Machine Learning: Assignment 2
Linear Regeression

Filippo Gandolfi 4112879

04/11/2019

Abstract

This assignment is based on linear regression algorithm. There are different type of algorithm in order to adapt to different cases. The requirement of this assignment was to discover these algorithms, one-dimensional problem with or without intercept, multi-dimensional problem on a full multi-column dataset. At the end, it was required also to discover the average of the Mean Square Error of a part of the database (10% and 90%).

[1]

1 Introduction*

The goal of this lab assignment was to build three different types of linear regression models:

- one-dimensional linear regression without intercept;
- one-dimensional linear regression with intercept;
- multi-dimensional linear regression;

and to test them with two data sets.

2 Data set*

We have been given two different data sets to work with. The first one is about the variation of the MSCI Turkish index with respect to Standard and Poor's 500 return index; it is composed of two columns (SP500 and MSCI) and 536 observations. The second data set is about a survey on some car models that takes into account four variables: the miles-per-gallon (mpg), the displacement (disp), the horse-power (hp) and the weight; it is composed of four columns, one for each variable, and 32 observations.

3 Linear regression*

The goal of regression is to predict the value of one or more target variables t given the value of a D-dimensional vector X of input variables comprising of N observations:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_N \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & ... & x_{1,D} \\ x_{2,1} & x_{2,2} & ... & x_{2,D} \\ & & \cdot & \\ & & \cdot & \\ & & \cdot & \\ x_{N,1} & x_{N,2} & ... & x_{N,D} \end{pmatrix}, \qquad t = \begin{pmatrix} t_1 \\ t_2 \\ \cdot \\ \cdot \\ \cdot \\ t_N \end{pmatrix}$$

[1]* Written with: Francesca Canale

so given a training data set x_{n}, where n = 1,...,N, together with corresponding target values t_{n}, the goal is to predict the value of t for a new value of x.
The simplest linear model for regression is one that involves a linear combination of the input variables:

$$y(x, w) = w_0 + w_1 x_1 + ... + w_{\mathrm{D}} x_{\mathrm{D}} \tag{1}$$

This is often simply known as linear regression. **?**
Since generally it is not possible to find values of w_{i}, with i = 1,...,D, that are good for all points of a data set, it is sufficient to choose their values that minimize the cost of a loss function. A common choice of loss function in regression problems is the squared error loss given by:

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

that has the interesting properties of being even, of growing more that linearly, so giving heavier weight to a larger error, and of being differentiable with respect to the model output. The objective function (or cost function) that we want to minimize represents the mean value of the loss over the whole data set:

$$J_{\mathrm{MSE}} = \frac{1}{N} \sum_{l=1}^{N} \lambda_{\mathrm{SE}}(t_{\mathrm{l}}, y_{\mathrm{l}}) = \frac{1}{N} \sum_{l=1}^{N} (t_{\mathrm{l}} - y_{\mathrm{l}})^2 \tag{3}$$

and it is called *mean square error objective*. It is a quadratic function and hence its minimum always exists, but may not be unique. [2]

3.1 One-dimensional linear regression

3.1.1 Without intercept

We have already loaded the data-sets, we start working on them. The first task is to use the *one-dimensional problem without intercept* on the Turkish stock exchange data. x is the value of the variation of SP500 return index on a given day. The target is t, the value of the variation of MSCI on the same day. In order to do this, we build the first function $linear_regression$. We use the formula

$$w = \frac{\sum_{l=1}^{N} x_l t_l}{\sum_{l=1}^{N} x_l^2} \tag{4}$$

To obtain our y we simply multiply the obtained w by x, and we obtain the results visible on Fig. 1.

3.1.2 With intercept

In this task, we develop and algorithm that introdue an intercept, offeset w_0. This means we do not shrink this component towards zero. To do this we introduce two new observations $\overline{x}$ and target $\bar{t}$, defined as:

$$\overline{x} = \frac{1}{N} \sum_{l=1}^{N} x_l \tag{5}$$

$$\bar{t} = \frac{1}{N} \sum_{l=1}^{N} t_l \tag{6}$$

w_0, the our offset, is:

$$w_0 = \bar{t} - w_1 \overline{x} \tag{7}$$

w_1, is the slope and it is:

$$w = \frac{\sum_{l=1}^{N} (x_l - \overline{x})(t_l - \bar{t})}{\sum_{l=1}^{N} (x_l - \overline{x})^2} \tag{8}$$

[2]* Written with: Francesca Canale

We define our final y as:

$$y = w_1 x + w_0 \tag{9}$$

Our model is no more linear and is more refine.

3.2 Multi-dimensional linear regression

We have previously described J_{MSE}, In order to setting $\Delta J_{MSE} = 0$ we get

$$X^T X w = X^T t \tag{10}$$

To obtain the closed-form,

$$w = (X^T X)^{-1} X^T t \tag{11}$$

Where X is (Nd) and $X^T X$ is square $(d+1)(d+1)$, in our dataset is the last three columns and T is again our target, in this specific case it the first column. As before, to compute y:

$$y = Xw \tag{12}$$

3.3 Mean Square Error

In the last task of this assignment We have to compute the mean square error on the training data. However, we don't use the whole dataset, but only the 10% and 90% of it. We also start by computing (in the smaller dataset) the same task as before (linear regression with and without intercept and the multidimensional regression), and then we use the slope found in the objective function of the largest database. At the end we compute it several times and write the value in a table.

4 Test and Results

4.1 One-dimensional problem without intercept on the Turkish stock exchange data

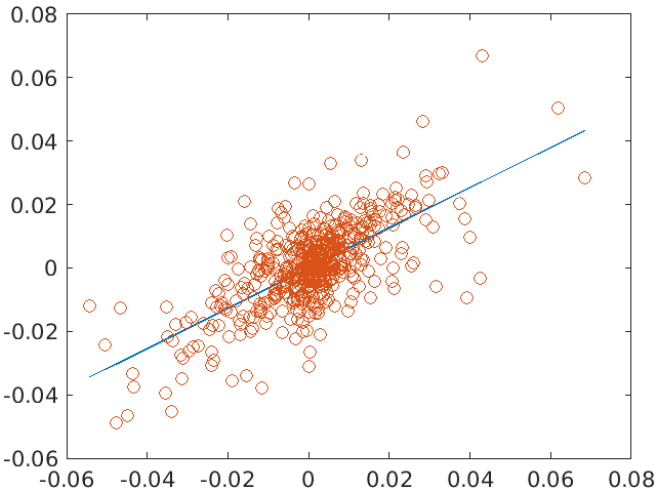

Figure 1: Linear regression without intercept

4.2 Graphical comparison between the solution obtained on different random subsets and the whole data set.

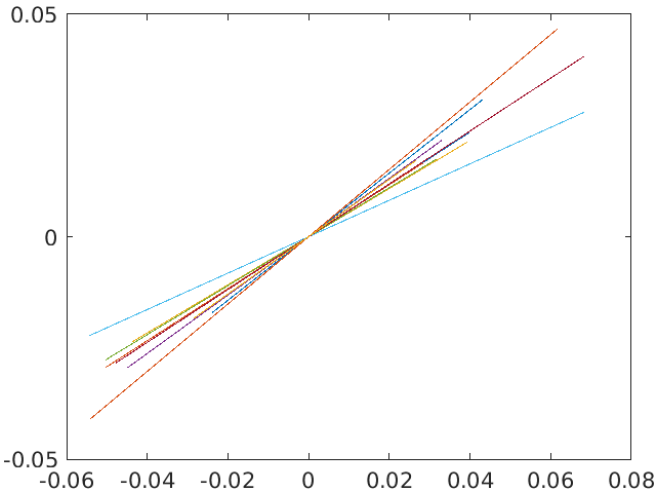

Figure 2: Graphical comparison

4.3 One-dimensional problem with intercept on the Motor Trends car data

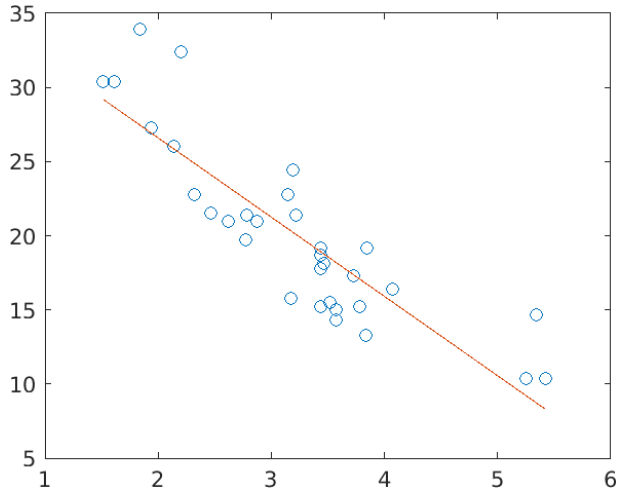

Figure 3: Linear regression with intercept

4.4 Multi-dimensional problem on the complete MTcars data

In this task we create a table that shows the difference between the predicted and the effective ones.

Mpg real (t)	Mpg estimated (y)
21	17.7300
21	20.8063
22.8000	19.4427
21.4000	13.4556
18.7000	7.0986
18.1000	20.0486
14.3000	11.7344
24.4000	24.0573
22.8000	25.7103
19.2000	27.3040
17.8000	27.3040
16.4000	24.7577
17.3000	20.6560
15.2000	21.2592
10.4000	17.1604
10.4000	21.1001
14.7000	23.1416
32.4000	20.2359
30.4000	12.9155
33.9000	16.6769
21.5000	19.9532

Figure 4: Multidimensional problem

4.5 Task 3, objective function one dimensional without intercept

For the third task, again we made different tables

	Dataset	Percentage	MSE
1	Train set	10%	7.2305e-09
2	Test set	90%	8.6657e-09

Figure 5: objective function one dimensional without intercept

4.6 Task 3, objective function one dimensional with intercept

	Dataset	Percentage	MSE
1	Train set	10%	0.0291
2	Test set	90%	0.0034

Figure 6: objective function one dimensional with intercept

4.7 Task 3, objective function multi-dimensional problem

For the third task, again we made different tables

	Dataset	Percentage	MSE
1	Train set	10%	0.0115
2	Test set	90%	1.4822

Figure 7: objective function multi-dimensional