Introduction to Big Data and Machine Learning
A real-life machine learning problem

Dr. Mihail

August 22, 2019

Linear Regression

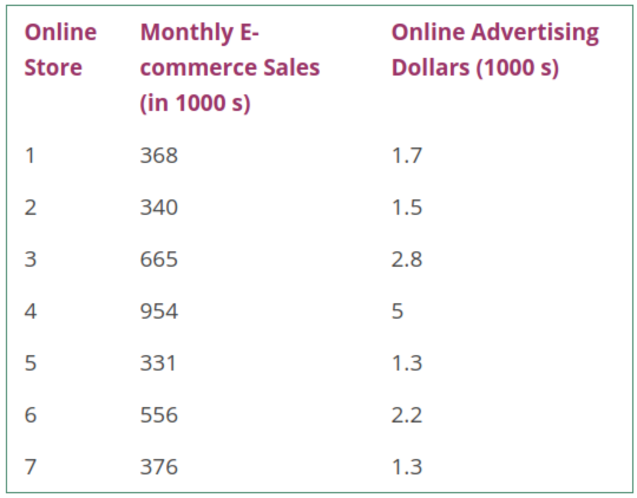
Problem statement

- You have to study the relationship between the monthly e-commerce sales and the online advertising costs.
- You have the survey results for 7 online stores for the last year.
- Your task is to find the equation of the straight line that fits the data best.

Linear Regression

Data

The following table represents the survey results from the 7 online stores.

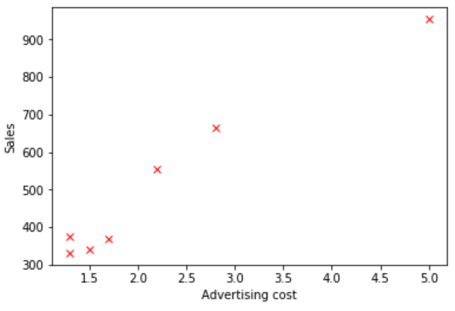


Online Store	Monthly E-commerce Sales (in 1000 s)	Online Advertising Dollars (1000 s)
1	368	1.7
2	340	1.5
3	665	2.8
4	954	5
5	331	1.3
6	556	2.2
7	376	1.3

Linear Regression

Modeling

- The "model" is a theoretical set of rules that real data were generated from
- In our case, we will assume there is a linear relationship between the variables
- In some cases, visualizing data can help with model intuition



Linear Regression

Mathematical model

- $Y = B_0 + B_1 X$

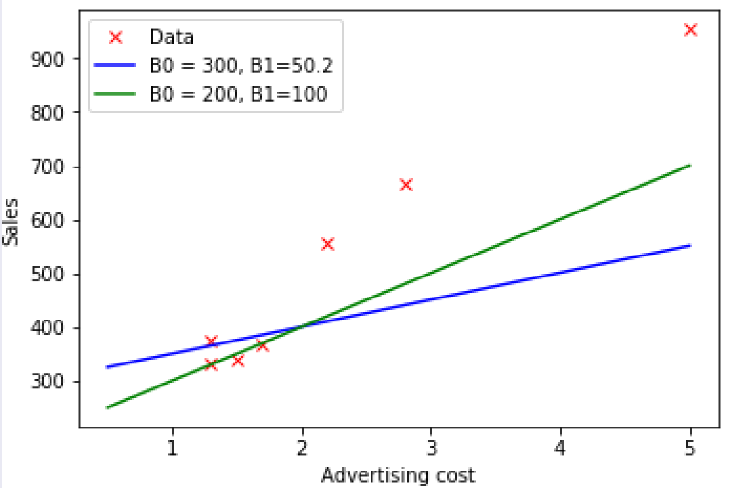
Terms

- Y: the dependent variable (sales), what we're trying to model
- X: the independent variable (cost to advertise)
- B_0 and B_1: model parameters that we're trying to estimate from the data

Linear Regression

Estimating model parameters

Guesswork:



Linear Regression

Optimization

- In order to “best” fit the data, we need an **objective**
- The objective is a function of the model parameters (B_0, B_1)
- Objective is at a minimum, when the model fits the data “better”
- We will call the objective “loss”, and attempt to minimize it
- $\mathcal{L}$

Loss

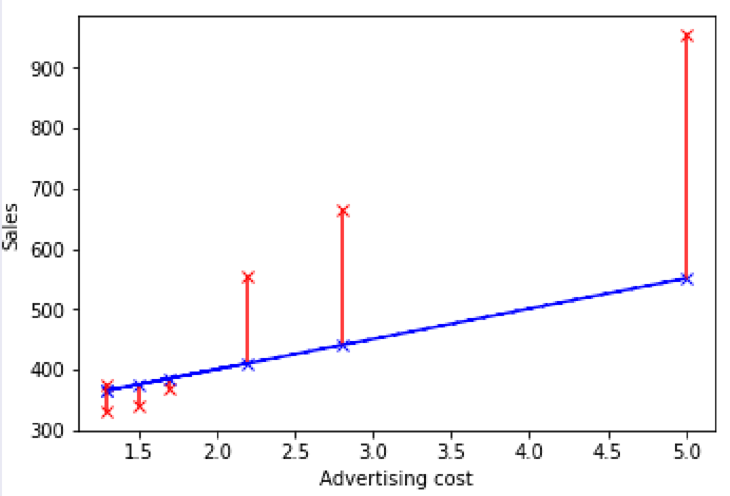
$$\mathcal{L}(B_0, B_1, Y, X) = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - (B_0 + \hat{x}_i B_1))^2$$

- N number of data points
- $\hat{x}$ and $\hat{y}$ input data pairs

Linear Regression

Estimating model parameters

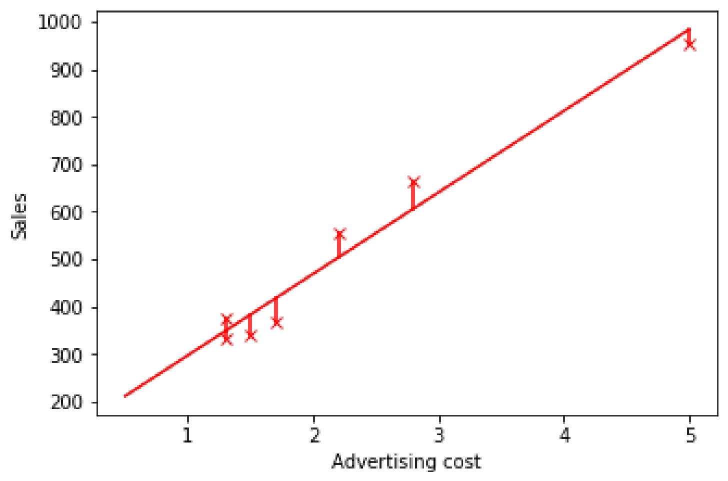
Before optimization, $\mathcal{L}$ is the sum of the lengths of red lines:



Linear Regression

Estimating model parameters

After optimization:


Sales
Advertising cost