CSE574-ASSIGNMENT 2 – GROUP 52
CLASSIFCATION AND REGRESSION

April 12, 2017

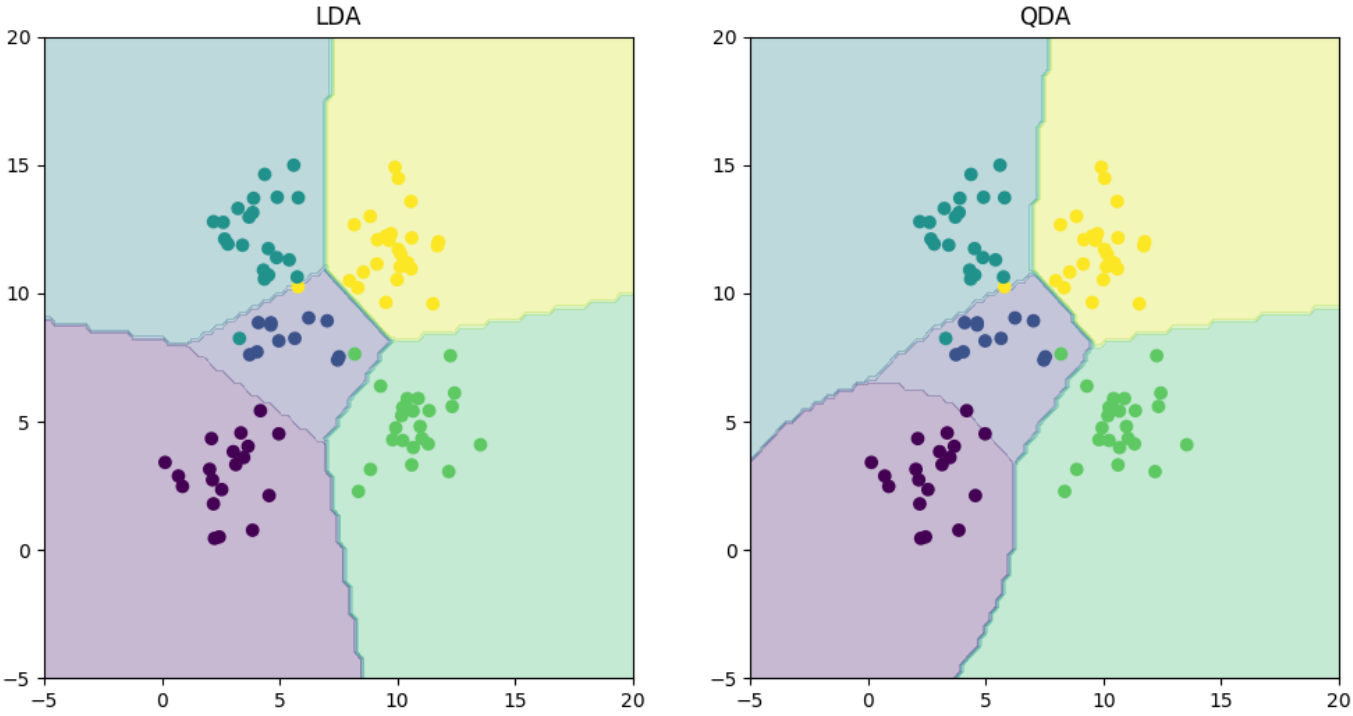
Name	UBIT Name	Person #
Sourav Puri	Souravpu	50206918
Shaleen Somani	Shaleens	50207380
Amardeep Virdy	Avirdy	50208831

Problem 1: Experiment with Gaussian Discriminators

LDA Accuracy = 97.0

QDA Accuracy = 96.0

Discriminating boundary for LDA and QDA



Reason for difference in the two boundaries:

LDA and QDA are boundary discriminant methods, which aim to find boundaries that separate groups or classes of samples. LDA obtains linear boundaries, where a straight line or hyperplane divides the variable space into regions whereas QDA obtains quadratic boundaries, where a quadratic curve divides the variable space into regions.

In case of LDA, we assume a single variance–covariance matrix over all classes, which means we do not take into account different variance structures for different classes. On the other hand, for QDA we assume different variance–covariance matrices for each class, that helps in discriminating classes which have significantly different class-specific covariance matrices and forms a separate variance model for each class.

Linear discriminant function is:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log(\pi_k)$$

Quadratic discriminant function is:

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$

We observe that discriminant function for QDA is very much like that of LDA except that because the covariance matrix (Σ_k), is not identical, we cannot throw away the quadratic terms. This discriminant function is a quadratic function and will contain second order terms.

Problem 2: Experiment with Linear Regression

Training data:

MSE without intercept: **19099.44684457**

MSE with intercept: **2187.16029493**

Test data:

MSE without intercept: **106775.36155789**

MSE with intercept: **3707.84018132**

Which one is better?

MSE calculated using intercept is better for both, training data as well as test data. We can observe that value of MSE drastically decreases when intercept is used. In the case of training data MSE decreases by a factor of 8.7 when intercept is used, while in the case of test data MSE decreases by a factor of 28.7917. It is also observable that the MSE decrease in case of test data is sharper and more drastic upon the use of intercept, than the MSE decrease in training data.

In summary it is worthy to note that use of intercept is better because it decreases the MSE observed drastically.

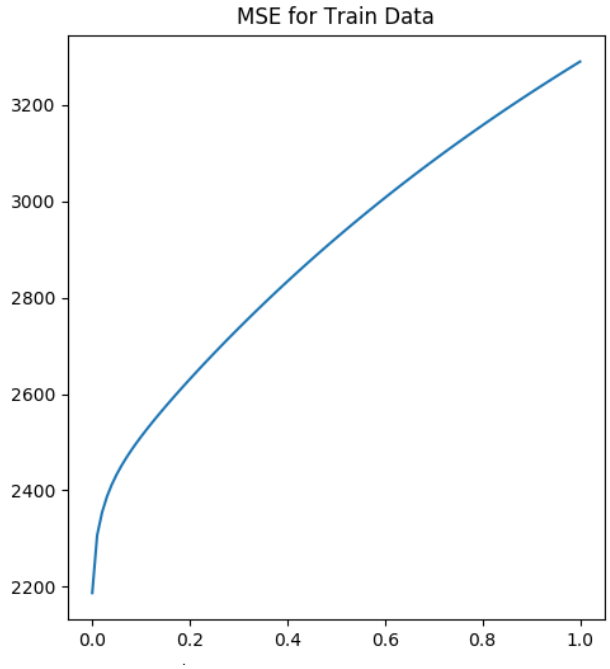
Problem 3: Experiment with Ridge Regression

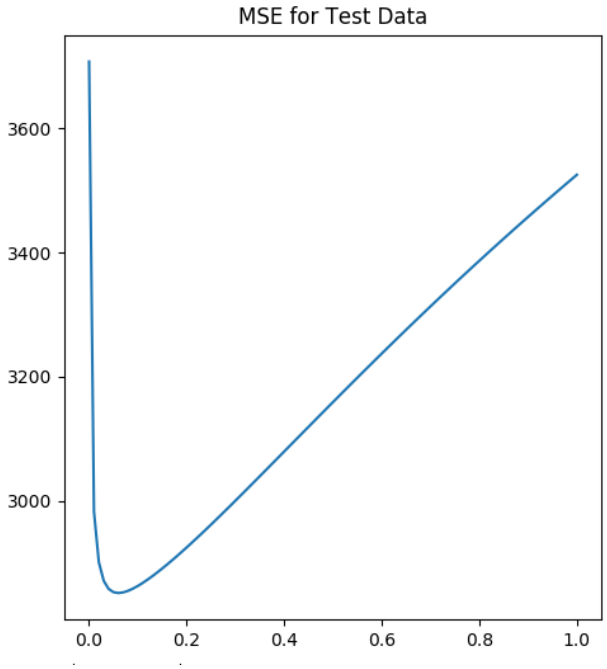
Below are the computed MSE for training and test data:

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

Lambda	Training Data	Test Data	Lambda	Training Data	Test Data
0.52	2940.827193	3174.813291	0.78	3143.468045	3372.339896
0.53	2949.331065	3182.688908	0.79	3150.567979	3379.590137
0.54	2957.772777	3190.547215	0.8	3157.621831	3386.812661
0.55	2966.153041	3198.387318	0.81	3164.630117	3394.007386
0.56	2974.472563	3206.208382	0.82	3171.593342	3401.174246
0.57	2982.732039	3214.009633	0.83	3178.512005	3408.313184
0.58	2990.93216	3221.790346	0.84	3185.3866	3415.424154
0.59	2999.073611	3229.549851	0.85	3192.21761	3422.507124
0.6	3007.157067	3237.287523	0.86	3199.005514	3429.562069
0.61	3015.183199	3245.002781	0.87	3205.750782	3436.588973
0.62	3023.152668	3252.695087	0.88	3212.453878	3443.587832
0.63	3031.066127	3260.363943	0.89	3219.115258	3450.558648
0.64	3038.924224	3268.008886	0.9	3225.735372	3457.50143
0.65	3046.727598	3275.629488	0.91	3232.314665	3464.416198
0.66	3054.476879	3283.225355	0.92	3238.853573	3471.302975
0.67	3062.172691	3290.796124	0.93	3245.352525	3478.161794
0.68	3069.81565	3298.341459	0.94	3251.811947	3484.992692
0.69	3077.406362	3305.861052	0.95	3258.232255	3491.795713
0.7	3084.945428	3313.354623	0.96	3264.613861	3498.570906
0.71	3092.43344	3320.821913	0.97	3270.95717	3505.318324
0.72	3099.870981	3328.262686	0.98	3277.262582	3512.038029
0.73	3107.258627	3335.676731	0.99	3283.53049	3518.730082
0.74	3114.596946	3343.063853	1	3289.761281	3525.394553
0.75	3121.886499	3350.423878			
0.76	3129.127838	3357.75665			
0.77	3136.321508	3365.062031			

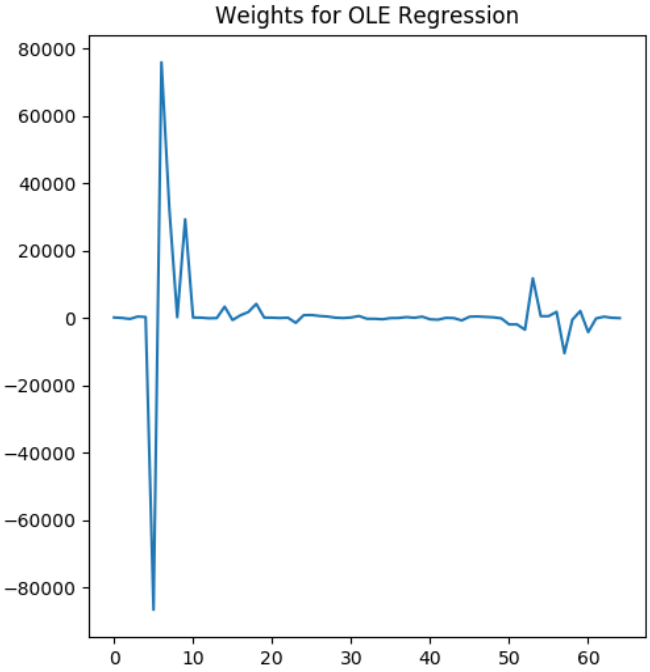
Errors on test and train data for different lambda values

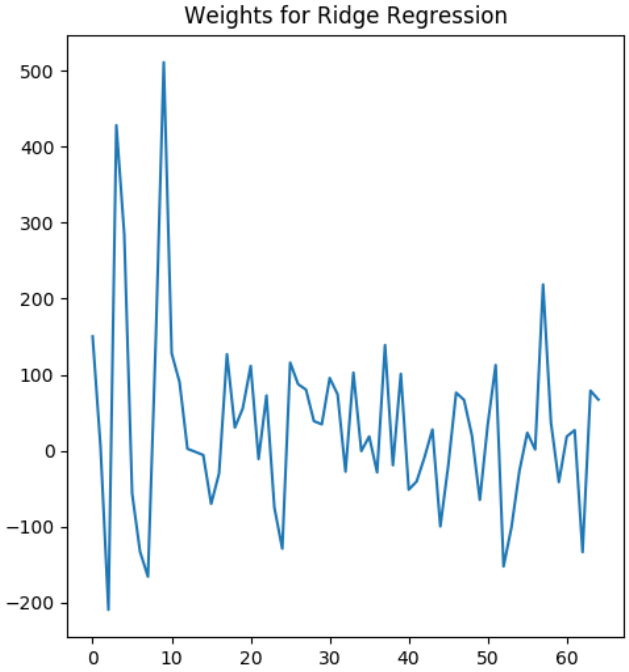




The optimal value for **lambda is 0.05** because we obtain the least error at this point, as can be seen from the plot above

*Taking **lambda=0.05**, we obtain the following weights for ridge regression which have been plotted below alongside weights obtained through OLE regression.*





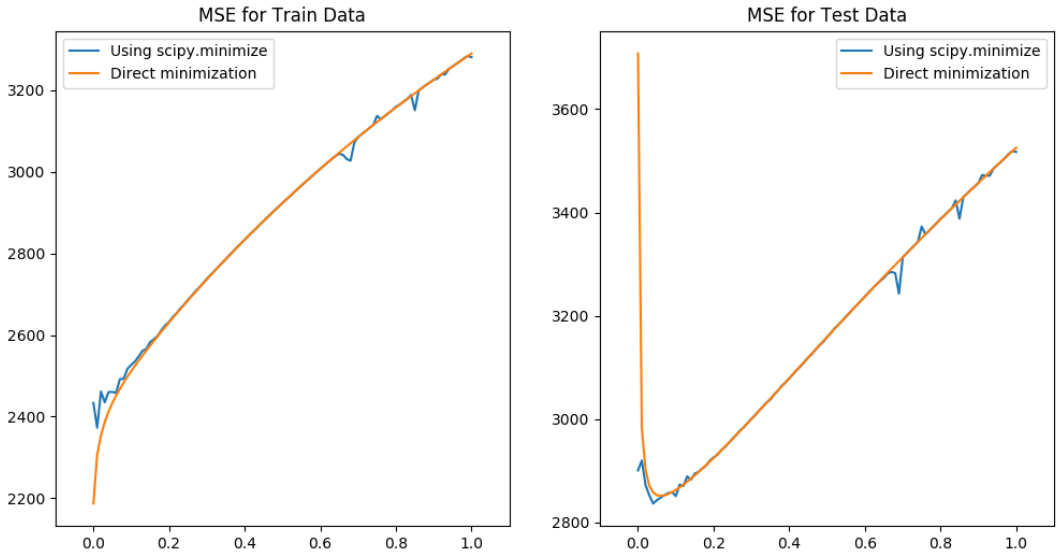
Using numpy.linalg.norm(), we have computed the below values:

The 2 norm value for OLE weights is **124531.526519**

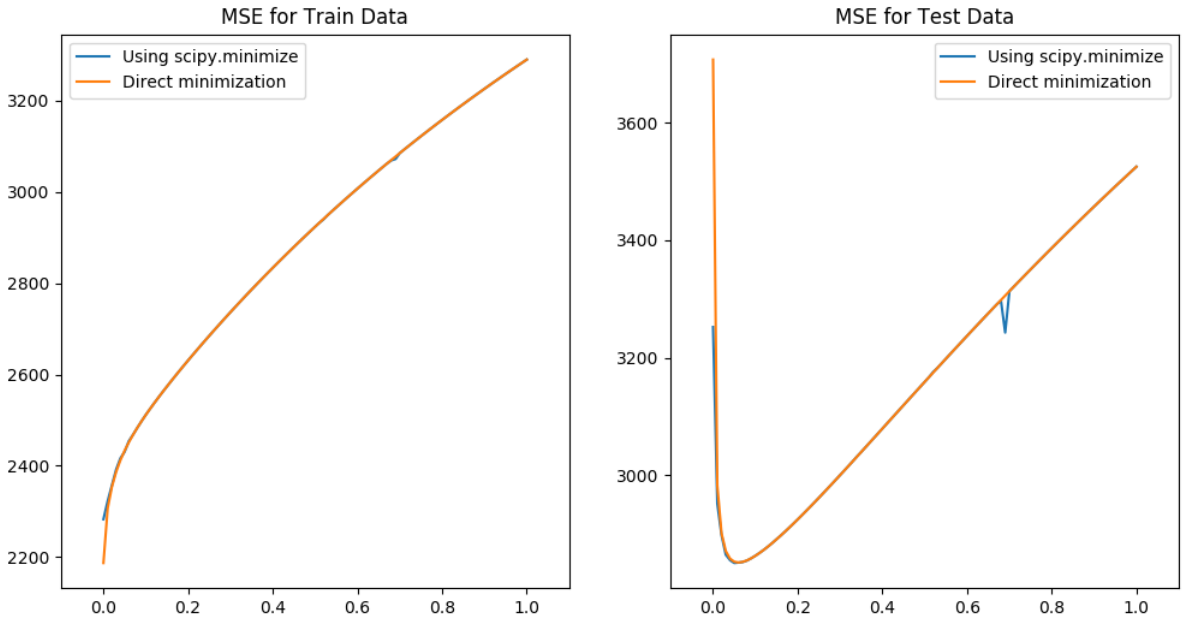
The 2 norm value for ridge regression weights is **1000.70393822**

Problem 4: Gradient Descent for Ridge Regression Learning

Iterations: 20

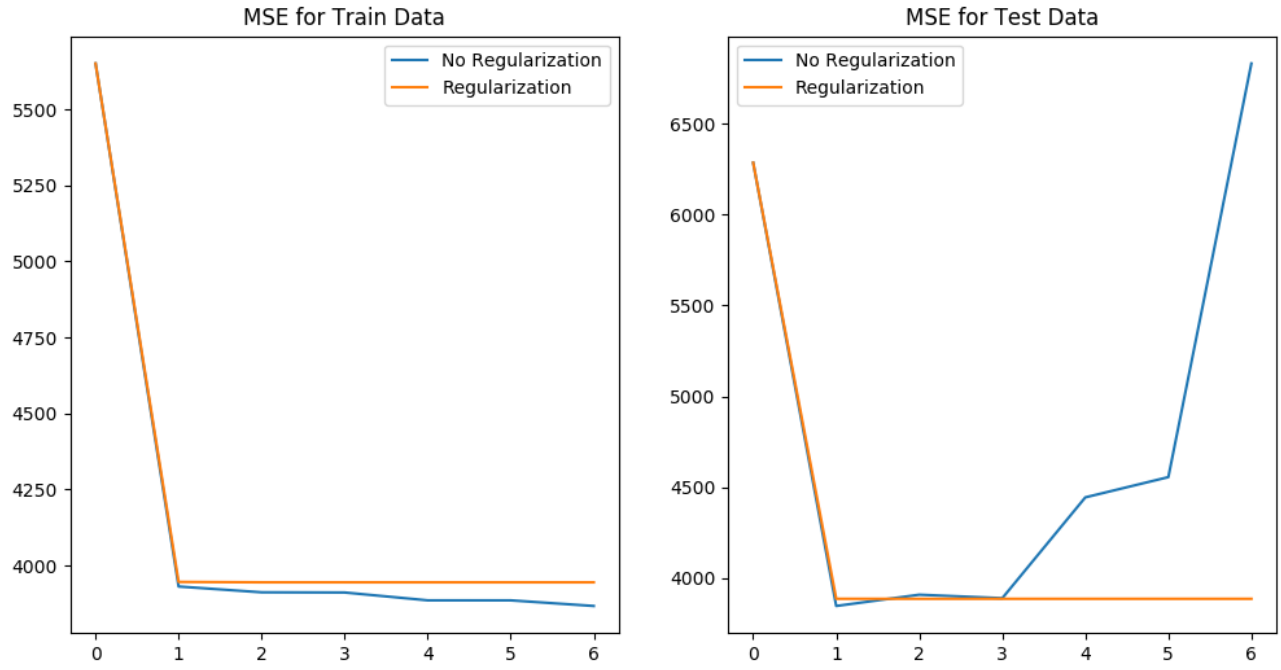


Iterations: 50



From the graphs plotted above, it can be seen that there is no much difference between the MSE values plotted for Ridge Regression with and without gradient descent. There are some spikes in between the data but it could be for some small outliers or error.

Problem 5: Non-linear Regression



Train data:

p-value	MSE for lambda=0 (No regularizaion)	MSE for lambda=0.05 (With regularization)
0	5650.710539	5650.711489
1	3930.915407	3945.994834
2	3911.839671	3944.677681
3	3911.188665	3944.673053
4	3885.473068	3944.672833
5	3885.407157	3944.672831
6	3866.883449	3944.672831

Test data:

p-value	MSE for lambda=0 (No regularizaion)	MSE for lambda=0.05 (With regularization)
0	6286.404792	6286.802264
1	3845.03473	3884.695623
2	3907.128099	3884.545679
3	3887.975538	3884.546642
4	4443.327892	3884.546636
5	4554.830377	3884.546637
6	6833.459149	3884.546637

In terms of test error, we can observe from the table above, that MSE value for **p=1** is minimum. Therefore**, p=1** is the optimal value.

Problem 6:

From our experiments in problem 2, it becomes clear that the use of intercept is beyond doubt essential due to the fact that it contributes heavily towards decreasing the Mean square error observed. Hence intercept should always be used for classification.

As we can see from the graphs of the weights, it is clearly visible that in OLE regression the value of weights computed, range from **-80000 to +80000**, whereas in Ridge regression the value of weights ranges from **-200 to 500**. Comparison of weights after 2 norm normalization further proves the fact that the weights obtained in Ridge regression are far lower than the weights observed during OLE regression. Thus, it is feasible to say that use of ridge regression is better in terms of computation costs due to lower values of weights in comparison to OLE regression.

Upon computation of MSE, the lowest value of MSE for test data for OLE regression is **3707.84** (with intercept) & the lowest value of MSE for test data for Ridge regression is **2856 (lambda = 0.05).**
It is also observable that the lowest value of MSE for training data for OLE regression is **2187 (with Intercept)** & the lowest value of MSE for training data for Ridge regression is **2185 (lambda = 0.05).**
We can now conclude that Ridge regression is more accurate than OLE regression due to the fact that Ridge regression has lower MSE (Mean square error).

Hence, it is safer to say that Ridge regression is better that OLE regression in terms of computation cost as well as the accuracy of the results.

Let us recall that in Ridge regression, solving for weight (w) requires inverting D × D **matrices $\mathbf{X^TX}$ or $\mathbf{(X^TX + \lambda I)}$.** Matrix inversion can be expensive if data dimensionality D is large.
Thus, when working on larger datasets it is also worth noting we can also use Gradient Descent to obtain similar results.
However, use of Gradient Descent results in loss of accuracy. Accuracy of ridge regression using Gradient Descent can be improved by increasing the number of iterations.

References:

1. *https://onlinecourses.science.psu.edu/stat857/node/75*
2. *https://www.cs.utah.edu/~piyush/teaching/6-9-slides.pdf*
3. *http://www.sciencedirect.com/science/article/pii/S0169743916303318*