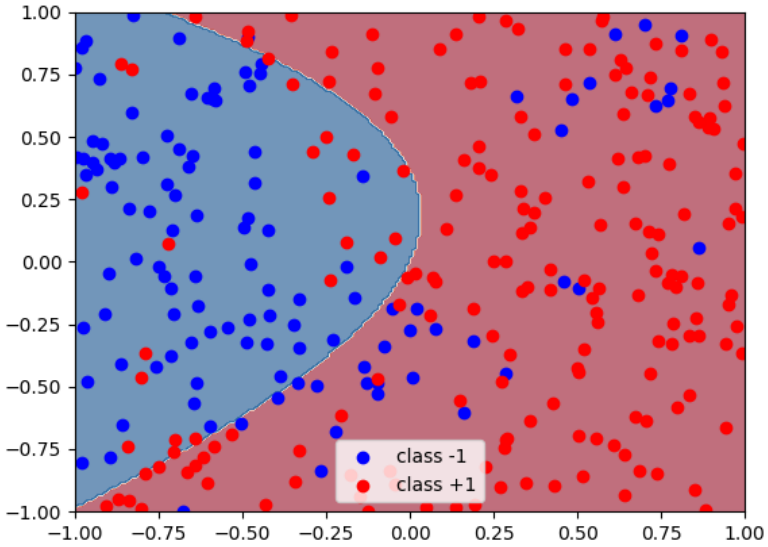
Q1

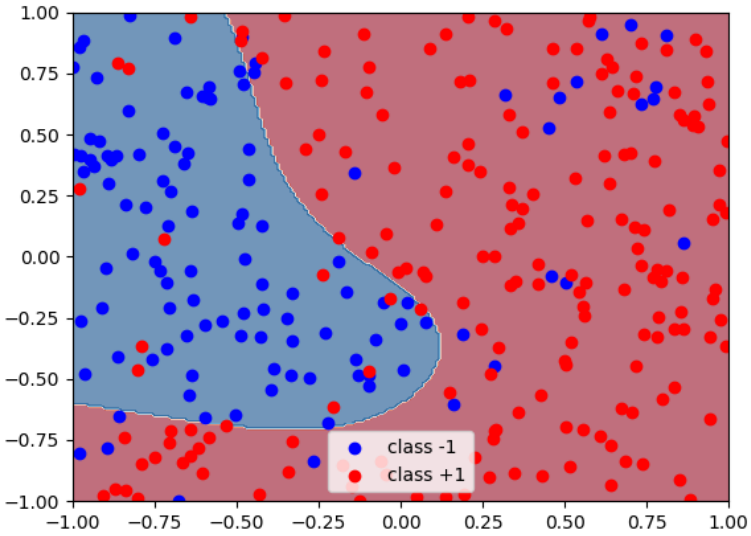
1. Polynomial:

Training error: 0.183
Validation error: 0.170



RBF:

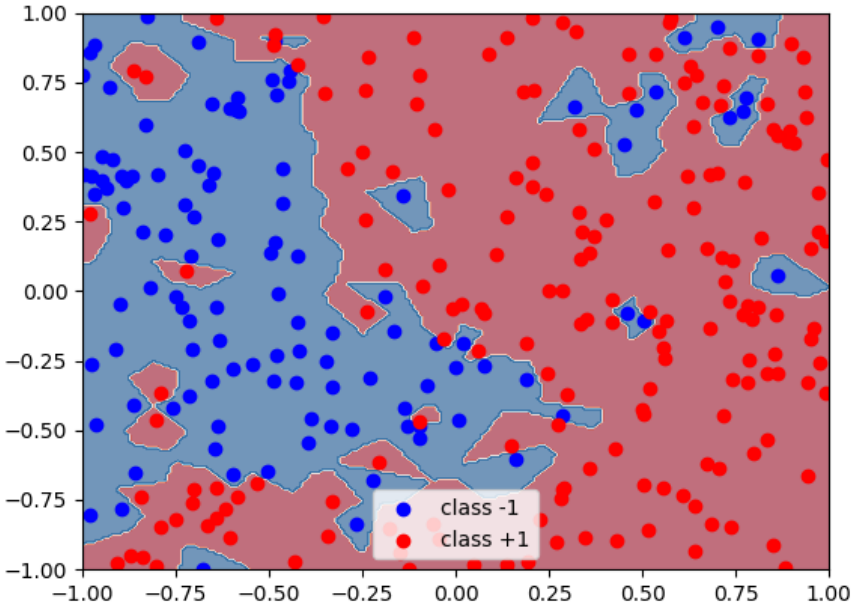
Training error: 0.127
Validation error: 0.090



2. Training error:

Sigma value: 0.01

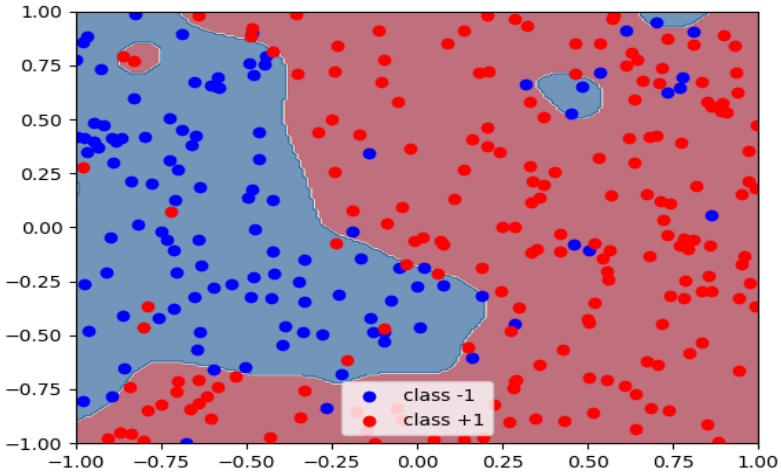
Lamda value: 0.0001



Validation error:

Sigma value: 0.1

Lamda value: 1.0



3. The training error was minimized by the least possible value of sigma and lambda which causes overfitting which is also present in the graph/plot above

   The validation error was minimized by the highest possible value of sigma and lambda which does not overfit as visible in the graph/plot above

Q2

1. $||Xw - y||_1 + \frac{1}{2\sigma^2}\left||w|\right|^2$
2. $\frac{1}{2}(Xw - y)\Sigma^{-1}(Xw - y) + \lambda||w||_1$
3. $\left(\frac{v+1}{2}\right) sum_{i=1}^{N} log(1 + \left(\frac{w^T x_i - y_i}{v}\right)^2) + \frac{\lambda}{2}\left||w|\right|^2$
4. $\sum_{i=1}^{n}(\exp(w^T x_i) - y_i w^T x_i)$

Q3

1. The first column is equal to the second column if we take mean of the second column and subtract it and then multiply the result by -2. Hence, $w = \frac{1}{\sqrt{2}}$

2. To calculate error, need to break it and form it again

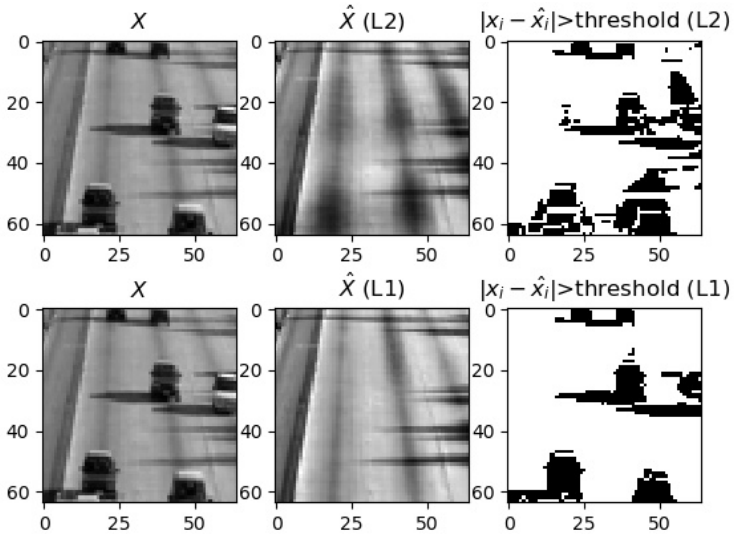
$$-3 - 0 * w + 2.5 - 1 * w$$
$$= -1.5w = -\frac{3}{2\sqrt{2}}$$
$$= -\frac{3}{2\sqrt{2}} * w + 0 = -\frac{3}{4}$$
$$= -\frac{3}{2\sqrt{2}} * w + 1 = \frac{1}{4}$$
$$= \sqrt{\left(-\frac{3}{4} - 3\right)^2 + \left(\frac{1}{4} - 2.5\right)^2} = \frac{\sqrt{306}}{4}$$

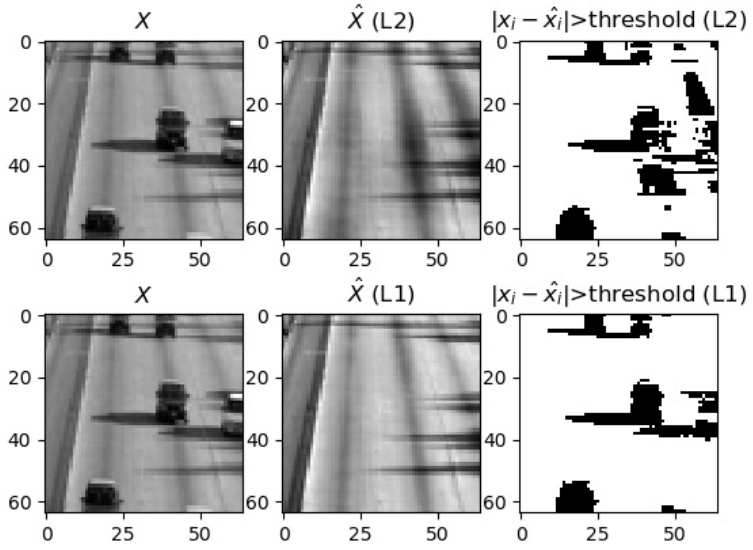
3. To calculate error, need to break it and form it again

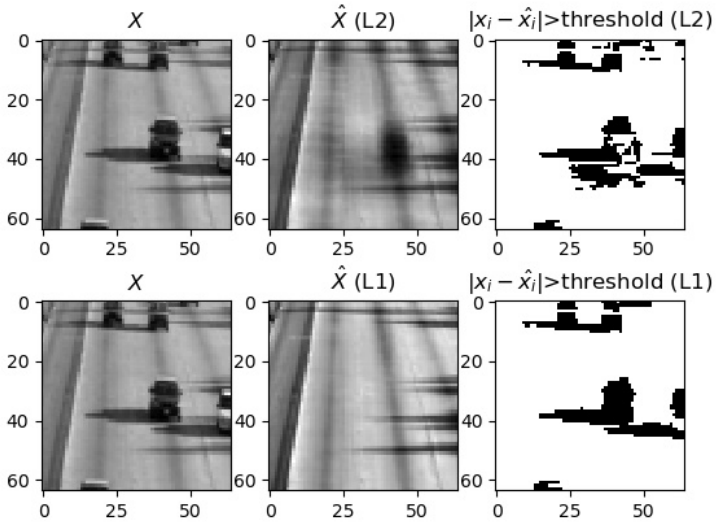
$$-3 - 0 * w + 2 - 1 * w$$
$$= -1w = -\frac{1}{\sqrt{2}}$$
$$= -\frac{1}{\sqrt{2}} * w + 0 = -\frac{1}{2}$$
$$= -\frac{1}{\sqrt{2}} * w + 1 = \frac{1}{2}$$
$$= \sqrt{\left(-\frac{1}{2} - 3\right)^2 + \left(\frac{1}{2} - 2.5\right)^2} = \frac{\sqrt{65}}{2}$$

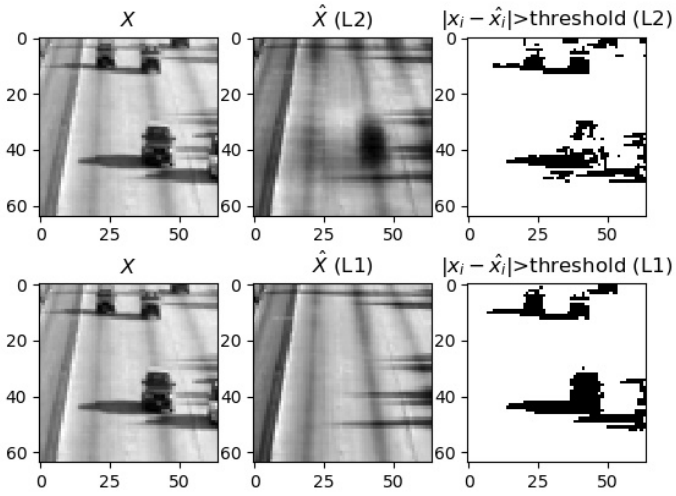
Q4

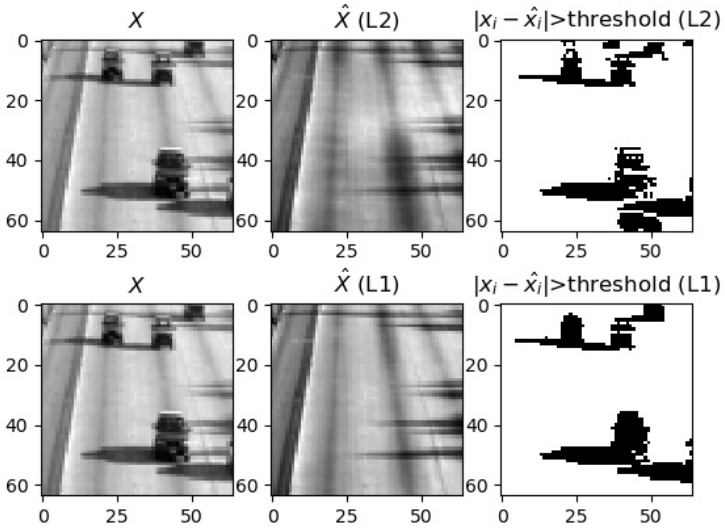
Q4.1

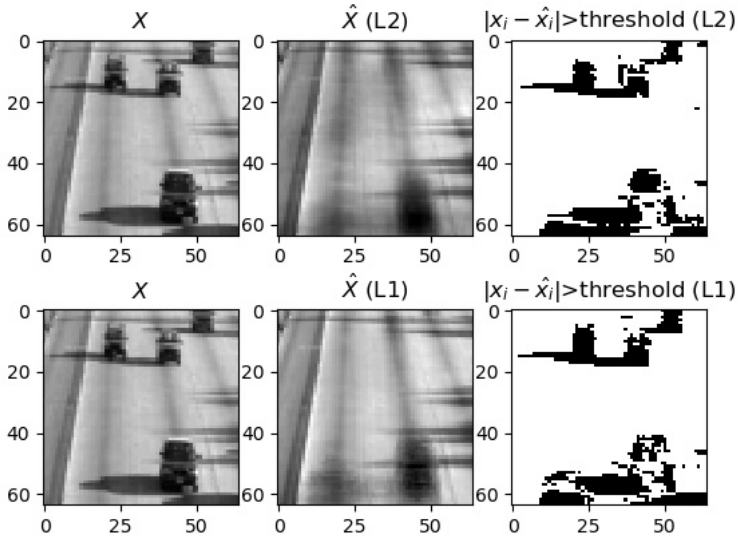

X
$\hat{X}$ (L2)
$|x_i - \hat{x}_i|$>threshold (L2)
X
$\hat{X}$ (L1)
$|x_i - \hat{x}_i|$>threshold (L1)

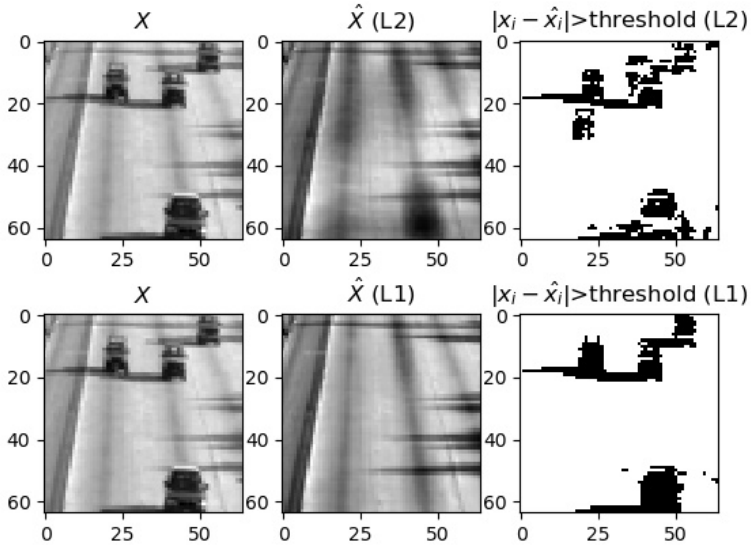

X
$\hat{X}$ (L2)
$|x_i - \hat{x}_i|$>threshold (L2)
X
$\hat{X}$ (L1)
$|x_i - \hat{x}_i|$>threshold (L1)

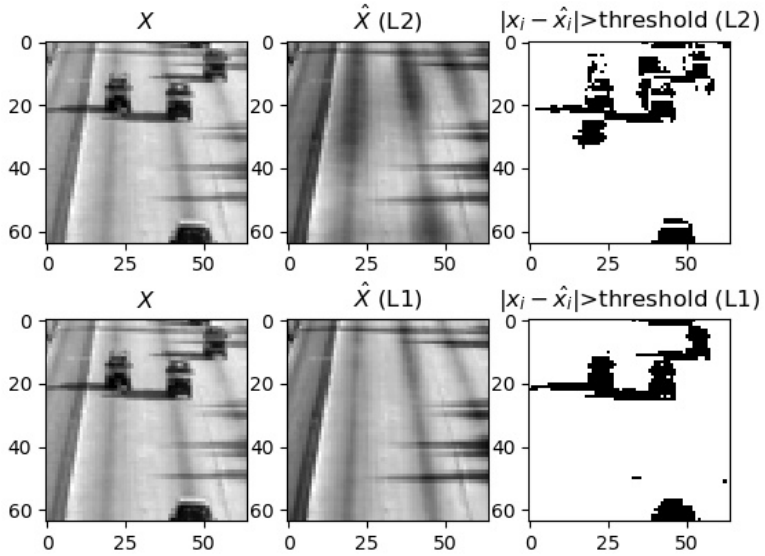

X
$\hat{X}$ (L2)
$|x_i - \hat{x}_i|$>threshold (L2)
X
$\hat{X}$ (L1)
$|x_i - \hat{x}_i|$>threshold (L1)

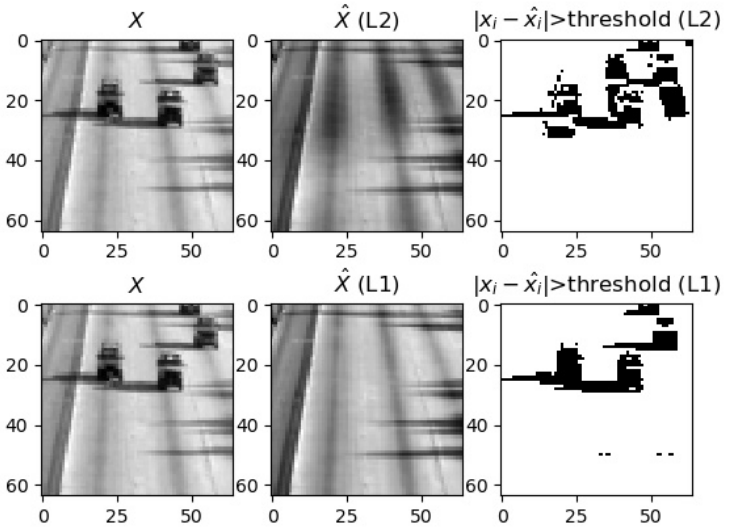

X
$\hat{X}$ (L2)
$|x_i - \hat{x}_i|$>threshold (L2)
X
$\hat{X}$ (L1)
$|x_i - \hat{x}_i|$>threshold (L1)

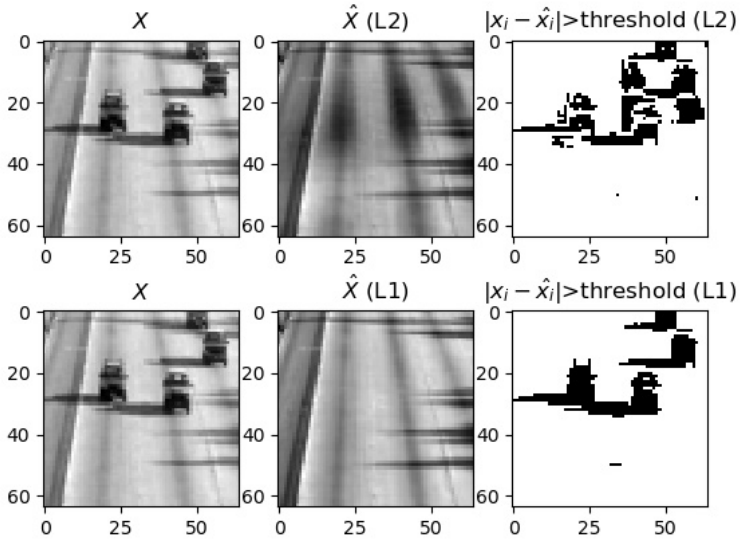

X
$\hat{X}$ (L2)
$|x_i - \hat{x}_i|$>threshold (L2)
X
$\hat{X}$ (L1)
$|x_i - \hat{x}_i|$>threshold (L1)


X
$\hat{X}$ (L2)
$|x_i - \hat{x}_i|$>threshold (L2)
X
$\hat{X}$ (L1)
$|x_i - \hat{x}_i|$>threshold (L1)


X
$\hat{X}$ (L2)
$|x_i - \hat{x}_i|$>threshold (L2)
X
$\hat{X}$ (L1)
$|x_i - \hat{x}_i|$>threshold (L1)


X
$\hat{X}$ (L2)
$|x_i - \hat{x}_i|$>threshold (L2)
X
$\hat{X}$ (L1)
$|x_i - \hat{x}_i|$>threshold (L1)





Q4.2

1. L1 loss is more suitable for this task compared to L2 as this produces a sparse matrix that is, it has many zero entries

2. N represents the rows of the frames, d represents columns of the frame and k represents the size of each frame

3. If the threshold is changed, then it might result in the L1 loss wrongly recognizing objects in the frame and removing them

Q5

1. An advantage of using other normal equations is that we can work with constraints of non-negativity

2. Advantages of using kernels in K-means clustering is that the model would identify the non-linear structures and that the model is better for real life data

3. MLE finds the w that gives the largest conditional probability of our data given w. MAP finds the w with the largest conditional probability given our data. MAP has regularization in the form of a prior, and is equivalent to MLE with a flat prior

4. A generative model learns the probability distribution p(x,y) while a discriminative model learns p(y,x)

5. Yes, it is possible to increase the loss if k is increased, as more dimensions are taken into account

6. Label switching in context of PCA means that the span of n vectors would be same if the vectors were switched

7. It does not make sense to do PCA with k>d because PCA creates a subspace in k dimension from the original d dimension hence, we cannot create an additional dimension which is inferred with k>d

8. They would be stored in W matrix

9. Advantage of using stochastic gradient is that it is faster as only one random example whereas a disadvantage is that SVD gets rid of redundant data

10. Constant value does not converge to a stationary point due to the erratic behavior hence, (a), (b), (c) converge to a stationary point

11. An advantage of using global features is that they refer to the entire dataset, typically refereeing to large senders and an advantage of using local features is that they mimic behavior of specific user.