CS 337-335 : Assignment 2 - 180050068

1.1

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

Likelihood of data — Prior on weights

$\hat{y}_i = x_i \cdot w$

$\varepsilon_i = \hat{y}_i - y_i \sim N(0, \sigma^2)$

$P(x|w) \propto \prod_{i=1}^{n} e^{\frac{-(y_i - x_i \cdot w)^2}{2\sigma^2}} = \exp\left(\sum_{i=1}^{n} \frac{-(y_i - x_i \cdot w)^2}{2\sigma^2} \right)$

$P(w) \propto \prod_{i=1}^{d} e^{\frac{-|w_i|}{b}} = \exp\left(\sum_{i=1}^{d} \frac{-|w_i|}{b} \right) = \exp\left(-\frac{1}{b} \|w\|_1 \right)$

(0 mean Laplace prior)

$P(w|x) \propto \exp\left(\sum_{i=1}^{n} \frac{-(y_i - x_i \cdot w)^2}{2\sigma^2} - \frac{1}{b} \|w\|_1 \right)$

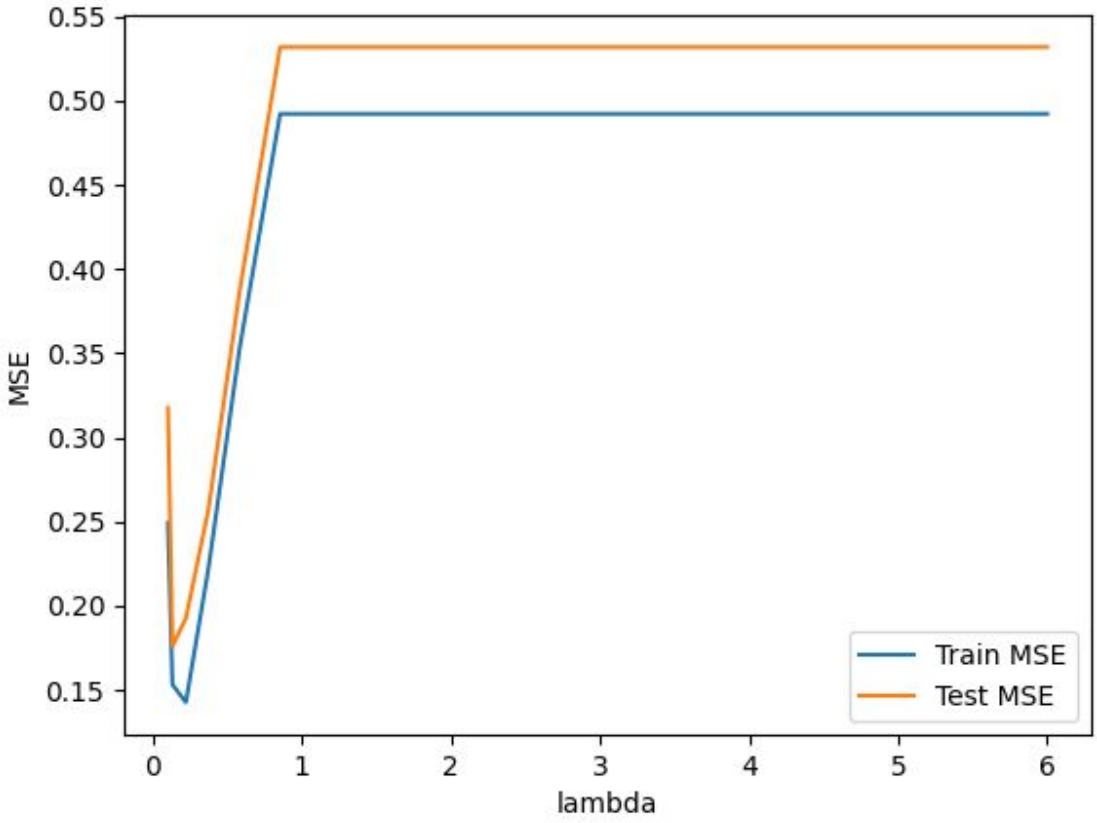
$= \exp\left(-\frac{n}{\sigma^2} \left(\frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i \cdot w)^2 + \frac{\sigma^2}{bn} \|w\|_1 \right) \right)$

$= \exp\left(\frac{-n}{\sigma^2} L(w) \right) \quad \left\{ \lambda = \frac{\sigma^2}{bn} \right\}$

Maximizing posterior is the same as minimizing loss of LASSO.

∴ both these methods give the same solution

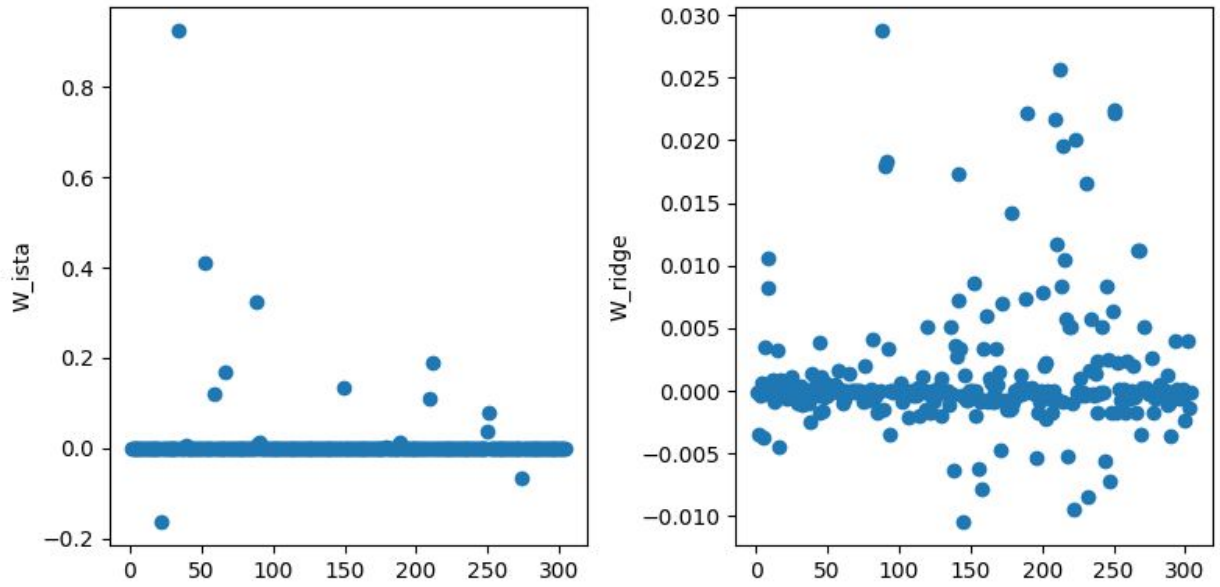
1.2
(b)



Empirically, the optimal value of λ that minimizes the Test MSE seems to be 0.2.

MSE first decreases then increases, then remains constant. The loss function of LASSO has a L1 regularizer part to it. Therefore, increasing λ beyond a certain point forces all the weights to be 0, in order to minimize the loss.

(c)



In ISTA, a lot of the weights are 0 as compared to Ridge regression. And this was expected as Lasso enforces sparsity much more strictly than Ridge regression. This is because Lasso regularizes L1 norm of weights in the loss function whereas Ridge regression regularizes L2 norm of weights.
(Intuition: $3^2 + 4^2 = 5^2 + 0^2$ but $3 + 4 < 5 + 0$, hence Lasso enforces stricter sparsity of weights compared to Ridge)

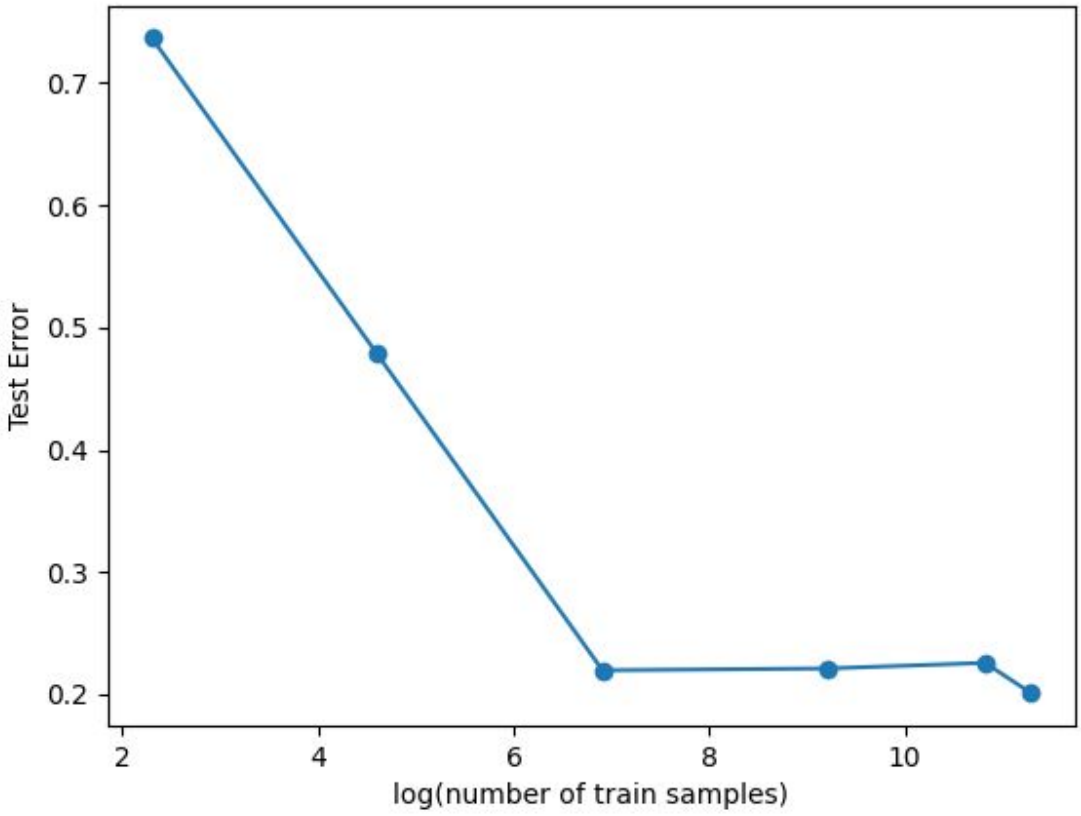
2.1

- Advantage of multiple 1-vs-1 perceptrons: This would much more accurate as some decision boundaries might not be possible to learn for the 1-vs-rest perceptron in the linear subspace
- Advantage of 1-vs-rest: much more computationally cheaper, else we would require ${}^{n}C_{2}$ number of 1-vs-1 perceptrons, where n is the number of classes

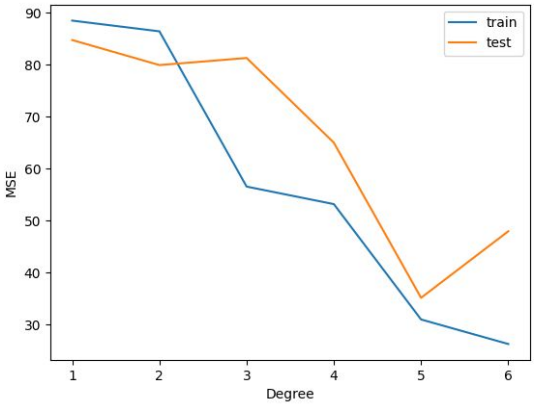
3.1

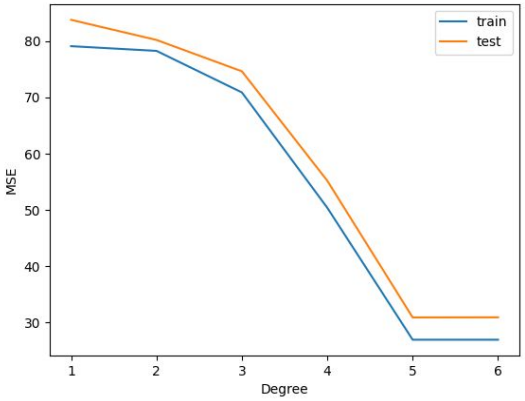
(a) Increasing λ would lead to decrease in variance as the model won't overfit to the data. But it may increase the bias as a higher than optimal λ would also lead to the case of underfitting.
(b) Adding more number of training examples would increase the bias as there might be more number of features/deviations to fit. But it would decrease the variance as the likelihood of encountering any vastly unseen data in the test set decreases.
(c) Adding more features might tend to overfit. As a result, leading to decrease in bias and increase in variance.

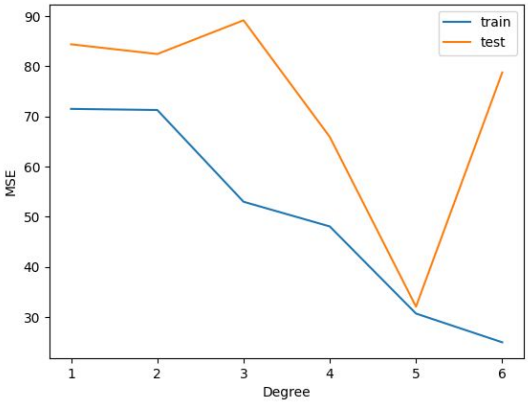
3.2

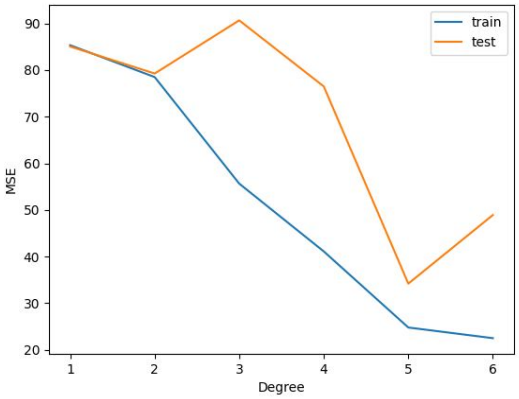


Test error strictly decreases on increasing the number of training samples. This suggests that Perceptron is more susceptible to variance than bias.







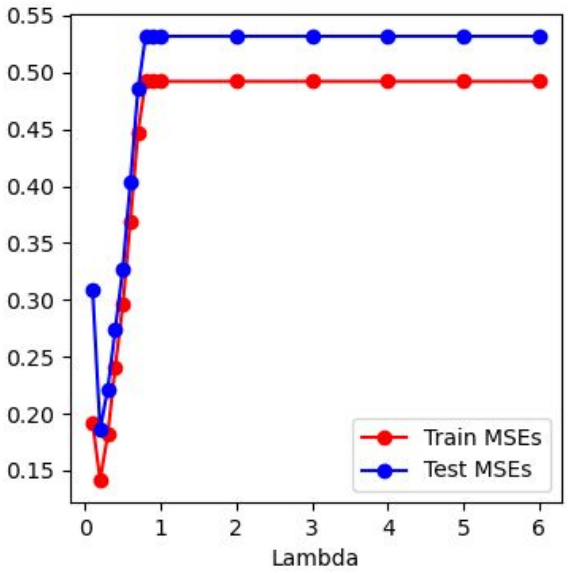


Theoretically, higher degree of polynomial means our model will be able to better fit the training data. Hence, train mse is lowest at degree 6.

As we can see, empirically, test mse seems to minimize at degree 5. Bias decreases with increase in degree but variance increases. Degree 5 seems to the optimal point. Degree 6 seems to be overfitting.

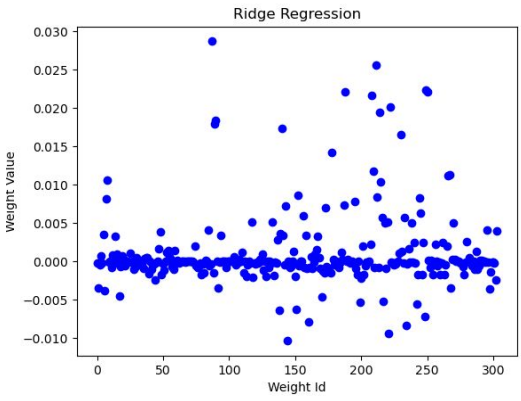
Assignment 2 Graphs: Niraj Mahajan: 180050069

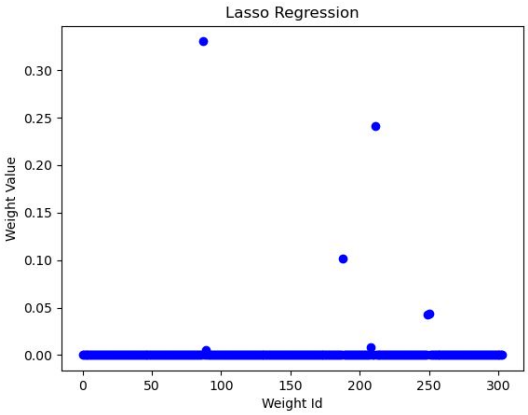
Q1.2 b)



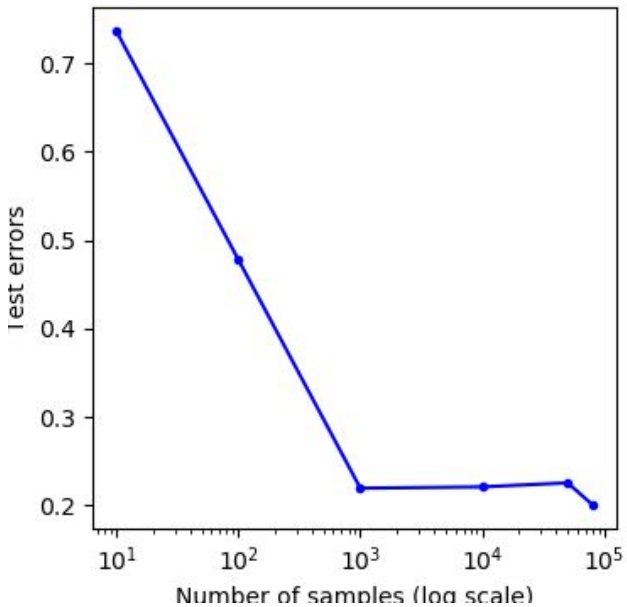
The optimal lambda in the above figure is 0.2. The explanation is provided in the written in the scans.

Q1.2 c]

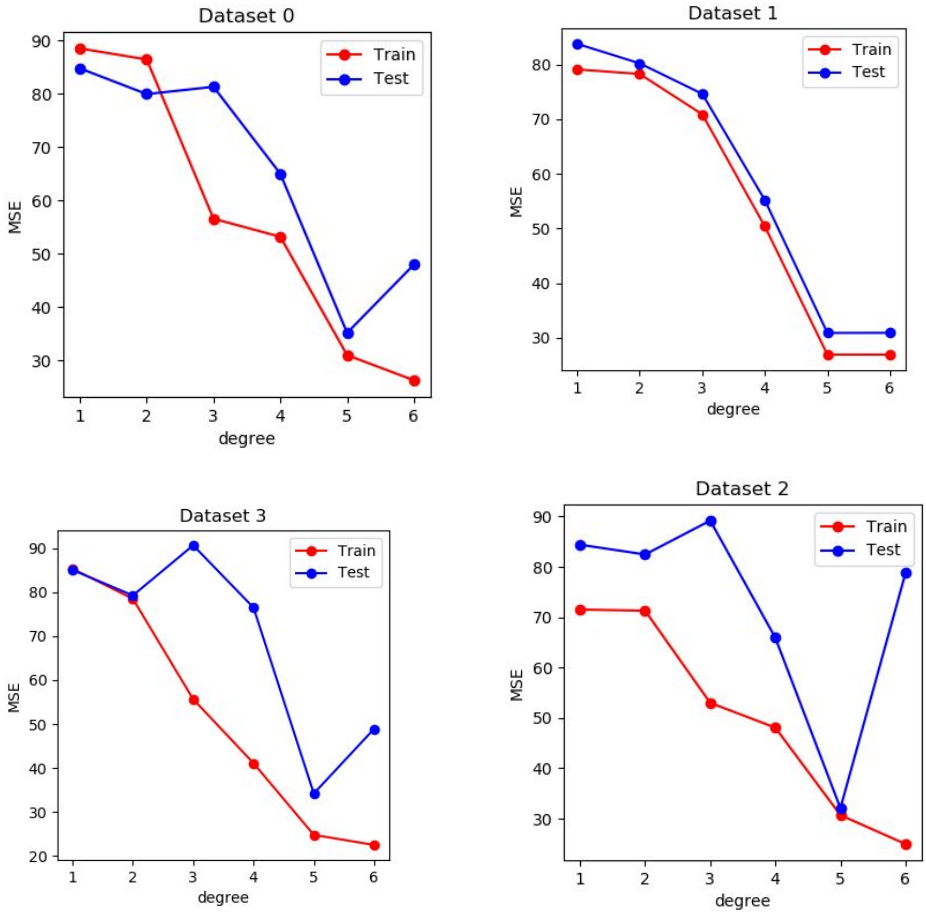




Q 3.2] a]


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 2
Dataset 3
Train
Test
MSE
degree

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