Simple Linear Regression Assignment

CMPS 499/CSCE 598: A. Maida

August 28, 2018

1 Introduction

The assignment is to fit a straight line to a dataset. When one tries to fit a function to a set of data points, it is known as a regression problem. Since we are trying to fit a straight line to the data points, it is a linear regression problem.

We will use the dataset from the website below:[1]

`http://people.sc.fsu.edu/~jburkardt/datasets/regression/x03.txt`

This dataset contains $n = 30$ samples of a person's blood pressure and age. We want to predict blood pressure as a function of age. Since blood pressure tends to increase with age, it's reasonable to try to fit the data with a linear function.

Here are the first five lines of the dataset. Each line has four columns. The first column holds the sample number, the third column holds the person's age, and the fourth column holds the person's systolic blood pressure. Ignore column two for the moment.

```
Table 1
==============
 1   1   39   144
 2   1   47   220
 3   1   45   138
 4   1   47   145
 5   1   65   162
```

Let the blood pressure be represented by the variable y and the person's age by the variable x_1. Our straight line will take the form below.

$$\hat{y} = w_1 \cdot x_1 + b \tag{1}$$

$\hat{y}$ represents the predicted value of y after fitting the straight line to the training data. To do the linear regression, we need to find the best values of w (which controls the slope) and b (which controls the intercept).

[1] More generally: `http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html`.

Following the textbook, in section 5.1.4, we can find the best fitting straight line by minimizing the following mean-squared error (MSE).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)_i^2 \tag{2}$$

As seen in the equation, MSE is just the average of the squared prediction error.

We can characterize the regression function as

$$\hat{y}(x; w_1, b) \tag{3}$$

which says that $\hat{y}$ is some function of x and w_1 and b are parameters of the function. The goal is to search for optimal values of the parameters.

To take steps toward putting this in a more general framework, let us rewrite Eqn. 1 as follows.

$$\hat{y} = w_0 \cdot 1 + w_1 \cdot x = \mathbf{w}^T\mathbf{x} \tag{4}$$

In the above, $w_0 = b$, $\mathbf{w}^T = [w_0, w_1]$, and $\mathbf{x} = [1, x]^T$. This allows us to frame the problem in terms of a single weight vector $\mathbf{w}$ rather than as a separate combination of w and b. Now we can rewrite Expression 3 as

$$\hat{y}(x; \mathbf{w}) \tag{5}$$

To put the equation into a more compact form, let's re-express this equation as the average of the L_2 distance measure. L_2 just means squared Euclidean distance.

$$\text{MSE} = \frac{1}{n}||\hat{\mathbf{y}} - \mathbf{y}||_2^2 \tag{6}$$

Notice that we have eliminated the summation by packing the data values for all of the training samples into a vector. The expression $||\hat{\mathbf{y}} - \mathbf{y}||$ denotes some distance measure between $\hat{\mathbf{y}}$ and $\mathbf{y}$. The subscript 2 specifies that this is the L_2 distance measure. The superscript denotes the usual squaring operation.

There is a derivation in the textbook that shows we can solve for the optimal weights using the formula:

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{7}$$

What is $\mathbf{X}$? Assume that the dataset has only the five samples shown in Table 1. Then $\mathbf{X}$ is given by the matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 39 \\ 1 & 47 \\ 1 & 45 \\ 1 & 47 \\ 1 & 65 \end{bmatrix} \tag{8}$$

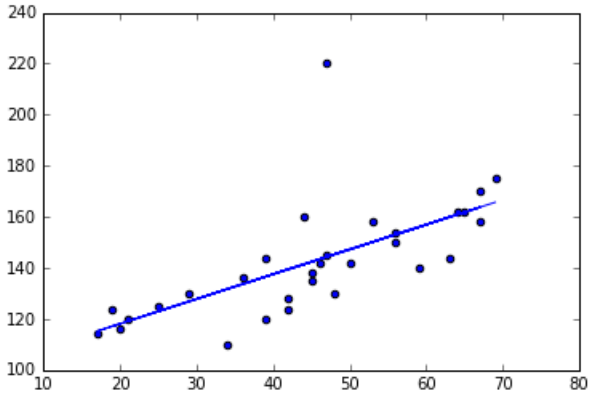

Figure 1: Raw data points with best fitting line. x-axis is age and y-axis is systolic blood pressure.

Each row consists of one input sample. Now we can understand the reason that there are one's in column 2 of Table 1. These are the coefficients that go with w_0 in Eqn. 4. A matrix in which each row consists of one input sample is called a *design matrix*.

The column vector $\mathbf{y}$ consists of the target values for the dataset, as shown below.

$$\mathbf{y} = \begin{bmatrix} 144 \\ 220 \\ 138 \\ 145 \\ 162 \end{bmatrix} \tag{9}$$

Plug the values for $\mathbf{X}$ and $\mathbf{y}$ into Formula 7 and you will obtain the optimal weights for this 5–sample dataset.

2 Write a Python Program

Your assignment is to write a Python program to read the dataset and then perform the computation given in Formula 7. This will give you weights for the optimal regression line. Now create a plot showing the original data points and the optimal regression line superimposed. You can see an example in Figure 5.1 (left) in the textbook.

Your program should read a data file named: `raw_bloodpressure_data.txt`. It is formatted as shown in Table 1, except it has 30 lines instead of 5.

You should print the optimal weights as output and duplicate the plot shown in Fig. 1. The assignment is due in class Tues, Sept. 11. Turn in the source code program with output.