Linear Regression: An Introductory Course

Chandresh Kumar Maurya, PhD

Assistant Professor in CSE deptt at IIT Indore

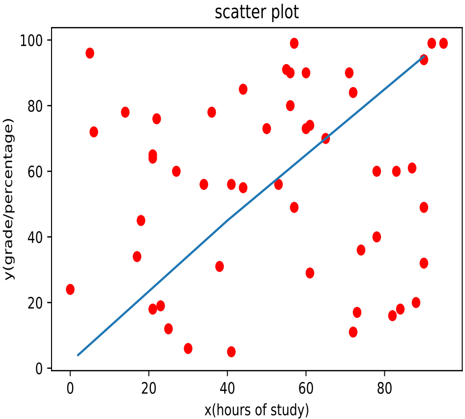
February 3, 2021

Outline

1. Introduction
2. The Objective Function
3. The Solution

An example

Suppose we have been given some data about how many hours of study you do and it determines the grade or percentage score that you will get in the final exam. Denote the number of hours studied by x and grade or percentage obtained is y.



The model

Our model here is a line which we use to predict y given x. Recall the equation of a line from high school geometry class.

$$y = mx + c \tag{1}$$

Here x is called the predictor variable and y the response variable. Given $\{x, y\}$ pairs, our objective is to find the line which represents the data in the best possible way.
The Question: what is the best possible way?

The model

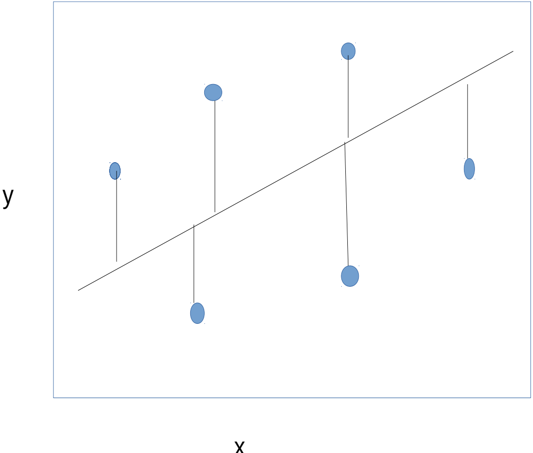
A line which goes through most of the points will be our best line.
Next quesstion:
How to find such a line?
finding a line is to estimate the coefficients m and c in (1).
How to do that?
Recall when we need to find two unknowns, we need two equations at least.

$$\begin{aligned} y_1 &= mx_1 + c \\ y_2 &= mx_2 + c \end{aligned}$$

Two equations means we need two data points (x_1, y_1) and (x_2, y_2). This is called simple linear regression in one variable. A variable is also called feature.

The model

In reality there are more data points and a line should ideally go through all of them. But, this is impossible when all the data points don't lie on the same line. See fig. 6. So, what to do in this case?



The Objective function

Let's say our prediction is $\hat{y}$. So want to minimize the distance $(\hat{y} - y)$. We have total n data points $\{(x_i, y_i)\}_{i=1}^{n}$ from which we want to estimate our parameters (m, c) . Hence, we want to minimize the average squared error or mean squared error (MSE).

$$
\begin{aligned}
J(m,c) &= \frac{1}{2n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \\
&= \frac{1}{2n}\sum_{i=1}^{n}((mx_i + c) - y_i)^2
\end{aligned}
$$

The Solution: Calculus way

Two solution techniques:

- Using calculus
- Using optimization

Using calculus, we can differentiate the objective function wrt m and c. So, we get

$$\begin{aligned} m &= \frac{n\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ &= \frac{Cov(X,Y)}{Var(X)} \\ c &= \frac{\sum_{i=1}^{n} y_i - m\sum_{i=1}^{n} x_i}{n} \end{aligned}$$

The Solution: Optimization way

We can use gradient descent to optimize the objective function.

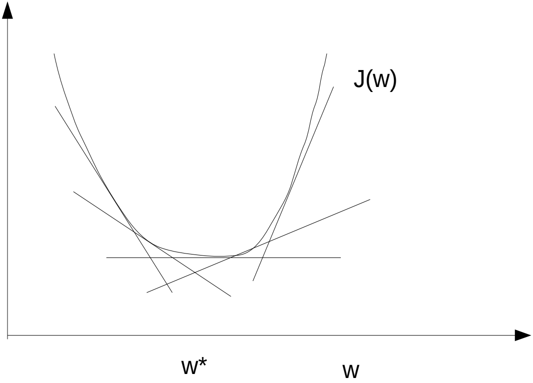


Figure: Gradient descent intuition

The Solution: Optimization way

We can use gradient descent to optimize the objective function.

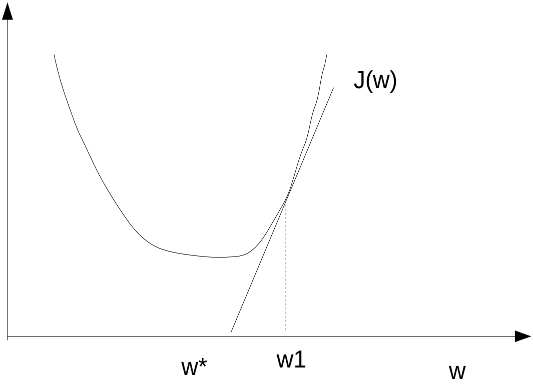


Figure: Gradient descent intuition

The Solution: Optimization way

linear approx (infact linear plus a quadratic term equal to $\frac{L}{2}(w - w_1)^2)J(w)$ around w_1 can be written as:

$$J(w) = J(w_1) + h\nabla J(w_1) \tag{2}$$

To minimize (2), we can set $h = -\alpha\nabla J(w)$. To see this, diff (2) wrt h. Thus update equation for w is given by:

$$w_{t+1} := w_t - \alpha\nabla J(w) \tag{3}$$

$t \in \{1, \ldots, T\}$

The Solution: Gradient Descent

Large learning rate α can diverge and overshoot the minimum.

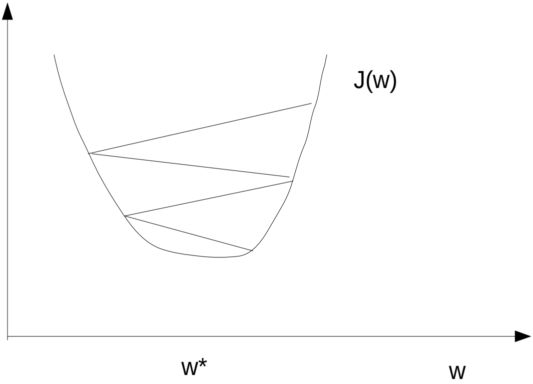


Figure: Gradient descent intuition

The Solution: Gradient Descent

Small learning rate α can be too slow to converge.

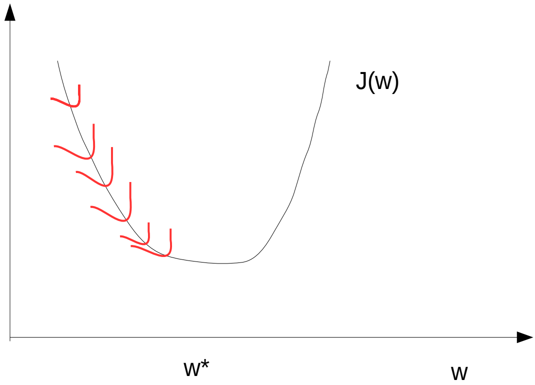


Figure: Gradient descent intuition

Bibliography I