CS F320

FODS: ASSIGNMENT-II

NAME	ID
T.Praneeth Bhargav	2020A7PS1299H
S.V.S.Rahul	2020A7PS0204H

1. Description of Model

Linear Regression of degree 1 is the model. Then we split the data into training and testing sets. Both datasets are normalized for faster learning in the case of training sets. Gradient descent is the algorithm used to find the optimal weights of the chosen regression model. The linear regression model in Machine Learning solves the problem in which the target variable is continuous. The learning rate is 0.01, and 1000 iterations have been chosen for the Gradient descent algorithm. Various feature selection techniques have been applied to the model and compared based on the results they produce.

Implementation

1. Correlation coefficients(Pearson)

1. We have implemented the Pearson Correlation Coefficient using the NumPy library of python
2. We have calculated the correlation between each feature and the target variable and stored it in a list
3. We then loop to select the features from 1 to 26; when we need to take 5 features to train the model, we choose the ones with the highest correlation.
4. We select the top k features with the highest correlation to the target attribute
5. We then train the regression model with these features and then calculate the training error and testing error
6. The One with the least testing error is considered the best model

2. Principal Component Analysis

1. We have implemented the PCA using the inbuilt libraries of python
2. We have used the sklearn library of python to perform PCA
3. PCA is a dimensionality reduction technique that uses Eigen values and Eigen Vectors of the Covariance Matrix of the Data
4. We have decided the no of features to give to the regression model at each iteration, and we have incremented the loop from 1 to 26
5. The PCA model inherently selects the top k features for the kth iteration where we want to use k features for training
6. We have calculated the Eigen Values and the Eigen Vectors
7. We have also calculated the training and the testing errors for the top k features we have selected using PCA.

3. Greedy Forward

1. We have implemented the Forward Greedy Method by finding the best feature at every given stage of program
2. We have initially taken a single feature that gave the least testing error and then have taken all the combinations of the first best feature with all the remaining features
3. To achieve this, we have maintained a global list such that at the end of each iteration, we add the best feature we found in iteration
4. Let's suppose we initially have an empty list, and we find that feature 1 has the least testing error when trained, then we add 1 to list
5. Next, we take all the possible combinations of the features with 1 like (12), (13)… and again find the combination with the least testing error
6. Then we add that feature to the list, and we continue this until we get the 26 features, and we select the feature combination which has the least testing error as the best error
7. It is not always guaranteed that the model proposed by the Greedy Forward Method provides the most optimal model

4. Greedy Backward

1. We have implemented the Backward Greedy Method by finding the feature we need to remove to decrease the testing error.
2. We have initially taken all the features then by removing each feature in an iteration and then calculating the testing errors at each iteration, and the feature whose removal gives the least testing error is removed
3. To achieve this, we have maintained a global list initially with all the features such that at the end of each iteration, we remove the feature whose removal decreases the testing error
4. Let's suppose we initially have the whole list, and we find that feature 1, whose removal gives the least testing error, then we remove it from list
5. Next, we repeat the above step and remove features at the end of iterations until we have a single feature
6. We conclude that the best model is the one whose combination has the least testing error
7. It is not always guaranteed that the model proposed by the Greedy Backward Method provides the most optimal model

2. Tabulation of training and testing errors

a. correlation coefficients

Out[26]:

	Features	Training Error	Testing Error
0	1	5392.077838	4967.930218
1	2	5388.817170	4955.096793
2	3	5386.477544	4955.532293
3	4	5374.603599	4945.543095
4	5	5372.142246	4937.695277
5	6	5341.200797	4873.496885
6	7	5290.486931	4882.054223
7	8	5250.898773	4841.766744
8	9	5087.954803	4711.029068
9	10	5005.559605	4633.737856
10	11	4985.806139	4626.265794
11	12	4979.278451	4634.360710
12	13	4966.753785	4620.246786
13	14	4953.167378	4612.116252
14	15	4954.133253	4616.016509
15	16	4954.327582	4617.048301
16	17	4918.645908	4649.192995
17	18	4897.575689	4641.765784
18	19	4897.269555	4637.304700
19	20	4881.692440	4633.158679
20	21	4877.540613	4618.963567
21	22	4877.076297	4615.497933
22	23	4875.804754	4613.530932
23	24	4849.888983	4598.229723
24	25	4849.704087	4597.406853
25	26	4849.703079	4597.400973

26 is the optimal number of features

b. PCA

Out[64]:

	Features	Training Error	Testing Error
0	1	5475.654371	5052.185416
1	2	5475.654088	5052.226084
2	3	5475.343905	5054.847665
3	4	5475.321471	5055.106285
4	5	5361.063818	4951.623241
5	6	5359.051785	4960.170946
6	7	5353.334265	4950.168118
7	8	5353.168384	4950.721449
8	9	5305.646486	4947.182698
9	10	5230.810173	4865.377282
10	11	5221.317611	4888.913963
11	12	5057.877857	4741.329357
12	13	5057.724318	4743.807773
13	14	5023.144201	4728.293035
14	15	5008.116754	4717.158540
15	16	5006.103374	4735.454617
16	17	5005.847330	4737.714898
17	18	5005.381150	4737.970348
18	19	4897.447238	4661.250749
19	20	4896.776604	4660.895721
20	21	4869.733779	4629.228632
21	22	4769.811015	4530.983447
22	23	4739.814340	4494.301650
23	24	4724.684332	4488.707547
24	25	4722.525682	4488.507057
25	26	4722.663446	4488.305689

26 is the optimal number of features

c. Greedy Forward

Out[52]:

	Features	Added Feature	Training Error	Testing Error
0	1	20	5482.053103	5020.585228
1	2	1	5403.267402	4954.236047
2	3	15	5330.298066	4882.585512
3	4	21	5317.814918	4845.215091
4	5	13	5296.830274	4829.315934
5	6	4	5297.296178	4814.295236
6	7	16	5235.865175	4791.445786
7	8	9	5233.677447	4780.788891
8	9	3	5223.787689	4770.120917
9	10	7	5216.830253	4763.884513
10	11	10	5217.329298	4760.967309
11	12	8	5205.541069	4757.245830
12	13	19	5203.059050	4754.763763
13	14	23	5199.747071	4752.395350
14	15	17	5190.837055	4750.447642
15	16	6	5190.508803	4749.739174
16	17	14	5190.331571	4748.934027
17	18	24	5190.259507	4748.627572
18	19	12	5183.422791	4748.444945
19	20	25	5183.411924	4748.353295
20	21	22	5182.456472	4748.741377
21	22	11	5182.391433	4750.284562
22	23	0	5181.872131	4751.965734
23	24	18	5183.782803	4755.524179
24	25	2	5178.633281	4757.968493
25	26	5	5155.639681	4770.305081

Optimal features = 20

d. Greedy Backward

Out[41]:

	Features	Removed Feature	Training Error	Testing Error
0	26	None	5155.639681	4770.305081
1	25	6	5178.633281	4757.968493
2	24	3	5183.782803	4755.524179
3	23	19	5181.872131	4751.965734
4	22	1	5182.391433	4750.284562
5	21	12	5182.456472	4748.741377
6	20	23	5183.411924	4748.353295
7	19	25	5183.422791	4748.444945
8	18	13	5190.259507	4748.627572
9	17	26	5190.331571	4748.934027
10	16	15	5190.508803	4749.739174
11	15	7	5190.837055	4750.447642
12	14	20	5192.484023	4752.350698
13	13	24	5195.408186	4754.792493
14	12	18	5205.541069	4757.245830
15	11	9	5217.329298	4760.967309
16	10	11	5216.830253	4763.884513
17	9	8	5223.787689	4770.120917
18	8	4	5233.677447	4780.788891
19	7	10	5235.865175	4791.445786
20	6	14	5257.917574	4810.920546
21	5	17	5320.501148	4834.097242
22	4	5	5317.814918	4845.215091
23	3	22	5330.298066	4882.585512
24	2	16	5403.267402	4954.236047
25	1	2	5482.053103	5020.585228

Optimal features = 20

3. Best Model

a. Correlation coefficients

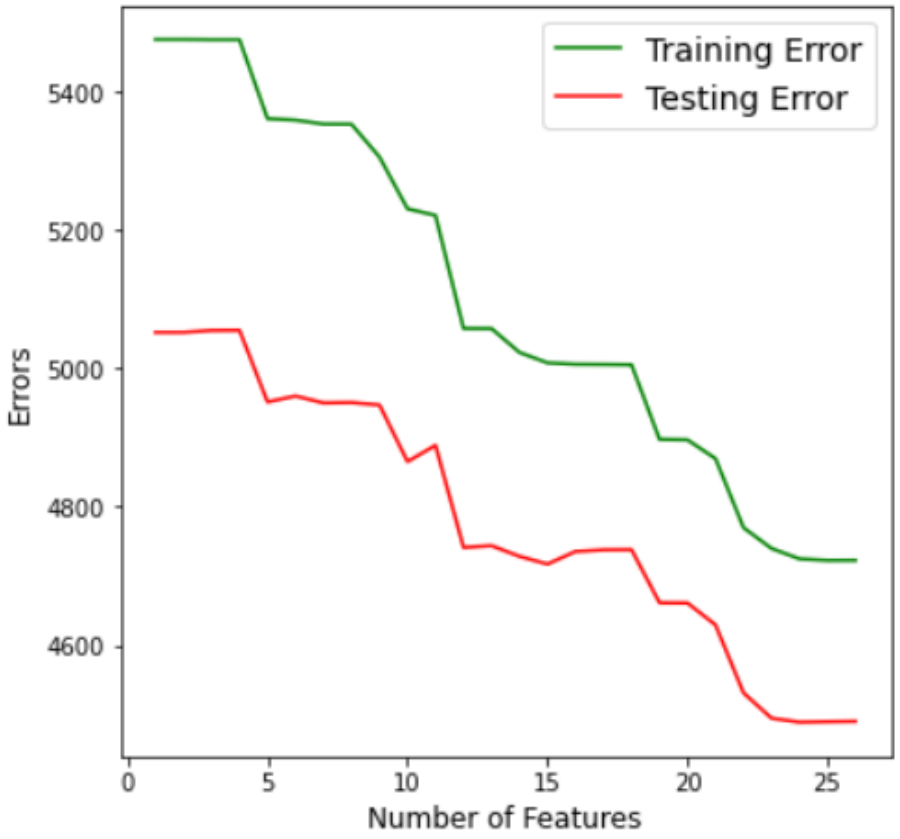
The Best Model obtained using the Pearson correlation coefficient method is of 26 dimensions

20	21	4877.540613	4618.963567
21	22	4877.076297	4615.497933
22	23	4875.804754	4613.530932
23	24	4849.888983	4598.229723
24	25	4849.704087	4597.406853
25	26	4849.703079	4597.400973

b. PCA

The Best Model obtained using the PCA method is of 26 dimensions.

18	19	4897.447238	4661.250749
19	20	4896.776604	4660.895721
20	21	4869.733779	4629.228632
21	22	4769.811015	4530.983447
22	23	4739.814340	4494.301650
23	24	4724.684332	4488.707547
24	25	4722.525682	4488.507057
25	26	4722.663446	4488.305689



c. Greedy Forward

The Best Model obtained using the Greedy Forward method is 20 features.

16	17	14	5190.331571	4748.934027
17	18	24	5190.259507	4748.627572
18	19	12	5183.422791	4748.444945
19	20	25	5183.411924	4748.353295
20	21	22	5182.456472	4748.741377
21	22	11	5182.391433	4750.284562
22	23	0	5181.872131	4751.965734
23	24	18	5183.782803	4755.524179

d. Greedy Backward

The Best Model obtained using the Greedy Backward method is 20 features.

2	24	3	5183.782803	4755.524179
3	23	19	5181.872131	4751.965734
4	22	1	5182.391433	4750.284562
5	21	12	5182.456472	4748.741377
6	20	23	5183.411924	4748.353295
7	19	25	5183.422791	4748.444945
8	18	13	5190.259507	4748.627572
9	17	26	5190.331571	4748.934027
10	16	15	5190.508803	4749.739174

4. Results of Best Models

4.1 PCA

Here, we got the optimal model using 26 features in linear regression. The error for this optimal model is 4488.30.

Below are the eigen values for the optimal model with 26 dimension.
[1172.1303587214952, 411.8630332563255, 157.06750057816086, 134.9085424906583, 100.74932487064656, 67.07988186414605, 47.70234589529474, 11.442906900532654, 7.7512578092019035, 7.4590453700775745, 4.004486113095089, 3.404491795456807, 2.6816306764317703, 2.5598612292849197, 1.3054523196434697, 0.9156756105354825, 0.7149523269979322, 0.6277132659155278, 0.485737425823727, 0.40559705592752804, 0.2952284740061571, 0.2145916056343997, 0.13468612846638411, 0.09807723904497495, 0.07167246041201834, 6.330836179564316e-30]

4.2 Correlational Coefficient

Here, we got the optimal model using 20 features in linear regression. The error for this optimal model is 4597.35

4.3 Greedy Forward:

Here, we got the optimal model using 20 features in linear regression. The features are 20,1,15,21,13,4,16,9,3,7,10,8,19,23,7,6,14,24,12,25,22. The error for this optimal model is 4748.35

4.4 Greedy Backward:

Here, we got the optimal model, again using 20 features in linear regression. The features are 6,3,19,1,12,23. The error for this optimal model is 4748.35

Overall best model is obtained using PCA with the testing errors as 4480.30

Note: Here our testing error values are too high because the metric we have used to determine the testing error is directly Cost function.

Comparative Analysis

Method	No of Features	Train Error	Test Error
PCA	26	4722.66	4488.30
Pearson correlation	26	4849.70	4597.40
Greedy Forward	20	5183.411	4748.35
Greedy Backward	20	5183.411	4748.35

Conclusions

- We can clearly see that PCA with 26 features has the least training error and testing error compared to all other models.
- From the Pearson correlation technique, among all 26 feature sets, the feature set with 26 features has the least training error and testing errors. This is the second-best model.

- Using the Greedy forward technique, we can conclude that the feature set, which has 20 features in it gave the least training and testing errors.
- Using the Greedy backward technique, we can conclude that the feature set which has 20 features in it gave the least training and testing errors.