Supervised ML Algorithms - Regression

Linear Regression: Introduction

dφ

Democratizing Data Science Learning

DPhi

Learning Objectives

Dependent and Independent Variables

Equation of a Straight Line

Linear Regression

DPhi

Dependent and Independent Variables

- So far you've been studying input and output/target variables. Commonly, the input variable is known as independent variable and target variable is known as dependent variable.

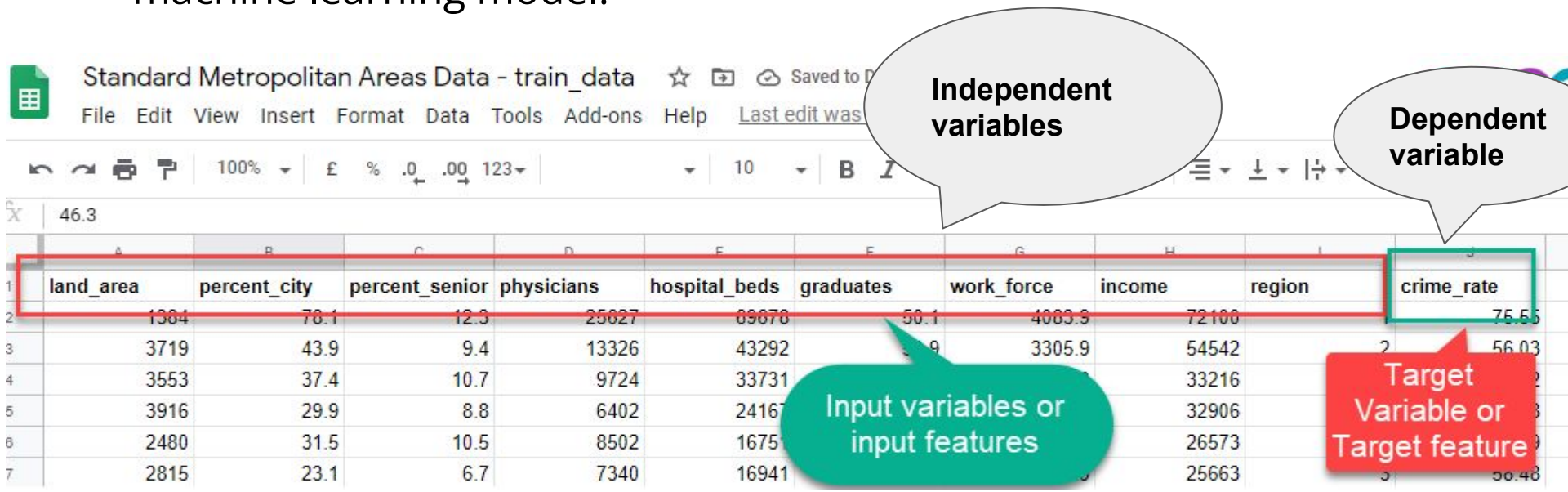
- In nutshell, our target variable is nothing but a dependent variable. Why dependent? Because the values of this variable are dependent on other variables (i.e. input variables)

- And, our input variables are known as independent variables. Here the values of these variables are not dependent on any other variables.

Let's look at some examples to learn more about them!

Dependent and Independent Variables

- Look at the Standard Metropolitan Areas Data below. In the dataset **we might be curious to predict "crime_rate" in future**, so that becomes our target variable **(dependent variable)** and rest of the variables become input variables (**independent variables**) for building a machine learning model.



land_area	percent_city	percent_senior	physicians	hospital_beds	graduates	work_force	income	region	crime_rate
1384	78.1	12.3	25627	89678	50.1	4083.3	72100	1	75.55
3719	43.9	9.4	13326	43292	9	3305.9	54542	2	56.03
3553	37.4	10.7	9724	33731			33216		
3916	29.9	8.8	6402	24167			32906		
2480	31.5	10.5	8502	16751			26573		
2815	23.1	6.7	7340	16941			25663	3	58.48

DPhi

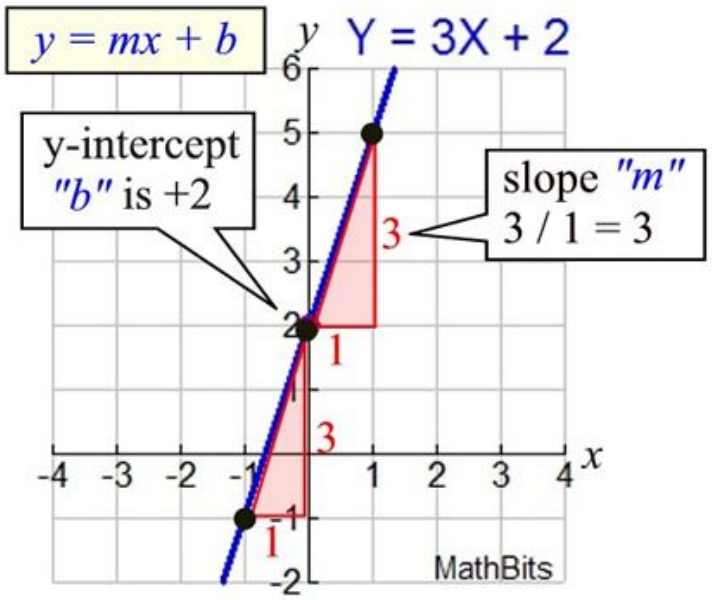
Another example

- A scientist wants to see if the brightness of light has any effect on a moth being attracted to the light.
- The brightness of the light is controlled by the scientist. This would be the independent variable.
- How the moth reacts to the different light levels (distance to light source) would be the dependent variable.

Equation of a Straight Line

- In algebra, a linear equation (equation of a straight line) typically takes the form y = mx + b, where m and b are constants, **x** is the **independent variable**, **y** is the **dependent variable**.

- Basically, the value of y is being calculated using x whereas x has no dependence on value of y.



➔ y = how far up
➔ x = how far along
➔ m = Slope or Gradient (how steep the line is)
➔ b = value of y when x=0

- **How do you find "m" and "b"?**

  - b is easy: just see where the line crosses the Y axis.
  - m (the Slope) needs some calculation:

$$m = \frac{\text{Change in Y}}{\text{Change in X}}$$

DPhi

Synonyms Recap

Too many synonyms to memorise? Let me put them all down at one place for better understanding:

- Variables = Features
- Input Variables = Attributes = Predictor = Independent Variables
- Target Variables = Labels = Outcomes = Dependent Variables

What is linear regression? - an example

Suppose you are thinking of selling your home. And, various houses around you with different sizes (area in sq.ft) around you have sold for different prices as listed below:

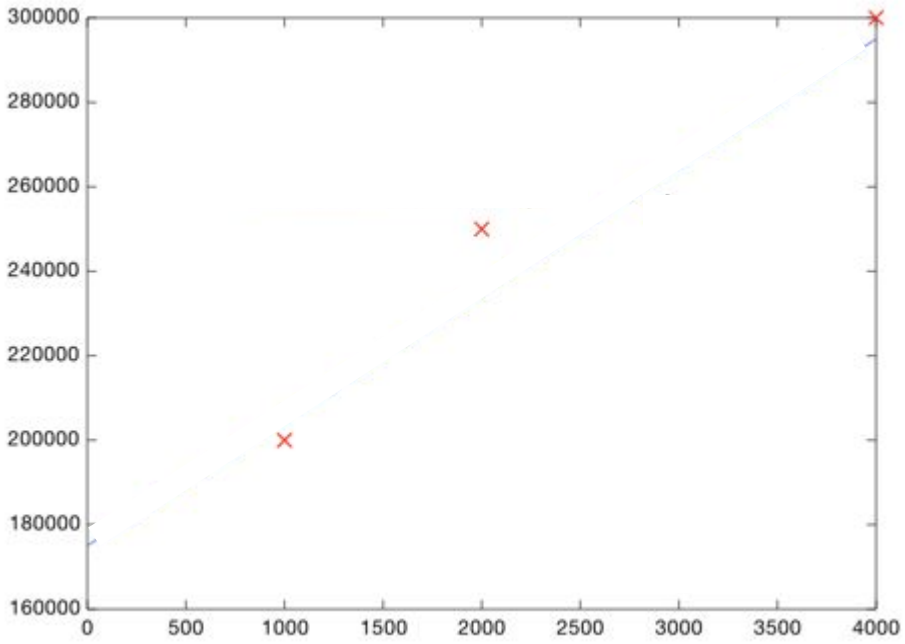
And considering, **your home is 3000 square feet**. How much should you sell it for?

Well! You have to **look at the existing price patterns (data) and predict a price for your home.** This is called **linear regression.**

DPhi

What is linear regression? - an example

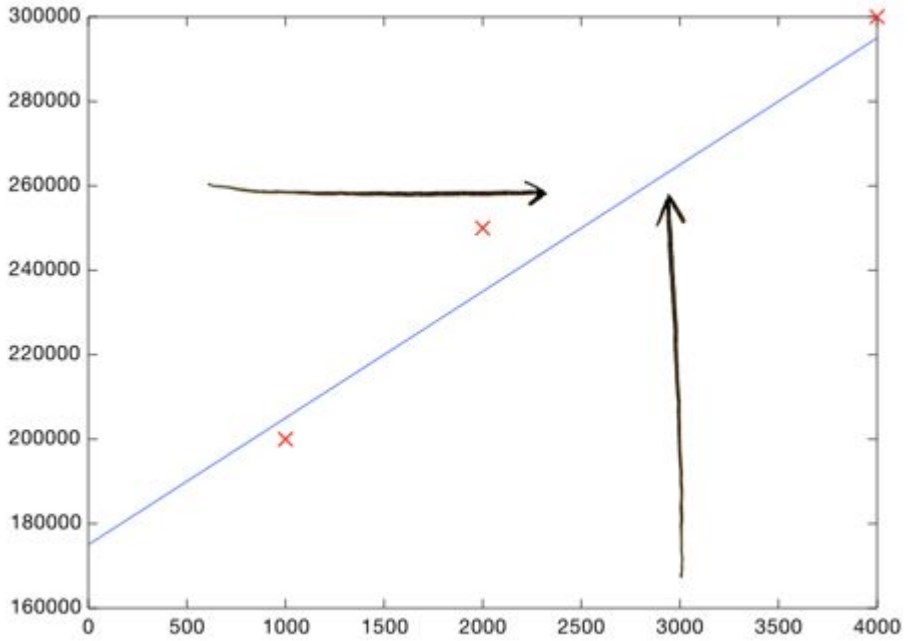
Here's an easy way to do it. Plotting the 3 data points we have so far:



Each point represents one home.

What is linear regression? - an example

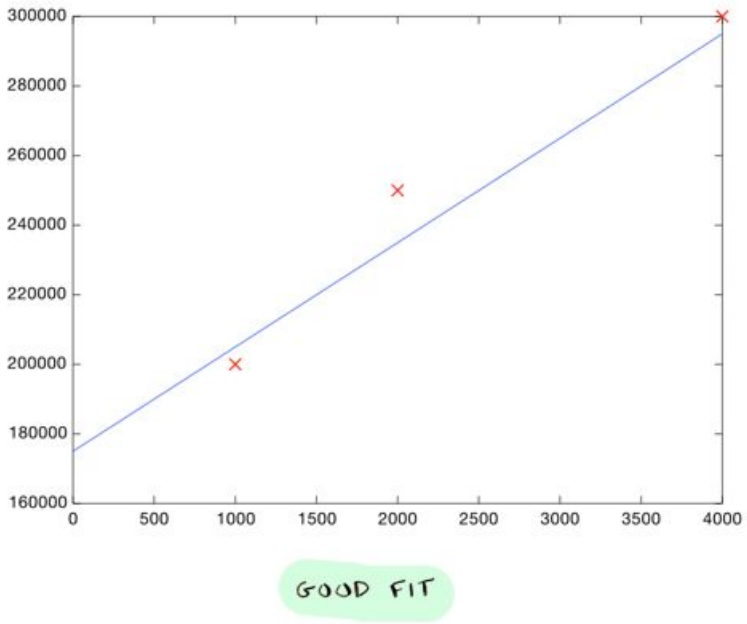
Now you can eyeball it and roughly draw a line that gets pretty close to all of these points. Then look at the price shown by the line, where the square footage is 3000:

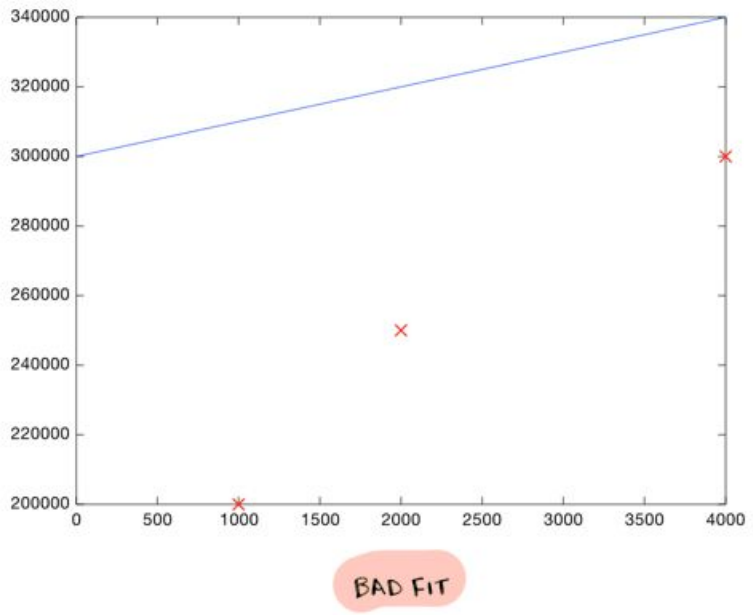


Boom! Your home should sell for $260,000.

What is linear regression? - an example

That's all! You plot your data, make a rough line, and use the line to make predictions. You need to make sure your line fits the data well:





But of course we don't want to roughly make a line, we want to compute the exact line that best "fits" our data. That's where machine learning comes into play!

DPhi

What is linear regression?

- Linear regression is a linear model i.e. a model that assumes a **linear relationship** (straight-line relationship) between the input variables (x) and the single output variable (y).

- When there is a single input variable (x), the method is referred to as **simple linear regression** or just linear regression. **Eg:** Salary dataset given here. There is only one target variable and one input variable where we are predicting the salary of individual using their years of experience.

- When there are multiple input variables, it is often referred to as **multiple linear regression**. **Eg:** Smart Metropolitan areas data set, we have multiple input variables

References

- http://adit.io/posts/2016-02-20-Linear-Regression-in-Pictures.html

DPhi

Slide Download Link

You can download these slides from the below link:

https://docs.google.com/presentation/d/1TSYSfCtbesru2CoWGOf_78VyymisyfYl66z5nWWZEc8/edit?usp=sharing

DPhi

That’s it for this unit. Thank you!

Feel free to post any queries on Discuss.

DPhi