Intro to ML

Sameer Dhiman

July 12, 2020

1 Description

This a meant to be for those who are new to ML and want a motivation to get started. We will start from basics and build the foundation from there. During these sessions you can use any programming launguage. If any quesiton needs to done using progamming it will be explicitly mentioned. If you have any other concerns feel free to contact me.

2 Regression

2.1 Intro

Guess the missing numbers in the sequence

a. 1 2 3 4 5 6 7 _ _ _

b. 10 20 _ 40 50 _ _ 80 90 100

c. 2 4 8 _ 32 _ _ 256 _ _

What you just did here that is basically what machine leaning is about, finding the missing values that we don't know using some pattern that appear in the data. In the *a.* part you saw that all the numbers differ by 1 so you used this knowledge that you learned by observing the data to guess the next numbers in the sequence.

Now, How can we represent this sequence as a function? Input for all functions an be $x = \{1, 2, 3, 4, 5, \dots, 10\}$

a. $f(x) = x$

b. $f(x) = 10 * x$

c. $f(x) = 2^x$

This was easy, right? We won't be here if it was this easy, Now find the function for these values input is the same $x = \{1, 2, 3, 4, 5, \dots, 10\}$:

1. 42, 74, 106, 138, 170, 202, 234, 266, 298, 330
2. 26, 29, 32, 35, 46, 41, 52, 63, 50, 53
3. 3, -25, -41, -49, 67, 103, 119, 119, 131, 167

If you were able to guess the funcions then congrats, otherwise don't worry we are not here to do manual labor. We will make the computer do all the work for us. We will see it later.

Btw fuctions are:

1. $f(x) = 32 * x + 10$
2. $f(x) = x^2 + 4 * x + 21$
3. $f(x) = x^4 - x^3 + 2 * x^2 - 20 * x + 12$

2.2 Error

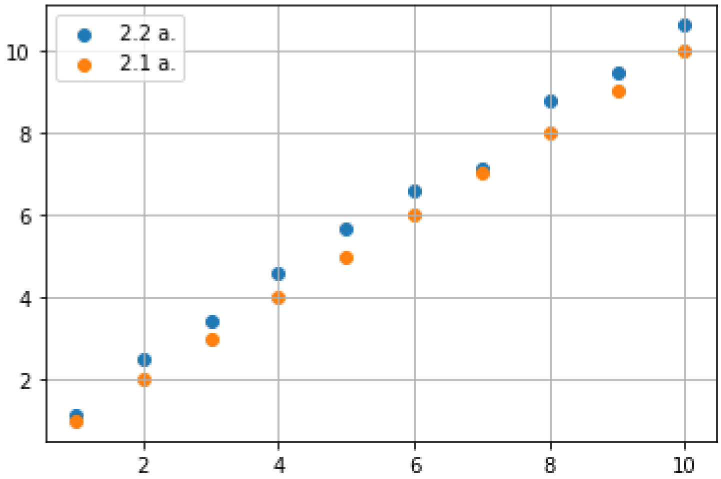
Guess the function for given inputs and outputs.

a. 1.136, 2.518, 3.428, 4.601, 5.646, 6.582, 7.105, 8.759, 9.466, 10.607

b. 10.512, 24.143, 34.197, 43.986, 50.352, 61.258, 71.592, 84.839, 94.502, 104.009

c. 6.474, 8.359, 11.588, 19.33, 34.602, 68.35, 129.084, 260.18, 515.72, 1025.247

Now we seem to have run into some problems here. These seems to follow similar pattern as above question but there is some deviation from above parts. Let us try to plot these then we maybe able to figure out what is happening (Use any programming language to plot these and see for yourself).



The blue one is our 2.2 a. data and orange one is our 2.1 a. data. As you can see it almost follows the same line, so if we use the same function then there will be some error in our calculations. Let us check the error. Here error will be difference in our precdiction and actual values, our precdiction is $f(x) = x$ and actual values are 1.136, 2.518, 3.428, 4.601, 5.646, 6.582, 7.105, 8.759, 9.466, 10.607

$$Error = \{-0.136, -0.518, -0.428, -0.601, -0.646, -0.582, -0.105, -0.759, -0.466, -0.607\}$$

for respective values of x

Total error will be sum of all these error values

$$Total\ error = -4.848$$

Try other parts yourself, use any programming language to plot the graphs and see for yourself how they deviate.

Now problem with deinfing error this way is that is error is positive for some value and negative for other value thye will cancel each other out. So to solve this problem we can either take modulus of error or square it. We will prefer to square each error values. We will see reason for doing this later. So now

$$TotalError = \frac{\sum Error^2}{Total\ number\ of\ input\ values}$$

This is called Mean Squared Error. For above problem

$$Mean\ Squared\ Error = 2.76$$

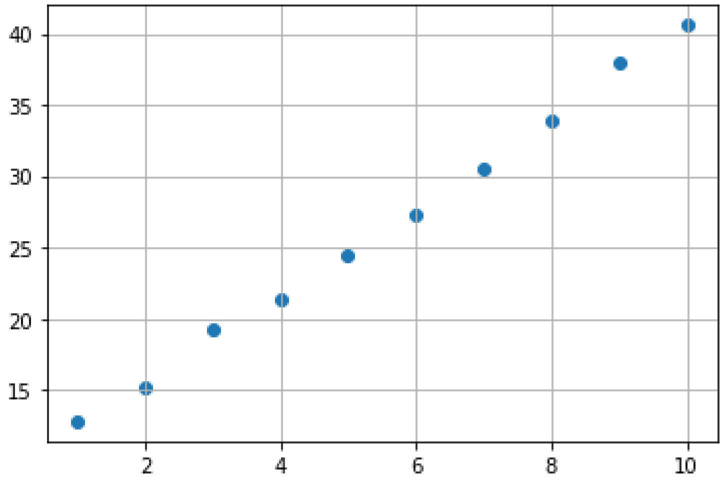
Similarly in real life problems data will not always follow a nice function so there will be some error in your precdiction. So our main task is to minimize this error.

2.3 Regression

Now the question is How to make the computer do the work for you? Well Let us analyze one way, we can approach this problem.

Let's say you have a data

$$\{12.8, 15.2, 19.2, 21.3, 24.5, 27.3, 30.5, 33.9, 38.0, 40.6\}$$

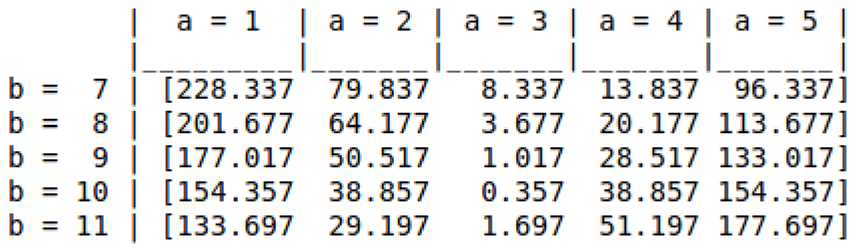


Now after plotting this data you realize it kind of follows a linear path so we can try a linear funciton. What does a linear function looks like?

$$f(x) = a.x + b$$

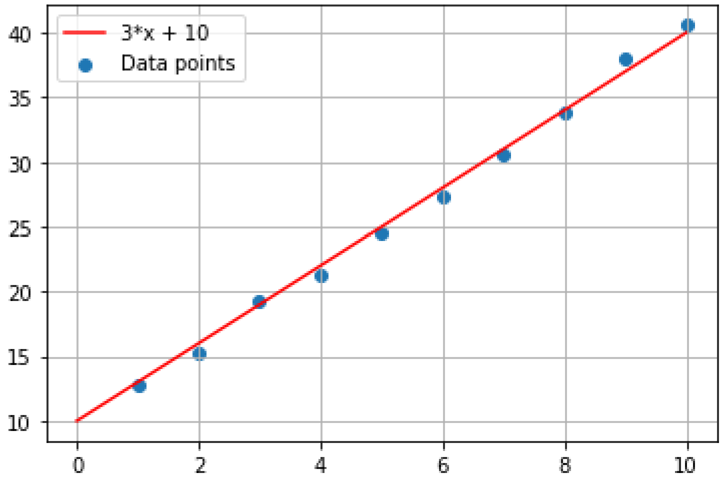
Here a and b are cofficients that we have to find.

After some trial and error these are the error values we found:



```
            |   a = 1   | a = 2 | a = 3 | a = 4 | a = 5 |
            |___________|_______|_______|_______|_______|
b =  7 | [228.337  79.837   8.337  13.837  96.337]
b =  8 | [201.677  64.177   3.677  20.177 113.677]
b =  9 | [177.017  50.517   1.017  28.517 133.017]
b = 10 | [154.357  38.857   0.357  38.857 154.357]
b = 11 | [133.697  29.197   1.697  51.197 177.697]
```

Error is minimum for a = 3 and b = 10 so we will choose those as our cofficients. Note: I checked only the natural numbers for simplicity. Let us plot the line and check.



There are better ways to do it than trial and error. We will look at those in the next session.

Congrats you made it to the end. If you would like to share your feedback, you can contact me directly. Did you like this format or you would like to change something?