Assignment 2 — Machine Learning

Tarush Goyal — 180050110

September 12, 2020

1 LASSO AND ISTA

1.1

Laplacian distribution :

$$P(x) = \frac{1}{2b} exp(-\frac{|x-\mu|}{b})$$

MAP estimate of Linear regression subject to the Laplacian prior:

$$\begin{aligned} w^* &= argmax\left(Pr(w|D)\right) \\ &= argmax\left(Pr(D|w)Pr(w)\right) \\ &= argmax\left(LL(w) + log(Laplacian(w))\right) \\ &= argmax\left(log(exp(-\tfrac{||Xw-y||^2}{2\sigma^2})) + log(exp(-\tfrac{|w-0|}{b}))\right) \\ &= argmin\left(\tfrac{||Xw-y||^2}{2\sigma^2} + \tfrac{|w|}{b}\right) \end{aligned}$$

Estimation using Lasso

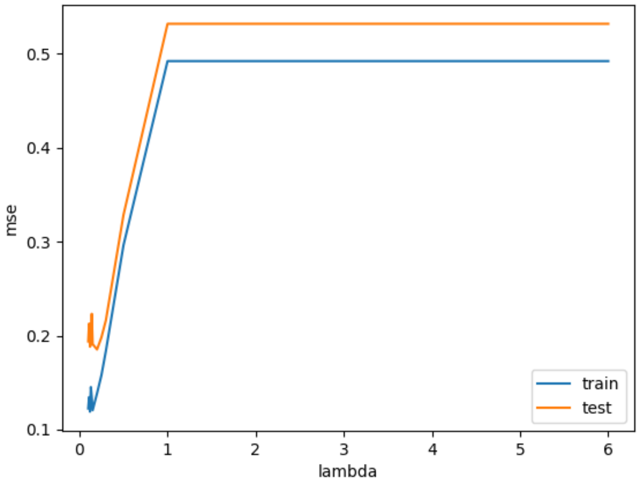
$$\begin{aligned} w^{**} &= argmin\left(||Xw-Y||^2 + \lambda||w||\right) \\ &= argmin\left(\tfrac{||Xw-y||^2}{2\sigma^2} + (\tfrac{b\lambda}{2\sigma^2})\tfrac{|w|}{b}\right) \end{aligned}$$

Thus $w^* = w^{**}$ for $\lambda = \frac{2\sigma^2}{b}$

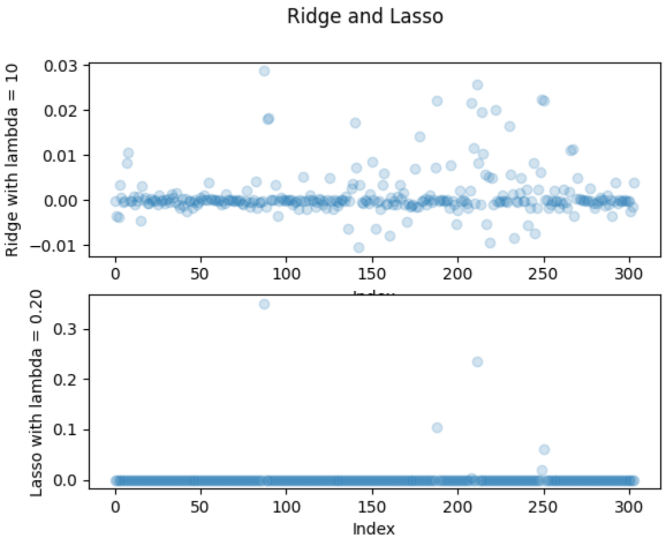
1.2

(b) The graph can be observed in 2 parts:
1) Initially, the test mse reduces and the train mse increases. Here, as we increase λ we are reducing overfitting by clipping the values of w_i and hence decreasing test loss
2) Later on, both the test and the train mse increase. Here as we increase λ we underfit the model with respect to the test data and hence the increasing test loss

Thus $\lambda_{optimal}$ is the λ where we see a tradeoff between overfitting and overfitting. It is approximately = 0.20



(c) Clearly the weights for lasso are much more sparse than ridge. This is because of ista algorithm. In ridge regression the second term of gradient is proportional to w whereas in lasso regression we are pushing the w_i to 0 with a constant value (λ)



2 Multi-class Classification using 1 vs rest Perceptron

2.1

Advantages of One-One :
1) With sufficient data, they can outperform one-all perceptron model because of greater parameters and complexity
2) It is less effected by imbalance of data wrt to labels as compared to One-All, because each perceptron in one-one predicts probability depending on only its data.
Disadvantages of One-One:
1) One-One uses ${}^{n}C_2$ different perceptrons. This makes it very slow to train and predict as we have to train/check every pair separately
2) Also, One-One does not use the knowledge of one 'pair' for other making it very data hungry.

3 Bias Variance Trade-off

3.1

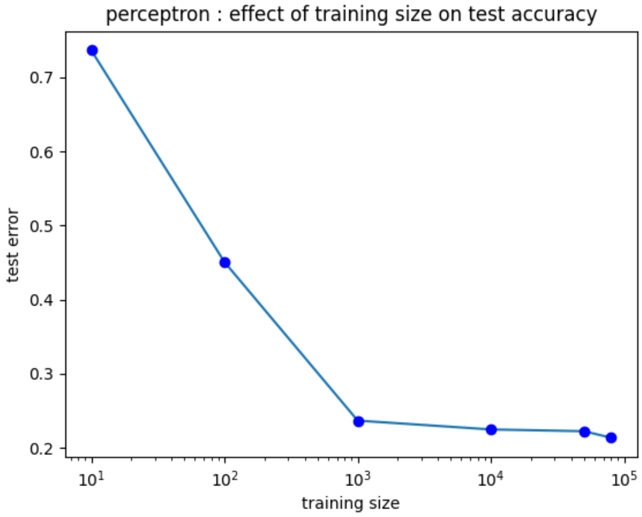
(a) λ is a measure of weight given to regularisation term over mse. The lasso loss function is a trade off between minimising mse (and hence bias) and minimising weights of W (and hence variance). Thus increasing λ **decreases variance but increases bias**.

(b) Increasing 'sample size' or number of training examples would let the data be a better representative of the distribution. This would hence **decrease the variance** of the predicted model over possible training samples. However it **should not effect bias** much as mean of predicted function over training sample would not change much.

(c) Adding extra features which are linear combination of other features adds no additional information about the distribution and its mapping to labels to the model. Thus there is **no effect on bias and variance**. It just adds an extra dependent column to feature matrix which does not effect the minimum mse

3.2

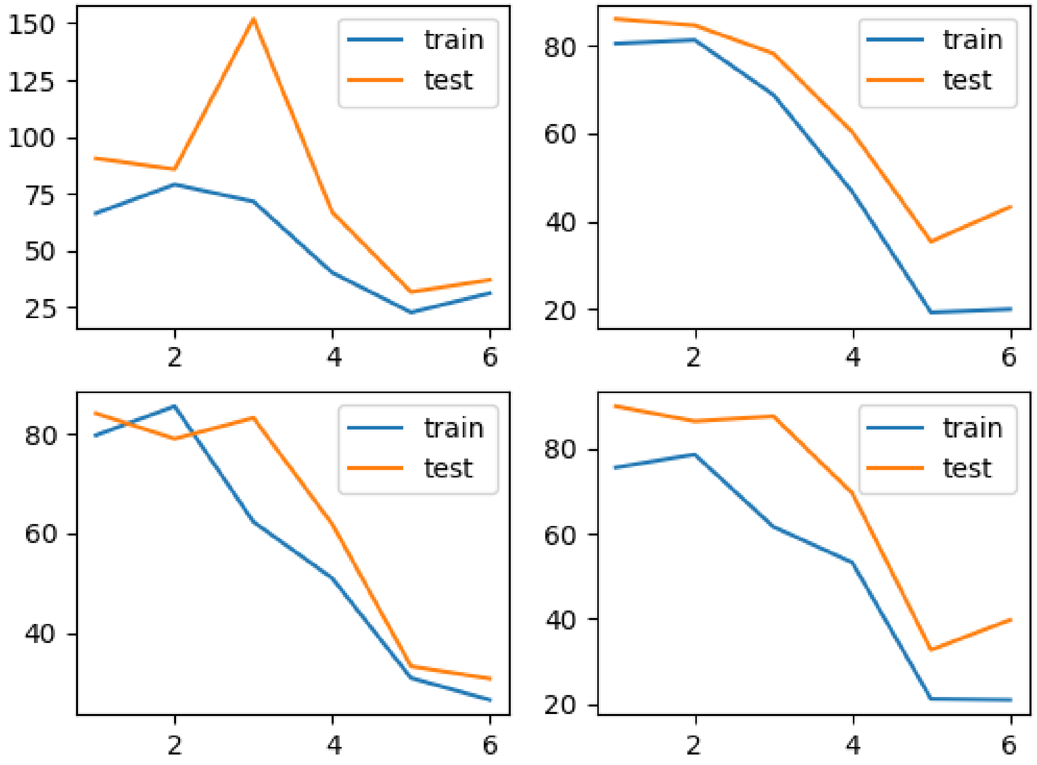
The figure below represents test error on increasing training size. As we can see test error decreases with increase in training size. However, the rate of change decreases in general because the amount of 'extra information' carried by same amount of data decreases.
The variance of the prediction decreases as the training size increases and the bias remains constant.



Theoretically, we will obtain least training error for highest degree of polynomial i.e. 6 because higher the degree, greater the flexibility of the model and hence the model can adapt to the distribution more.

The figure below shows effect of degree of kernel on train and test errors of 4 samples. We have the following observations:
1) In general, training error reduces as degree of polynomial increases due to lower bias.
2) In general, test error reduces till degree = 5 and then increases again. This is because bias (and hence $bias^2$ is decreasing and variance is increasing as degree increases. So there sum first decreases (bias factor is dominant) and later on increases (variance factor is dominant), resulting in a minima at degree = 5.

3) So, degree = 5 is the optimal value of degree and expected mse for a test sample is minimum.


train
test
train
test
train
test
train
test