CS-337/335 || Assignment-2

Niraj Mahajan
Roll No- 180050069

1 Lasso and ISTA:

1.1 Assume the weights w_i $\forall$ $i \in [0, d)$ be IID from Laplace Distribution, with 0 mean.

$$P(w_i/\sigma) = \frac{1}{2\sigma} \cdot e^{\frac{-|w_i|}{\sigma}}$$

$$\therefore P(w/\sigma) = \prod_i P(w_i/\sigma) = \frac{1}{(2\sigma)^d} e^{-\frac{\sum_{i=0}^{d} |w_i|}{\sigma}}$$

$$\boxed{P(w/\sigma) = \frac{1}{(2\sigma)^d} e^{\frac{-\|w\|_1}{\sigma}}} \quad \text{— (1)}$$

Now, For $P(D/w)$,

→ we know for any sample x_i, y_i

$$P(y_i/x_i, w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{[y_i - w^T\phi(x_i)]^2}{2\sigma^2}\right)}$$

Hence, $$\boxed{P(D/w) = \prod_{i=0}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[y_i - w^T\phi(x_i)]^2}{2\sigma^2}\right)} \quad \text{— (2)}$$

Hence, our Posterior.

$$P(w|D) \propto P(D|w) \cdot P(w)$$

Hence forth, we will ignore constants.

$$P(w|D) \propto \exp\left(\sum_{i=0}^{n} \left\{\frac{[y_i - w^T\phi(x_i)]^2}{2\sigma^2}\right\}\right) \cdot \exp\left(\frac{-\|w\|_1}{\sigma}\right)$$

$$P(w|D,\sigma) \propto \exp\left(-\sum_{i=0}^{n} \left\{\frac{[y_i - w^T\phi(x_i)]^2}{2\sigma^2}\right\} - \frac{\|w\|_1}{\sigma}\right).$$

Taking the log-likelihood:-

$$LL(D|w,\sigma) = -\sum_{i=0}^{n} \left\{\frac{[y_i - w^T\phi(x_i)]^2}{2\sigma^2}\right\} - \frac{\|w\|_1}{\sigma}$$

Hence our MAP estimate $\hat{w} = \underset{w}{\operatorname{argmax}}\ LL(D|w,\sigma)$

$$= \underset{w}{\operatorname{argmin}}\ -LL(D|w,\sigma).$$

$$\therefore \hat{w} = \underset{w}{\operatorname{argmin}} \sum_{i=0}^{n} \left\{[y_i - w^T\phi(x_i)]^2\right\} + \lambda\|w\|_1$$

This is exactly what's LASSO Regression.

classmate

Date

Page

Q 1.2] Plots for partb, part C attached at the end. (Thanks for your cooperation :))

b)

Trends observed in train/test plot:- (Train MSE vs lambda OR Test MSE vs lambda)

→ Decreases till $\lambda = 0.2$, then increases, then remains constant.

⇒ Explanation for Test MSE.

↳ Increase in λ for Lasso leads to w_i's becoming '0'. Hence, initially, the Test MSE decreases as the w_i's that were resulting into overfitting are ~~get~~ forced to '0'.

→ Then, the Test MSE increases as after a moment, excess number of w_i's are forced to zero, leading to underfitting.

→ Then, after a particular lambda, all w_i's become '0' leading to constant Test MSE.

~~⇒ For Train MSE, the initial degree of freedom of the model is so high that the~~

⇒ For Train MSE, the initial decrease in MSE due to increase in lambda can be explained as follows.

→ Initially, the degree of freedom of the model is very high and the iterations are not sufficient to train the data. As lambda increases, the w_i's get forced to zero and hence [This can be proved by re plotting the graphs with increased max_iter

⇒ The explanations for the increase and the subsequent plateau of the train MSE is the same as that for Test MSE

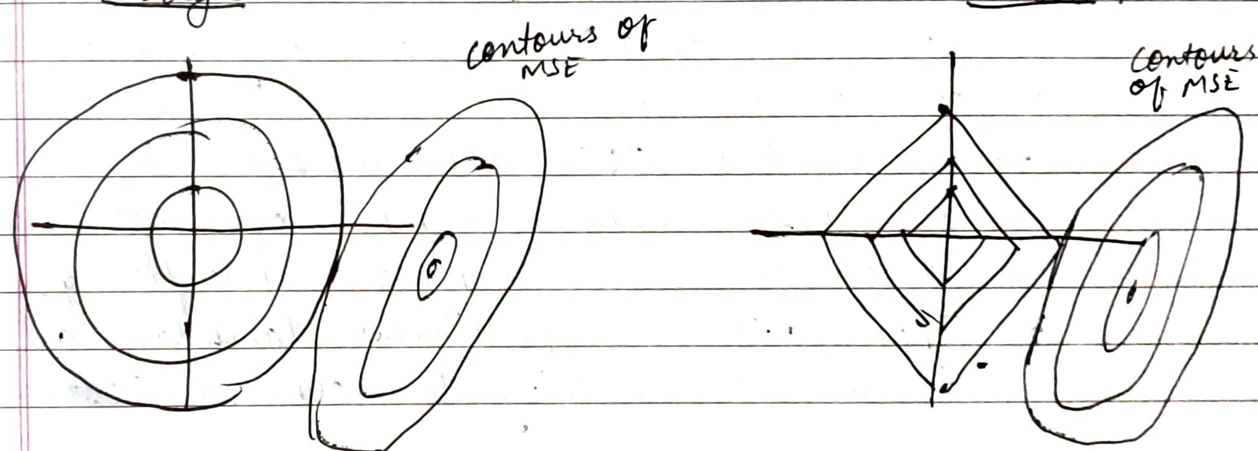
Q-1. 5]

c) · Observation:- Most of the values of w_i's in Lass Regression are '0' while in Ridge Regression, most of the values are non zero (though they are near 0).

⇒ To enplain this, consider Lasso and Ridge $\Omega(w)$ functions in a 2-d space.

Plotting $\Omega(w) = C$.

Ridge   Lasso.



As we can see, since the Lass plots are diamonds in the n-D space, it is most likely that the MSE contours first touch the Lasso at the corners, (where w_i's are zero). But, since Ridge Regression contours are hyper spheres, the MSE contours are most likely to touch ~~at~~ tangentially at some other point.

→ Hence, the weights in the Lasso solution are mostly 0s but the weights in the ~~MSE~~ Ridge Solution are, although small numbers, non zero.

Q 2.1] In one vs one perceptron, we have a classifying plane for each pair of classes, while, in case of one vs many perceptron, we have a single classifying plane for each class.

1 vs rest.

Advantages.

↳ Computationally efficient as compared to 1 vs 1 as only 'n' planes are to be trained.

~~Easier~~

Disadvantages.

↳ Consider the following Distribution:
Here each circle is a cluster of points belonging to class c_i.
It is clear that a single plane will not be able to seperate C_5 from the rest classes.

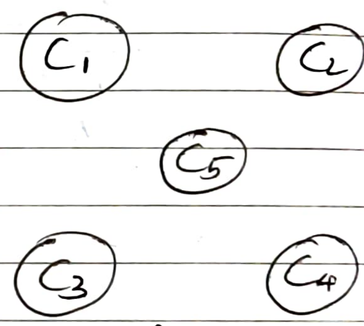


fig -1.

→ 1 vs 1

Advantages.

↳ If pairwise classes are linearly separable, then this method guarantees good results, irrespective of the distribution of the clusters in the subspace.

Disadvantages.

↳ Computationally costly as ${}^{n}C_2$ planes are to be trained.

↳ Even after training, for a input x, and class 'c', we have $(n-1)$ scores corresponding to whether x belongs to class c ($\forall c \in C_i$'s) {$(n-1)$ ~~because one~~ corresponding to each 1 vs 1 planes}

Q. 3.[1]

Definations:-

Bias :- Error arrising due to erroneous assumption made in model hyperparameters and architecture.

Variance :- Variability in the test error arising due to small variations in train data.

(Source :- Wikipedia)

a) Increase in λ will force more weights to 0 in Lasso Regression. Hence the degree of polynomial will decrease.

Hence, Increase in λ will lead to:

→ Increase in Bias.

→ Decrease in Variance.

b) Adding higher number of training examples in perceptron:-

→ No change in Bias.

→ Decrease in variance.

Reason:- If we look at the defination of bias, no "assumption" or "hyperparameter" is changing on adding more test data. Hence, there is no reason for bias to change.

On adding more data, Using Law of Large Numbers our estimated solution will start coinciding with the true solution and hence, variability in the test accuracy will decrease as our prediction become more and more ideal. Hence, there is a decrease in variance.

Date
Page

c) • Adding more featurs (Redundant).
→ No change in Bias.
→ No change in Variance.
Assuming :- Data is normalised.
→ On adding redundant dimensions in Lasso, the algorithm will reduce the weight of the ~~corresp~~ redundant dimension to 0. Hence the solution obtained before and after are the exact same. Hence, there is no change in bias or variance.

Q.3.2] Plots attached to the end. Thank you for your co-operation :) !

a) → Plot 1 on test error vs number of samples.
As explained in par 3.1] b, Adding more training examples will decrease variance but have no change on bias ~~Lasso~~. ~~(Because~~
Reason in [3.1 b].

Hence, we can see that the test error is decreasing upto some extent.
But after a point, the variance has decreased greatly and hence the variation in test error plateaus down.

3.2] b) Plot at the end (.. — ' — —) :)

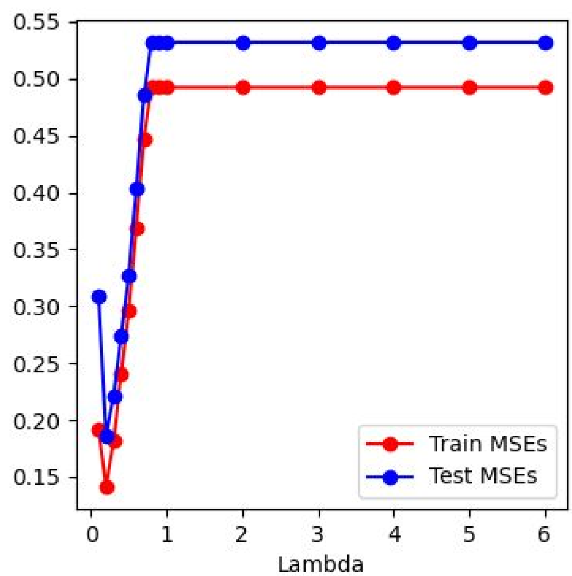
⇒ Theoritically, the least train error should be obtained on degree = 6, because we are giving (largest possible degree)
our model more and more features to overfit on.
(And indeed this is observed in the plots).

⇒ On analysing the plots, we see that the test error decreases till degree = 5, and then increases. Hence degree = 5 is the optimal degree. Here, as we increas degree, we are ~~do~~ reducing bias and increasing variance, and at degree = 5 we attain a manima. After degree = 5, the model starts overfitting.

⇒ I also have added plots of how the ~~data~~ model fits on the train data for various degrees. ~~the optimal is~~. As we can see, the fit gets better and better as the degree increases.

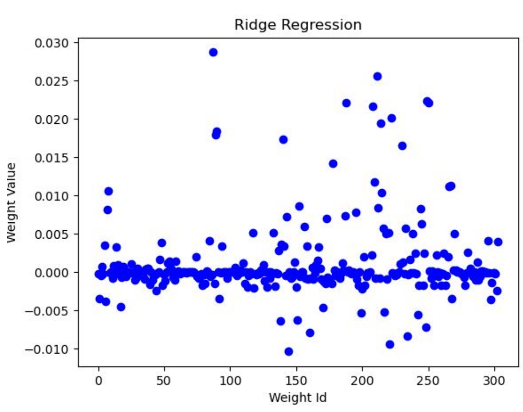
Assignment 2 Graphs: Niraj Mahajan: 180050069

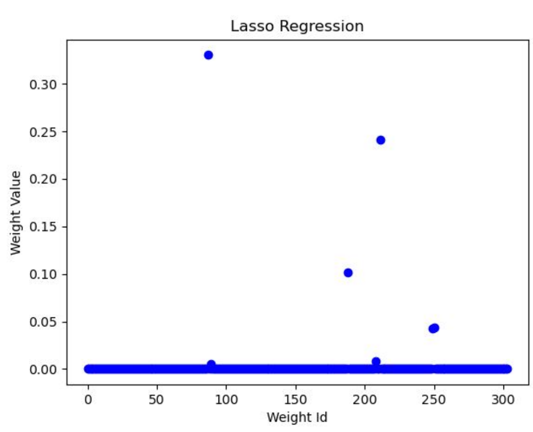
Q1.2 b)



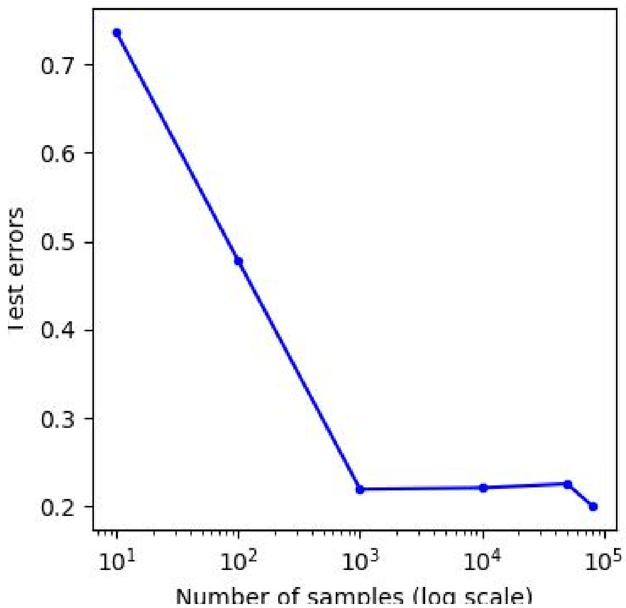
The optimal lambda in the above figure is 0.2. The explanation is provided in the written in the scans.

Q1.2 c]



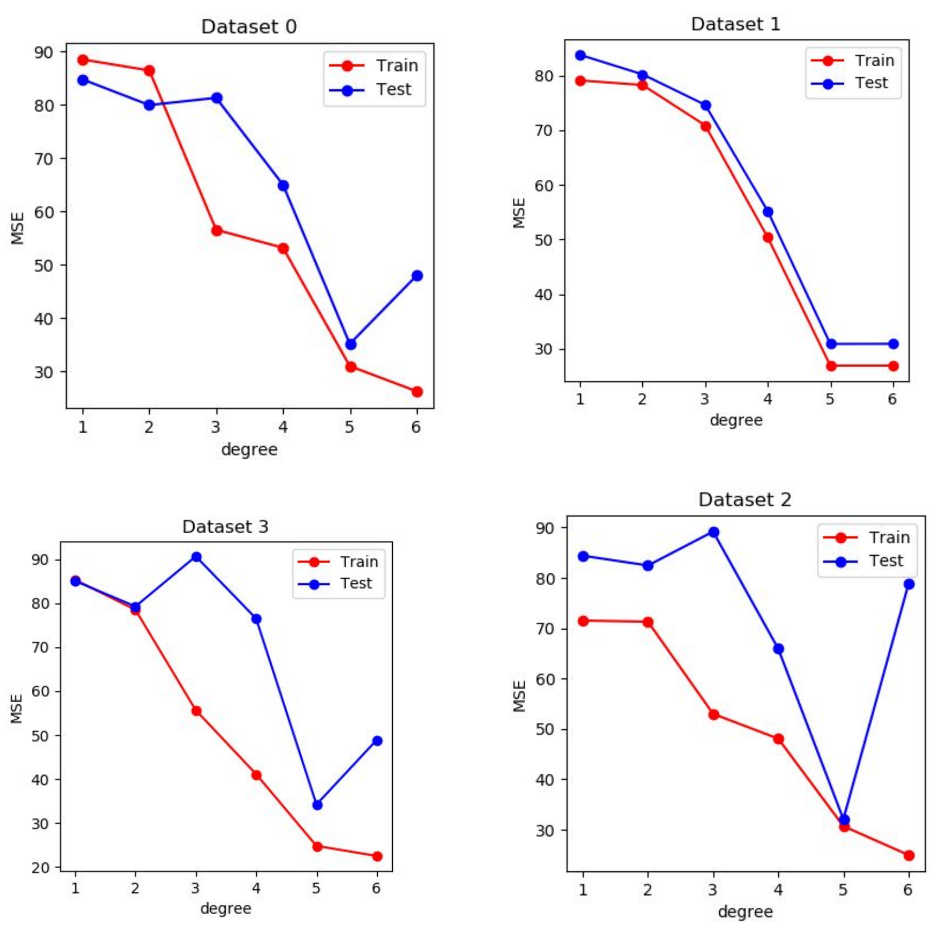


Q 3.2] a]



Test errors
0.7
0.6
0.5
0.4
0.3
0.2
10^1
10^2
10^3
10^4
10^5
Number of samples (log scale)

Q3.2]b]



Dataset 0
Dataset 1
Dataset 3
Dataset 2
Train
Test
MSE
degree

Plots of fitting model for various degrees. (order is as follows)
1,2
3,4
5,6