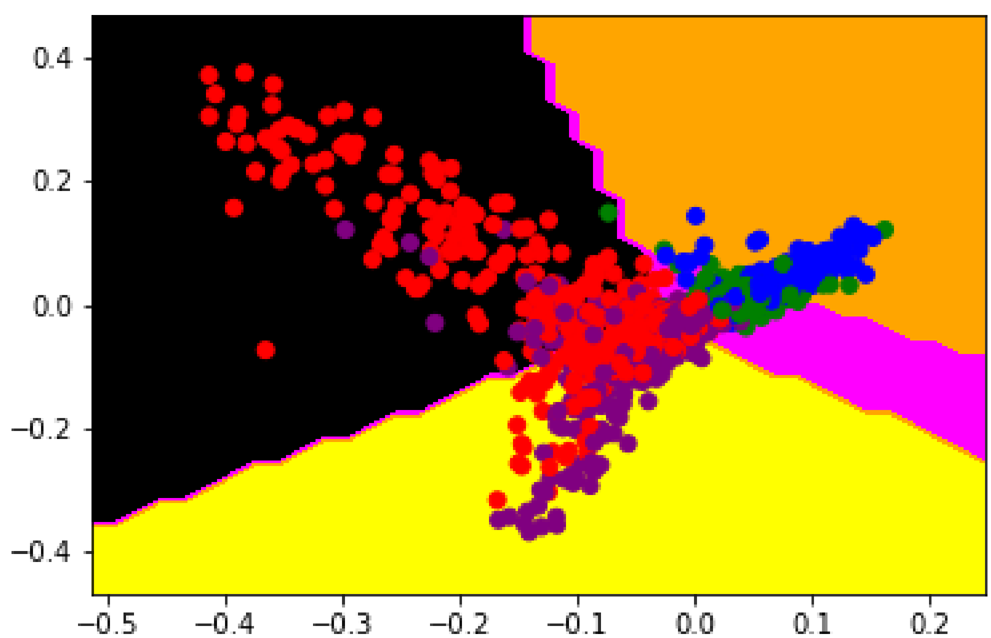



Receiver operating characteristic example_group3
True Positive Rate
False Positive Rate
ROC curve (area = 0.69)
0.0
0.2
0.4
0.6
0.8
1.0


Receiver operating characteristic example_group1
True Positive Rate
False Positive Rate
ROC curve (area = 0.76)
0.0
0.2
0.4
0.6
0.8
1.0



Decision Boundary and data points Scatter plot:



Case-3:

Confusion Matrix:

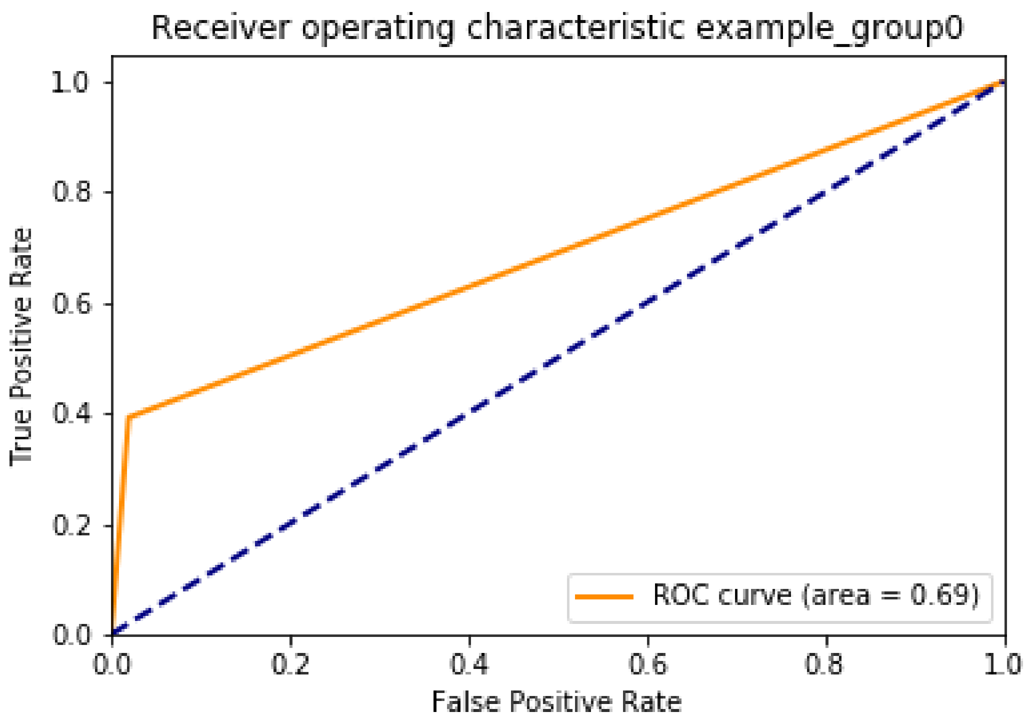
[[20 3 0 28]

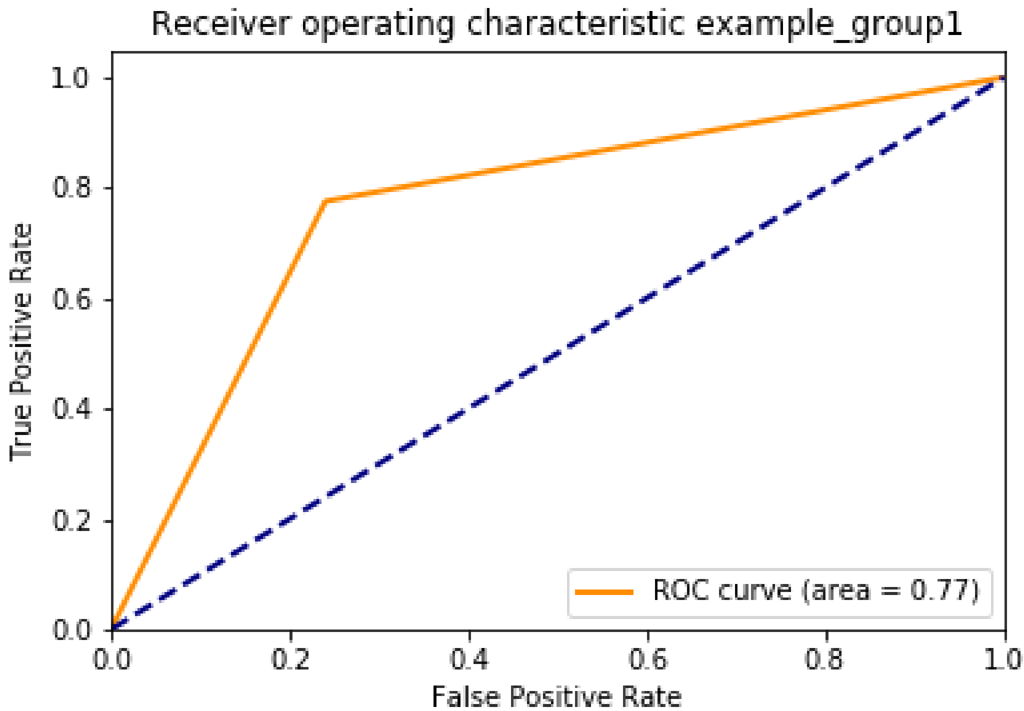
[0 38 11 0]

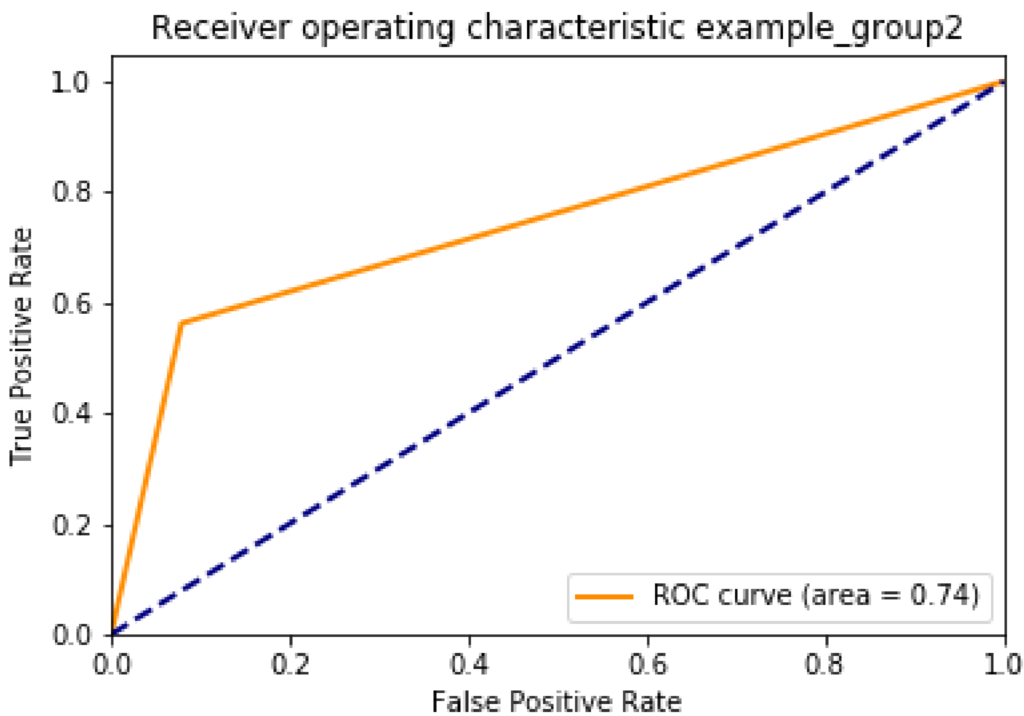
[0 28 36 0]

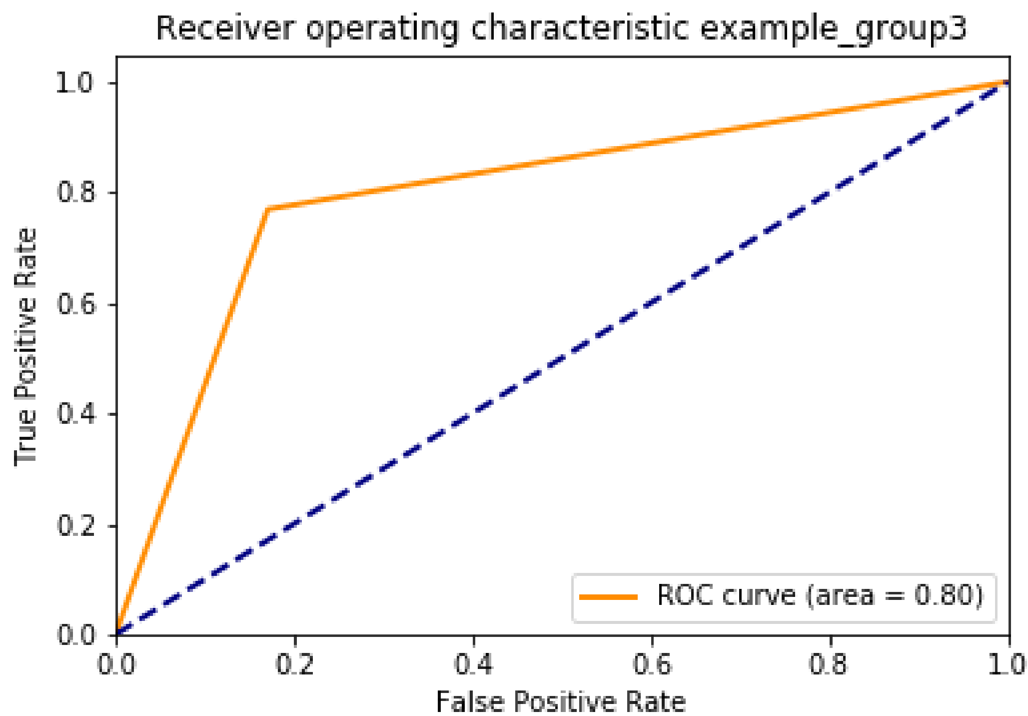
[3 6 0 30]]

ROC Curves:

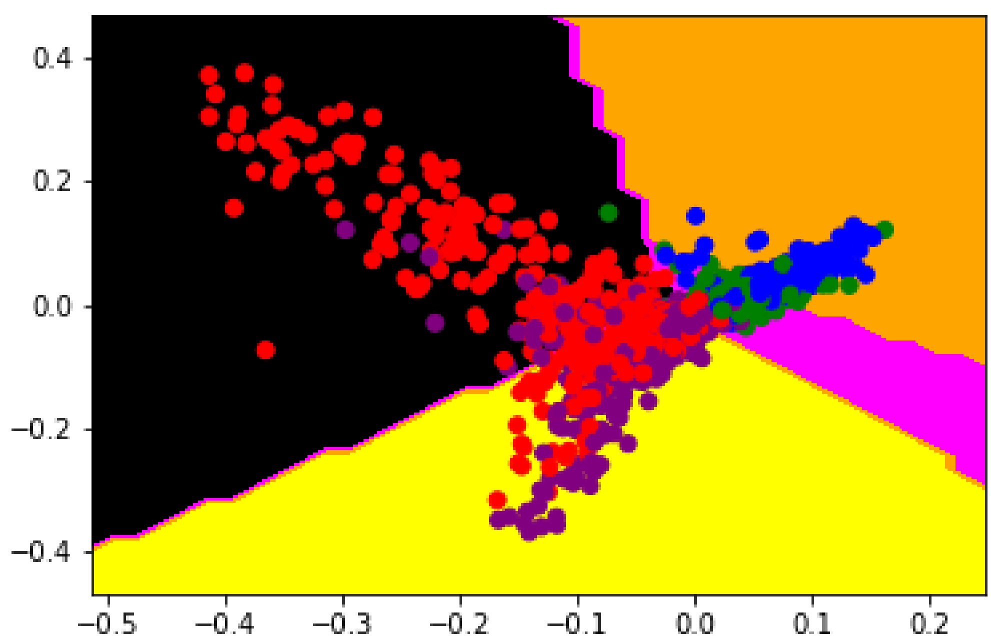



Receiver operating characteristic example_group1
True Positive Rate
False Positive Rate
ROC curve (area = 0.77)
0.0
0.2
0.4
0.6
0.8
1.0


Receiver operating characteristic example_group2
True Positive Rate
False Positive Rate
ROC curve (area = 0.74)
0.0
0.2
0.4
0.6
0.8
1.0



Decision Boundary and data points Scatter plot:



Case-4:

Threshold = 0.2

Confusion Matrix:

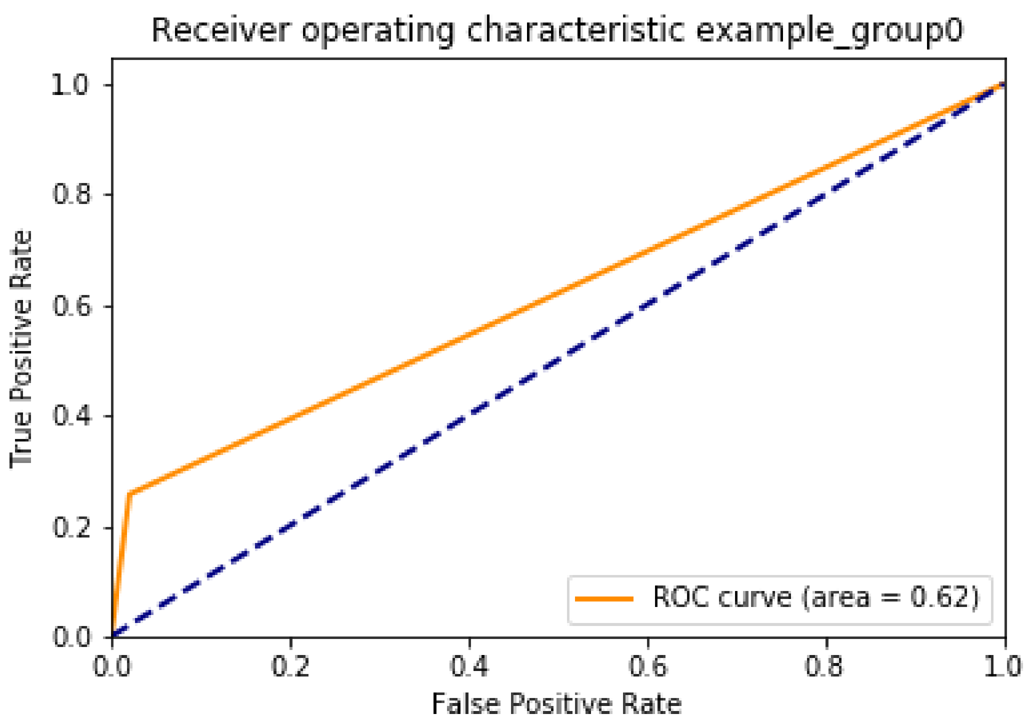
[[9 3 0 23]

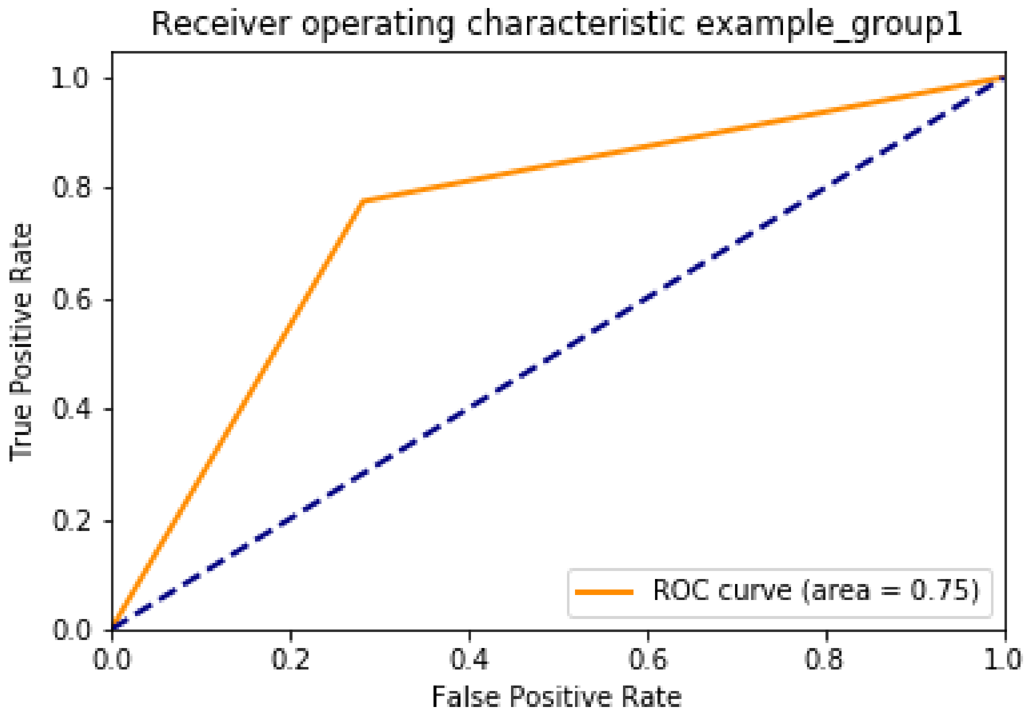
 [0 38 11 0]

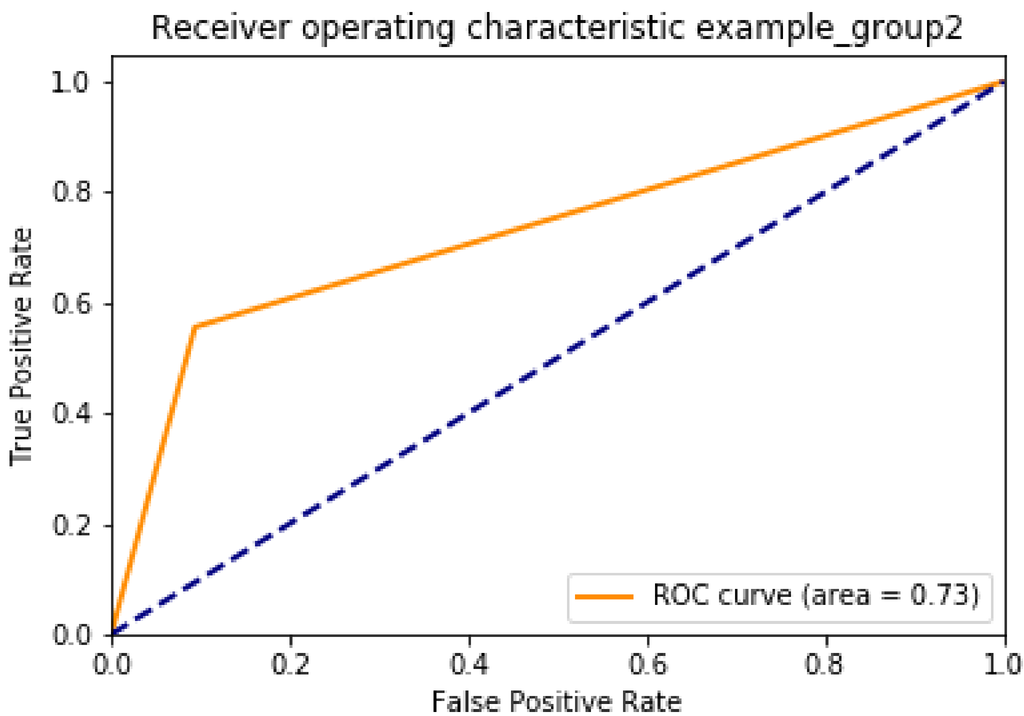
 [0 28 35 0]

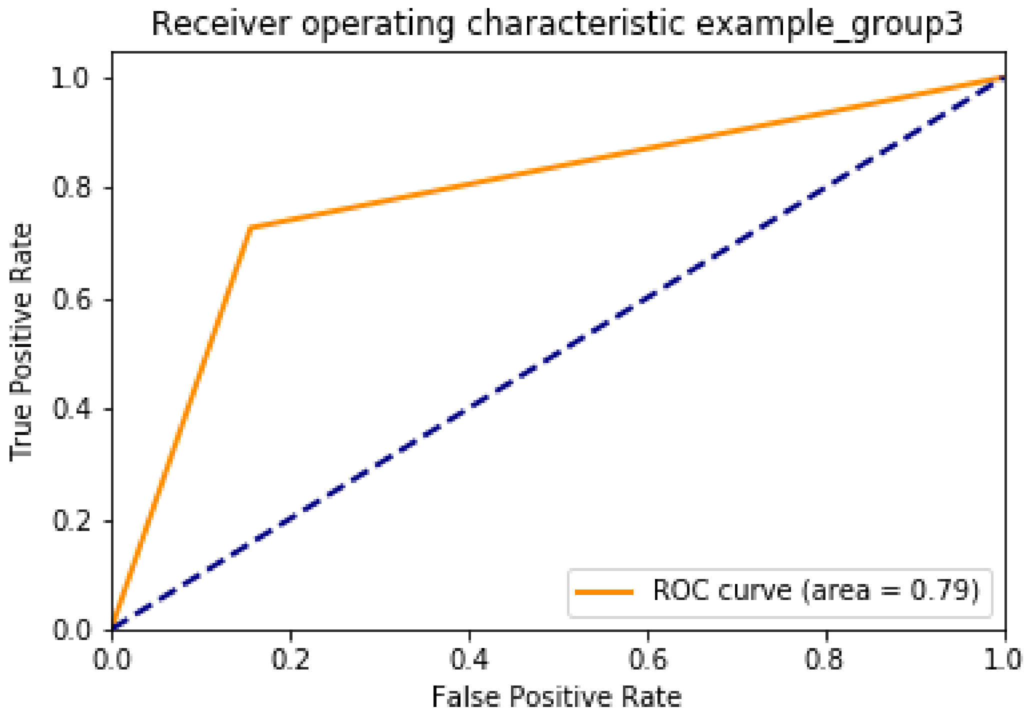
 [3 6 0 24]]

ROC Curves:

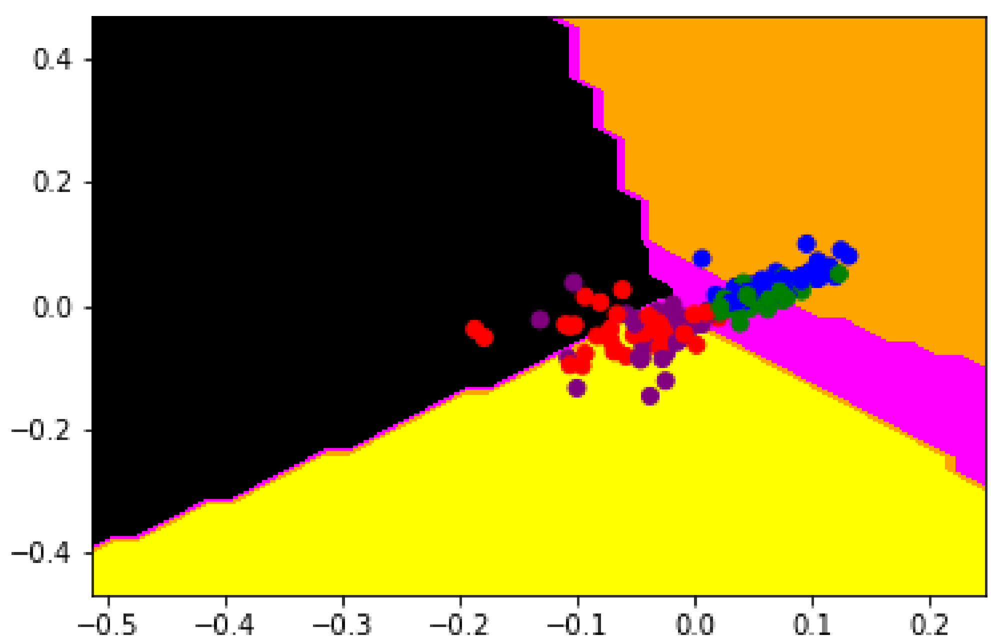



Receiver operating characteristic example_group1
True Positive Rate
False Positive Rate
ROC curve (area = 0.75)
0.0
0.2
0.4
0.6
0.8
1.0


Receiver operating characteristic example_group2
True Positive Rate
False Positive Rate
ROC curve (area = 0.73)
0.0
0.2
0.4
0.6
0.8
1.0



Decision Boundary and data points Scatter plot:



LDA

Mean, Gamma, Beta etc values are shown in the LDA_student.ipynb file

When the reduce factor is 0.005

For LDA

Percentage of Accuracy: 88.17442719881744

Number of Points predicted successfully 1193.0

For KNN

Percentage of Accuracy is 78.122691

LDA performed better compared to KNN

When the reduce factor is 0.03

For LDA

Percentage of Accuracy: 89.20916481892091

Number of Points predicted successfully 1201.0

For KNN

Percentage of Accuracy is 84.0702882483

LDA performed better compared to KNN

Accuracy Percentage is increased when the reduce factor is increased from 0.005 to 0.03