Università degli Studi di Genova

Robotics Engineering

Sara Romano

Report of Machine Learning course no.2

Linear Regression

November 2020

Contents

Chapter 1

Introduction

The goal of this assignment is to implement in MATLAB the Linear Regression algorithms and adapt them to different cases, considering 2 given data set (Turkish stock exchange data and MT cars data).

1.1 Linear Regression

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or Target 't') and one or more explanatory variables (or Observations 'x'): the goal is to predict the variation of the target t, described by a linear model function 'y(x)', by observing some given variables (the observation variables).

$$y = w \cdot x$$

It is typical supervised problem (having both input and output).

For instance, a single parameter of the linear model can not be good for all points, although, basing on the given data, we can estimated an unknown one which is generally accepted for most of the points, with an expected low error.

The measure of the error is a **square loss function**, which is differentiable with respect to the target and always gives a positive contribution.

$$\lambda_{SE} = (t - y)^2$$

The main idea of the linear regression problem is to make this measure as small as possible on average; hence defining the mean square error objective function J_{MSE}, the goal is to find the minimum value of it.

$$J_{MSE} = \frac{1}{N} \cdot \sum_{l=1}^{N} (t_l - y_l)^2$$

Notice that the square loss is a function of the two arguments: y and t. In this case, having the whole data set and knowing the fixed target as well, the objective depends only on the model parameters w.

In this assignment, we deal with three different types of linear regression problem:

1) One-dimensional linear regression problem;

2) One-dimensional linear regression problem with offset (or intercept);

3) Multi-dimensional linear regression problem.

At the end, the mean square error on the 5 per cent of the data is computed and tested.

1.1.1 One-dimensional linear regression problem

The observation x is just one vector. Minimizing J_{MSE} we get:

$$w = \frac{\sum_{l=1}^{N} x_l t_l}{\sum_{l=1}^{N} x_l^2}$$

The linear regression model y is obtained multiplying w for x.

1.1.2 One-dimensional linear regression problem with intercept

As well as the previous case, the observation x is one vector. Although we deal with another more flexible model:

$$y = w_1 \cdot x + w_0$$

where

a) w_1 = slope;

b) w_0 = intercept, offset.

Minimizing J_{MSE} we get:

$$w_1 = \frac{\sum_{l=1}^{N}(x_l - x_M)(t_l - t_M)}{\sum_{l=1}^{N}(x_l - x_M)^2}$$

$$w_0 = t_M - w_1 x_M$$

where x$_M andt_M arethemeanvalue.$

1.1.3 Multi-dimensional linear regression problem

The data is now composed of d-dimensional vectors.

Therefore we get a Nxd Observation matrix.

The vector w depends on the type of X matrix. Usually it is a flat matrix, then the closed-form solution is computed:

$$W = (X^T X)^{-1} X^T t$$

As before, the multi-dimensional linear regression model y is computer multiplying X for W.

Rovetta A.y. 2020/2021 *Linear Regression* 2020

Chapter 2

Description

On MATLAB the main code is called "LinearRegression.m", which uses some functions to compute our tasks.

2.1 Task 1: Get Data

Two data set have been given:

1) Turkish stock exchange data, composed by two columns (respectively the observation x and the target t) and 263 rows (number of observations).

2) MT cars data, composed by four columns (four variables) and 32 rows (number of observations).

2.2 Task 2: Fit a linear regression model

2.2.1 Task 2a: One-dimensional problem without intercept on the Turkish stock exchange data

Using the first column of the Turkish data set as observation vector x and the second column as the target t, the parameter w is computed.

On MATLAB "one-dim-lin-regres.m" function is called to performed it.

We have got the following result.

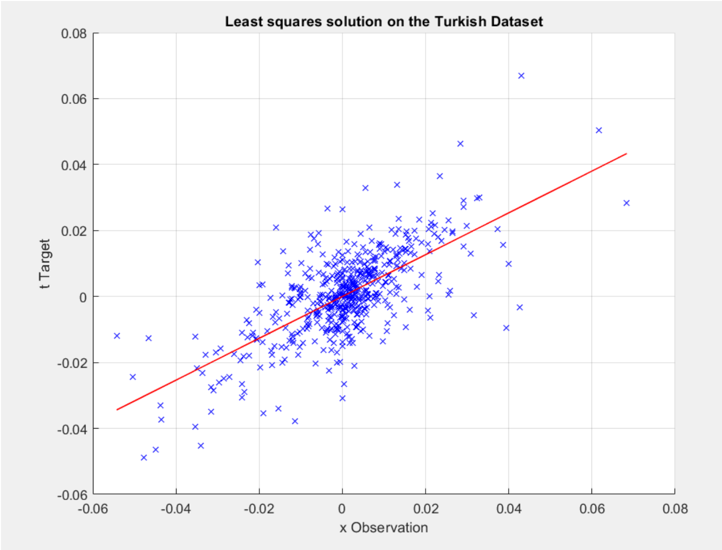

Figure 2.1: One-dimensional linear regression model

2.2.2 Task 2b: Compare graphically the solution obtained on different random subsets (10 percent) of the whole data set

The same MATLAB function is applied on a random subset (10 percent) of the data set.

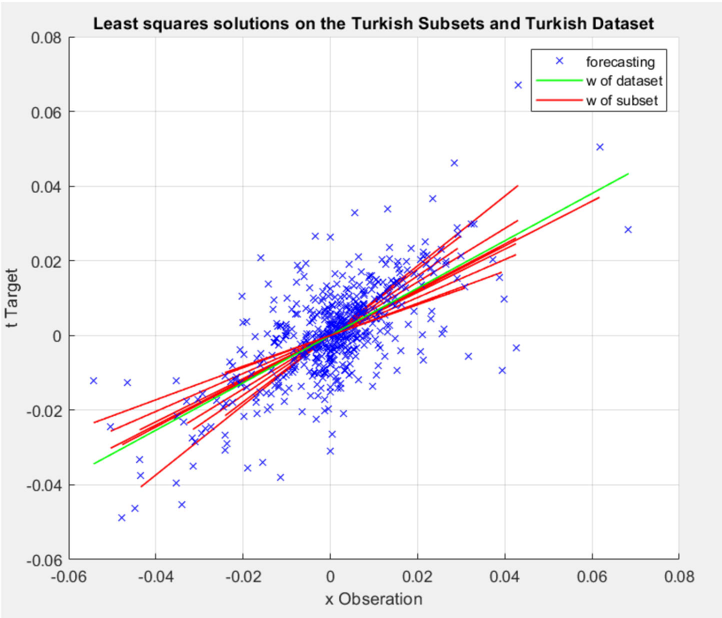

Figure 2.2: Comparation

2.2.3 Task 2c: One-dimensional problem with intercept on the Motor Trends car data, using columns mpg and weight

Using the first column of the MTcar data set as observation vector x and the last column as the target t, the parameters w_1 and w_0 are computed.

On MATLAB "one-dim-lin-regres-intercept.m" function is called to performed them.

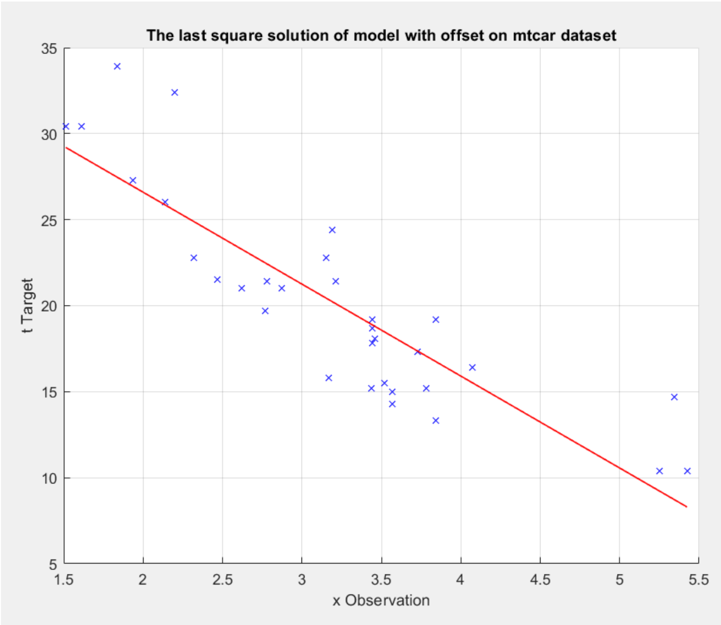

Figure 2.3: One-dimensional linear regression model with offset

2.2.4 Task 2d: Multi-dimensional problem on the complete MTcars data, using all four columns (predict mpg with the other three columns)

Using the last column as the target t and the first three columns for getting X matrix, W is computed.

On MATLAB "multi-dim-lin-regres.m" fuction is called to performed it.

The comparison between the real target and the predicted one is showed in the following table.

	Real Target t	Predicted Target y
1	21	17.7300
2	21	20.8063
3	22.8000	19.4427
4	21.4000	13.4556
5	18.7000	7.0986
6	18.1000	20.0486
7	14.3000	11.7344
8	24.4000	24.0573
9	22.8000	25.7103
10	19.2000	27.3040
11	17.8000	27.3040
12	16.4000	24.7577
13	17.3000	20.6560
14	15.2000	21.2592
15	10.4000	17.1604
16	10.4000	21.1001
17	14.7000	23.1416
18	32.4000	20.2359
19	30.4000	12.9155
20	33.9000	16.6769
21	21.5000	19.9532
22	15.5000	11.8763
23	15.2000	12.4870
24	13.3000	16.1603
25	19.2000	7.3100
26	27.3000	17.0039
27	26	15.7461
28	30.4000	12.0911
29	15.8000	8.7933
30	19.7000	24.1410
31	15	22.5733
32	21.4000	24.1740

Figure 2.4: Comparison between real and predicted target

2.3 Task 3: Test regression model

Considering only the 5 per cent of the two data set, firstly the slope of each type of model is computed.

The Mean Square Error is performed using the "MSE.m" function on MATLAB and it is computed and repeated for 10 times on both the training and test set.

The results are the following:

	Training	Test
Model 1	8.5285e-05	9.8117e-05
Model 2	2.4770e-29	134.0966
Model 3	4.1319e-21	7.6221e+04

Figure 2.5: objective function with respect to the 3 models

References

Linear Regression (2020). Wikipedia. URL: https://en.wikipedia.org/wiki/Linear_regression (visited on 10/20/2020).

Rovetta, Stefano (A.y. 2020/2021). *ml-2020-21-03*. URL: https://2020.aulaweb.unige.it/pluginfile.php/276344/mod_folder/content/0/ml-2020-21--03.pdf?forcedownload=1.