Data Analysis Club

08/09/2017

Hi guys,
It's been few days since we met last time and you might have forgotten many things. So Let's refresh everything.
Here I am giving you a problem called as Normal Equation. Don't worry about the term now. Just try to implement the formula.

Data:

1.From attachment Download file data.txt. Open file in text editor. The file has data of housing prices. The data.txt file has two column, first column contains size of House(sq.m) and second column

contains price of the house.Note values of two columns are separated by comma.

2.Now open python shell(type python3 in terminal) and import matplotlib.pyplot and numpy.

3. Now load the data using command:

data= numpy.loadtxt("data.txt", delimiter= ',')

Delimiter is used because value between two columns are separated By ,(comma)

4. Visualizing data:

A.our X will be size of house, i.e first column of data;

So, X= ..... (write command)

B. Y will be price of house, i.e. second column of data.

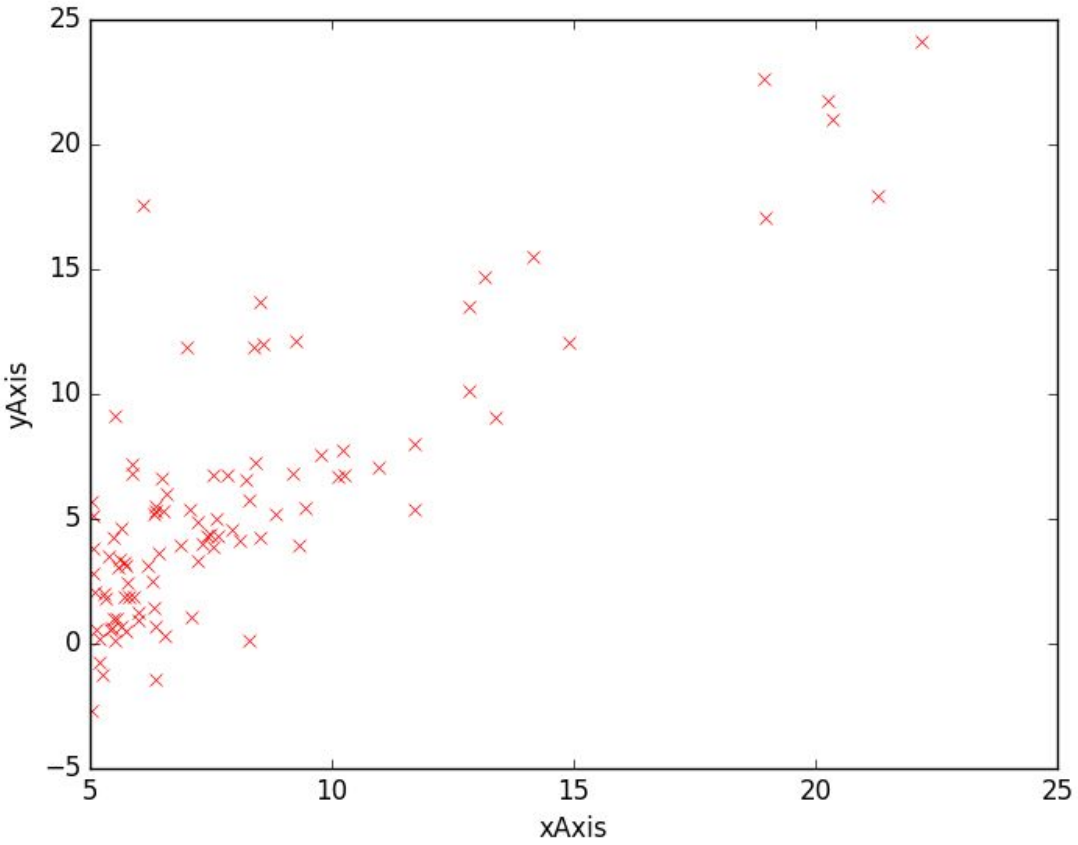
S0, Y= ..... (write command)

C. now plot the data

plt.plot(X,Y, 'rx')

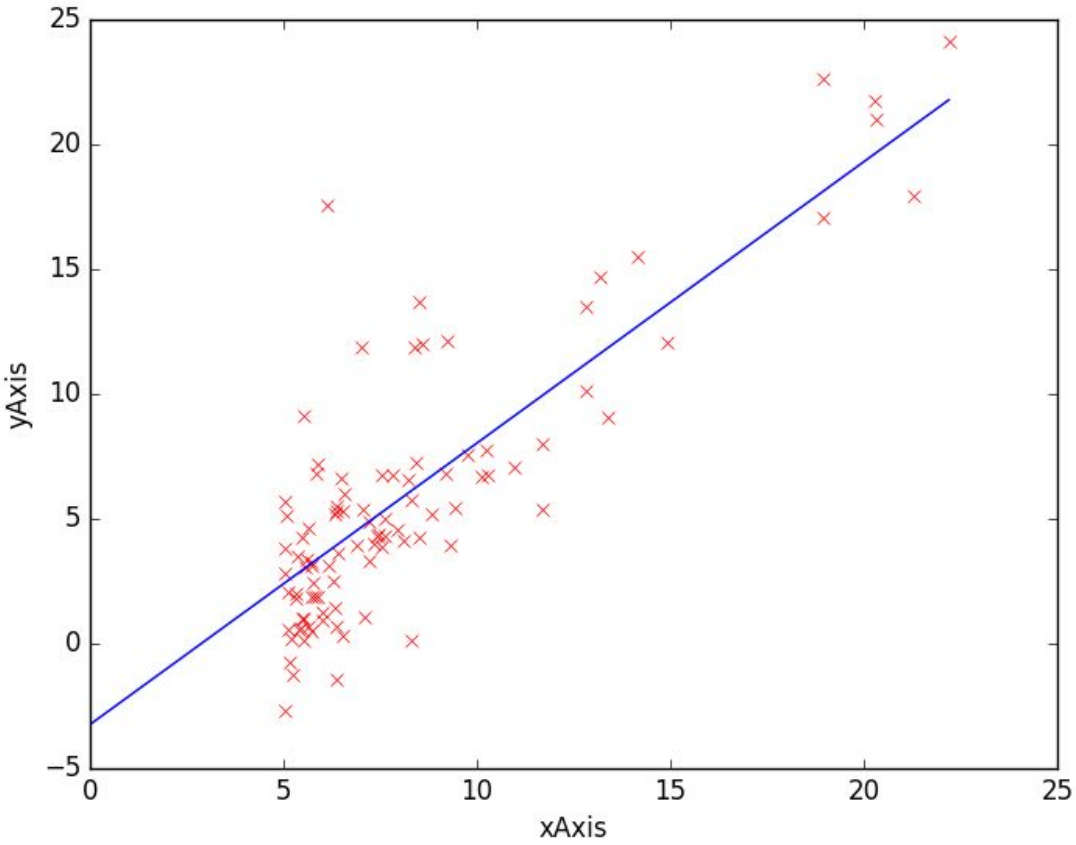
plt.show()

5.Now you can visualize data, it should look like:



6. Aim: Our aim is to draw a fittest line through this data. i .e. To find the slope and y-intercept of fittest line.

Below figure shows that line:



THEORY:

1.Notations:

$x^{(i)}$: size of i^{th} house

$y^{(i)}$:price of i^{th} house

m: total number of house

Y:matrix of price of houses (m X 1)

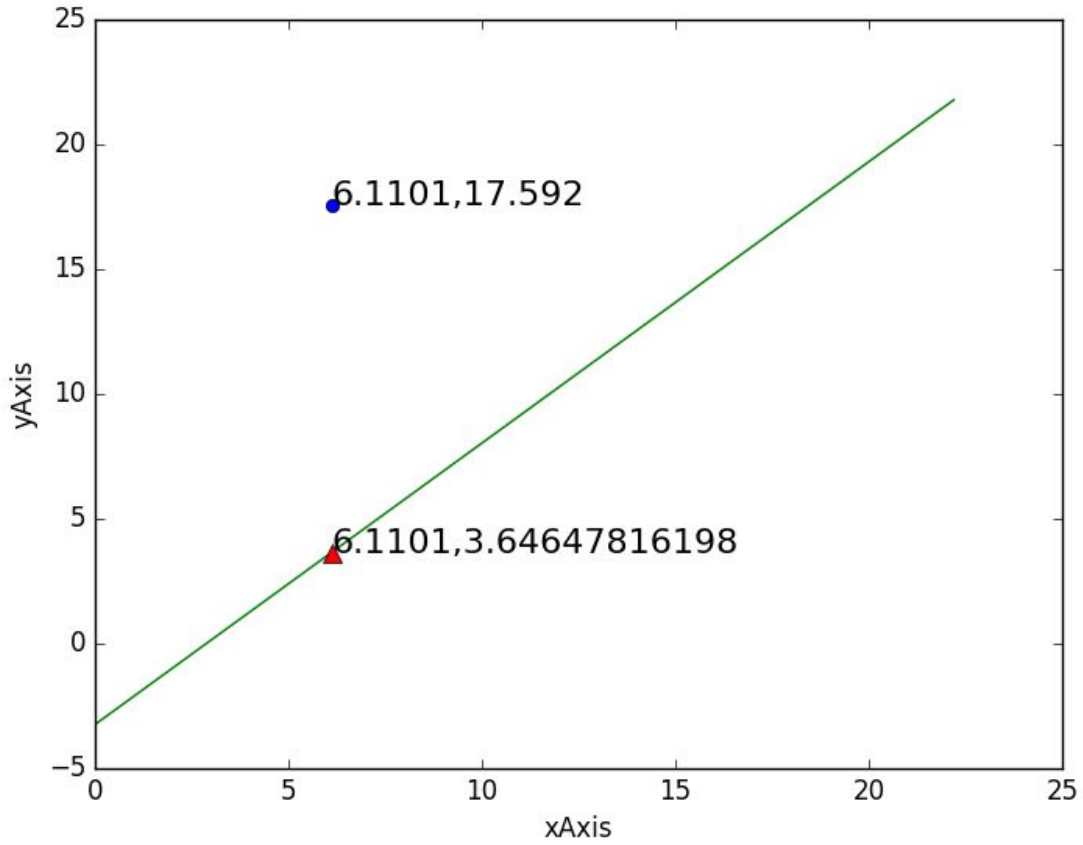
X: matrix of size of houses (m X 1)

2.Hypothesis: Hypothesis is predicted value of y(price of house), represented as h.

Let us see what is hypothesis. Consider the data of first house.

$y^{(0)}$.(price of house) =17.592

$x^{(0)}$: size of house = 6.1101



In above figure green line is our fittest line. Blue point is data of first house(0th house). And red triangle is our prediction.the value 3.6464(h),i.e y coordinate of red triangle is hypothesis.

So, $h^{(i)} = theta0 + theta1 * x^{(0)}$; (here'*' is multiplication of two number)

Where theta1 is slope of fittest line and theta0 is y-intercept.

A.Define theta= [theta0 theta1]

B. X: matrix of size of houses is a m X 1 matrix, where m is number of house, take a matrix **X0= numpy.ones(m,1).**

And stack(column stack) it at the front of X. Use the command **numpy.column_stack(X0, X).** now print(X) to see how X looks now.

C.then our hypothesis is given by h = X * theta; (here'*' is multiplication of two matrix)

D. the slope and intercept of fittest line is given by formula

Normal Equation:

$$theta = (X^TX)^{-1}(X^TY)$$

Where X^T is transpose of X.

X^TX is matrix multiplication of X^T with X

$(X^TX)^{-1}$ is inverse of matrix X^TX

Numpy command to find inverse of a matrix is:

```
from numpy.linalg import inv
```

Then to find inverse of a matrix A just type

```
inv(A)
```

1.Even if you don't understand hypothesis section, leave it.

Only Find $theta = (X^TX)^{-1}(X^TY)$ using numpy command.

where

```
data= numpy.loadtxt("data.txt", delimiter= ',')
Y = data [:,1]
X= data[:,0]
m= X.shape[0]   //number of house
X0 = np.ones((m,1))
X= np.column_stack(X0, X)
```

2.Now you have theta which is a 2 X 1 matrix, where first element

is intercept fittest line and second element is slope of fittest line. Using matplotlib.pyplot, try to plot this line.

If you have doubt contact us. We can meet somewhere if required(e.g. In mess).