CSE - 574 Introduction to Machine Learning

Classification & Regression

Assignment - 2

Group 29

Prashanth Seralathan(pseralat)
Shreya Ravi Hegde(shegde3)
Shreya Ravi Kumar(sr264)

Introduction

In this programming assignment we performed different classification and regression techniques on the Sample dataset and Diabetes dataset. The various experiments performed were as follows :

1. Gaussian Discriminators - Linear Discriminant Analysis(LDA) and Quadratic Discriminant Analysis(QDA)
2. Linear Regression
3. Ridge Regression
4. Using Gradient descent for Ridge regression
5. Non Linear regression

We have discussed about the various strategies used in classification that has been mentioned above, the corresponding results obtained and interpretation of the results.

1. Experiment with Gaussian Discriminators - LDA and QDA

The accuracy results obtained for Linear Discriminant Analysis and Quadratic Discriminant Analysis were as follows :

LDA Accuracy : 97%

QDA Accuracy : 96%

The plots obtained for LDA and QDA are as below:

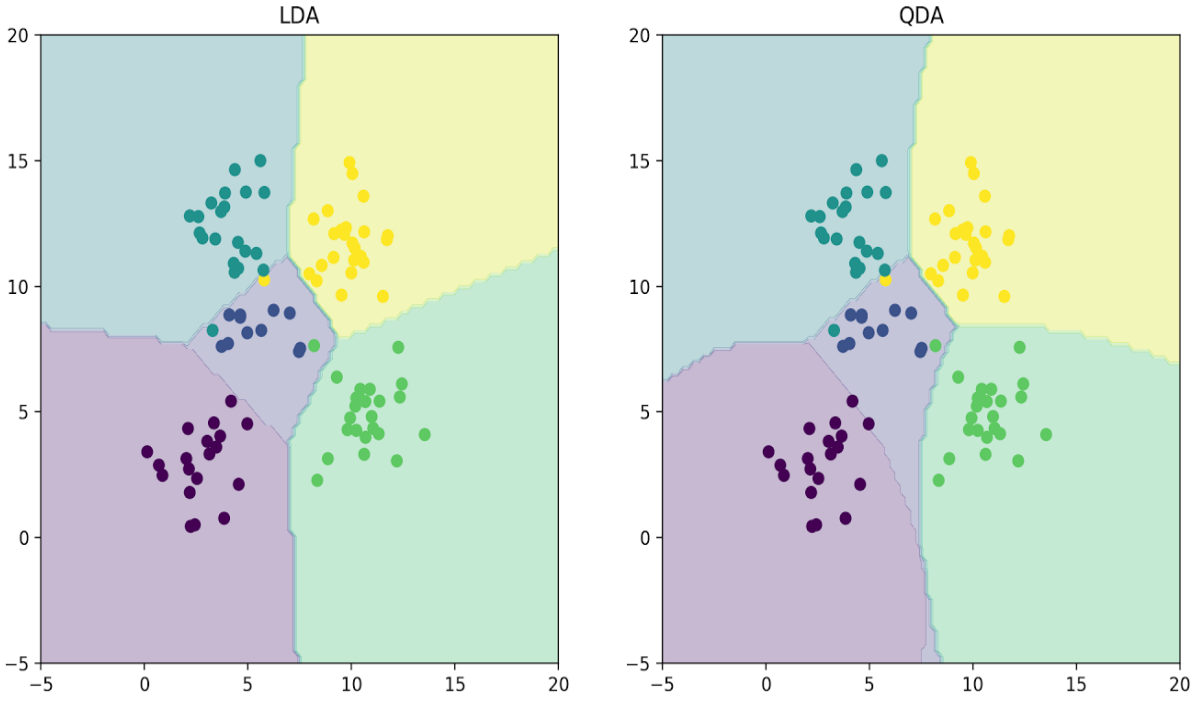

Fig. 1

In LDA, we assume that the covariance matrix is the same for all classes , i.e.,

$$\Sigma_k = \Sigma, \quad \forall k.$$

Hence by making this assumption, the classifier becomes linear. Since the covariance matrix determines the shape of the gaussian density, in LDA the gaussian densities for different classes have the same shape but are shifted versions of each other because of the varying mean vectors.

Whereas for QDA, the decision boundary is determined by a quadratic fraction and we calculate the covariance matrix Σ_k for each class k=1,2,3,4....K. In QDA, because it allows more flexibility for the covariance matrix tends to fit the data

better than LDA, but it has more parameters to estimate and becomes computationally expensive. LDA tends to perform better than QDA if there are relatively few training observations and data seems to fit the linear curve, QDA is recommended if the training data is very large so that the variance of the classifier is not a major concern.

Since the decision boundaries are a quadratic estimate for QDA and linear for LDA, there is the difference in the boundaries obtained as seen from the above graph.

2. Experiment with Linear Regression

It can be seen that that the linear regression model **performs better with an intercept.** This is due to the reason that adding a bias(intercept) makes the linear regression predictor line to align more closely with the actual data. Without the bias, the predictor line is forced to pass through the origin and thus it does not fit well with the actual data. Since most of the real world data is distributed randomly, It is hard to generate an accurate model without the bias.
The large values obtained from error function is because of the calculation of square of errors and it can be seen that the Linear model performs better in Test Data possibly because of less number of outliers and minimal deviation in the data.

Values obtained for the Mean Squared Error(MSE) for the training and test data are as follows :

	Test Data	**Training Data**
MSE without intercept	106775.361558	19099.4468446
MSE with intercept	3707.84018132	2187.16029493

It is evident from the above data that **Mean Square Errors calculated with intercepts is much lesser when compared to using a model without the intercepts.**

3. Experiment with Ridge Regression

In Ridge Regression,The best model that fits the data is generated by varying the regularization parameter λ. The regularization works as a error tuning parameter that is fed back to adjust the weights in such a way that it reduces the Mean Square Error. But there is a risk of overfitting if we train the model in such a way that it fits in perfectly with the training data.The ideal value of λ should be a point where there is a fine balance between overfitting and underfitting of data.It operates on the logic of L2 regularization, i.e., it adds penalty that is equal to the square of the magnitude of the coefficients.

MSE values of Train and test data for different values of λ arranged in ascending order of MSE values :

λ	MSE for train data	MSE for test data	λ	MSE for train data	MSE for test data
0.06	2451.528491	2851.330213	0.51	2932.260444	3166.921324
0.07	2468.077553	2852.349994	0.52	2940.827193	3174.813291
0.05	2433.174437	2852.665735	0.53	2949.331065	3182.688908
0.08	2483.365647	2854.879739	0.54	2957.772777	3190.547215
0.04	2412.119043	2858.00041	0.55	2966.153041	3198.387318
0.09	2497.740259	2858.444421	0.56	2974.472563	3206.208382
0.1	2511.432282	2862.757941	0.57	2982.732039	3214.009633
0.11	2524.600039	2867.637909	0.58	2990.93216	3221.790346
0.03	2386.780163	2870.941589	0.59	2999.073611	3229.549851
0.12	2537.3549	2872.962283	0.6	3007.157067	3237.287523
0.13	2549.776887	2878.645869	0.61	3015.183199	3245.002781
0.14	2561.924528	2884.626914	0.62	3023.152668	3252.695087
0.15	2573.841288	2890.85911	0.63	3031.066127	3260.363943
0.16	2585.559875	2897.306659	0.64	3038.924224	3268.008886
0.02	2354.071344	2900.973587	0.65	3046.727598	3275.629488
0.17	2597.105192	2903.941126	0.66	3054.476879	3283.225355
0.18	2608.4964	2910.739372	0.67	3062.172691	3290.796124
0.19	2619.748386	2917.682164	0.68	3069.81565	3298.341459
0.2	2630.872823	2924.753222	0.69	3077.406362	3305.861052
0.21	2641.878946	2931.938544	0.7	3084.945428	3313.354623

0.22	2652.774126	2939.22593	0.71	3092.43344	3320.821913
0.23	2663.564301	2946.604624	0.72	3099.870981	3328.262686
0.24	2674.254297	2954.065056	0.73	3107.258627	3335.676731
0.25	2684.848078	2961.598643	0.74	3114.596946	3343.063853
0.26	2695.348935	2969.197637	0.75	3121.886499	3350.423878
0.27	2705.759629	2976.855001	0.76	3129.127838	3357.75665
0.01	2306.832218	2982.44612	0.77	3136.321508	3365.062031
0.28	2716.082507	2984.564321	0.78	3143.468045	3372.339896
0.29	2726.319587	2992.319722	0.79	3150.567979	3379.590137
0.3	2736.47263	3000.115809	0.8	3157.621831	3386.812661
0.31	2746.543191	3007.947616	0.81	3164.630117	3394.007386
0.32	2756.532665	3015.810555	0.82	3171.593342	3401.174246
0.33	2766.442316	3023.700386	0.83	3178.512005	3408.313184
0.34	2776.273307	3031.613181	0.84	3185.3866	3415.424154
0.35	2786.026719	3039.545297	0.85	3192.21761	3422.507124
0.36	2795.703568	3047.493351	0.86	3199.005514	3429.562069
0.37	2805.30482	3055.454198	0.87	3205.750782	3436.588973
0.38	2814.831398	3063.424913	0.88	3212.453878	3443.587832
0.39	2824.284191	3071.402772	0.89	3219.115258	3450.558648
0.4	2833.664063	3079.385238	0.9	3225.735372	3457.50143
0.41	2842.971855	3087.369947	0.91	3232.314665	3464.416198
0.42	2852.208389	3095.354694	0.92	3238.853573	3471.302975
0.43	2861.374474	3103.337424	0.93	3245.352525	3478.161794
0.44	2870.470905	3111.316218	0.94	3251.811947	3484.992692
0.45	2879.498467	3119.289287	0.95	3258.232255	3491.795713
0.46	2888.457936	3127.254961	0.96	3264.613861	3498.570906
0.47	2897.350077	3135.211679	0.97	3270.95717	3505.318324
0.48	2906.17565	3143.157988	0.98	3277.262582	3512.038029
0.49	2914.935407	3151.09253	0.99	3283.53049	3518.730082
0.5	2923.630092	3159.014036	0	2187.160295	3707.840182

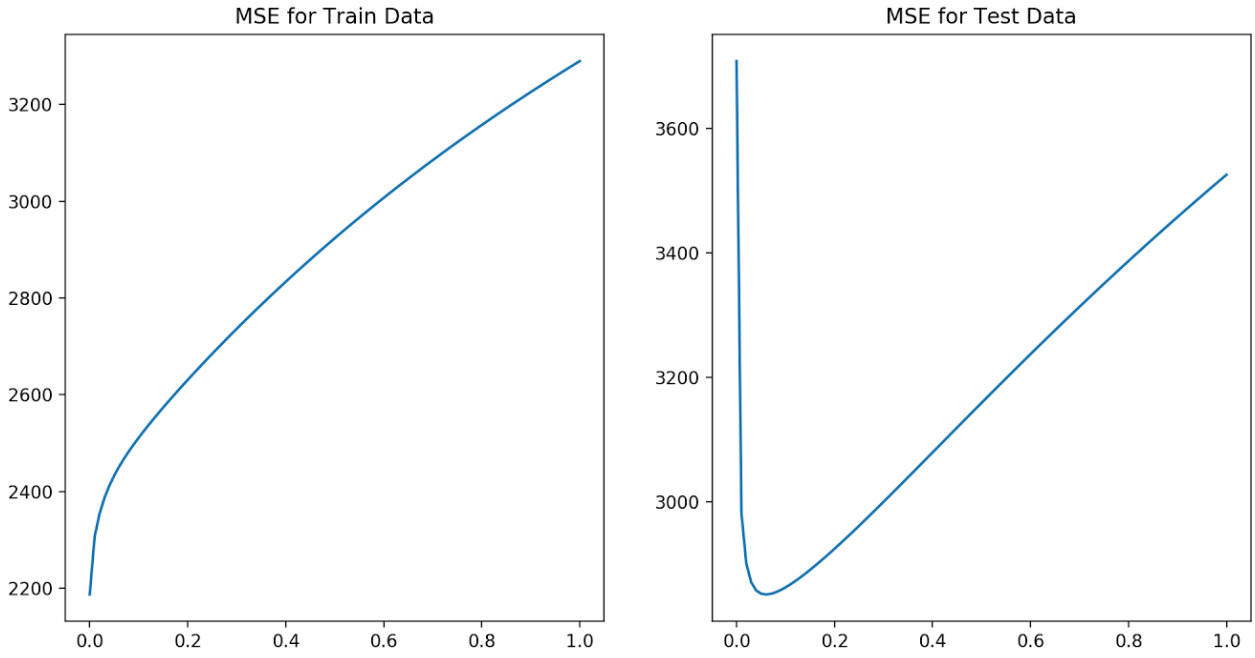

Fig. 2

MSE plots for Training data and Test data plotted by varying values of Regularization parameter(λ) is as shown in the above graph.

It can be interpreted from the results that the Mean Square Error(MSE) values increases as the value of regularization goes up, this explains the increase in MSE after certain value for regularization.

The comparison of weights learnt between linear regression(OLE) and Ridge regression can be plotted as follows :

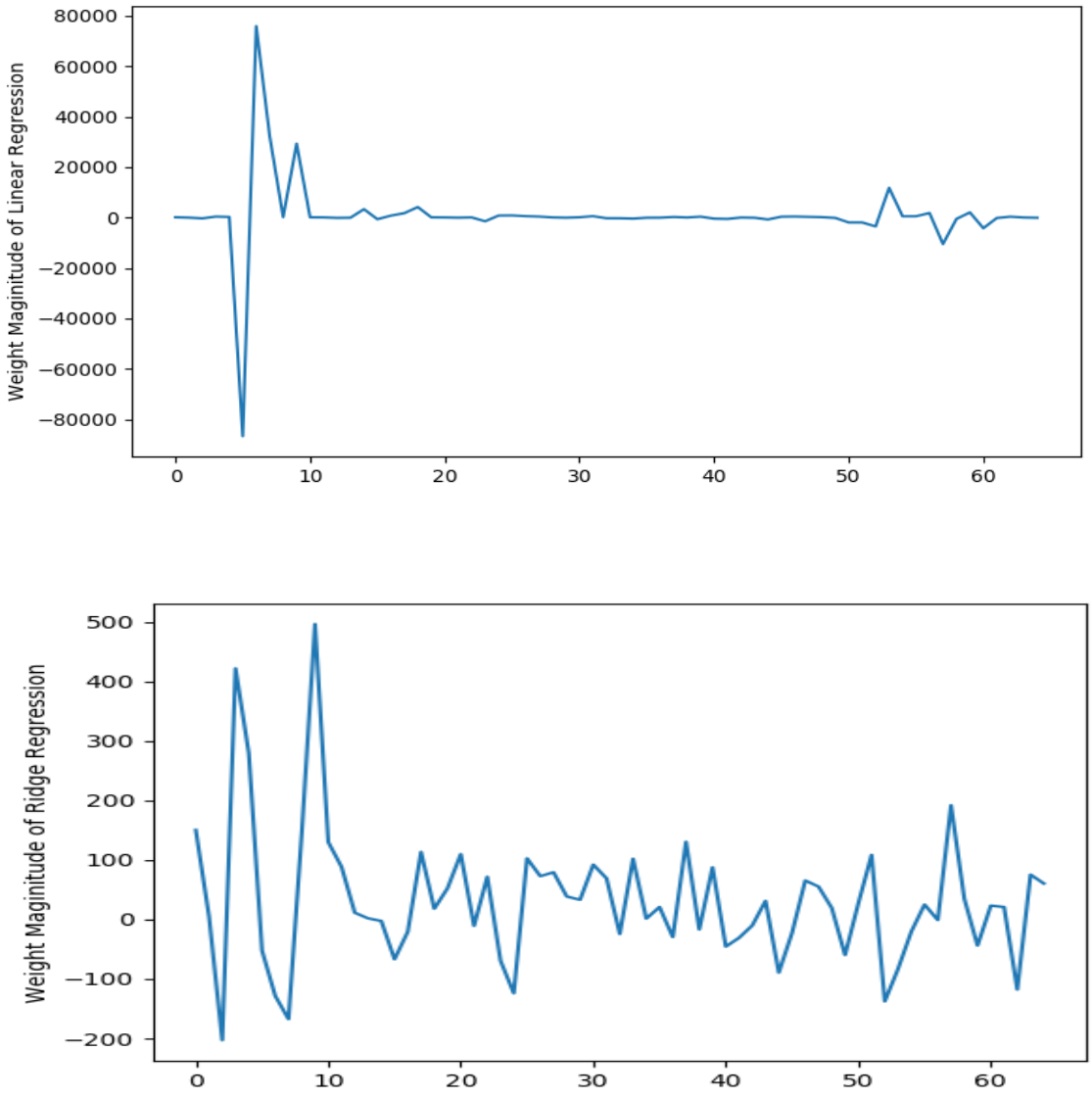

Fig. 3

Mean Value of Weights obtained in Linear Regression: 882.807624977
Mean Value of Weights obtained in Ridge Regression: 32.0451365211

It is evident from the above graph that the **ideal value for Regularization parameter(λ) is 0.06 as it contains the minimum MSE.**

4. Using Gradient Descent with Ridge Regression

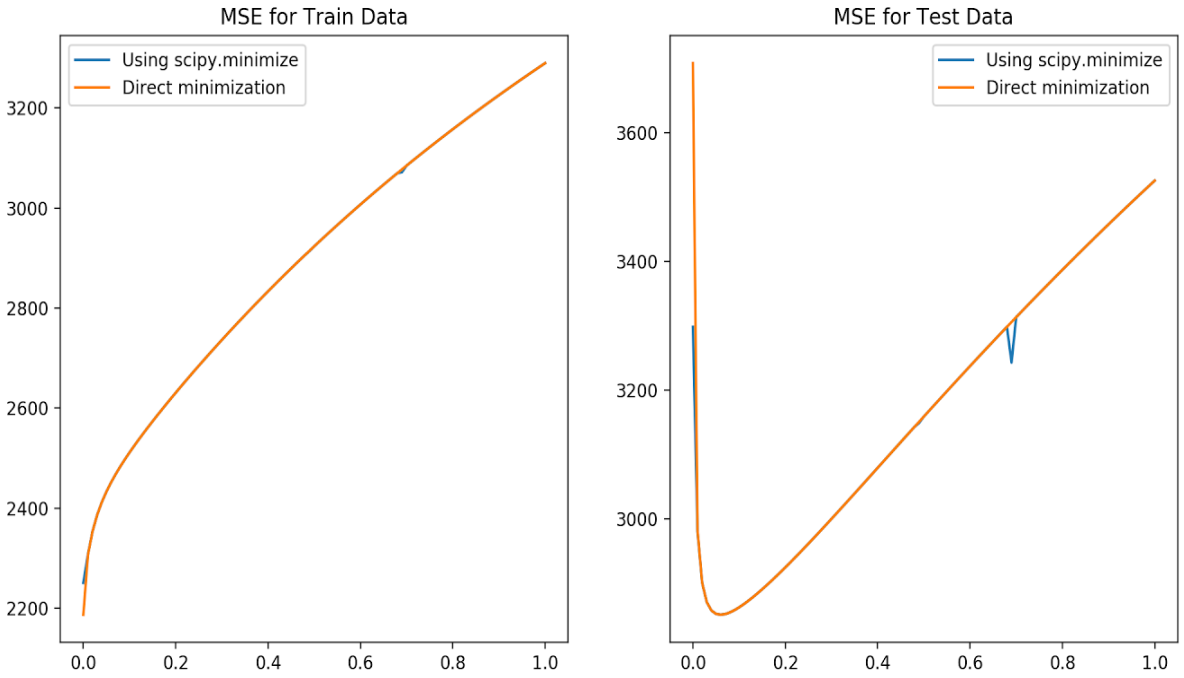

Fig. 3

The above graph plots MSE of Training Data and Test Data for varying values of regularization parameter (λ)

Comparing the results of Gradient Descent with the results obtained using Ridge Regression we can see that the MSE values are almost similar. But if we observe closely we can see that there is some distortions present when using Gradient Descent - The reason being that gradient descent uses the minimizer function.

5. Non - Linear Regression

Non - Linear regression is a type of regression where the regression function is modelled as a nonlinear function of the unknown parameters.

$$p(y|x, \theta) \sim N(w > \phi(x))$$

Where, $\phi(x) = [1, x, x^2, \ldots, x^p]$

The Mean Squared Error(MSE) values has been computed with two cases:

Without regularization , $\lambda = 0$.

With regularization test case, $\lambda = 0.06$.

Tabulation of polynomial powers(p) and the corresponding MSE values for λ=0 and optimal λ=0.06 obtained from ridge regression:

λ=0	Without Regularization			λ=0.06	With Regularization	
p	Test Data	Train Data		p	Test Data	Train Data
0	6286.40479168	5650.7105389		0	6286.88196694	5650.71190703
1	3845.03473017	3930.91540732		1	3895.85646447	3951.83912356
2	3907.12809911	3911.8396712		2	3895.58405594	3950.68731238
3	3887.97553824	3911.18866493		3	3895.58271592	3950.68253152
4	4443.32789181	3885.47306811		4	3895.58266828	3950.6823368
5	4554.83037743	3885.4071574		5	3895.5826687	3950.68233518
6	6833.45914872	3866.88344945		6	3895.58266872	3950.68233514

From the above table, we can infer the following results :

λ = 0,

The most optimal results for **Training Data** is 3866.88344945 obtained for p = 6.

The most optimal results for the **Test Data** is 3845.03473017 obtained for p = 1.

λ = 0.06,

The most optimal results for **Training Data** is 3950.68233514 obtained for p=6.

The most optimal results for the **Test Data** is 3895.58266828 obtained for p=4.

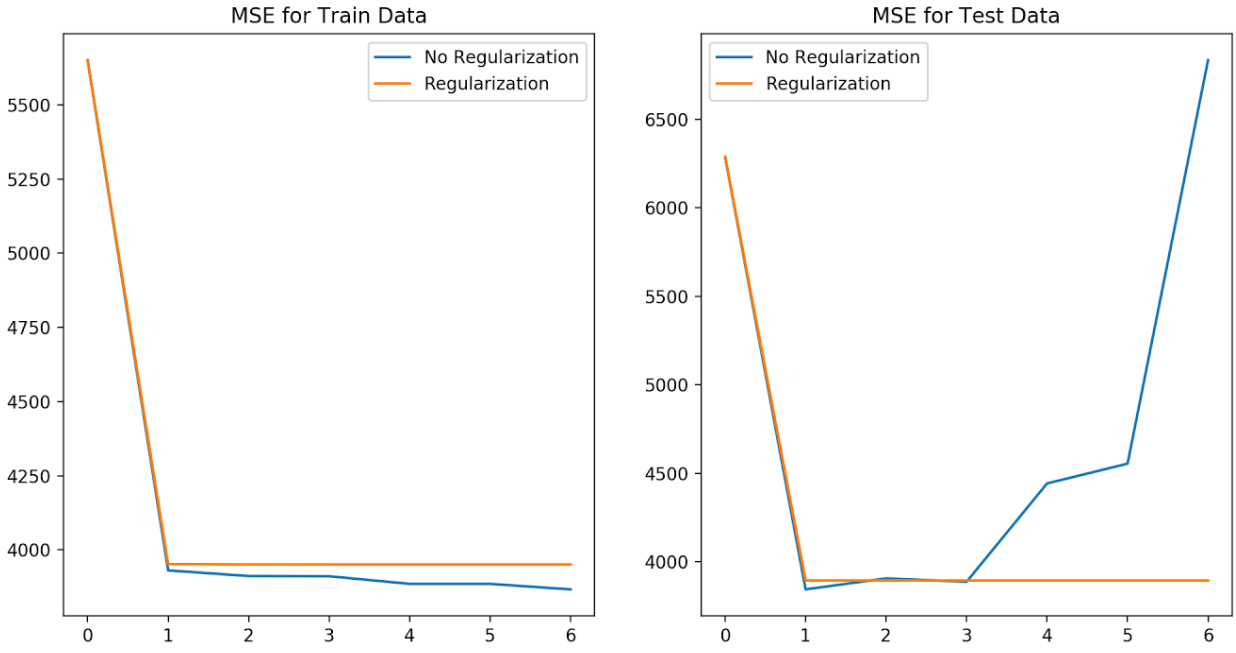

Fig. 4

6. Interpretation of Data:

Models	Training MSE	Test MSE
Linear Regression - Without Intercept	19099.4468446	106775.361558
Linear Regression - With Intercept	2187.16029493	3707.84018132
Ridge Regression($\lambda = 0.06$)	2451.528491	2851.330213
Gradient Descent	Similar to Ridge Regression	Similar to Ridge Regression
Non- Linear Regression - Without regularization ($\lambda = 0$)	3866.88344945	3845.03473017
Non- Linear Regression - Without regularization ($\lambda = 0.06$)	3950.68233514	3895.58266828

Conclusion:

It can be seen that generally the regression models works based on the concept of minimizing the errors between prediction and the actual model, i.e it has to have a **low Mean Squared Error(MSE) value**. It is evident from the above data that Ridge Regression performs a better job in classification of the diabetes data. The best metric to choose would be based on lower value of MSE, Ridge Regression does a better performance compared to Linear Regression and Non linear regression. It is also notable that ridge regression had a better performance since it was dealing with smaller set of data.

In cases where we would have to deal with large amount of data we can choose Gradient descent along with Ridge Regression as it provides almost similar accuracy and it is not that computationally expensive.