① {Simple linear regression}

It is a foundational algorithm, and very easy to understand.

It comes under the supervised machine learning algorithms.

In supervised machine learning algorithms, we further have 2 types.

- Regression Algorithms
- Classification Algorithms

② {Types of Linear Regression}

- Linear Regression
  - Simple Linear Regression
  - Multiple LR
  - Polynomial LR
  - Regularization

③ {Simple linear regression}

In simple LR, we have a single input feature and a single output feature. For example, CGPA vs package data.

CGPA	Package
7.1	3.5
4.7	1.2
8.9	4.2
8.1	3.9
...	...

Suppose we have the data of 5000 students. So we have (x_i, y_i) $i \in \{1, \ldots, 5000\}$ where

$x_i \triangleq$ CGPA of i^{th} student

$y_i \triangleq$ package of i^{th} student

We need to find a line $y = mx + b$, the (sum of) distance of which from each point in our dataset is minimum.

Let (x_i, y_i) be a point of our dataset. Then the error made by the line is

$$E(x_i, y_i) = mx_i + b - y_i$$

We then define the error function $E : \mathbb{R}^2 \to \mathbb{R}$ as

$$E(m, b) \triangleq \frac{\sum_{i=1}^{n} (mx_i + b - y_i)^2}{n}$$

Our task is to find m, b such that $E(m, b)$ is minimum.

(4) {How to find m and b}

In order to find the values of m and b, we have 2 approaches.

- → closed form solution
- → Non closed form solution

But what is a closed form solution?

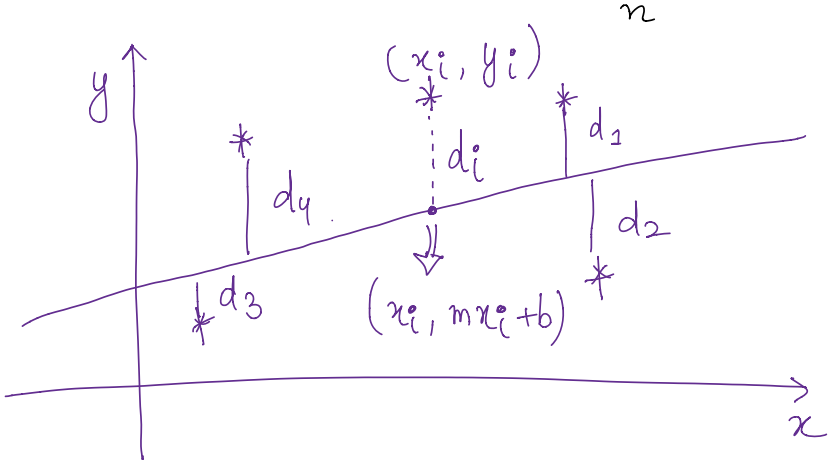
In mathematics, an expression is in closed form if it is formed with constants, variables and a finite set of basic functions connected by arithmetic operators $(+, -, *, /)$ and function composition

The ordinary least square gives us the closed form solution of a LR problem. However, in higher dimensions the OLS method is quite expansive according to time complexity. In that case, we use the non-closed solution like the method of gradient descent. This method is an approximation method which is used to find the minima of the loss function in the higher dimensions. (n sufficiently large)

- Closed form Solution
  - Ordinary least square [OLS]
- Non Closed form Solution
  - gradient descent [GD]

⑤ { Deriving the equation }

$$E(m,b) \triangleq \frac{\sum_{i=1}^{n} (y_i - mx_i - b)^2}{n}$$



$$E \triangleq d_1^2 + \dots + d_n^2 = \sum_{j \in \{1,\dots,n\}} d_j^2$$

We define $\hat{y}_j \triangleq mx_j + b$ for the pt (x_j, y_j)

then $d_j \triangleq \hat{y}_j - y_j = mx_j + b - y_j$

Thus,

$$E(m,b) \triangleq \sum_{j=1}^{n} (mx_j + b - y_j)^2$$

It is easy to see that E is a function from $\mathbb{R}^2$ to $\mathbb{R}$. We need to find (m, b) $\ni$ $E(m, b)$ attains minimum value.

For finding minima, $\left(\frac{\partial E}{\partial m}\right) = 0 = \left(\frac{\partial E}{\partial b}\right)$

Now, $$\frac{\partial E}{\partial b} = \frac{\partial}{\partial b} \sum_{j \in \{1, \ldots, n\}} (y_j - mx_j - b)^2$$

$$= 2 \sum_{j \in \{1, \ldots n\}} (y_j - mx_j - b)(0 - 0 - 1)$$

$$= 2 \sum_{j \in \{1, \ldots, n\}} (mx_j + b - y_j)$$

Equating this to 0, gives,

$$\sum_{j \in \{1, \ldots, n\}} (mx_j + b - y_j) = 0$$

$$\Rightarrow \quad m \sum_{j=1}^{n} x_j + nb - \sum_{j=1}^{n} y_j = 0$$

$$\Rightarrow \quad \frac{m \sum x_j}{n} + b - \frac{\sum y_j}{n} = 0$$

$$\Rightarrow \quad m\left(\frac{\Sigma x_j}{n}\right) + b - \left(\frac{\Sigma y_j}{n}\right) = 0$$

$$\Rightarrow \quad m\bar{x} + b - \bar{y} = 0$$

$$\Rightarrow \quad \boxed{b = \bar{y} - m\bar{x}} \qquad -(1)$$

Now, $E(m, b) = \sum (y_j - mx_j - b)^2$

$$= \sum (y_j - mx_j - \bar{y} + m\bar{x})^2$$

$$\frac{\partial E}{\partial m} = 2 \sum (y_j - mx_j - \bar{y} + m\bar{x})(0 - x_j + 0 + \bar{x})$$

$$= -2 \sum (y_j - mx_j - \bar{y} + m\bar{x})(x_j - \bar{x})$$

Equating this to 0, we get

$$\sum (y_j - mx_j - \bar{y} + m\bar{x})(x_j - \bar{x}) = 0$$

$$\Rightarrow \quad \sum (y_j - \bar{y} - m(x_j - \bar{x}))(x_j - \bar{x}) = 0$$

$$\Rightarrow \quad \sum (y_j - \bar{y})(x_j - \bar{x}) - m \sum (x_j - \bar{x})^2 = 0$$

$\Rightarrow$

$$m = \frac{\sum (x_j - \bar{x})(y_j - \bar{y})}{\sum (x_j - \bar{x})^2}$$

⑥ { Regression Metrics }

We have various kinds of metrics

ⓐ MAE (mean absolute error) → same unit, → Robust to outliers

$$MAE \triangleq \frac{\sum_{j \in \{1,\ldots,n\}} |d_j|}{n}$$

In other words,

$$MAE \triangleq \left(\frac{\sum_{j \in \{1,\ldots,n\}} |y_i - \hat{y}_i|}{n} \right)$$

→ Not differentiable ☹

ⓑ MSE (mean squared error) → differentiable

$$MSE \triangleq \frac{\sum d_j^2}{n}$$

In other words,

$$MSE \triangleq \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

Not have same unit as the original data

heavily penalizes outliers

Not robust to outliers

(c) RMSE $\left(\sqrt{MSE}\right)$ → same unit as data

$$RMSE \triangleq \sqrt{\frac{\sum (\hat{y}_i - y_i)^2}{n}}$$

Robust to outliers

These metrics are not absolute and depend on context. We have other metrics which are context independent like R2 score and adjusted R2 score.

⑦ { R2 score }

Consider the dataset $D \triangleq \{(x_j, y_j),\ j \in \{1, \ldots, 100\}\}$ where x_j denotes the CGPA of j^{th} student and y_j denotes the package of j^{th} student, where $j \in \{1 \ldots 100\}$.

In worst case scenario, if someone asks us to predict the package of a student, we would have used the mean package.

$$\bar{y} = \frac{\sum y_j}{n} = \text{average (mean) package.}$$

This is the worst case scenario. However, when we create our model M, the line which we get might perform better or worse than this line { the mean line }

R2 score measures the error made by our line relative to the line of mean.

$$\text{R2 - score} \triangleq 1 - \frac{\sum_{j \in \{1, \ldots, n\}} (\hat{y_j} - y_j)^2}{\sum_{j \in \{1, \ldots, n\}} (\bar{y} - y_j)^2}$$

If the error made by our line is the same as error made by the mean line, then $\sum_{j \in \{1,\dots,n\}} (\hat{y}_j - y_j)^2 = \sum_{j \in \{1,\dots,n\}} (\bar{y} - y_j)^2$ and

thus R2-score = 0.

This means that our input column was useless. and our model is worst.

If R2-score is 1, this means that our model did not make any error. Thus, it is the perfect model.

The R2-score is always less than 1. The closer we are to 1, the better it is. If the R2-score is less than 0, this means that our model is perfoming even worse than the mean line. [Doob कर मर जाओ]