Jupyter Notebook: <https://github.com/ben12385/DS4400-HW1>

Readme: In the Github

**Problem 1 [Average, variance, and correlation]**

(a) For each feature, write code to compute the average value, the min and max values, as well as its variance.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature | Average | Max | Min | Variance | Correlation Coefficient |
| Bedrooms | 3.37 | 33 | 0 | 0.86 | 0.3083 |
| Bathrooms | 2.11 | 8 | 0 | 0.59 | 0.5251 |
| Sqft\_living | 2079.90 | 13540 | 290 | 843494.65 | 0.7020 |
| Sqft\_lot | 15106.97 | 1651359 | 520 | 1715579393.30 | 0.0897 |
| Floors | 1.49 | 3.5 | 1 | 0.29 | 0.2568 |
| Waterfront | 0.01 | 1 | 0 | 0.01 | 0.2664 |
| View | 0.23 | 4 | 0 | 0.59 | 0.3973 |
| Condition | 3.41 | 5 | 1 | 0.42 | 0.0364 |
| Grade | 7.66 | 13 | 1 | 1.38 | 0.6674 |
| Sqft\_above | 1788.39 | 9410 | 290 | 685702.94 | 0.6056 |
| Sqft\_basement | 291.51 | 4820 | 0 | 195863.61 | 0.3238 |
| Yr\_built | 1971.01 | 2015 | 1900 | 862.76 | 0.0540 |
| Yr\_renovated | 84.40 | 2015 | 0 | 161338.75 | 0.1264 |
| Zipcode | 98077.94 | 98199 | 98001 | 2862.66 | -0.0532 |
| Lat | 47.56 | 47.7776 | 47.1559 | 0.02 | 0.3055 |
| Long | -122.21 | -121.315 | -122.519 | 0.02 | 0.0216 |
| Sqft\_living15 | 1986.55 | 6210 | 399 | 469739.50 | 0.5854 |
| Sqft\_lot15 | 12768.46 | 871200 | 651 | 745483731.37 | 0.0824 |

(b) Compute the correlation coefficient of each feature with the response. Which feature are positively correlated (i.e., have positive correlation coefficient) and which ones are negatively correlated with the response? Which features have highest correlation with the response (both positive and negative)?

Majority of the features are positively correlated with only zipcode having a negative correlation coefficient. The features with the highest correlation would be Sqft\_living for the positive and since there is only 1 feature with a negative correlation, zipcode has the highest negative correlation among all the features. However there are a couple more features like Grade, Sqft\_above which are quite close to Sqft\_living in terms of correlation.

**Problem 2 [Linear regression]**

(a)Use an existing package to train a linear regression model on the training set. Report the coefficients of the linear regression models and the 3 metrics of interest: MSE, RSE, and R2 .

|  |  |
| --- | --- |
| MSE: | 31119892883.73 |
| RSS: | 31119892883733.62 |
| R2: | 0.73 |

(b) Perform feature standardization so that each feature has mean 0 and variance of 1. Train again a linear regression model on the training data.

|  |  |
| --- | --- |
| MSE: | 31119892883.73 |
| RSS: | 31119892883733.62 |
| R2: | 0.73 |

(c) Evaluate both models on the testing set. Report the same metrics (MSE, RSE, and R2 ) on the testing set

Unnormalized with Test Set

|  |  |
| --- | --- |
| MSE: | 57161532843.15 |
| RSS: | 57161532843152.25 |
| R2: | 0.66 |

Normalized with Test Set

|  |  |
| --- | --- |
| MSE: | 57161532843.16 |
| RSS: | 57161532843158.33 |
| R2: | 0.66 |

(d) Interpret the results in your own words. Which features contribute mostly to the linear regression model? Is the model fitting the data well? How large is the model error?

The features that contributes most to the linear regression model is grade, yr\_built and lat because they have the largest coefficient after normalization. The model does not fit the data well as seen from the R2 score which is only 0.66 which causes the model error to be large. The results show that normalization does not affect the output of the package as the output is very similar.

**Problem 3 [Closed-form solution for linear regression]**

(c) Compare the models given by your implementation with those trained in Problem 2 by the R or Python packages. Report the MSE, RSE, and R2 metrics for the models you implemented. Compare the coefficients output by your model with the ones computed by the package.

Simple Linear Regression

|  |  |
| --- | --- |
| MSE: | 57947526161.29 |
| RSS: | 57947526161288.38 |
| R2: | 0.50 |

Multiple Linear Regression

|  |  |
| --- | --- |
| MSE: | 31119892883.73 |
| RSS: | 31119892883733.63 |
| R2: | 0.73 |

From Package

|  |  |
| --- | --- |
| MSE: | 31119892883.73 |
| RSS: | 31119892883733.62 |
| R2: | 0.73 |

The metrics outputted from the closed form for multiple linear regression compared to the one from the package is very similar, it is almost identical. However, the metric from the simple linear regression is not close to the metric from the package. This is because it only uses 1 feature which is just a linear correlation between sqft\_living and price and since it is not perfectly correlated, the error would be larger than if we used multiple features.

**Problem 4 [Gradient descent]**

(b) Vary the value of the learning rate (5 different values) and the number of iterations (5 different values)) and report the value of θ for each of the 25 combinations, as well as the MSE metric on the training set. Report the MSE on the testing set.

**R2 Value**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Repetitions | 0.150 | 0.125 | 0.100 | 0.075 | 0.050 |
| 20 | 0.7292 | 0.7288 | 0.7279 | 0.7234 | 0.6884 |
| 30 | 0.7296 | 0.7294 | 0.7291 | 0.7284 | 0.7226 |
| 40 | 0.7297 | 0.7296 | 0.7295 | 0.7291 | 0.7276 |
| 50 | 0.7297 | 0.7297 | 0.7296 | 0.7294 | 0.7287 |
| 60 | 0.7297 | 0.7297 | 0.7297 | 0.7295 | 0.7291 |

**MSE and R2 Value with Test Set**

|  |  |  |
| --- | --- | --- |
| Repetitions | MSE | R2 |
| 0.15 | 57192370267.37 | 0.6570 |
| 0.125 | 57217071970 | 0.6568 |
| 0.1 | 57267160612 | 0.6565 |
| 0.075 | 57375197315 | 0.6559 |
| 0.05 | 57622261973 | 0.6544 |

**Learning Rate 0.150**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Repetition 20 | | | | | | |
| 519999.6 | | | -13225.8 | 16075.31 | 57202.34 | 5393.334 |
| 10451.09 | | | 62895.75 | 49477.3 | 13627 | 82537.42 |
| 49091.55 | | | 26626.53 | -64985.6 | 19434.37 | -21527.6 |
| 83899.51 | | | -12270.3 | 49875.57 | -6948.65 |  |
| MSE: | | 31183282289.22 | | | | |
| Repetition 30 | | | | | | |
| 520403.1 | | | -13923.2 | 17042.75 | 56931.47 | 7410.646 |
| 10794.33 | | | 63379.55 | 49233.84 | 12375.55 | 87336.65 |
| 48802.9 | | | 26598.9 | -69017.1 | 18129.93 | -23404.5 |
| 83437.26 | | | -11891.4 | 46364.28 | -8511.24 |  |
| MSE: | | 31137919123.15 | | | | |
| Repetition 40 | | | | | | |
| 520414.5 | | | -14105.5 | 17697 | 56712.55 | 8644.13 |
| 10923.24 | | | 63529.08 | 49327.12 | 11812.27 | 89880.88 |
| 48561.28 | | | 26591.17 | -70900.9 | 17518.14 | -23973.6 |
| 83111.94 | | | -11396.9 | 44290.76 | -9494.01 |  |
| MSE: | | 31125376926.67 | | | | |
| Repetition 50 | | | | | | |
| 520414.8 | | | -14158.3 | 18064.38 | 56585.21 | 9393.62 |
| 10949.98 | | | 63594.03 | 49435.14 | 11551.38 | 91258.21 |
| 48431.01 | | | 26568.68 | -71835 | 17219.68 | -24173.3 |
| 82939.72 | | | -11040.5 | 43113.86 | -10124.1 |  |
| MSE: | | 31121608083.94 | | | | |
| Repetition 60 | | | | | | |
| 520414.8 | | | -14175.2 | 18260.61 | 56521.06 | 9852.346 |
| 10944.62 | | | 63628.31 | 49510.28 | 11425.05 | 92007.27 |
| 48370.83 | | | 26547.79 | -72310.8 | 17068.95 | -24252.1 |
| 82850.45 | | | -10823.1 | 42449.86 | -10525.1 |  |
| MSE: | 31120440190.10 | | | | | |

**Learning Rate 0.125**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Repetition 20 | | | | | | |
| 518764.5 | | | -12582.5 | 15829.77 | 57253.48 | 4503.314 |
| 10308.5 | | | 62457.53 | 49860.92 | 14301.55 | 79982.27 |
| 49076.85 | | | 26753 | -62557.2 | 20165.87 | -19978.9 |
| 83872.22 | | | -12263 | 51423.92 | -6163.73 |  |
| MSE: | | 31224163324.77 | | | | |
| Repetition 30 | | | | | | |
| 520321.9 | | | -13672.4 | 16578.29 | 57070.36 | 6500.696 |
| 10648.74 | | | 63202.96 | 49281.74 | 12900.35 | 85260.52 |
| 48958.53 | | | 26599.72 | -67333.9 | 18679.02 | -22705.9 |
| 83672.27 | | | -12121.2 | 47944.28 | -7810.53 |  |
| MSE: | | 31153803854.56 | | | | |
| Repetition 40 | | | | | | |
| 520409.6 | | | -14002.9 | 17290 | 56850.33 | 7873.379 |
| 10850.55 | | | 63441.25 | 49257.87 | 12150.75 | 88315.7 |
| 48711.98 | | | 26598.44 | -69757.6 | 17888.56 | -23645.7 |
| 83312.49 | | | -11719.3 | 45576.62 | -8875.12 |  |
| MSE: | | 31132218642.72 | | | | |
| Repetition 50 | | | | | | |
| 520414.5 | | | -14115.9 | 17765.38 | 56688.7 | 8780.809 |
| 10929.23 | | | 63541.17 | 49346.39 | 11763.08 | 90136.47 |
| 48536.58 | | | 26587.47 | -71075.8 | 17462.17 | -24012.1 |
| 83079.73 | | | -11331.9 | 44073.45 | -9608.14 |  |
| MSE: | | 31124516095.68 | | | | |
| Repetition 60 | | | | | | |
| 520414.8 | | | -14156.9 | 18057.28 | 56587.86 | 9381.57 |
| 10948.72 | | | 63592.8 | 49433.68 | 11556.86 | 91232.3 |
| 48433.93 | | | 26568.77 | -71816.3 | 17225.72 | -24168.5 |
| 82943.14 | | | -11046.4 | 43135.08 | -10114.5 |  |
| MSE: | 31121657579.44 | | | | | |

**Learning Rate 0.100**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Repetition 20 | | | | | | |
| 514414.9 | | | -11408 | 15863.19 | 57240.8 | 3586.55 |
| 10224.15 | | | 61659.02 | 50548.84 | 15052.94 | 76709.63 |
| 48845.59 | | | 27130.02 | -59180 | 21076.01 | -17420.5 |
| 83342.76 | | | -12251.7 | 53000.63 | -5117.39 |  |
| MSE: | | 31327929673.60 | | | | |
| Repetition 30 | | | | | | |
| 519770.6 | | | -13150.2 | 16097.76 | 57196.55 | 5363.344 |
| 10450.41 | | | 62845.03 | 49532.56 | 13665.82 | 82383.18 |
| 49067.52 | | | 26650.39 | -64801.9 | 19481.97 | -21363.4 |
| 83858.4 | | | -12247.3 | 49927.74 | -6903.45 |  |
| MSE: | | 31185703166.20 | | | | |
| Repetition 40 | | | | | | |
| 520345.7 | | | -13754 | 16732.66 | 57025.32 | 6802.151 |
| 10696.79 | | | 63260.5 | 49267.16 | 12727.46 | 85945.55 |
| 48905.35 | | | 26599.12 | -67888 | 18498.12 | -22933.9 |
| 83593.77 | | | -12044 | 47421.64 | -8042.56 |  |
| MSE: | | 31147982334.11 | | | | |
| Repetition 50 | | | | | | |
| 520407.4 | | | -13994.7 | 17280.28 | 56854.04 | 7854.695 |
| 10845.55 | | | 63436.1 | 49261.83 | 12165 | 88266.89 |
| 48713.96 | | | 26597.16 | -69714.5 | 17902.86 | -23624.1 |
| 83317.96 | | | -11720.3 | 45611.43 | -8861.95 |  |
| MSE: | | 31132474420.79 | | | | |
| Repetition 60 | | | | | | |
| 520414 | | | -14097.5 | 17677.73 | 56717.97 | 8612.147 |
| 10916.57 | | | 63523.82 | 49327.52 | 11831.88 | 89804.21 |
| 48569.72 | | | 26590.6 | -70839 | 17538.54 | -23950.4 |
| 83122.2 | | | -11406.8 | 44350.09 | -9470.67 |  |
| MSE: | 31125642186.13 | | | | | |

**Learning Rate 0.075**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Repetition 20 | | | | | | |
| 500243.8 | | | -9104.5 | 16574.87 | 57135.09 | 2802.991 |
| 10276.89 | | | 60092.41 | 51629.37 | 15751.31 | 72407.82 |
| 48155.97 | | | 28133.45 | -54311.4 | 22136.72 | -13151.5 |
| 81490.17 | | | -12488.1 | 54307.32 | -3605.89 |  |
| MSE: | | 31852442887.29 | | | | |
| Repetition 30 | | | | | | |
| 516443.7 | | | -11969.8 | 15859.01 | 57243.77 | 4042.563 |
| 10268.73 | | | 62041.06 | 50219.28 | 14681.92 | 78310.13 |
| 48951.54 | | | 26953.45 | -60822 | 20630.08 | -18651.5 |
| 83584.42 | | | -12256.6 | 52215.5 | -5627.74 |  |
| MSE: | | 31270388040.51 | | | | |
| Repetition 40 | | | | | | |
| 519633 | | | -13111.3 | 16109.93 | 57190.31 | 5349.347 |
| 10450.22 | | | 62818.69 | 49559.85 | 13684.09 | 82305.67 |
| 49056.24 | | | 26664.7 | -64710.3 | 19505.15 | -21279.3 |
| 83834.79 | | | -12237.6 | 49951.61 | -6880.86 |  |
| MSE: | | 31187017377.94 | | | | |
| Repetition 50 | | | | | | |
| 520260.9 | | | -13631.2 | 16572.52 | 57068.92 | 6461.986 |
| 10639.13 | | | 63175.53 | 49312.46 | 12940.19 | 85130.23 |
| 48953.2 | | | 26606.21 | -67198.4 | 18720.36 | -22606.6 |
| 83668.15 | | | -12105.2 | 48013.5 | -7777.38 |  |
| MSE: | | 31155028726.08 | | | | |
| Repetition 60 | | | | | | |
| 520384.5 | | | -13888.3 | 17016.73 | 56936.45 | 7351.801 |
| 10777.7 | | | 63357.24 | 49255.78 | 12426.15 | 87172.02 |
| 48808.58 | | | 26598.76 | -68862.8 | 18180.41 | -23314.8 |
| 83450.68 | | | -11884.6 | 46473.05 | -8469.08 |  |
| MSE: | 31138988973.07 | | | | | |

**Learning Rate 0.050**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Repetition 20 | | | | | | |
| 457144.7 | | | -4254.49 | 18785.03 | 56860.97 | 2481.811 |
| 10480.28 | | | 56592.39 | 52782.93 | 16134.36 | 66463.42 |
| 46491.8 | | | 30509.65 | -46832.9 | 23045.74 | -6319.24 |
| 76182.89 | | | -13404 | 54616.12 | -1292.75 |  |
| MSE: | | 35881042095.27 | | | | |
| Repetition 30 | | | | | | |
| 498353.9 | | | -8930.18 | 16679.01 | 57125.21 | 2833.559 |
| 10289.61 | | | 59966.68 | 51649.64 | 15731.89 | 72290.18 |
| 48088.44 | | | 28232.31 | -54129.1 | 22134.59 | -12954.5 |
| 81262.42 | | | -12539.7 | 54243.22 | -3547.95 |  |
| MSE: | | 31940929115.44 | | | | |
| Repetition 40 | | | | | | |
| 512722.7 | | | -11202.3 | 15984.99 | 57221.94 | 3605.941 |
| 10249.86 | | | 61519.68 | 50628.67 | 15048.49 | 76508.46 |
| 48762.18 | | | 27242.25 | -58893.2 | 21110.18 | -17094.5 |
| 83108.72 | | | -12290.4 | 52960.35 | -5029.27 |  |
| MSE: | | 31358443507.91 | | | | |
| Repetition 50 | | | | | | |
| 517732.7 | | | -12397.7 | 15921.97 | 57231.64 | 4486.874 |
| 10326.19 | | | 62332.38 | 49969.3 | 14335.79 | 79719.03 |
| 49004.12 | | | 26837.42 | -62214.7 | 20233.75 | -19623.2 |
| 83705.6 | | | -12255.9 | 51447.02 | -6070.72 |  |
| MSE: | | 31234230525.07 | | | | |
| Repetition 60 | | | | | | |
| 519479.6 | | | -13071.4 | 16122.83 | 57185.87 | 5335.808 |
| 10450.34 | | | 62791.68 | 49586.99 | 13701.61 | 82228.26 |
| 49043.11 | | | 26678.98 | -64618.3 | 19527.97 | -21194.7 |
| 83809.34 | | | -12229.2 | 49974.37 | -6857.69 |  |
| MSE: | 31188419058.42 | | | | | |

(c) Tune your implementation to obtain results close to those obtained with the package. Write some observations: How does the objective change with different learning rates; how many iterations are needed, etc. 2

With different learning rates, the objective changes to finding the number of iterations to reach the minimum for smaller learning rates to decreasing the learning rate so that the adjustments will not keeping overshooting the minimum. The ideal value would be a learning rate that does not take many iterations to converge yet not too large such that it would oscillate between being above and below the average by a large value. The number of iterations would be when there is a small value of slope, however as seen in a) with a learning curve of 0.1 it is able to reach roughly the same optimal value after 60. Therefore by increasing the iterations to 100, it can obtain a close optimal value compared to the regression package. This will give a MSE of 31120170243.65 and R2 value of 0.729712 with the training set. The difference between the implementation and the package is 277,359.92.

MSE Difference between GD implementation and package

|  |  |
| --- | --- |
| GD Implementation | Package |
| 31120170243.65 | 31119892883.73362 |

Speed Comparison

|  |  |
| --- | --- |
| GD Implementation | Package |
| 0.29920077323913574 | 0.024931907653808594 |

My implementation is about 10 times slower than the package.

**Problem 5 [Ridge regression]**

(a) Write the derivation of the closed form solution for parameter θ that minimizes the loss function J(θ) in ridge regression.

(b) Modify your linear regression implementation to handle ridge regression. Compare the results of linear regression and ridge regression on the dataset. Take several values of the regularization parameter λ and output the MSE, RSE, and R2 metrics. Which model performs better? Interpret the results in your own words.

Training Data

|  |  |  |  |
| --- | --- | --- | --- |
| Lambda | MSE | RSS | R2 |
| 1 | 31120205709.02 | 31120205709020.75 | 0.72971 |
| 20 | 31239402982.09 | 31239402982092.66 | 0.72868 |
| 40 | 31576566573.88 | 31576566573878.92 | 0.72575 |
| 60 | 32103594460.09 | 32103594460085.35 | 0.72117 |
| 80 | 32797072257.42 | 32797072257421.34 | 0.71515 |
| 100 | 33636887975.12 | 33636887975118.42 | 0.70785 |

Test Data

|  |  |  |  |
| --- | --- | --- | --- |
| Lambda | MSE | RSS | R2 |
| 1 | 57168038846.69 | 57168038846686.68 | 0.65712 |
| 20 | 57407516299.16 | 57407516299157.16 | 0.65568 |
| 40 | 57874427842.70 | 57874427842696.78 | 0.65288 |
| 60 | 58532604624.53 | 58532604624534.44 | 0.64893 |
| 80 | 59357892754.97 | 59357892754968.65 | 0.64398 |
| 100 | 59357892754.97 | 59357892754968.65 | 0.64398 |

The normal linear regression performs better on the training data and test set as can be seen from the data above. As lambda increases, the fit becomes less and less perfect on the training data, this also applies to the test set. This is because as lambda increases, it forces all the values in lambda to decrease. This reduces the fit as it slowly progresses into a horizontal line when lambda approaches infinity.