Machine Learning

Linear Regression & gradient descent

Jesús Medina

Índice general

Capítulo 1

Linear Regression

Linear regression is a statistical and mathematical technique used to make predictions from a linear model. The goal is to find a function that describes, in the simplest possible way, the relationship between a dependent variable and one or more independent variables, based on a set of observed data.

Since observations are usually discrete and limited in practice, it is common not to have exact records for every possible value of the independent variables. In such cases, linear regression allows us to construct an approximate function whose trajectory closely resembles, as much as possible, the behavior of the data.

When working with a single independent variable, the resulting model is a line that tries to fit the observed points. This case is known as **simple linear regression**, and it is expressed with an equation of the form:

$$\hat{y} = \theta_0 + \theta_1 x$$

where θ_1 represents the slope of the line, θ_0 is the y-intercept, and $\hat{y}$ is the estimated value of the dependent variable. The model seeks to have this line pass as close as possible to all data points, minimizing some measure of error, such as the mean squared error.

If multiple independent variables are available, the model extends to what is known as **multiple linear regression**. In this case, a plane or hyperplane (depending on the number of variables) is fitted to represent the linear relationship between the dependent variable and the independent ones. The general model takes the form:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Simple Linear Regression

To use a simple linear regression model, we require a set X of n observations, where each data point consists of an independent variable $x^{(i)}$ and a dependent variable $y^{(i)}$, with $i = 1, \ldots, n$. The goal is to find a linear function that relates both variables so that the values of x allow us to estimate the corresponding values of y.

Ejemplo 1.1.1

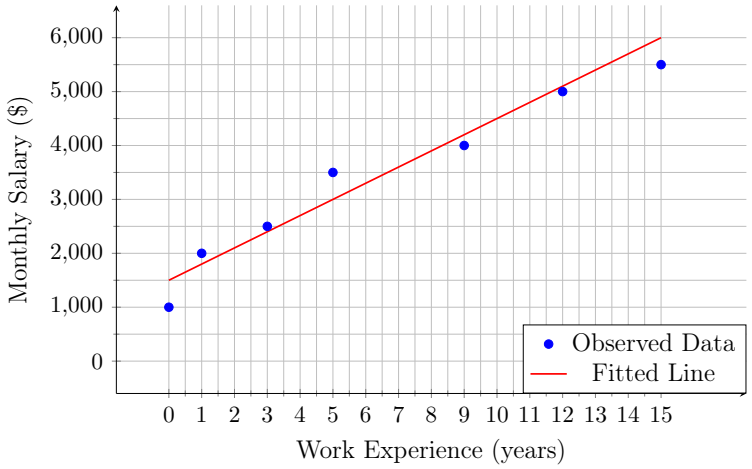
We have the salaries of 7 employees along with their years of work experience. With this data, we want to find a model that represents the dependency of salary on work experience. The observation table is as follows:

Monthly Salary ($)	1000	2000	2500	3500	4000	5000	5500
Work Experience (years)	0	1	3	5	9	12	15

In this case, $n = 7$, the independent variable $x^{(i)}$ corresponds to work experience in years, and the dependent variable $y^{(i)}$ to monthly salary in pesos.

For example, the fourth employee has $x^{(4)} = 5$ years of experience and a salary of $y^{(4)} = 3500$ pesos per month.

Graphically, the data would appear as follows in the plane, along with a sample line modeling its behavior.



Deriving a Formula

To find the line that best fits the data, we assume that there exists a linear function $h : X \to \mathbb{R}$ that models it. This function is known as the model's **hypothesis**.

Definición 1.1.1: Hypothesis

The hypothesis is a linear function that takes as input a value $x^{(i)}$ and returns a prediction for the dependent variable $y^{(i)}$, given by:

$$h_\theta(x^{(i)}) = \theta_0 + \theta_1 x^{(i)}$$

where the parameter vector is $\theta = \langle \theta_0, \theta_1 \rangle$.

By comparing the hypothesis output with the actual observed value $y^{(i)}$, we find a discrepancy. This difference between the prediction and the actual value is expressed as:

$$h_\theta(x^{(i)}) - y^{(i)}.$$

However, when we talk about *distance* or error, we are only interested in the magnitude of this difference, not its sign. To avoid positive and negative errors canceling each other out, we square the difference:

$$\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \geq 0.$$

Observación 1.1.1

Alternatively, we could use the absolute value function to measure the magnitude of the error. However, the absolute value is not differentiable at zero, which complicates the application of optimization techniques based on differential calculus. In contrast, the quadratic function is smoothly differentiable throughout its domain, making it a more convenient option for the model's purposes.

If we generate different hypotheses $h_\theta(x^{(i)})$ and measure the associated error for each, our goal is to find the one that minimizes the total error. However, each individual error measurement corresponds to a single point $\langle x^{(i)}, y^{(i)} \rangle$, and what we truly seek is a hypothesis that fits well to all points in the dataset, that is, for all $i = 1, \ldots, n$.

To achieve this, we sum the squared errors associated with each observation. This sum represents the overall error of the model. If we also divide by $2n$, we obtain a convenient expression that serves as an objective function to minimize. This is known as the **cost function**.

Definición 1.1.2: Cost Function

The cost function measures the average squared error of the hypothesis with respect to the observed data and is defined as:

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

The cost function is a strictly convex quadratic function with respect to the parameters θ_0 and θ_1. This means it has a single global minimum, which is reached at the point where its derivative (or gradient) is zero.

To find this minimum, we need to solve the system:

$$\nabla_\theta J(\theta) = 0$$

which is equivalent to solving the following equations:

$$\frac{\partial J}{\partial \theta_0} = 0 \quad \text{and} \quad \frac{\partial J}{\partial \theta_1} = 0$$

To find the minimum of the cost function, we compute its partial derivatives with respect to the parameters θ_0 and θ_1.

Observación 1.1.2 T

o simplify notation, we define the mean of the data $a^{(i)}$, with $i = 1, 2, \ldots, n$, as:

$$\bar{a} = \frac{1}{n} \sum_{i=1}^{n} a^{(i)}$$

Partial derivative with respect to θ_0

$$\begin{aligned}
\frac{\partial J}{\partial \theta_0} &= \frac{\partial}{\partial \theta_0} \left[\frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \right] \\
&= \frac{1}{2n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_0} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\
&= \frac{1}{2n} \sum_{i=1}^{n} 2 \left(h_\theta(x^{(i)}) - y^{(i)} \right) \cdot \frac{\partial}{\partial \theta_0} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \cdot \frac{\partial}{\partial \theta_0} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)} \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right)
\end{aligned}$$

Which when developed yields:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x^{(i)})-y^{(i)}\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(\theta_0+\theta_1 x^{(i)}-y^{(i)}\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\theta_0+\sum_{i=1}^{n}\theta_1 x^{(i)}-\sum_{i=1}^{n}y^{(i)}\right)\\
&= \frac{1}{n}\left(n\theta_0+\theta_1\sum_{i=1}^{n}\ x^{(i)}-\sum_{i=1}^{n}y^{(i)}\right)\\
&= \frac{n\theta_0}{n}+\theta_1\left(\frac{1}{n}\sum_{i=1}^{n}x^{(i)}\right)-\frac{1}{n}\sum_{i=1}^{n}y^{(i)}\\
&= \theta_0+\theta_1\bar{x}-\bar{y}
\end{aligned}$$

Formula to obtain θ_0

Since we try to cancel the partial derivative with respect to θ_0 we have:

$$\theta_0+\theta_1\bar{x}-\bar{y}=0$$

Solving for θ_0:

$$\theta_0=\bar{y}-\theta_1\bar{x} \tag{1.1}$$

Partial derivative with respect to θ_1

$$\begin{aligned}
\frac{\partial J}{\partial\theta_1} &= \frac{\partial}{\partial\theta_1}\left[\frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta(x^{(i)})-y^{(i)}\right)^2\right]\\
&= \frac{1}{2n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_1}\left(h_\theta(x^{(i)})-y^{(i)}\right)^2\\
&= \frac{1}{2n}\sum_{i=1}^{n}2\left(h_\theta(x^{(i)})-y^{(i)}\right)\cdot\frac{\partial}{\partial\theta_1}\left(h_\theta(x^{(i)})-y^{(i)}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x^{(i)})-y^{(i)}\right)\cdot\frac{\partial}{\partial\theta_1}\left(\theta_0+\theta_1 x^{(i)}-y^{(i)}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x^{(i)})-y^{(i)}\right)x^{(i)}
\end{aligned}$$

If we perform the same procedure as with θ_0 we then obtain:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x^{(i)})-y^{(i)}\right)x^{(i)} &= \frac{1}{n}\sum_{i=1}^{n}\left(\theta_0+\theta_1 x^{(i)}-y^{(i)}\right)x^{(i)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\bar{y}-\theta_1\bar{x}+\theta_1 x^{(i)}-y^{(i)}\right)x^{(i)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)}\bar{y}-\theta_1 x^{(i)}\bar{x}+\theta_1 x^{(i)}x^{(i)}-x^{(i)}y^{(i)}\right) \\
&= \frac{1}{n}\left(\bar{y}\sum_{i=1}^{n}x^{(i)}-\theta_1\bar{x}\sum_{i=1}^{n}x^{(i)}+\theta_1\sum_{i=1}^{n}x^{(i)}x^{(i)}-\sum_{i=1}^{n}x^{(i)}y^{(i)}\right) \\
&= \bar{y}\bar{x}-\theta_1\bar{x}^2+\theta_1\left(\frac{1}{n}\sum_{i=1}^{n}(x^{(i)})^2\right)-\left(\frac{1}{n}\sum_{i=1}^{n}x^{(i)}y^{(i)}\right) \\
&= \bar{y}\bar{x}-\left(\frac{1}{n}\sum_{i=1}^{n}x^{(i)}y^{(i)}\right)+\theta_1\left(\frac{1}{n}\sum_{i=1}^{n}(x^{(i)})^2-\bar{x}^2\right) \\
&= 0
\end{aligned}$$

Formula to obtain θ_1

Solving for θ_1:

$$\theta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}x^{(i)}y^{(i)}-\bar{y}\bar{x}}{\frac{1}{n}\sum_{i=1}^{n}(x^{(i)})^2-\bar{x}^2}$$

With this, we have obtained 2 equations called the **normal equations** which are used to obtain the linear regression parameters θ_0 and θ_1.

Multiple Linear Regression

When working with multiple independent variables, the scalar notation used in simple linear regression becomes cumbersome and inefficient. In these cases, it is preferable to use a **matrix representation**, which allows us to manipulate all data and operations in a compact and efficient manner.

Definición 1.2.1: Matrix hypothesis

The model hypothesis for the i-th observation is expressed as:

$$\hat{y}^{(i)} = h_\theta(x^{(i)}) = \theta_0 + \theta_1 x_1^{(i)} + \cdots + \theta_d x_d^{(i)}$$

where $\hat{y}^{(i)} \in \mathbb{R}$ is a scalar.
Meanwhile, the complete **prediction vector** is written in matrix form as:

$$\hat{y} = \begin{pmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(n)} \end{pmatrix} = X\theta$$

where:

- n is the number of observations (rows of the dataset),
- d is the number of independent variables (also called features),
- $x^{(i)} \in \mathbb{R}^{1\times(d+1)}$ is the row vector corresponding to the i-th observation (including a leading 1 for the intercept),
- $X \in \mathbb{R}^{n\times(d+1)}$ is the **design matrix**, where each row represents an observation:

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{pmatrix}$$

- $\theta \in \mathbb{R}^{(d+1)\times 1}$ is the **parameter vector**:

$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{pmatrix}$$

From scalar to matrix form of the cost function

In scalar notation, the cost function is defined as:

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

If we define the prediction vector as $\hat{y} = X\theta$, and the vector of actual outputs as $y \in \mathbb{R}^n$, then the **error vector** is:

$$\hat{y} - y = X\theta - y$$

This is a column vector of length n. When we multiply this vector by its transpose, we obtain the sum of squared errors:

$$(\hat{y} - y)^T(\hat{y} - y) = \sum_{i=1}^{n} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

Therefore, the cost function in matrix form is expressed as:

$$J(\theta) = \frac{1}{2n}(\hat{y} - y)^T(\hat{y} - y) = \frac{1}{2n}(X\theta - y)^T(X\theta - y)$$