CS F320 - FOUNDATIONS OF DATA SCIENCE

Semester I, 2021-2022

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI

HYDERABAD CAMPUS

ASSIGNMENT-2

Vineet Venkatesh - 2019A7PS0043H

Pavan Kumar Reddy Yannam - 2019A7PS0038H

Jagath Kaparthi - 2019A4PS0547H

Table of Contents

1.Introduction

In this assignment we performed linear regression in 3 different ways to predict the price of a house using 13 features.

1. Greedy forward feature selection along with preprocessing of the data
2. Greedy backward feature selection along with preprocessing of the data
3. Linear Regression Without any preprocessing and feature selection

A greedy feature selection algorithm is used to either select the best features one by one (forward selection) or remove the worst feature one by one (backward selection).

2. Dataset

The data set consists of 13 feature attributes, namely

1. ‘bedrooms’,
2. ’bathrooms’,
3. ’sqft_living’,
4. ‘sqft_lot’,
5. ‘floors’,
6. ‘waterfront’,
7. ‘view’,
8. ‘condition’,
9. ‘grade’,
10. ‘sqft_above’,
11. ‘sqft_basement’,
12. ‘sqft_living15’,
13. ‘sqft_lot15’,

which are used to predict the target attribute ‘price’. The data set consists of 1188 data points.

	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	sqft_living15	sqft_lot15	price
0	4	1.75	2120.0	7420	1.0	0	0	4	7	1060.0	1060	1540	7420	453000.0
1	4	2.50	2180.0	9861	2.0	0	2	3	8	2180.0	0	2390	9761	480000.0
2	3	1.50	1540.0	9800	1.0	0	0	3	7	1010.0	530	1600	8250	180500.0
3	3	3.50	2380.0	6250	2.0	0	3	3	8	1670.0	710	2540	4010	495000.0
4	4	2.50	2230.0	8500	2.0	0	0	3	8	2230.0	0	2270	8770	325000.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1183	4	2.50	2623.0	7184	2.0	0	0	3	8	2623.0	0	2010	4939	333000.0
1184	2	2.00	1730.0	4248	2.0	0	0	3	7	1730.0	0	1380	4000	450000.0
1185	4	2.50	3040.0	6425	2.0	0	0	3	8	3040.0	0	3040	7800	776000.0
1186	4	1.50	2150.0	11026	1.0	0	0	4	7	2150.0	0	1760	10283	400000.0
1187	3	1.00	1200.0	9194	1.0	0	0	4	7	1200.0	0	1330	8650	369500.0

1188 rows × 14 columns

3. Data Preprocessing

Before implementing greedy feature selection, we performed data preprocessing which included,

- Shuffling the data
- Handling missing values- we eliminated the data points in which at least one of the feature value was missing
- Standardizing the data
- Detecting outliers- we eliminated outliers which were detected based on z value (standardized value). If z value was greater than 3 or less than -3 they were detected as outliers.
- Creating a random 70-30 split of the data to aid in training and testing respectively.

4. Model

A Linear Regression is performed to predict the 'Price' value from the input features. Gradient Descent is used to optimize the model to best fit the data given and learn the relationship between features

4.1 Implementation

All computations of the Regression algorithm are done in a class named LinearRegression.

1. First, the fit method of the class, which trains the model on the training, is called with parameters - the learning rate and the number of epochs the gradient descent should run for. A variable w, representing the weights, is initialized randomly.
2. The model is then trained, and the weights updated using the Mean Square Error between the predicted and true output value. The weight update is done using Gradient Descent, where, for each datapoint in the training set, the error is calculated, multiplied by the learning rate and subtracted from the weights.
3. Once the dataset is trained and the error saturates, the evaluate method is called with the testing data as the parameters. The testing Loss/Error is calculated and reported.

5. Greedy Forward Feature Selection

A function is created for the task of Greedy Forward Feature Selection, in which all the computations are done.

5.1 Algorithm

1. The function iterates through the original feature set, and selects each feature for the regression task, one after the other.
2. The feature which gives the best RMSE error for the regression task, is stored in another feature set, which was initially empty.
3. The function is called again recursively, passing the new feature set each time.
4. Each time the feature set is updated with the feature that gives the least RMSE for the regression task.
5. Finally among all these feature sets of size 1,2,3…13, the feature set which gives the lowest RMSE overall, is chosen as the best feature set.

5.2 Results

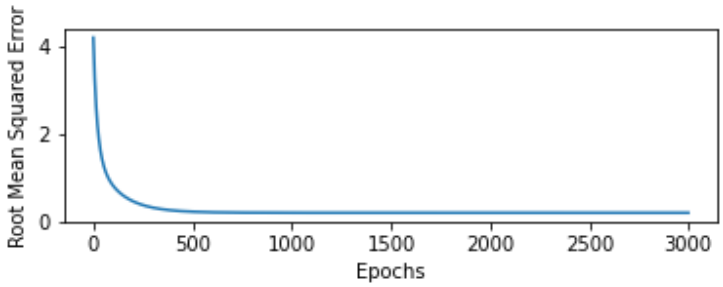
Each model is trained for 3000 epochs with a learning rate of 0.03.
Best subset of features that provided optimal model was {2,3,7,8,9,13}, which are as follows :

- ’bathrooms’
- ’sqft_living’
- ‘view’
- ‘condition’
- ‘grade’
- ‘sqft_lot15’

This combination of features gave

1. Minimum Training RMSE = 0.44347
2. Minimum Testing RMSE = 0.401522

The RMSE vs Epochs graph is shown below.



6. Greedy Backward Feature Selection

A function is created for the task of Greedy Backward Feature Selection, in which all the computations are done.

6.1 Algorithm

1. The original feature set is copied to another set, from which features are removed in each iteration of the function.
2. The function iterates through the original feature set, and selects each feature to be removed from the copied feature set, which is then used to perform the regression task.
3. The feature which, when removed, gives the minimum RMSE error for the regression task, is removed from the copied feature set.
4. The function is called again recursively, passing the updated feature set each time.
5. Each time the feature set is updated by removing the feature that gives the least RMSE, upon removal, for the regression task.
6. Finally among all these feature sets of size 13,12,11…1, the feature set which gives the lowest RMSE overall, is chosen as the best feature set.

6.2 Results

Each model is trained for 3000 epochs with a learning rate of 0.03.
Best subset of features that provided optimal model was {2, 3, 7, 8, 9, 12, 13}, which are as follows :

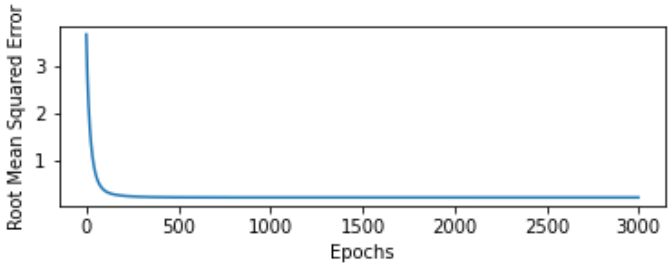
- ’bathrooms’
- ’sqft_living’
- ‘view’
- ‘condition’
- ‘grade’
- ‘sqft_living15’

- ‘sqft_lot15’

This combination of features gave

1. Minimum Training RMSE = 0.438582
2. Minimum Testing RMSE = 0.417648

The RMSE vs Epochs graph is shown below.



7. Linear Regression without Preprocessing and Feature Selection

The dataset was also used to perform Linear Regression without Preprocessing and Feature Selection. As expected, this raw data performed very poorly and gave the below results,

1. Minimum Training RMSE: 216772.2279
2. Minimum Testing RMSE: 363678.1257

Figure - GD

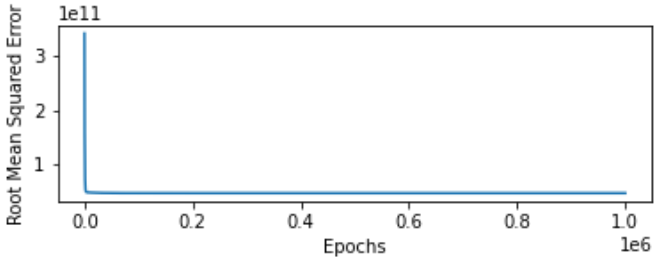
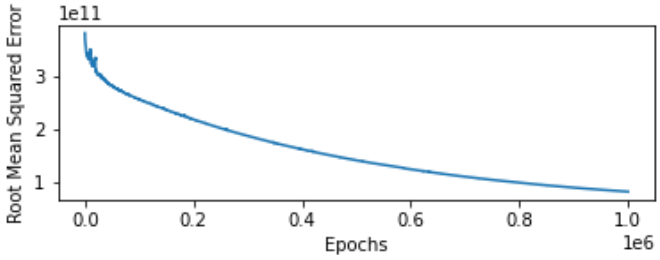


Figure - SGD



8. Results

Minimum training and testing error obtained from each of the method

Method	Number of features selected	Subset of Features	Training Error (RMSE)	Testing Error (RMSE)
Greedy Forward Feature Selection	6	● ’bathrooms’ ● ’sqft_living’ ● ‘View’ ● ‘condition’ ● ‘grade’ ● ‘sqft_lot15’	0.443477	0.401522
Greedy Backward Feature Selection	7	● ’bathrooms’ ● ’sqft_living’ ● ‘view’ ● ‘condition’ ● ‘grade’ ● ‘sqft_living15’ ● ‘sqft_lot15’	0.438582	0.417648
Without preprocessing and feature selection	13	● ‘bedrooms’ ● ’bathrooms’ ● ’sqft_living’ ● ‘sqft_lot’ ● ‘floors’ ● ‘Waterfront’ ● ‘view’ ● ‘condition’ ● ‘grade’ ● ‘sqft_above’ ● ‘sqft_basement’ ● ‘sqft_living15’ ● ‘sqft_lot15’	216772.2279	363678.1257