CSE – 574:
INTRODUCITON TO MACHINE LEARNING

Programming Assignment – 1

Classification and Regression

Datasets Used:

1. sample.pickle
2. diabetes.pickle

TA: Zhenyi Wang

Team Members (Group – 40):

50413342 - Bodhith Edara

50409091 - Teja Reddy Alla

50417093 - Venkata Krishna Sai Sakhamuri

PROBLEM-1

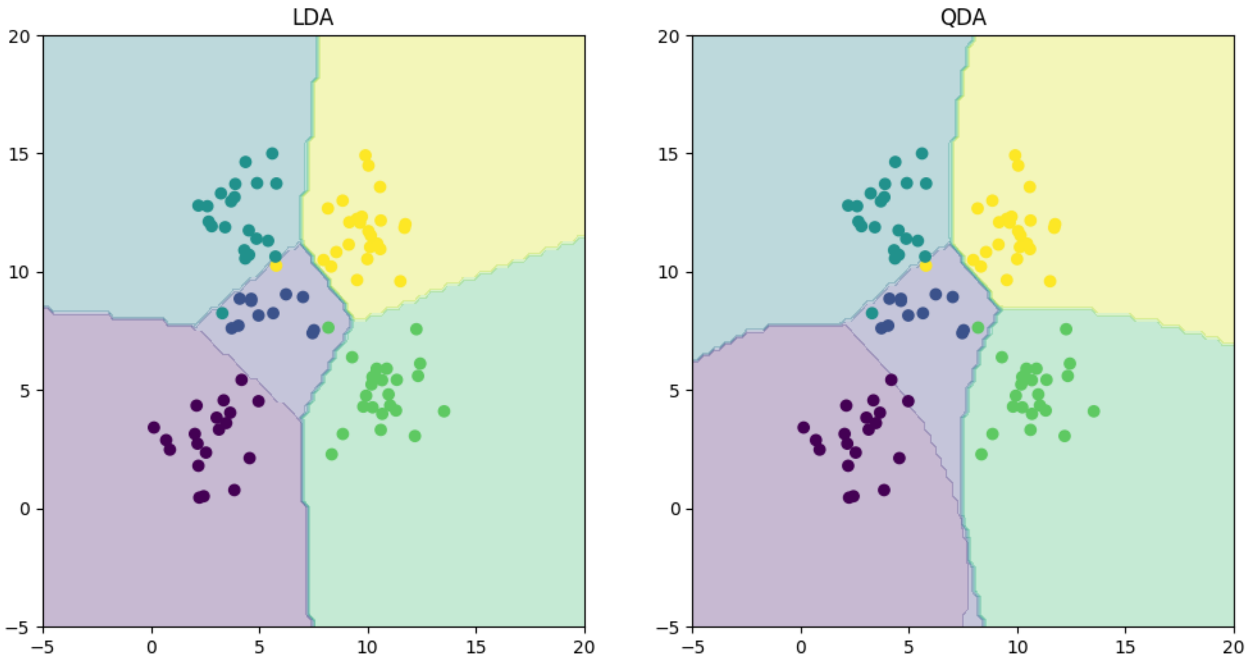
Experiment with Gaussian Discriminators

Q) Implement Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA).

i) Report the accuracy of LDA and QDA on the provided test data set (sample test)

Gaussian Discriminator	Accuracy
LDA	97%
QDA	96%

ii) Plot the discriminating boundary for linear and quadratic discriminators



iii) Explain why there is a difference in the two boundaries.

For LDA, the decision boundary is almost linear, whereas for QDA, it is non-linear. This is due to the fact that LDA considers the entire dataset (rather than just one label) when learning the covariance matrix i.e. the covariance is calculated by using the global mean of the entire data. As a result, the decision boundary for LDA is linear. However, the boundary for QDA is non-linear because it considers all of the dataset's unique labels i.e. the covariance matrix is calculated for each input dataset using the local mean of each unique input dataset.

In general, QDA is a flexible classifier than LDA as its nonlinear curve nature allows for much more flexible boundaries. However, LDA outperforms QDA when there are fewer training samples or the data is linearly separable. QDA, on the other hand, should be used when the training set is very large or the data cannot be separated linearly.

PROBLEM-2

Experiment with Linear Regression

Q) Implement ordinary least squares method to estimate regression parameters by minimizing the squared loss.

i) Calculate and report the MSE for training and test data for two cases:
a) Without using an intercept (or bias) term, and
b) With using an intercept.

	MSE without an Intercept	MSE with an Intercept
Training Data	19099.4468446	2187.1602949303897
Testing Data	106775.36145121799	3707.840180960783

ii) Which one is better?

Based on the above findings, we can conclude that Linear Regression (with intercept) produces significantly better results for both Training and Test data.

This is because when a linear regression line is modelled without an intercept, it must pass through the origin, which may not be the best model. If the Linear Regression line includes an integer, the model becomes more closely aligned with the actual data, yielding better results.

PROBLEM-3

Experiment with Ridge Regression

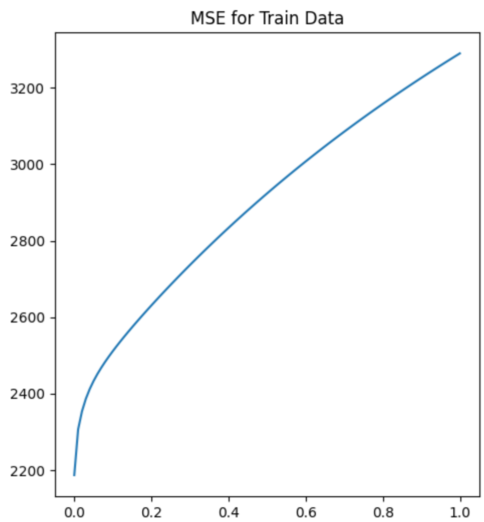
Q) Implement parameter estimation for ridge regression by minimizing the regularized squared loss

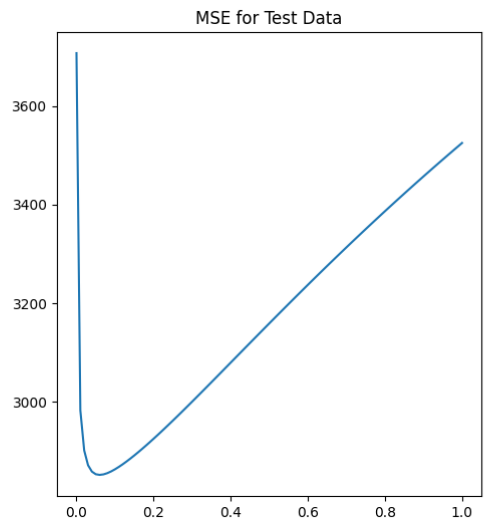
i) MSE values for training and test data:

Lambda(λ)	Train MSE	Test MSE	Lambda(λ)	Train MSE	Test MSE
0	2187.160295	3707.840181	0.31	2746.543191	3007.947616
0.01	2306.832218	2982.44612	0.32	2756.532665	3015.810555
0.02	2354.071344	2900.973587	0.33	2766.442316	3023.700386
0.03	2386.780163	2870.941589	0.34	2776.273307	3031.613181
0.04	2412.119043	2858.00041	0.35	2786.026719	3039.545297
0.05	2433.174437	2852.665735	0.36	2795.703568	3047.493351
0.06	2451.528491	2851.330213	0.37	2805.30482	3055.454198
0.07	2468.077553	2852.349994	0.38	2814.831398	3063.424913
0.08	2483.365647	2854.879739	0.39	2824.284191	3071.402772
0.09	2497.740259	2858.444421	0.4	2833.664063	3079.385238
0.1	2511.432282	2862.757941	0.41	2842.971855	3087.369947
0.11	2524.600039	2867.637909	0.42	2852.208389	3095.354694
0.12	2537.3549	2872.962283	0.43	2861.374474	3103.337424
0.13	2549.776887	2878.645869	0.44	2870.470905	3111.316218
0.14	2561.924528	2884.626914	0.45	2879.498467	3119.289287
0.15	2573.841288	2890.85911	0.46	2888.457936	3127.254961
0.16	2585.559875	2897.306659	0.47	2897.350077	3135.211679
0.17	2597.105192	2903.941126	0.48	2906.17565	3143.157988
0.18	2608.4964	2910.739372	0.49	2914.935407	3151.09253
0.19	2619.748386	2917.682164	0.5	2923.630092	3159.014036
0.2	2630.872823	2924.753222	0.51	2932.260444	3166.921324
0.21	2641.878946	2931.938544	0.52	2940.827193	3174.813291
0.22	2652.774126	2939.22593	0.53	2949.331065	3182.688908
0.23	2663.564301	2946.604624	0.54	2957.772777	3190.547215
0.24	2674.254297	2954.065056	0.55	2966.153041	3198.387318
0.25	2684.848078	2961.598643	0.56	2974.472563	3206.208382
0.26	2695.348935	2969.197637	0.57	2982.732039	3214.009633
0.27	2705.759629	2976.855001	0.58	2990.93216	3221.790346
0.28	2716.082507	2984.564321	0.59	2999.073611	3229.549851
0.29	2726.319587	2992.319722	0.6	3007.157067	3237.287523
0.3	2736.47263	3000.115809			

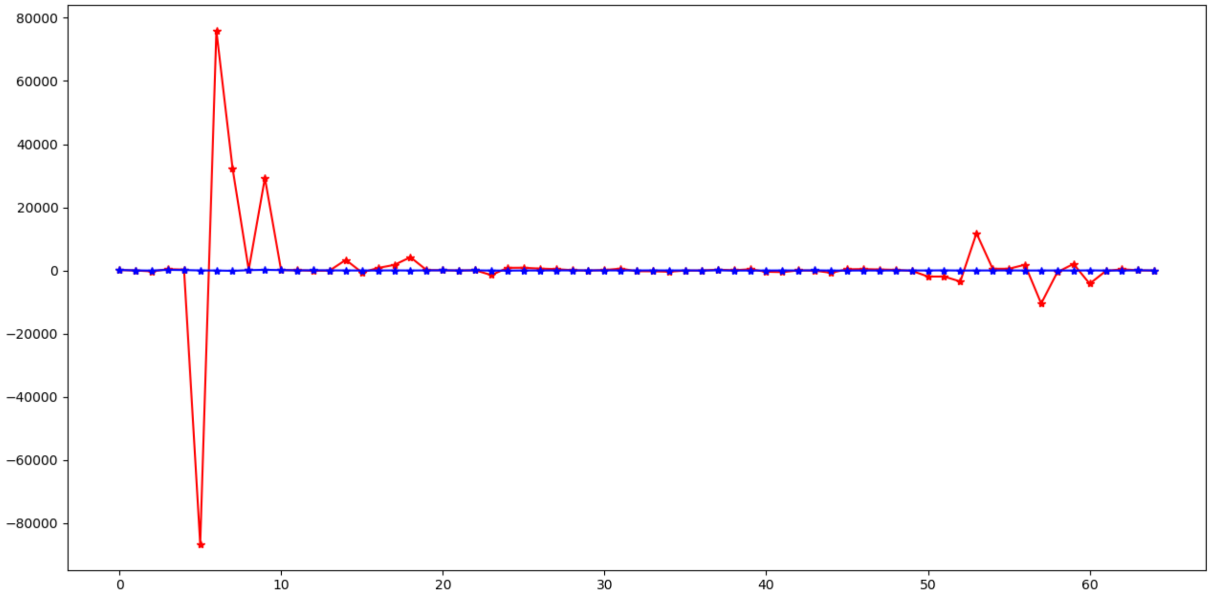
Lambda(λ)	Train MSE	Test MSE	Lambda(λ)	Train MSE	Test MSE
0.61	3015.183199	3245.002781	0.81	3164.630117	3394.007386
0.62	3023.152668	3252.695087	0.82	3171.593342	3401.174246
0.63	3031.066127	3260.363943	0.83	3178.512005	3408.313184
0.64	3038.924224	3268.008886	0.84	3185.3866	3415.424154
0.65	3046.727598	3275.629488	0.85	3192.21761	3422.507124
0.66	3054.476879	3283.225355	0.86	3199.005514	3429.562069
0.67	3062.172691	3290.796124	0.87	3205.750782	3436.588973
0.68	3069.81565	3298.341459	0.88	3212.453878	3443.587832
0.69	3077.406362	3305.861052	0.89	3219.115258	3450.558648
0.7	3084.945428	3313.354623	0.9	3225.735372	3457.50143
0.71	3092.43344	3320.821913	0.91	3232.314665	3464.416198
0.72	3099.870981	3328.262686	0.92	3238.853573	3471.302975
0.73	3107.258627	3335.676731	0.93	3245.352525	3478.161794
0.74	3114.596946	3343.063853	0.94	3251.811947	3484.992692
0.75	3121.886499	3350.423878	0.95	3258.232255	3491.795713
0.76	3129.127838	3357.75665	0.96	3264.613861	3498.570906
0.77	3136.321508	3365.062031	0.97	3270.95717	3505.318324
0.78	3143.468045	3372.339896	0.98	3277.262582	3512.038029
0.79	3150.567979	3379.590137	0.99	3283.53049	3518.730082
0.8	3157.621831	3386.812661	1	3289.761281	3525.394553

ii) Plot the errors on train and test data for different values of lambda λ





iii) Compare the relative magnitudes of weights learnt using OLE (Problem 2) and weights learnt using ridge regression



❖ Values used to plot the above graph:

S.No	W (Linear Regression)	W (Ridge Regression)	S.No	W (Linear Regression)	W (Ridge Regression)
0	148.154876	150.194724	33	-256.0714523	33.25175429
1	1.274852125	21.70175432	34	-385.1774006	4.638032286
2	-293.3835224	-39.06015175	35	-33.41767377	18.79589248
3	414.7254484	189.7687958	36	-10.73500661	-12.03993446
4	272.0891344	131.4363177	37	257.1071888	38.2074689
5	-86639.45657	12.87166763	38	59.95545924	-21.9866462
6	75914.46754	-12.64706671	39	383.7280423	-13.233832
7	32341.62261	-111.6774078	40	-404.1583898	-21.40648344
8	221.1012148	99.4380881	41	-514.2864344	9.528354748
9	29299.55101	203.314471	42	38.36366416	-3.041096104
10	125.2303603	112.5686029	43	-44.61028888	45.03265856
11	94.41108335	27.87307148	44	-729.6435305	-25.1457355
12	-93.86286321	55.6080895	45	377.4083364	-31.31497844
13	-33.72827999	35.94955735	46	439.7942902	8.845983391
14	3353.197762	9.320488703	47	308.5143733	-11.92001782

15	-621.0962656	-14.82227189	48	189.8596785	4.234694789
16	791.7365403	2.083102995	49	-109.773797	7.07637032
17	1767.760389	26.19273491	50	-1919.657062	-3.990913457
18	4191.674041	-11.76247315	51	-1924.633798	36.5751882
19	119.4381209	33.19690742	52	-3489.795276	-22.71617302
20	76.61034004	41.12190968	53	11796.96862	-1.905897348
21	-15.20012929	-1.629231709	54	530.6744147	8.069498614
22	82.24245936	37.81703094	55	543.3059231	25.84457843
23	-1456.662084	-26.61591914	56	1821.075179	-12.71125908
24	827.3867024	-46.27610924	57	-10463.98059	-7.177007087
25	869.2909522	0.338207547	58	-516.6276109	4.034333813
26	586.2344952	0.712858398	59	2064.359173	-17.36136688
27	427.0267265	30.42616757	60	-4199.413314	25.14496419
28	90.24676902	23.58120717	61	-140.4957052	-20.10395628
29	-17.88762245	9.146373001	62	374.1570899	-12.62145938
30	141.6967738	24.92991386	63	51.47574924	38.08617538
31	582.8193844	19.75759867	64	-46.44927305	22.55380091
32	-234.0375106	-0.17672983			

iv) Compare the two approaches in terms of errors on train and test data.

REGRESSION METHOD	TRAIN MSE	TEST MSE
Linear Regression (without Intercept)	19099.4468446	106775.36145121799
Linear Regression (with Intercept)	2187.1602949303897	3707.840180960783
Ridge Regression (for optimal value of λ)	2451.528491 (λ is 0)	2851.330213 (λ is 0.06)

When the two approaches are compared in terms of weights, we can see that weights learned using Linear Regression have a much higher magnitude than weights learned using Ridge Regression. This significant difference can be attributed to the regularization parameter (λ) used in Ridge Regression.

In conclusion, Ridge regression outperforms Linear regression on the test data (with and without the intercept).

V) What is the optimal value for lambda λ and why?

In case of test data Optimal value for λ is 0.06
In case of training data Optimal value for λ is 0

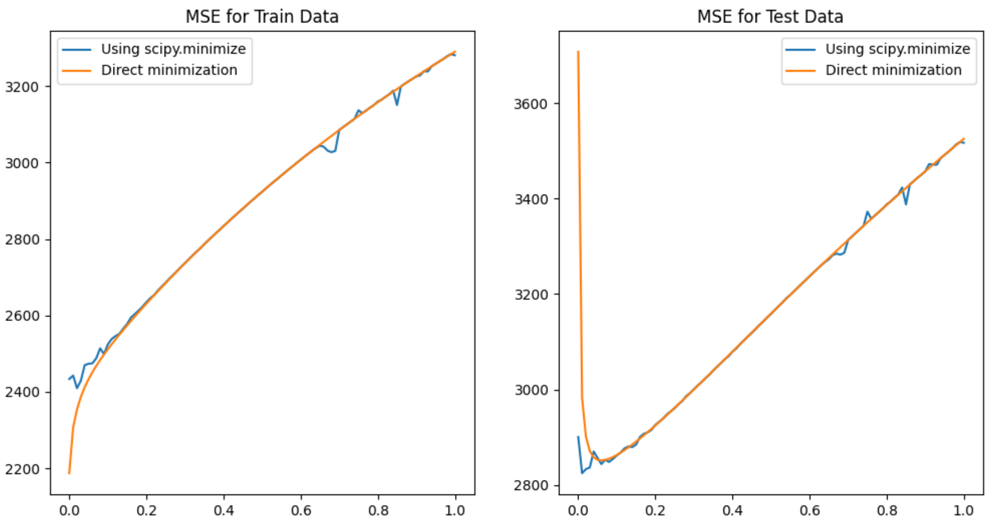
We observe that MSE is directly proportional to λ and we get the lowest MSE if λ =0 (no regularization). The high MSE value shows that there is a greater penalty for learning as the λ value rises.

We can see that MSE will be lowest when λ =0.06. After this value if we increase the λ value then MSE starts increasing which reflects poor fit for the testing dataset.

PROBLEM-4

Using Gradient Descent for Ridge Regression Learning

i) Plot the errors on train and test data obtained by using the gradient descent based learning by varying the regularization parameter λ



ii)Compare with the results obtained in Problem 3.

We may deduce from the plots above that the training and test error for Ridge regression using Gradient Descent and Ridge regression are nearly identical.

However, in the case of Ridge regression with Gradient descent, we can see that the graph is distorted for both small and large values of lambda. This is due to the usage of the minimize and inverse functions in gradient descent, which takes longer.

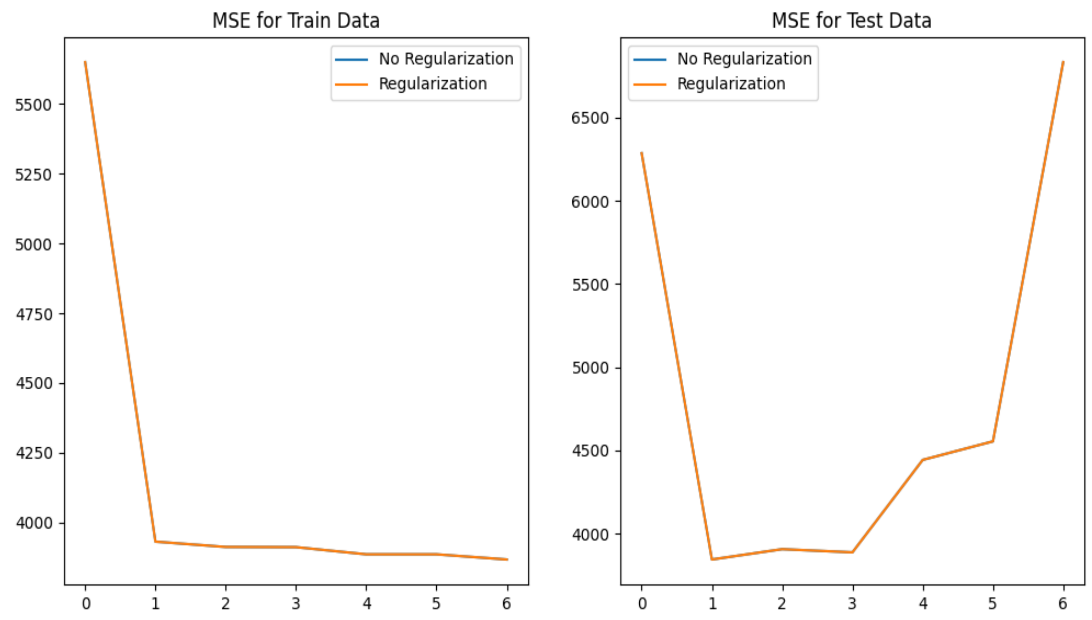
The inversion of matrices becomes costly as the matrices get larger. When compared to ridge regression, ridge regression using gradient descent is a better alternative in those instances.

PROBLEM-5

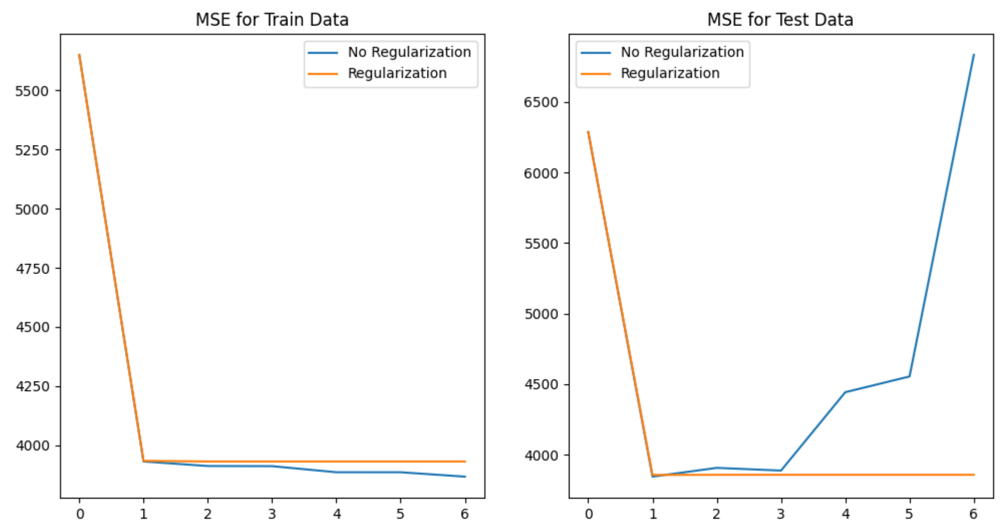
Non-linear Regression

i) Compute the errors on train and test data.

$\lambda = 0$



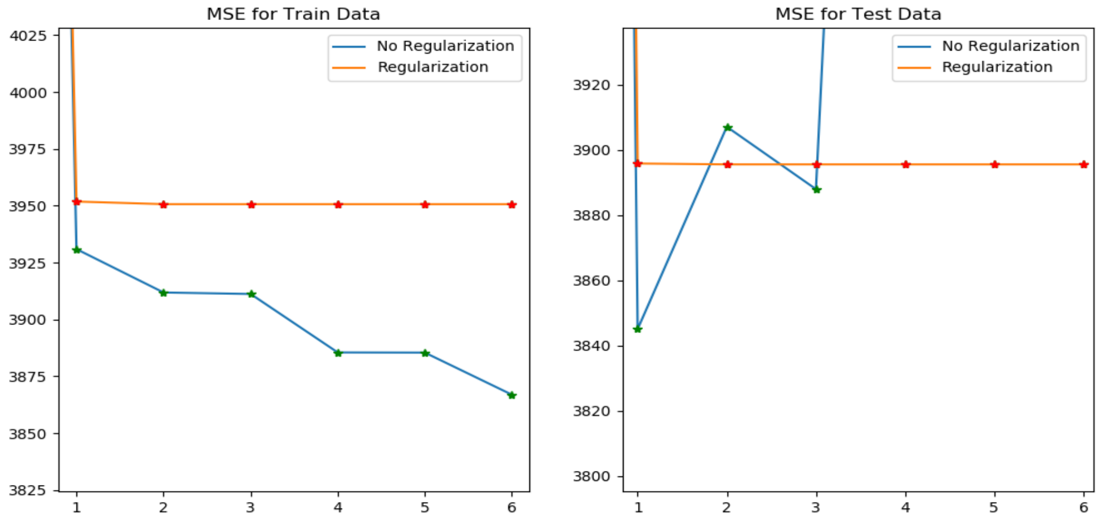
$\lambda = 0.06$



ii) Compare the results for both values of λ. What is the optimal value of p in terms of test error in each setting? Plot the curve for the optimal value of p for both values of λ and compare.

p	Non-Regularized Train MSE	Non-Regularized Test MSE	Regularized Train MSE	Regularized Test MSE
0	6286.404792	5650.710539	6286.881967	5650.711907
1	3845.03473	3930.915407	3895.856464	3951.839124
2	3907.128099	3911.839671	3895.584056	3950.687312
3	3887.975538	3911.188665	3895.582716	3950.682532
4	4443.327892	3885.473068	3895.582668	3950.682337
5	4554.830377	3885.407157	3895.582669	3950.682335
6	6833.459149	3866.883449	3895.582669	3950.682335

Zoomed in the p, MSE values graph



No Regularization:
In this case, optimal value of *p obtained for Train MSE is 6* and all the other values are almost similar for different values of p. But for *Test MSE*, the *optimal value for p = 1*.

Regularization:
In this situation, the best *p value achieved for Train MSE is 4* while *Test MSE is 5 or 6*. Also, for varying values of p, all of the other values are nearly identical. This is the benefit of regularization.

PROBLEM-6

Interpreting Results

i) Compare the various approaches in terms of training and testing error.

MSE		
MODEL	Training Data	Testing Data
Linear OLE Regression with intercept	2187.1602949303897	3707.840180960783
Linear OLE Regression without intercept	19099.4468446	106775.36145121799
Ridge Regression	2451.528491	2851.330213
Ridge Regression with gradient descent	2451.76715459	2824.06091919
Non-linear Regression with no regularization	3866.883449	3845.03473
Non-linear Regression with regularization	3950.682335	3895.582669

Recommendation: After considering all the classification and regression models, Ridge regression with gradient descent can be used to predict the diabetes level as it has the lowest test MSE in comparison to other models which means that the accuracy of the prediction will be more.

ii)What metric should be used to choose the best setting?

The best metric that can be used to choose the best settings is accuracy (calculated for all the models used). Some other parameters that influence the outcome of the model are size of the dataset, test MSE for the test dataset.

As seen in the above table, the MSE value for Linear Regression with Intercept is very high and will result in a large number of errors, making it an infeasible method if accuracy is our metric whereas Ridge Regression method gives us considerably lower MSE values, implying that the error obtained from this method is less in comparison. In addition, the MSE values for Ridge Regression with Gradient Descent and Ridge Regression are nearly identical.

The MSE value is at its lowest for the optimal value of λ (with regularization) which implies accurate output for this model. Moreover, the MSE values for the Non-linear Regression model, both with and without regularization, are slightly higher than those for the Ridge Regression approach, thus we will not prefer this approach.

If accuracy were the sole metric, we will always choose the Ridge Regression with Gradient Descent method (with regularization using the optimal value for lambda) as it gives the best results.