

Mouhammadou Dabo

February 25, 2021

CS 1675: Intro to Machine Learning

Professor Milos Hauskrecht

Problem Assignment 4

Problem 1. Linear regression

Part 3.1 Exploratory data analysis

a.

There is 1 binary attribute in the data set.

The attribute is CHAS (Charles River dummy variable)

b.

The correlations between the first 13 attributes and the target attribute are as listed:

- Attribute 1 (CRIM) = -0.3883
- Attribute 2 (ZN) = 0.3604
- Attribute 3 (INDUS) = -0.4837
- Attribute 4 (CHAS) = 0.1753
- Attribute 5 (NOX) = -0.4273
- Attribute 6 (RM) = 0.6954
- Attribute 7 (AGE) = -0.3770
- Attribute 8 (DIS) = 0.2499
- Attribute 9 (RAD) = -0.3816
- Attribute 10 (TAX) = -0.4685
- Attribute 11 (PTRATIO) = -0.5078
- Attribute 12 (B) = 0.3335
- Attribute 13 (LSTAT) = -0.7377

The attribute with the highest positive correlation is Attribute 6 (RM)

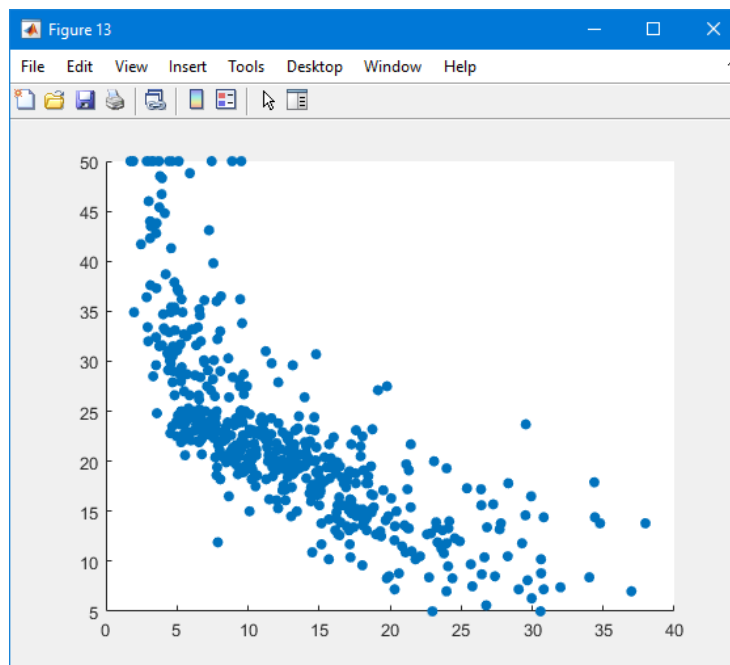
The attribute with the highest negative correlation is Attribute 13 (LSTAT)

c.

The scatter plots that look the most linear are Attribute 6 (RM) (as x increases, y generally increases) and Attribute 13 (LSTAT) (as x increases, y generally decreases)

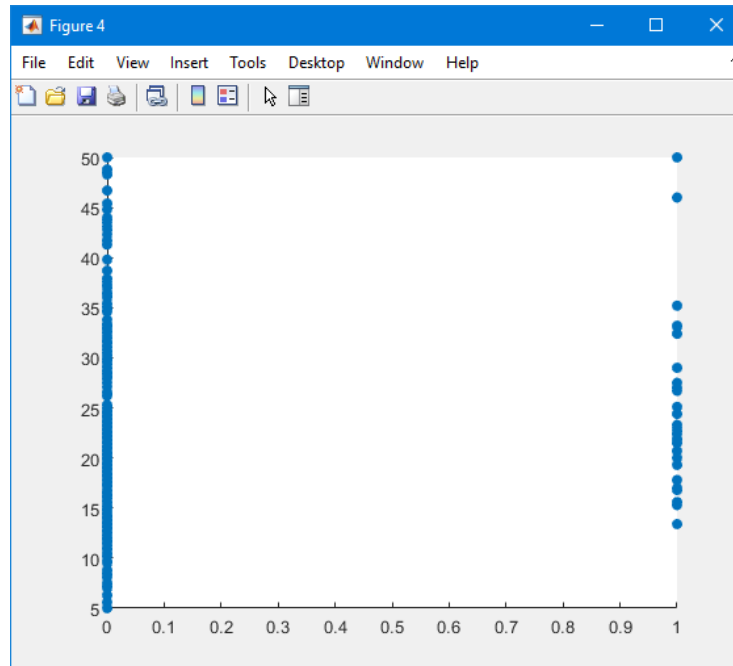


Attribute 6 (RM) Scatter Plot



Attribute 13 (LSTAT) Scatter Plot

The scatter plot that looks the most nonlinear is Attribute 4 (CHAS); there are values spanned across the y axis (from 0 to 50) at $x = 0$ and $x = 1$.



Attribute 4 (CHAS) Scatter Plot

d.

The two attributes that have the largest mutual correlation in the dataset are Attribute 9 (RAD) and Attribute 10 (TAX) with a correlation of 0.910228

Part 3.2. Linear regression

- a. *LR_solve.m*
- b. *LR_predict.m*
- c. *main_LR.m*
- d.

The resulting weights are:

- Attribute 1 (CRIM) = -0.0979
- Attribute 2 (ZN) = 0.0490
- Attribute 3 (INDUS) = -0.0254
- Attribute 4 (CHAS) = 3.4509
- Attribute 5 (NOX) = -0.3555
- Attribute 6 (RM) = 5.8165
- Attribute 7 (AGE) = -0.0033
- Attribute 8 (DIS) = -1.0205
- Attribute 9 (RAD) = 0.2266
- Attribute 10 (TAX) = -0.0122
- Attribute 11 (PTRATIO) = -0.3880
- Attribute 12 (B) = 0.0170
- Attribute 13 (LSTAT) = -0.4850

The MSE for the training set is 24.4759

The MSE for the testing set is 24.2922

The MSE for the testing set is better, since the error is slightly lower than that of the training set.

Part 3.3. Online (stochastic) gradient descent

a.

The final set of weights are:

- Attribute 1 (CRIM) = -0.4908
- Attribute 2 (ZN) = 0.0483
- Attribute 3 (INDUS) = -1.9363
- Attribute 4 (CHAS) = 0.1447
- Attribute 5 (NOX) = -0.1715
- Attribute 6 (RM) = 2.3339
- Attribute 7 (AGE) = 1.0010
- Attribute 8 (DIS) = 2.1825
- Attribute 9 (RAD) = -2.1388
- Attribute 10 (TAX) = -1.8027
- Attribute 11 (PTRATIO) = -1.1864
- Attribute 12 (B) = 2.2053
- Attribute 13 (LSTAT) = -1.5533

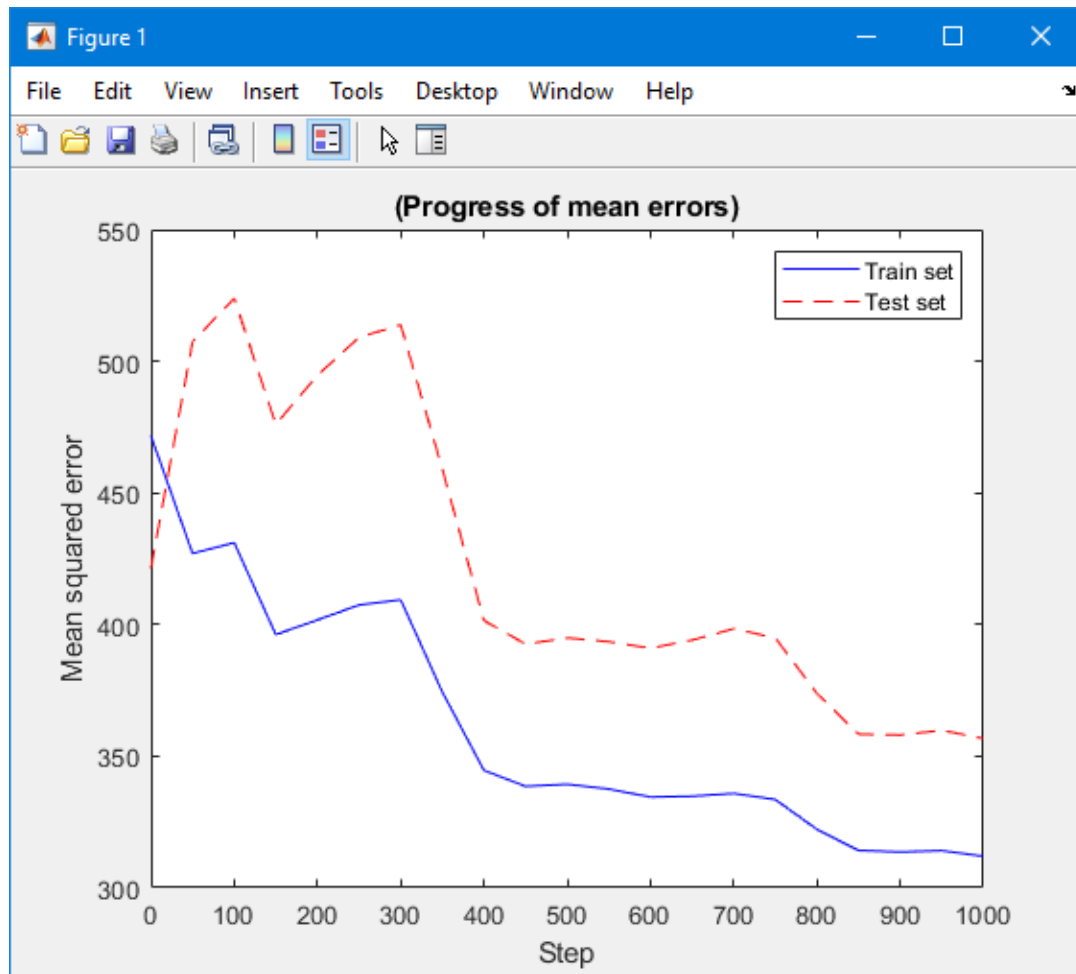
The MSE for the training set is 311.9926 and for the testing set is 356.6929

The result of the MSE is worse than solving the linear regression problem exactly.

b.

After running the program from part a with un-normalized data, I observed that the weights were too large to be computed by MATLAB due to the variance in values for each attribute. MATLAB therefore returned NaN when trying to calculate the weights and MSE on the training and test sets.

c.



Mean Squared Errors Graph

d.

Experimenting with the gradient descent procedure with different trials:

Steps	500	500	500	1000	1000	1000	3000	3000	3000
Learning Rate	0.05	0.01	$2/\sqrt{n}$	0.05	0.01	$2/\sqrt{n}$	0.05	0.01	$2/\sqrt{n}$
(1) CRIM	-0.75	-1.54	2.09	0.96	-1.06	3.47	16.18	-0.91	2.16
(2) ZN	-0.30	-0.13	-0.84	-1.53	0.15	-0.25	0.74	0.44	0.23
(3) INDUS	-0.03	1.03	0.01	1.67	0.31	0.10	-0.41	0.28	-0.39
(4) CHAS	-0.40	1.86	-0.22	-1.10	0.33	-0.29	0.43	1.26	-0.15
(5) NOX	0.09	0.99	-0.02	-1.97	-0.41	-0.84	-5.36	-1.87	-3.40
(6) RM	1.11	1.34	-0.46	2.09	1.71	-0.77	-7.61	0.82	-2.04
(7) AGE	0.25	0.17	0.14	1.38	0.06	0.56	8.99	0.51	1.96
(8) DIS	0.39	-0.15	-0.15	-1.27	-0.68	-0.02	-0.03	-2.61	0.19
(9) RAD	-2.89	-3.03	-0.15	-2.84	-1.48	-0.21	-2.04	0.84	-1.28
(10) TAX	-0.44	-1.67	0.55	-1.59	-0.68	-0.06	-3.62	-0.72	-0.39
(11) PTRATIO	-0.77	-0.82	0.37	-0.49	-0.65	0.11	-1.22	-1.32	-0.49
(12) B	0.35	1.72	-1.58	0.94	0.90	-0.15	0.85	0.38	0.29
(13) LSTAT	-4.90	-5.06	-0.19	-3.49	-5.13	-0.51	-5.89	-5.53	2.50
MSE Train	70.28	68.19	1.6e+73	49.32	39.42	1.2e+75	388.5	27.44	2.8e+73
MSE Test	64.56	81.60	2.7e+73	41.63	38.34	1.6e+75	392.0	23.26	3.3e+73

Trials Table

After experimenting with different values for the learning rate and number of steps, generally the smaller the learning rate, the error tends to be low as well. With an example such as $2/\sqrt{n}$, the error at first gets smaller but then grows exponentially meaning the regression never approaches the optimum. Also, as the number of steps increases with a fixed learning rate, the error tends to get smaller as well; there is definitely a threshold where this is not true anymore.