CS 6140: Machine Learning Spring 2018

Assignment 2

Linear and Ridge Regression

2. Gradient Descent for Linear Regression

2.1 Learning Regression Coefficients using Gradient Descent

5.

Housing Dataset - Gradient Descent				
Fold	**Training SSE**	**Training RMSE**	**Test SSE**	**Test RMSE**
1	10160.96886	4.725653362	1442.49723	5.318294783
2	9756.604383	4.630668046	1759.616144	5.873863918
3	10323.13093	4.763213184	1167.82376	4.785238249
4	10482.18507	4.799767639	981.7026897	4.387376477
5	10432.16738	4.788302454	1079.61773	4.600975451
6	10247.25203	4.745675193	1258.875506	4.968282747
7	10588.37448	4.818725918	796.120846	3.99029033
8	10419.71108	4.780192841	1017.163188	4.510350736
9	10287.02885	4.749660387	1222.853597	4.945409178
10	9975.591866	4.677210488	1580.325863	5.621967384
Mean	10267.30149	4.747906951	1230.659655	4.900204925
Standard Deviation	236.8106854	0.0546438847	277.6580545	0.5473672913

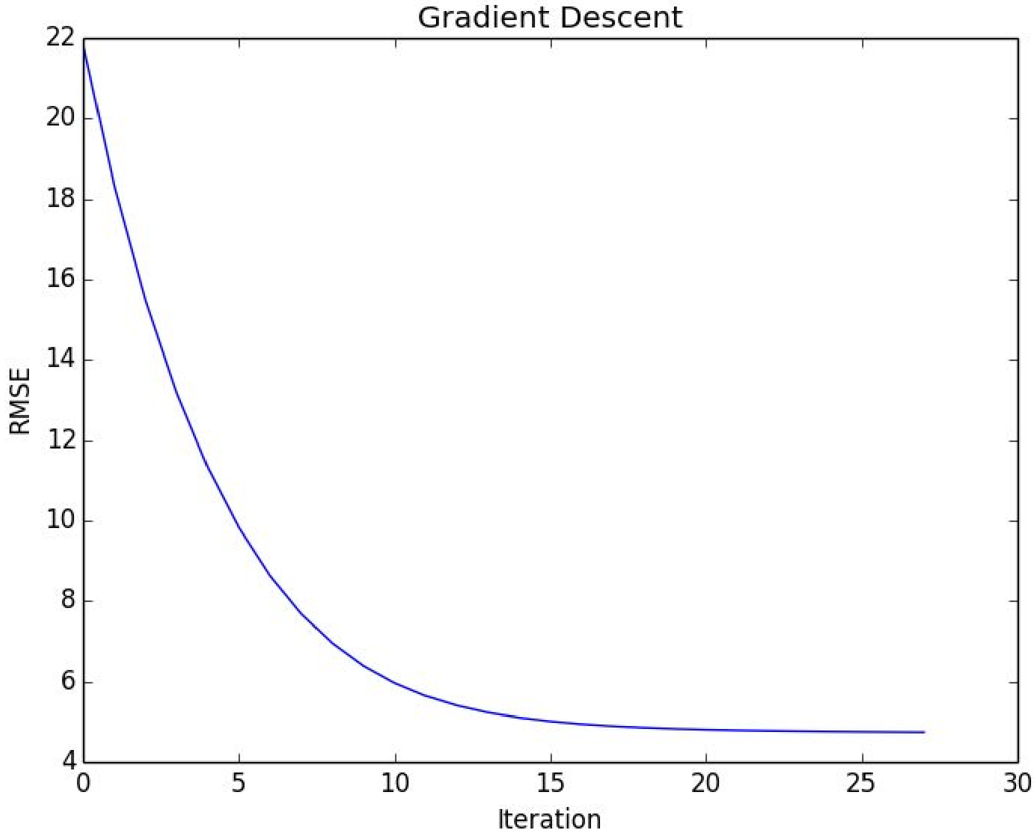
Yacht Dataset - Gradient Descent				
Fold	**Training SSE**	**Training RMSE**	**Test SSE**	**Test RMSE**
1	22247.35622	8.961884032	1975.846355	7.983544275
2	22410.96376	8.994776615	1847.200554	7.719269049
3	21887.58669	8.889125853	2346.270862	8.699779125
4	20635.85279	8.631202899	3650.152839	10.85112279
5	22346.01896	8.981734165	1912.241572	7.85399338
6	21381.54754	8.78576703	2828.193698	9.551547245

7	21144.12895	8.736852721	3050.335068	9.919571836
8	21378.28337	8.785096374	2841.684106	9.574300455
9	21994.59955	8.894788618	2205.606083	8.574392269
10	21597.96078	8.814221827	2639.907787	9.380667687
Mean	21702.42986	8.847545013	2529.743892	9.010818811
Standard Deviation	547.8188477	0.1116251266	549.3224121	0.9647844545

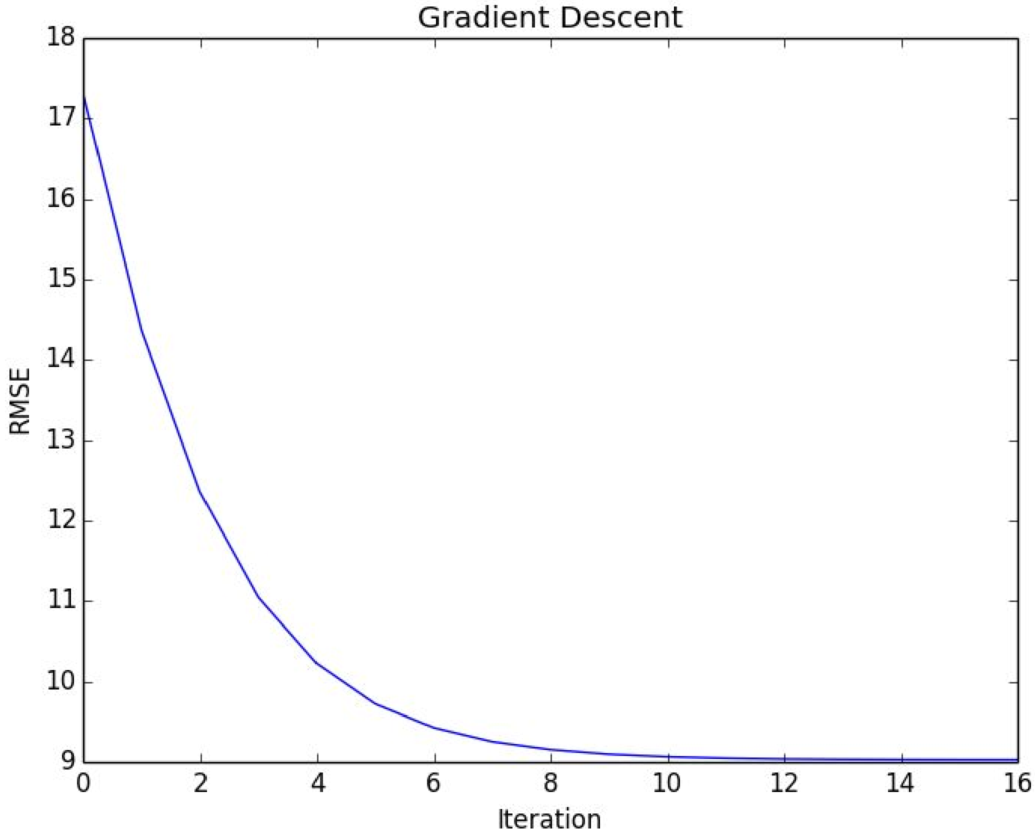
Concrete Dataset - Gradient Descent				
Fold	**Training SSE**	**Training RMSE**	**Test SSE**	**Test RMSE**
1	97074.11186	10.36257049	8728.653608	9.296360199
2	95015.54584	10.25210673	10893.40675	10.38535085
3	94423.63335	10.22012338	11755.07262	10.78827408
4	93773.72593	10.18489066	12072.91717	10.93315295
5	92260.9827	10.10240611	13746.82713	11.66649902
6	95903.59907	10.29421331	10001.1115	10.00055573
7	96612.55191	10.33219246	9332.320124	9.660393431
8	97055.93678	10.35587413	9138.656959	9.559632293
9	95165.37361	10.25451652	10748.30679	10.36740411
10	94341.47688	10.21003065	11827.61807	10.87548531
Mean	95162.69379	10.25689244	10824.48907	10.3533108
Standard Deviation	1465.617395	0.0778813124	1487.296687	0.698241035

6.

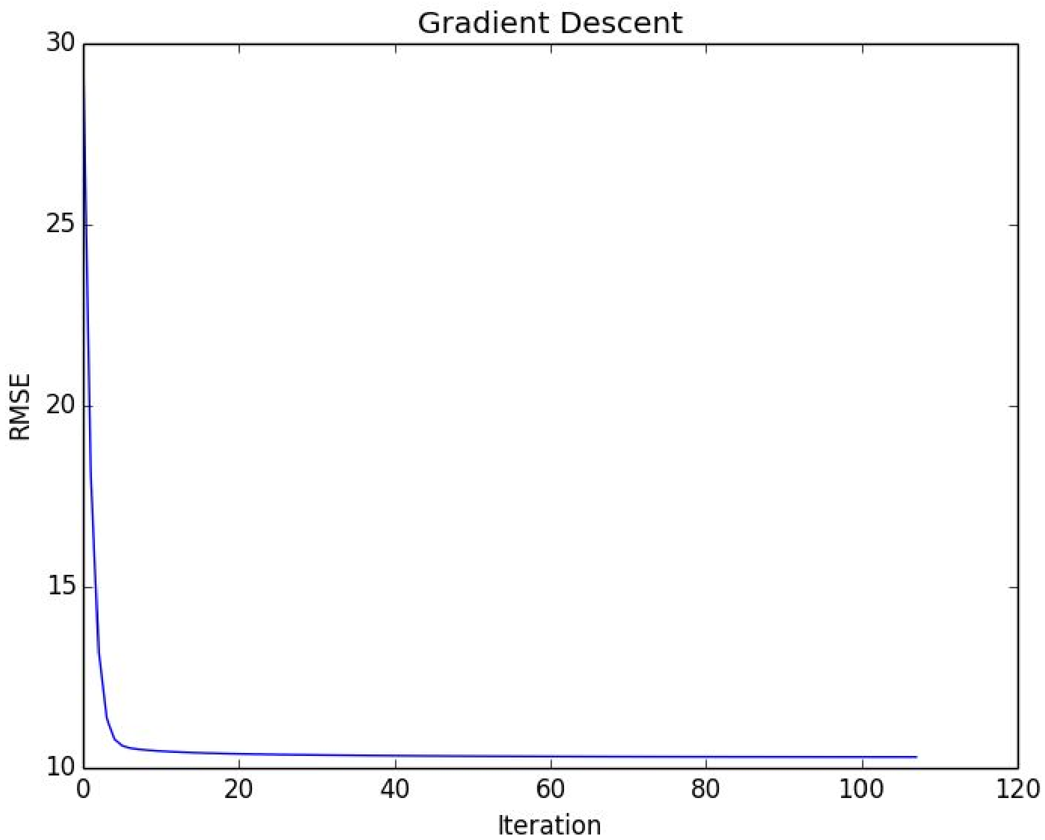
Housing Dataset - Gradient Descent - Fold 1



Yacht Dataset - Gradient Descent - Fold 2


Gradient Descent
RMSE
Iteration
18
17
16
15
14
13
12
11
10
9
0
2
4
6
8
10
12
14
16

Concrete Dataset - Gradient Descent - Fold 3



2.2 Interpreting the results

1. Changes in the initial weights have an effect on the initial gradient descent but after a couple of iterations the learning rate has more weightage on choosing the next gradient and thus any initial weights mostly end up with the same final coefficients and RMSE.

2. Yes tolerance parameter has an effect on the results. If a smaller tolerance level is used then it means that the gradient descent is closer to the minima thus decreasing the RMSE of the regression and in turn improving the prediction.

3. If a large learning rate is used then theres a high change of skipping the optimal solution and if it is too small then the regression takes too many steps to converge to the best solution.

3. Least Squares Regression using Normal Equations

Housing Dataset - Normal Equation				
Fold	**Training SSE**	**Training RMSE**	**Test SSE**	**Test RMSE**
1	10077.31765	4.706160949	1037.434228	4.510193745

2	9597.581874	4.592775472	1598.899977	5.599194579
3	10374.98157	4.775160447	740.2946427	3.809931919
4	10103.57528	4.712288187	1015.543028	4.46235458
5	9994.298907	4.686735762	1137.326755	4.722343127
6	9588.720788	4.590654814	1513.320359	5.447288152
7	9673.498219	4.605845404	1435.620933	5.358397024
8	9296.55083	4.515215706	1878.485419	6.12941338
9	10073.48049	4.700102763	1038.26277	4.55689098
10	10520.35568	4.803223443	579.2201346	3.403586739
Mean	9930.036129	4.668816295	1197.440825	4.799959423
Standard Deviation	362.9193589	0.08567884048	382.1414086	0.7956236901

Yacht Dataset - Normal Equation				
Fold	**Training SSE**	**Training RMSE**	**Test SSE**	**Test RMSE**
1	22578.8877	9.028412373	1656.559753	7.310089697
2	20499.74955	8.602692372	3783.195537	11.04710667
3	22213.46699	8.955055645	2003.864212	8.039949068
4	21308.81821	8.77081191	2885.108944	9.64717733
5	20188.74303	8.537186237	4125.078168	11.53546868
6	21353.92065	8.780089194	2930.265527	9.722381108
7	22860.29796	9.084500618	1361.78442	6.627860113
8	22434.92421	8.999583667	1820.150977	7.662541852
9	22055.95222	8.90718573	2132.896845	8.431877698
10	21279.90154	8.749080433	2937.396786	9.895111228
Mean	21677.46621	8.841459818	2563.630117	8.991956344
Standard Deviation	849.5820596	0.1741065432	873.0275764	1.545205382

Housing Dataset - Gradient Descent vs Normal Equation				
Fold	**Training RMSE - Gradient Descent**	**Training RMSE - Normal Equation**	**Test RMSE - Gradient Descent**	**Test RMSE - Normal Equation**
1	4.725653362	4.706160949	5.318294783	4.510193745
2	4.630668046	4.592775472	5.873863918	5.599194579
3	4.763213184	4.775160447	4.785238249	3.809931919
4	4.799767639	4.712288187	4.387376477	4.46235458
5	4.788302454	4.686735762	4.600975451	4.722343127
6	4.745675193	4.590654814	4.968282747	5.447288152
7	4.818725918	4.605845404	3.99029033	5.358397024

8	4.780192841	4.515215706	4.510350736	6.12941338
9	4.749660387	4.700102763	4.945409178	4.55689098
10	4.677210488	4.803223443	5.621967384	3.403586739
Mean	4.747906951	4.668816295	4.900204925	4.799959423
Standard Deviation	0.0546438847	0.08567884048	0.5473672913	0.7956236901

Yacht Dataset - Gradient Descent vs Normal Equation				
Fold	**Training RMSE - Gradient Descent**	**Training RMSE - Normal Equation**	**Test RMSE - Gradient Descent**	**Test RMSE - Normal Equation**
1	8.961884032	9.028412373	7.983544275	7.310089697
2	8.994776615	8.602692372	7.719269049	11.04710667
3	8.889125853	8.955055645	8.699779125	8.039949068
4	8.631202899	8.77081191	10.85112279	9.64717733
5	8.981734165	8.537186237	7.85399338	11.53546868
6	8.78576703	8.780089194	9.551547245	9.722381108
7	8.736852721	9.084500618	9.919571836	6.627860113
8	8.785096374	8.999583667	9.574300455	7.662541852
9	8.894788618	8.90718573	8.574392269	8.431877698
10	8.814221827	8.749080433	9.380667687	9.895111228
Mean	8.847545013	8.841459818	9.010818811	8.991956344
Standard Deviation	0.1116251266	0.1741065432	0.9647844545	1.545205382

4. Deriving Normal Equations for Univariate Regression

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$y = w_0 + w_1 x$$

Minimizing $SSE \Rightarrow \frac{\partial SSE}{\partial w_0} = 0$ & $\frac{\partial SSE}{\partial w_1} = 0$

$$SSE = \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

$$\frac{\partial SSE}{\partial w_0} = -2 \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i) = 0$$

$$\sum_{i=1}^{N} y_i = w_0 \sum_{i=1}^{N} 1 + w_1 \sum_{i=1}^{N} x_i$$

$$N\bar{Y} = w_0 N + w_1 N \bar{X}$$

$$\bar{Y} = w_0 + w_1 \bar{X}$$

$$\boxed{w_0 = \bar{Y} - w_1 \bar{X}}$$

$$\frac{\partial SSE}{\partial w_1} = -2 \sum_{i=1}^{N} x_i (y_i - w_0 - w_1 x_i) = 0$$

$$\sum_{i=1}^{N} x_i y_i = w_0 \sum_{i=1}^{N} x_i + w_1 \sum_{i=1}^{N} x_i^2$$

$$\sum_{i=1}^{N} x_i y_i = w_0 N \bar{x} + w_1 \sum_{i=1}^{N} x_i^2$$

$$\sum_{i=1}^{N} x_i y_i = (\bar{Y} - w_1 \bar{x}) N \bar{x} + w_1 \sum_{i=1}^{N} x_i^2$$

$$\sum_{i=1}^{N} x_i y_i = \bar{Y} N \bar{x} - w_1 N \bar{x}^2 + w_1 \sum_{i=1}^{N} x_i^2$$

$$\sum_{i=1}^{N} x_i y_i - \bar{Y} N \bar{x} = w_1 \left(\sum_{i=1}^{N} x_i^2 - N \bar{X}^2 \right)$$

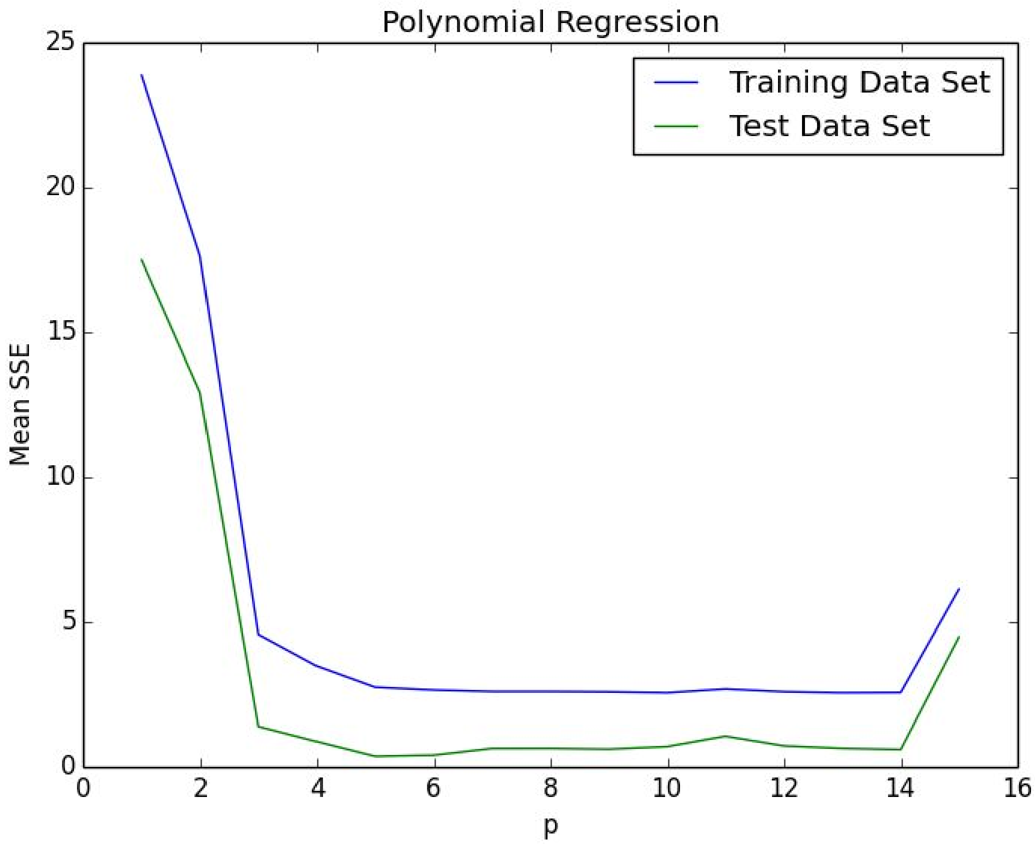
$$w_1 = \frac{\sum_{i=1}^{N} x_i y_i - \bar{y} N \bar{x}}{\sum_{i=1}^{N} x_1^2 - N \bar{x}^2}$$

$$\boxed{w_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_1 - \bar{x})^2}}$$

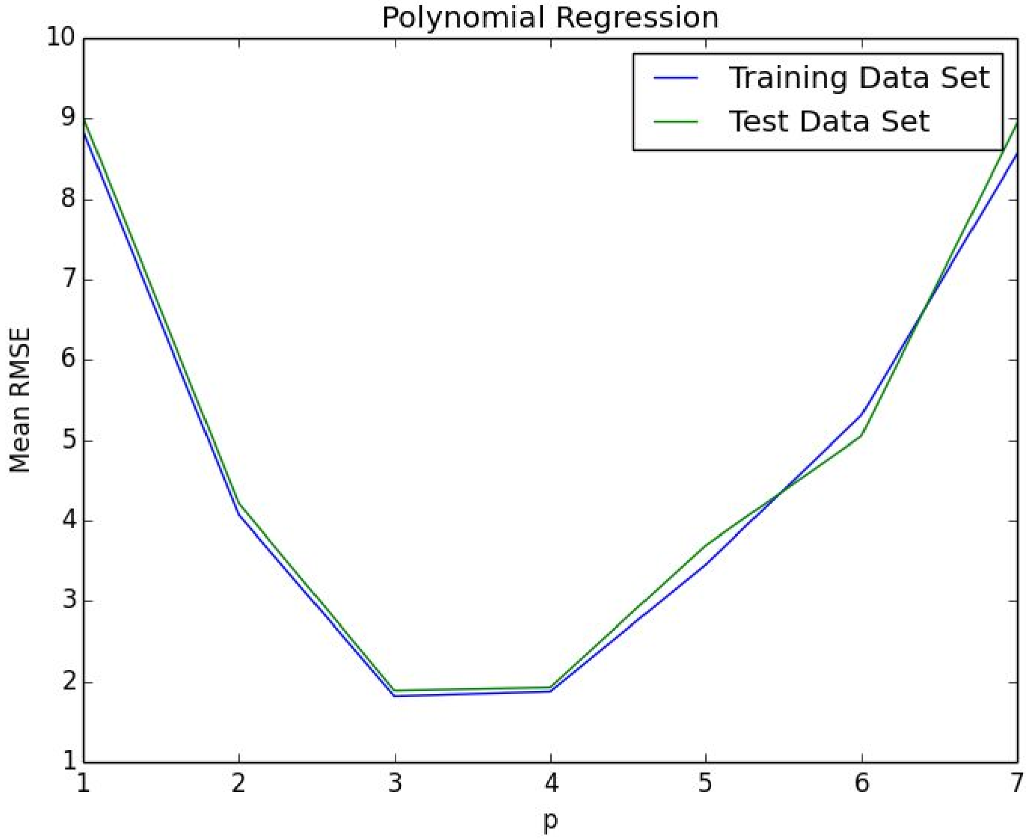
5. Polynomial Regression

5.1 Polynomial Regression using Normal Equations

Sinusoid Dataset - Polynomial Regression



Yacht Datset - Polynomial Regression



5.2 Interpreting the results

1. Yes the addition of new features to a certain polynomial level reduces the RMSE for the datasets.

2. We could add new features for the higher order polynomials and then try to remove the least significant ones from the list of features and have only the ones that contribute more towards reducing the RMSE. This might work but is not very efficient way considering the possibility of cross-terms.

6. The Hat Matrix

Hat Matix $H = X(X^T X)^{-1} X^T$

1. Symmetric $H^T = H$

$$H^T = \left(X (X^T X)^{-1} X^T\right)^T$$

$$(ABC)^T = C^T B^T A^T$$

$$\Rightarrow H^T = X\left((X^T X)^{-1}\right)^T X^T$$

$$= X\left((X^T X)^T\right)^{-1} X^T$$

$$= X (X^T X)^{-1} X^T$$

$$= H$$

2. Idempotent $H^2 = H$

$$H^2 = \left(X (X^T X)^{-1} X^T\right)\left(X (X^T X)^{-1} X^T\right)$$

$$H^2 = X (X^T X)^{-1} (X^T X) (X^T X)^{-1} X^T$$

$$A A^{-1} = I$$

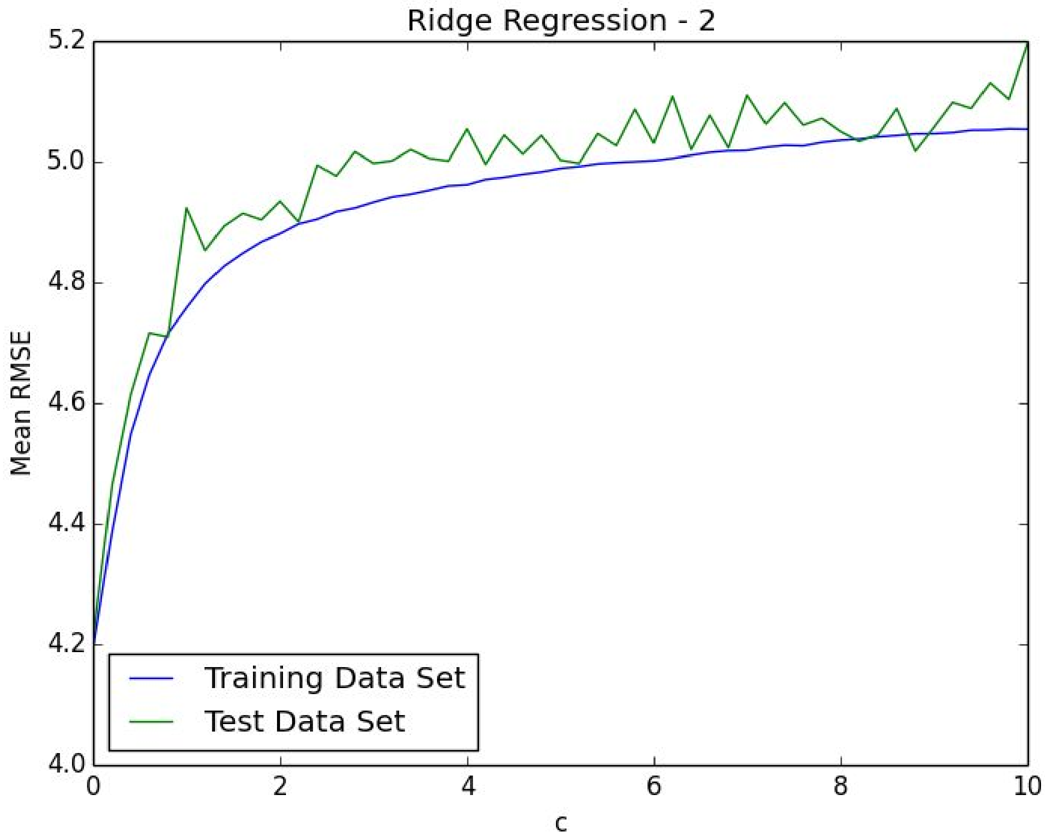
$$\Rightarrow H^2 = X (X^T X)^{-1} X^T$$

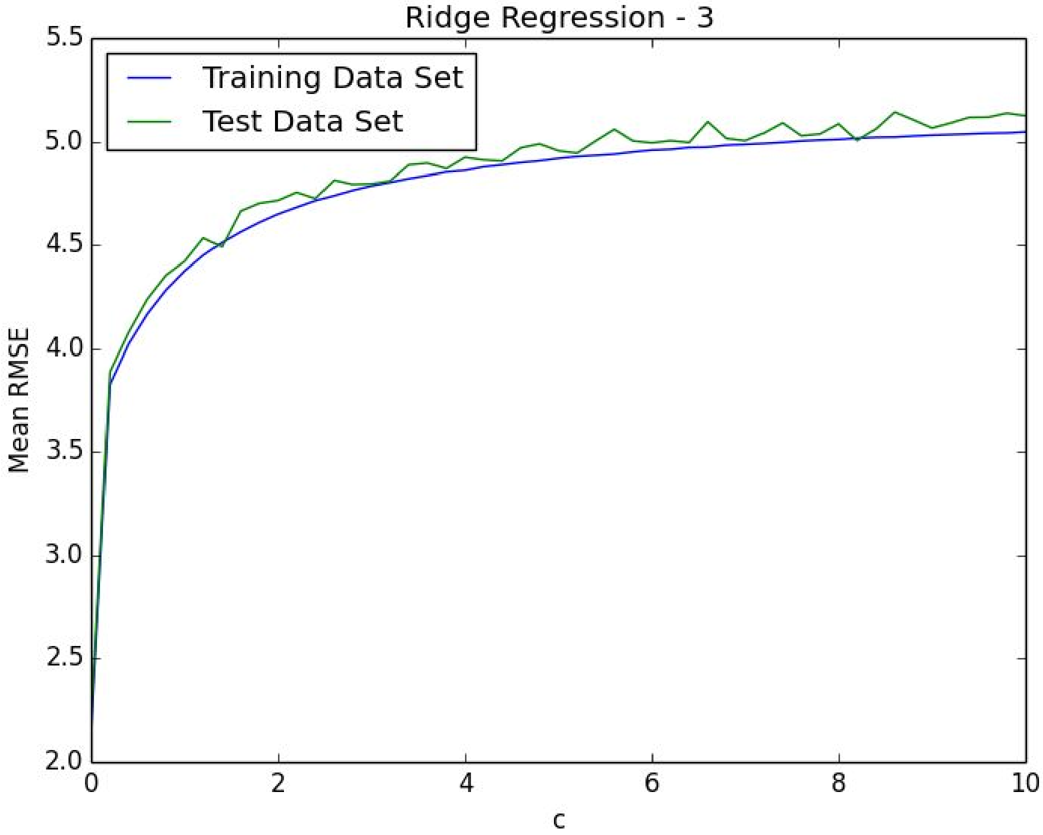
$$H^2 = H$$

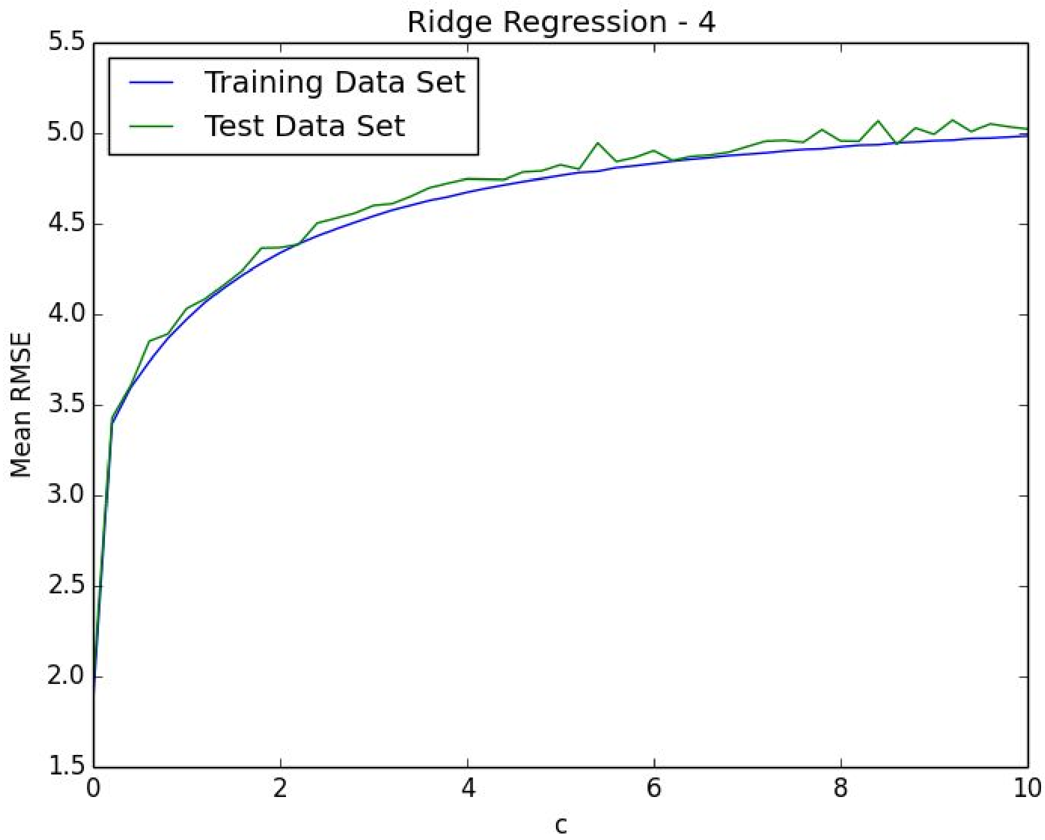
7. Programming Ridge Regression

1.


Ridge Regression - 1
Training Data Set
Test Data Set
Mean RMSE
c
5.15
5.10
5.05
5.00
4.95
4.90
4.85
0
2
4
6
8
10

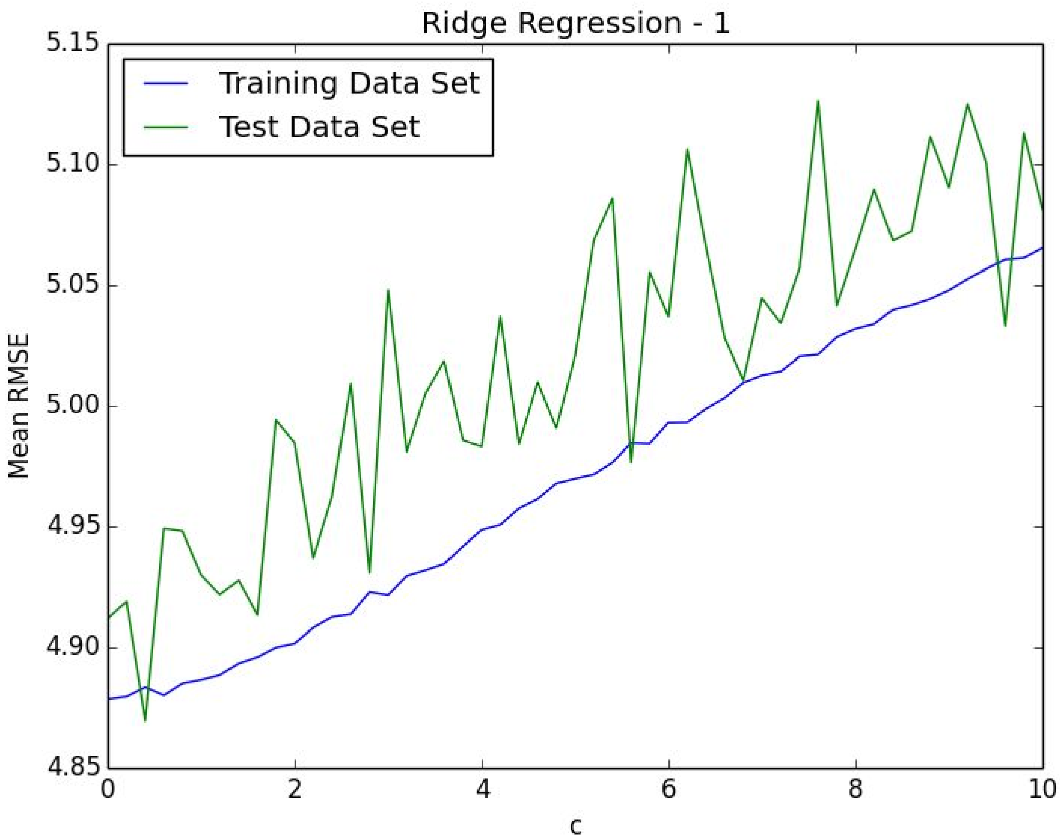

Ridge Regression - 2
Training Data Set
Test Data Set
Mean RMSE
c
5.2
5.0
4.8
4.6
4.4
4.2
4.0
0
2
4
6
8
10


Ridge Regression - 3
Training Data Set
Test Data Set
Mean RMSE
c
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
0
2
4
6
8
10


Ridge Regression - 4
Training Data Set
Test Data Set
Mean RMSE
c
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
2
4
6
8
10

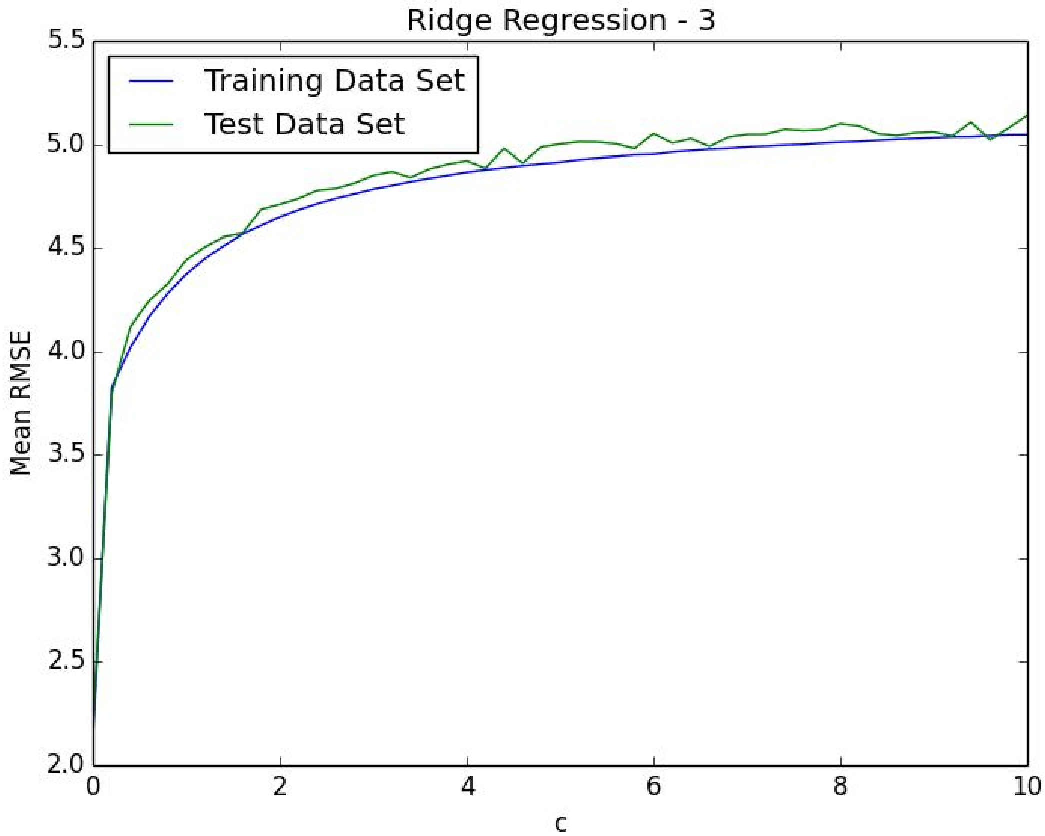

Ridge Regression - 5
Training Data Set
Test Data Set
Mean RMSE
c
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
2
4
6
8
10

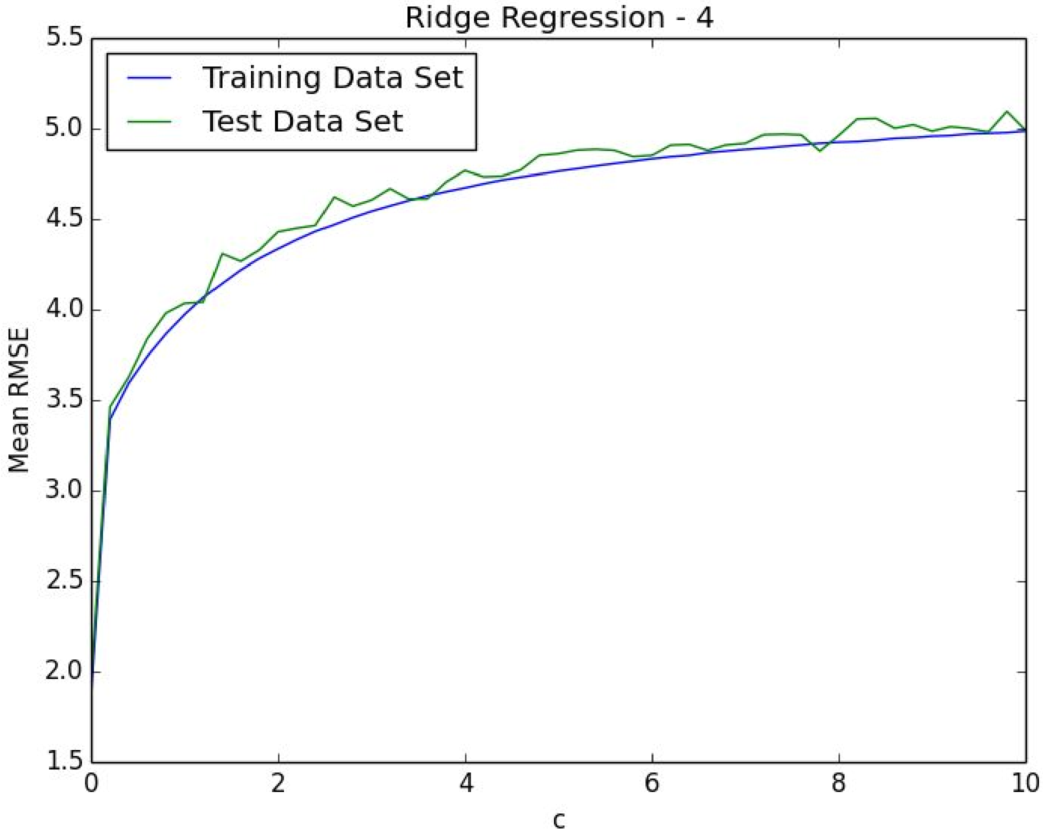
2.

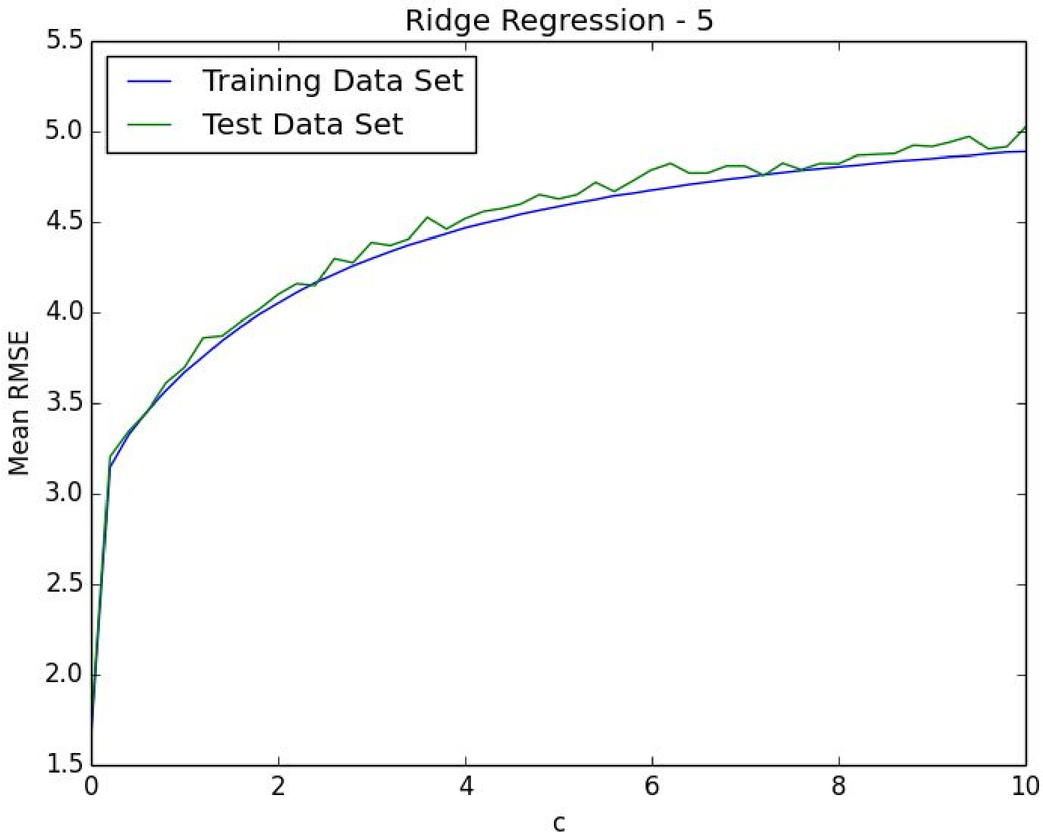


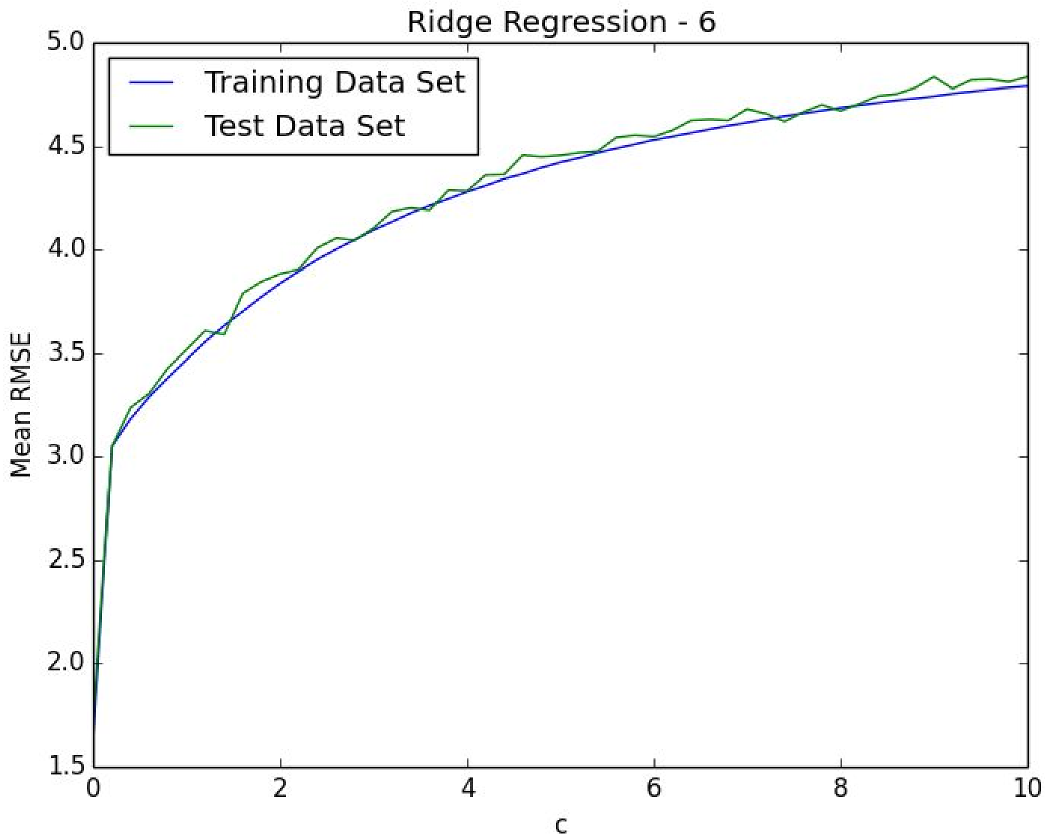
Ridge Regression - 1
Training Data Set
Test Data Set
Mean RMSE
c
5.15
5.10
5.05
5.00
4.95
4.90
4.85
0
2
4
6
8
10

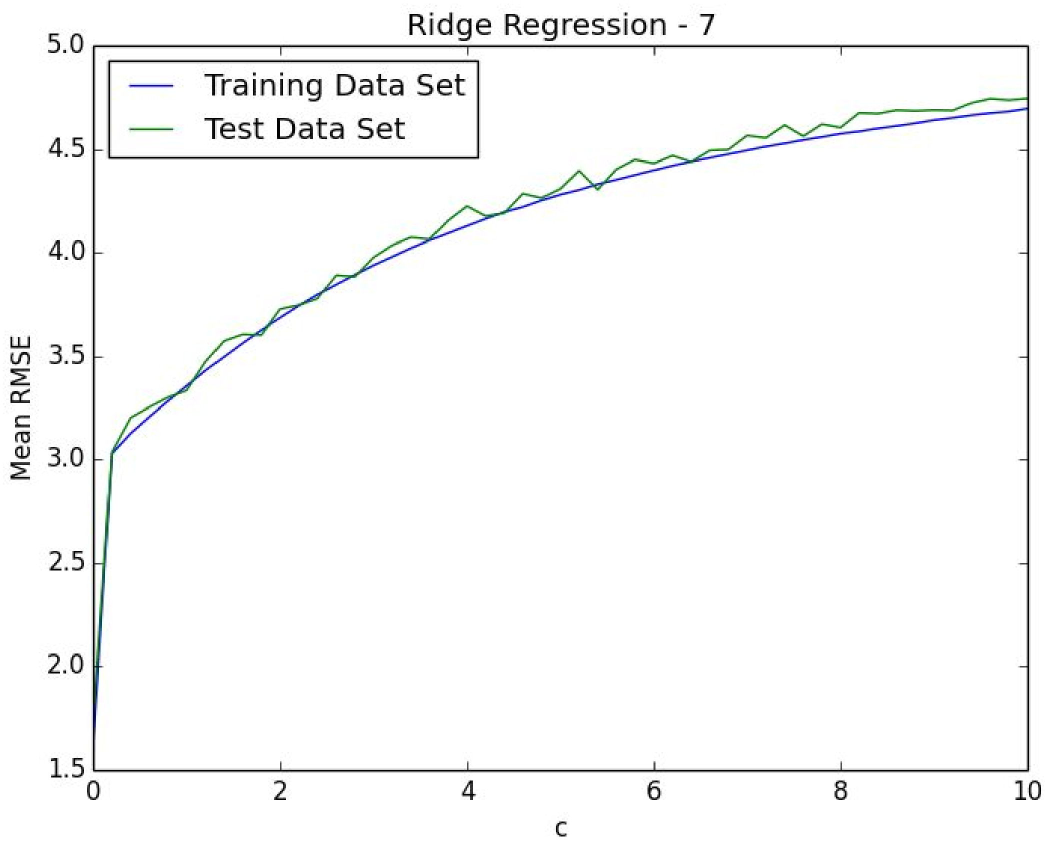

Ridge Regression - 2
Mean RMSE
c
Training Data Set
Test Data Set
5.2
5.0
4.8
4.6
4.4
4.2
4.0
0
2
4
6
8
10

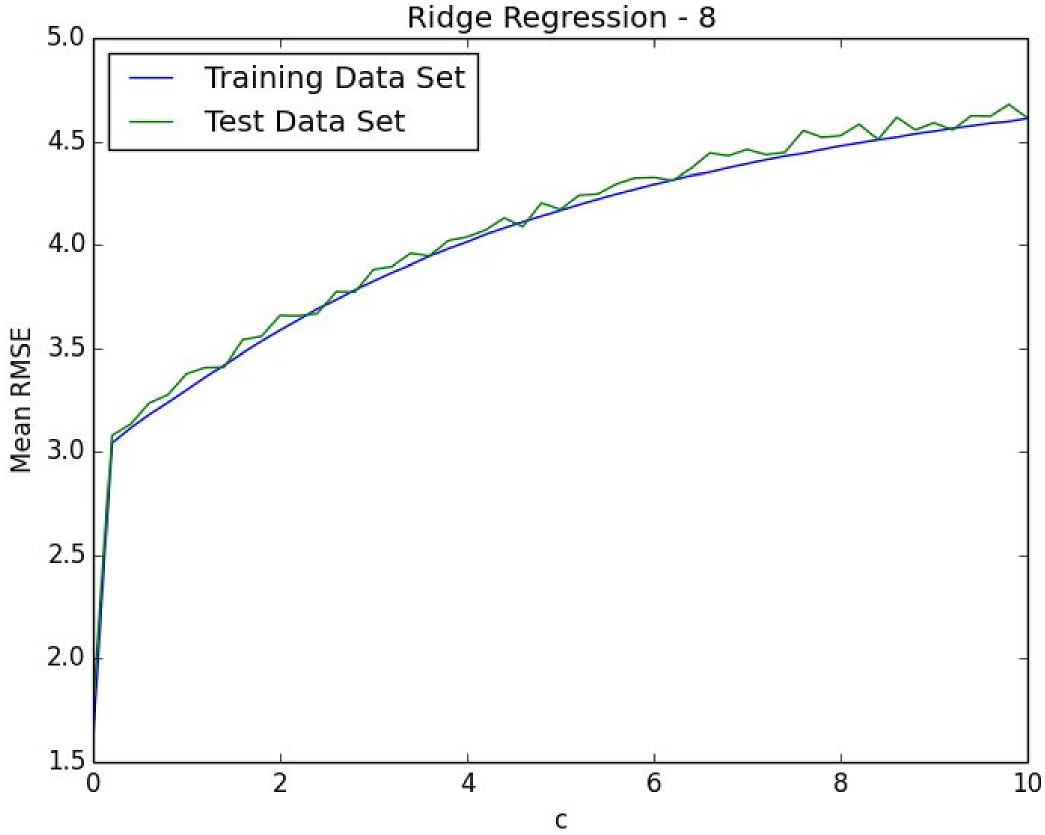

Ridge Regression - 3
Mean RMSE
c
Training Data Set
Test Data Set
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
0
2
4
6
8
10

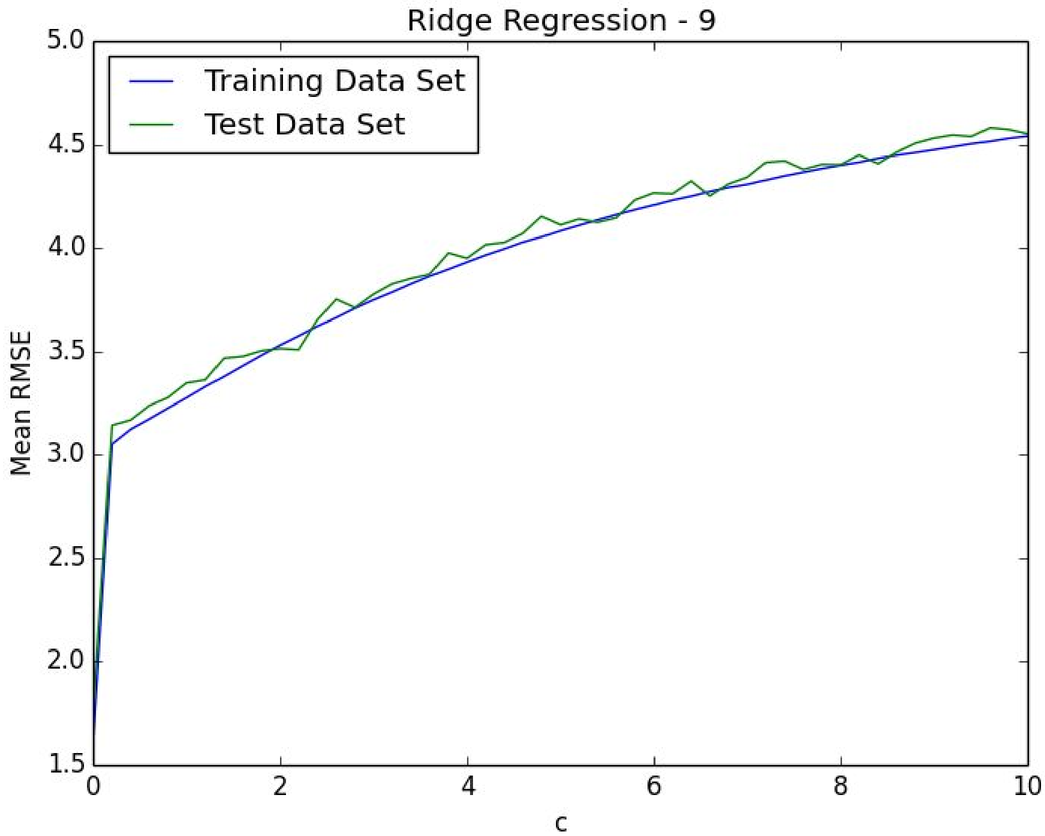

Ridge Regression - 4
Training Data Set
Test Data Set
Mean RMSE
c
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
2
4
6
8
10


Ridge Regression - 5
Training Data Set
Test Data Set
Mean RMSE
c
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
2
4
6
8
10


Ridge Regression - 6
Training Data Set
Test Data Set
Mean RMSE
c
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
2
4
6
8
10


Ridge Regression - 7
Training Data Set
Test Data Set
Mean RMSE
c
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
2
4
6
8
10


Ridge Regression - 8
Training Data Set
Test Data Set
Mean RMSE
c
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
2
4
6
8
10


Ridge Regression - 9
Training Data Set
Test Data Set
Mean RMSE
c
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
2
4
6
8
10

7.1 Interpretation

Observing the different polynomial regressions for multiple lambda values the RMSE increases when lambda values are increased. For the given dataset the linear regression using normal equation seems more efficient than polynomial ridge regression.

8. Maximum Likelihood For Univariate Normal

$$P(x/\mu) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\mu_{ML} \log \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2}$$

$$\frac{\partial}{\partial U} \ln P(x/\mu) = 0 \Rightarrow \frac{\partial}{\partial \mu} \log \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2} = 0$$

$$\frac{\partial}{\partial \mu}\left[\sum_{n=1}^{N} \log(2\pi\sigma^{-2})^{1/2} - \sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2}\right] = 0$$

$$\frac{1}{2\sigma^2} \sum_{n=1}^{N} \frac{\partial}{\partial U}(x_n-\mu)^2 = 0 \Rightarrow \sum_{n=1}^{N} 2(x_n-\mu)(-1) = 0$$

$$N\mu = \sum_{n=1}^{N} x_n$$

$$\mu_1 = \frac{1}{N}\sum_{n=1}^{N} x_n$$

8.1 Extra-credit

$$\arg\max_{\sigma^2} = \log P(x/\sigma^2)$$

$$\frac{\partial}{\partial \sigma^2} \log P(x/\sigma^2) = 0$$

$$\frac{\partial}{\partial \sigma^2} \log \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2} = 0$$

$$\frac{\partial}{\partial \sigma^2} \sum_{n=1}^{N} \log (2\pi\sigma^2)^{1/2} - \frac{\partial}{\partial \sigma^2} \sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2} = 0$$

$$N \frac{\partial}{\partial \sigma^2} \left[\log (2\pi\sigma^2)\right] \frac{1}{2} - \sum_{n=1}^{N} (x_n-\mu)^2 \frac{1}{2} \left(\frac{-1}{\sigma^2}\right)^2 = 0$$

$$-\frac{N}{2} \frac{1}{2\pi\sigma^2} 2\pi + \frac{1}{2} \frac{1}{(\sigma^2)^2} \sum_{n=1}^{N} (x_n-\mu)^2 = 0$$

$$\frac{N}{\sigma^2} = \sum_{n=1}^{N} (x_n-\mu)^2$$

$$\sigma^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n-\mu)^2$$