CS1675 - Assignment 4

Zachary M. Mattis

February 14, 2019

I. Problem 1 - Data Analysis

a. Binary Attributes

There is only 1 binary attribute, #4 CHAS. It is a Charles River dummy variable indicating if the housing tract bounds the river.

b. Correlations

The attribute that demonstrated the highest positive correlation was #6, RN, with a value of 0.6954. The attribute that demonstrated the highest negative correlation was #13, LSAT, with a value of -0.7377.

Attribute	Correlation
CRIM	-0.3883
ZN	0.3604
INDUS	-0.4837
CHAS	0.1753
NOX	-0.4273
RM	0.6954
AGE	-0.3770
DIS	0.2499
RAD	-0.3816
TAX	-0.4685
PTRATIO	-0.5078
B	0.3335
LSTAT	-0.7377

Table 1: Correlations

c. Linear / Non-Linear

The most correlated non-linear graph appears to be attribute 13, LSTAT. The data points are very closely packed, while the shape is not necessarily linear. The shape appears to be a quadratic function. It also has the highest negative correlation coefficient.

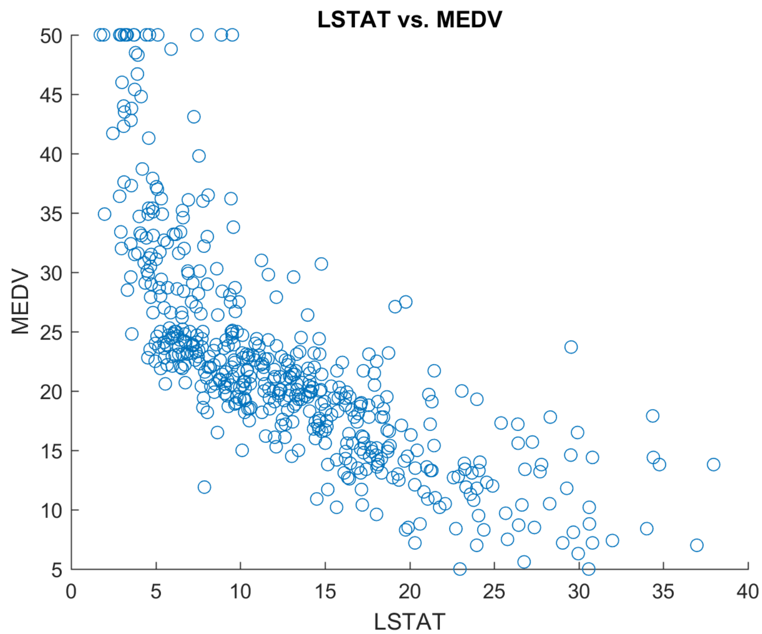

Figure 1: Non-Linear

The most correlated linear graph appears to be attribute 6, RM. In this case, there are a few outliers, but many points are located very close to a line of best fit. It also has the highest positive correlation coefficient.

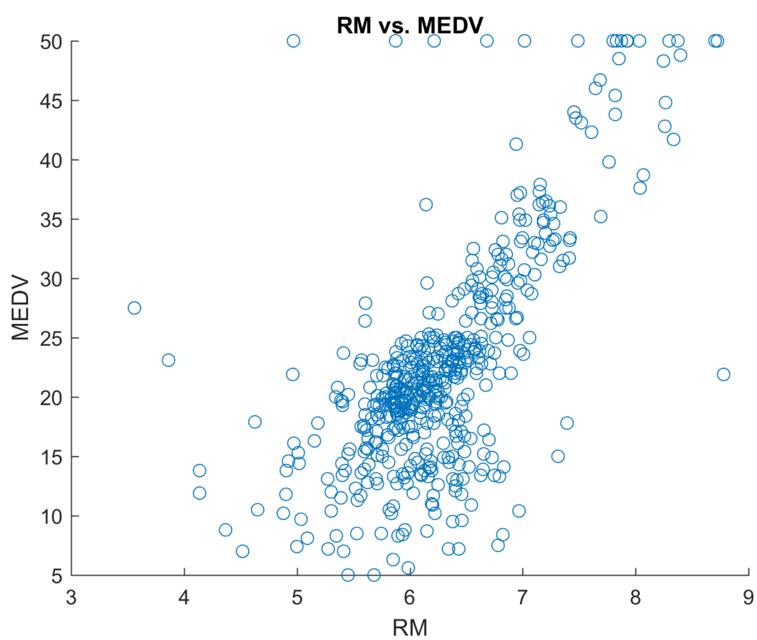

Figure 2: Linear

d. Mutual Correlation

Attribute #9, RAD, and attribute #10, TAX, have the highest mutual correlation with a coefficient of 0.910228.

II. Problem 2 - Linear Regression

a. Solve

```
function [ w ] = LR_solve( X, y )
    w = X\y;
end
```

b. Predict

```
function [ y ] = LR_predict( X, w )
    y = X*w;
end
```

c. Training / Testing Data

$$MSE = \frac{1}{N}\sum(f_i - y_i)^2$$

d. Weight / Mean Squared Error

Attribute	Weight
CRIM	-0.0979
ZN	0.0490
INDUS	-0.0254
CHAS	3.4509
NOX	-0.3555
RM	5.8165
AGE	-0.0033
DIS	-1.0205
RAD	0.2266
TAX	-0.0122
PTRATIO	-0.3880
B	0.0170
LSTAT	-0.4850

Table 2: LR Weights

The mean squared error of the training data was 24.4759, while the MSE of the testing data was 24.2922. Because the MSE of the testing data was lower, this appears to be the better set, although they are both very close.

III. Problem 3 - Online Gradient Descent

a. Regression Coefficients

`online_gradient_descent.m`

b. Gradient Procedure

The mean squared error of the training data was 34.2648, while the MSE of the testing data was 68.2374. In this particular case, the end result was worse than be solving the problem exactly.

Attribute	Weight
CRIM	0.2343
ZN	2.0068
INDUS	0.4363
CHAS	-0.3487
NOX	8.5354
RM	1.358
AGE	-5.4657
DIS	1.0777
RAD	-4.7237
TAX	0.4721
PTRATIO	2.4667
B	2.6942
LSTAT	-0.3014

Table 3: OGD Weights

c. No Normalization

With un-normalized data, the weights approached infinity instantly. Therefore, the weights and errors were not valid.

d. Data Observation

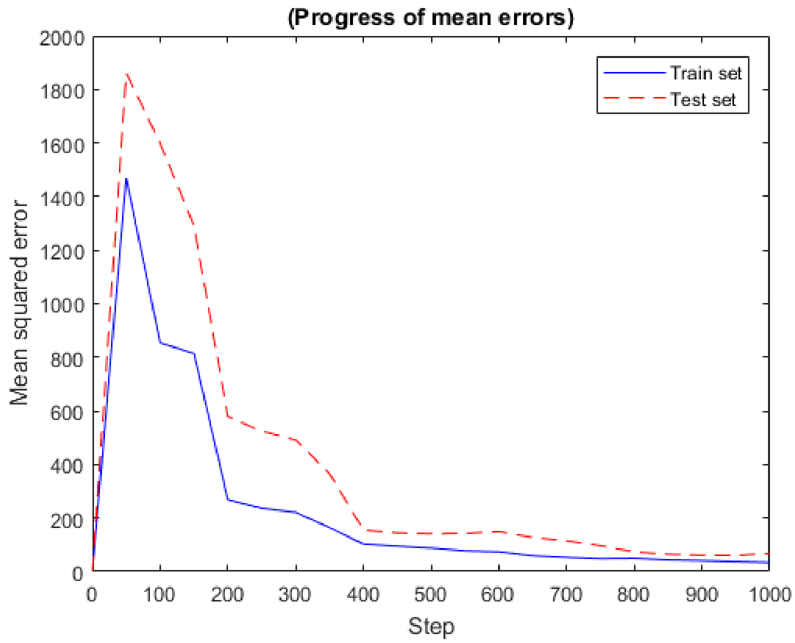

Figure 3: MSE

e. Trials

Steps	1000	1000	1000	1000	1000	1000	10000	100000
Learning Rate	0.01	0.05	$1/\sqrt{t}$	$2/t^2$	$2/t^3$	$2/t$	2/t	2/t
MSE Train	0.2895	0.9875	3.66E+55	0.8247	0.5577	34.2648	5.3167	1.5551
MSE Test	0.1909	0.8354	4.27E+55	0.9502	0.5534	68.2374	8.4973	2.1502
CRIM	-0.0639	0.4199	-6.4E+27	-0.1014	-0.0684	0.2343	0.3515	0.1312
ZN	0.0383	-0.1958	9.07E+26	-0.1454	-0.0823	2.0068	0.2848	0.0628
INDUS	0.0038	0.2318	-5.8E+25	0.2605	0.0154	0.4363	-0.6689	-0.6397
NOX	0.2284	0.2222	1.34E+27	1.0100	0.5398	1.3580	1.1946	0.7857
RM	-0.0005	0.2016	-1.4E+27	0.1962	-0.1313	-5.4657	-0.1862	-0.7714
AGE	-0.1767	-0.2337	6.56E+26	-0.1858	0.0170	1.0777	0.5802	0.2096
DIS	0.1175	0.0543	1.58E+27	-0.3084	-0.1616	-4.7237	-2.6102	-1.6003
TAX	-00507	-0.2419	4.12E+26	-0.2002	-0.1405	0.4721	0.9293	0.9607
PTRATIO	-0.1213	-0.0677	-1.4E+26	-0.2226	-0.1121	2.4667	1.3450	0.5477
B	0.0385	0.0768	-5.9E+25	0.0840	0.0449	2.6942	0.5061	0.2595
LSTAT	-0.4389	-0.3127	7.67E+26	-0.2942	-0.3686	-0.3014	-0.2189	-0.2919

Table 4: Trial Analysis

Keeping the steps constant while changing the learning rate affects error the most. The smaller the learning rate, the lower the error becomes. It is likely that there is a threshold where a lower rate will have worse error (since the regression never approaches the optimum) if the number of steps are not enough. If the learning rate is stalled, then the change becomes less noticeable as the function becomes "smaller and smaller" (such as $2/t^2 => 2/t^3$). Now, if the learning function remains the same while the steps increase, the errors go down with diminishing returns.