Yerbol Aussat

Ridge and Lasso Regression

PART 1.

Error

λ	Training set	Validation set	Test set
0	9.69	14.71	370.22
10	10.52	***13.50***	98.71
20	10.92	13.87	111.95
30	11.38	14.45	127.01
40	11.90	15.16	139.58
50	12.46	15.93	149.22
60	13.04	16.75	156.24
70	13.63	17.57	161.14
80	14.21	18.40	164.37
90	14.79	19.21	166.29
100	15.36	20.01	167.23

λ^*=10

Percentage of non-zeros in w

λ	Training set	Validation set	Test set
0	100.0	100.0	100.0
10	100.0	100.0	100.0
20	100.0	100.0	100.0
30	100.0	100.0	100.0
40	100.0	100.0	100.0
50	100.0	100.0	100.0
60	100.0	100.0	100.0
70	100.0	100.0	100.0
80	100.0	100.0	100.0
90	100.0	100.0	100.0
100	100.0	100.0	100.0

PART 2.

Error

λ	Training set	Validation set	Test set
0	27.34	675355477001.58	3395.21
10	27.99	***575898539962.09***	1716.58
20	28.61	576341193232.24	1185.21
30	29.35	578055457633.07	862.68
40	30.14	579452905639.71	652.89
50	30.96	580496119567.17	511.21
60	31.78	581292803424.63	412.79
70	32.58	581939740789.35	342.96
80	33.35	582506456225.78	292.61
90	34.11	583039841067.83	255.75
100	34.83	583570600224.09	228.48

λ^*=10

Percentage of non-zeros in w

λ	Training set	Validation set	Test set
0	100.0	100.0	100.0
10	100.0	100.0	100.0
20	100.0	100.0	100.0
30	100.0	100.0	100.0
40	100.0	100.0	100.0
50	100.0	100.0	100.0
60	100.0	100.0	100.0
70	100.0	100.0	100.0
80	100.0	100.0	100.0
90	100.0	100.0	100.0
100	100.0	100.0	100.0

PART 3.

Error

λ	Training set	Validation set	Test set
0	0.00	7408659205.26	1957980.66
10	0.01	44.94	112.94
20	0.03	44.78	112.20
30	0.06	44.63	111.51
40	0.10	44.51	110.87
50	0.16	44.39	110.25
60	0.22	44.29	109.68
70	0.28	44.20	109.13
80	0.36	44.12	108.62
90	0.44	44.05	108.12
100	0.52	***43.99***	107.66

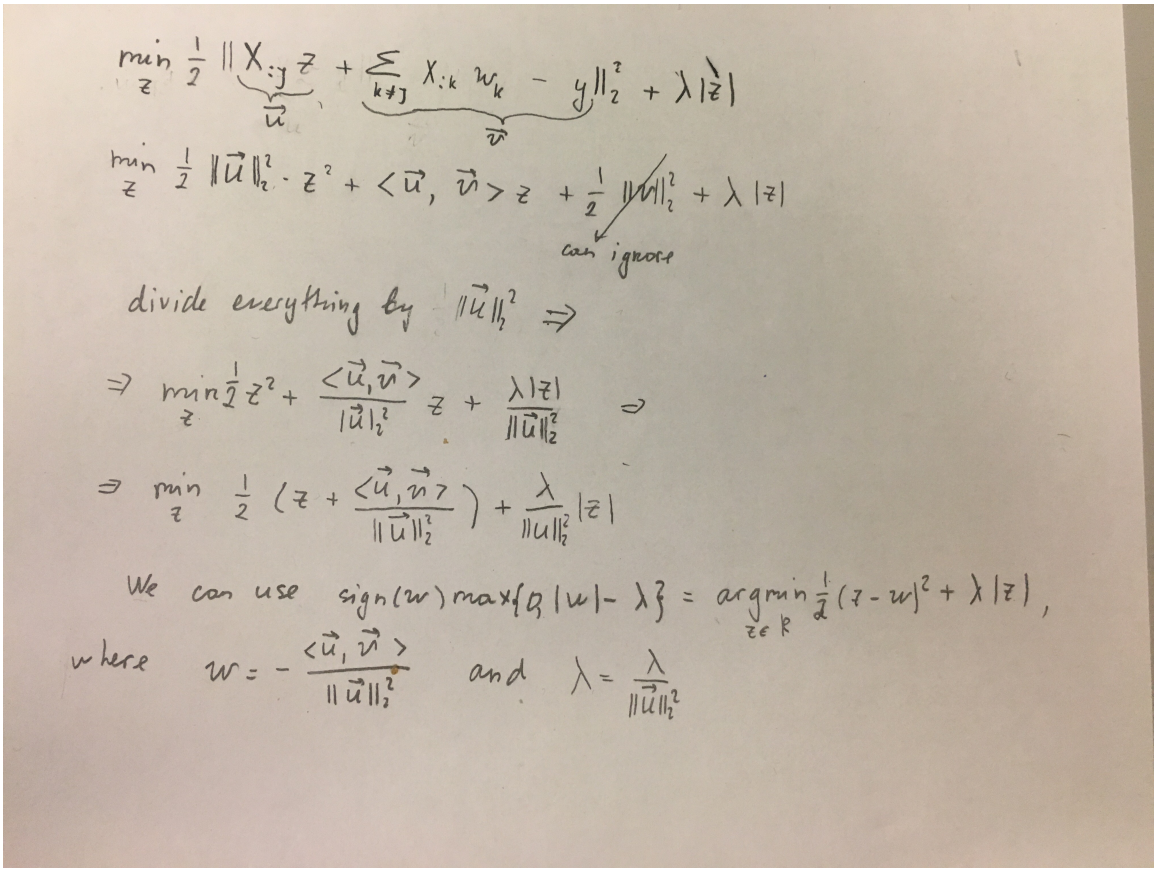
λ^*=100

Percentage of non-zeros in w

λ	Training set	Validation set	Test set
0	100.0	100.0	100.0
10	100.0	100.0	100.0
20	100.0	100.0	100.0
30	100.0	100.0	100.0
40	100.0	100.0	100.0
50	100.0	100.0	100.0
60	100.0	100.0	100.0
70	100.0	100.0	100.0
80	100.0	100.0	100.0
90	100.0	100.0	100.0
100	100.0	100.0	100.0

PART 4.

Derivation:



Error

λ	Training set	Validation set	Test set
0	145.89	104.95	328.43
10	142.89	100.56	103.30
20	142.70	100.26	112.85
30	142.17	99.86	113.98
40	141.22	99.34	113.86
50	140.66	98.84	114.96
60	140.12	98.35	116.11
70	139.59	97.87	117.34
80	139.06	97.40	118.64
90	138.55	96.94	120.00
100	138.45	***96.49***	123.18

λ^*=100

Percentage of non-zeros in w

λ	Training set	Validation set	Test set
0	100.0	100.0	100.0
10	92.8571428571	88.5714285714	92.8571428571
20	71.4285714286	72.1428571429	71.4285714286
30	71.4285714286	69.2857142857	71.4285714286
40	71.4285714286	68.5714285714	71.4285714286
50	71.4285714286	65.7142857143	71.4285714286
60	64.2857142857	64.2857142857	64.2857142857
70	64.2857142857	64.2857142857	64.2857142857
80	64.2857142857	62.8571428571	64.2857142857
90	64.2857142857	61.4285714286	64.2857142857
100	64.2857142857	60.7142857143	64.2857142857

As we can see from the Mean Square Error table above, λ^***=100** is the best value of regularization constant. However, one can see that the error on the validation set kept monotonously decreasing from 0 to 100. Therefore, we can guess that the optimal value of regularization constant is greater than 100.

To show this we ran the experiment again, but with λ**=[1; 100] * dim(w).**
The obtained result is summarized below.

Error

λ**/ dim(w)**	Training set	Validation set	Test set
0	145.89	104.95	328.43
10	136.45	94.91	123.55
20	130.10	90.57	125.86
30	124.84	***88.39***	128.64
40	120.67	88.41	131.35
50	117.72	90.02	138.15
60	115.73	91.67	146.11
70	114.86	94.12	160.29
80	114.61	97.00	178.05
90	115.02	97.52	200.06
100	114.95	96.95	210.00

λ^*=30 * dim(w) = 30*14 = 420

Percentage of non-zeros in w

λ/ dim(w)	Training set	Validation set	Test set
0	100.0	100.0	100.0
10	57.1428571429	58.5714285714	57.1428571429
20	57.1428571429	55.0	57.1428571429
30	57.1428571429	55.7142857143	57.1428571429
40	50.0	50.7142857143	50.0
50	50.0	49.2857142857	50.0
60	42.8571428571	42.8571428571	42.8571428571
70	42.8571428571	42.8571428571	42.8571428571
80	42.8571428571	40.0	42.8571428571
90	35.7142857143	35.7142857143	35.7142857143
100	35.7142857143	35.7142857143	35.7142857143

Therefore, it can be concluded that actual optimal regularization constant is greater than 100. Specifically, it is around **30 * dim(w) = 420.**

PART 5.

The lasso algorithm with lambda values in the range of 0 and 100 failed to converge.

We conducted another experiment with **λ=[1; 100] * dim(w).**

Error

λ/ dim(w)	Training set	Validation set	Test set
0	155.62	654.93	47713.81
10	88.74	92.47	142.02
20	86.08	90.38	144.52
30	86.05	***89.83***	143.75
40	86.07	89.86	142.75
50	86.10	89.91	141.76
60	86.14	89.98	140.78
70	86.20	90.06	139.81
80	86.28	90.17	138.86
90	86.37	90.29	137.92
100	86.48	90.43	136.99

λ^*=30 * dim(w) = 30*14 = 420

Percentage of non-zeros in w

λ/ dim(w)	Training set	Validation set	Test set
0	98.5207100592	98.5207100592	98.5207100592
10	0.0	0.0	0.0
20	0.0	0.0	0.0
30	0.0	0.0	0.0
40	0.0	0.0	0.0
50	0.0	0.0	0.0
60	0.0	0.0	0.0
70	0.0	0.0	0.0
80	0.0	0.0	0.0
90	0.0	0.0	0.0
100	0.0	0.0	0.0

From this experiment, it was observed that it takes Lasso algorithm very long to converge to 0.0001. Ridge regression from problem 2.3 was much quicker.

Ridge regression had lower values of mean square errors for regularization constants greater than 0. In ridge regression, when lambda = 0, the mean square error was very high, implying very strong overfitting due to a big number of irrelevant features, and not enough data. Lasso algorithm also had higher mean square error in the beginning, but the error dropped as the regularization constant increased.

For ridge regression the number of non-zero elements was always 100, confirming that ridge regression weight is always dense. Lasso algorithm quickly yielded the vector w is many zero-elements. In fact, even small values of lambda are able to produce very sparse weight vectors. Therefore, lasso algorithm is very good for feature selection.