→ "Linear Regression" (Actual Regression Technique) ①

It is an actual regression technique, where given a dataset.

$D = \langle x_i, y_i \rangle_{i=1}^{n}$ $\begin{Bmatrix} x_i \in \mathbb{R}^n \\ y_i \in \mathbb{R} \end{Bmatrix}$

Let's say we have a dataset 'D' comprising of 'n points'.

where $\{x_i \in \mathbb{R}^n\}$ → n dimensional real space
$\{ \& \ y_i \in \mathbb{R} \}$ → Real nos.

if $y_i \in \mathbb{R}$, it is a regression problem

& if $y_i \in \{-1, +1\}$, it is a classification problem.

Let's now understand the geometric interpretation of "Linear Regression"

Let's take a simple example to explain Geometry behind "Linear Regression"

Given some features, we want to predict height of a person.

features could be {weight, gender, ethnicity, hair color}

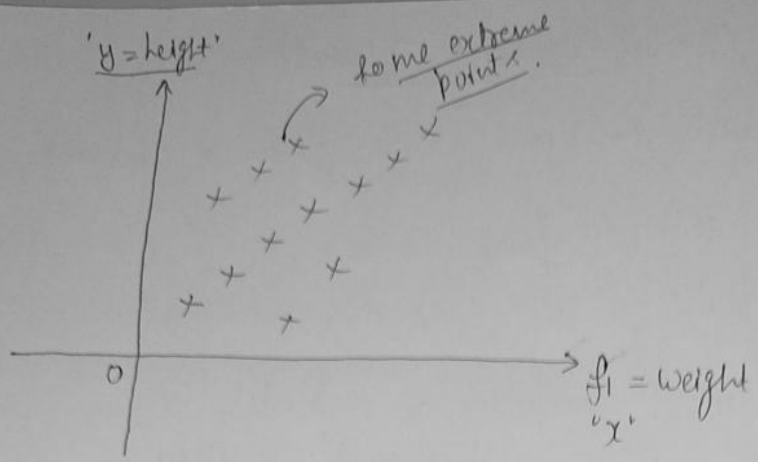
Suppose we have bunch of features like these & we want to predict height.

Note :> 'Height' is a real-valued number.

Let's pick weight as a feature & let's try to predict height. We know typically as weight increases, height also increases, exceptions are there also.

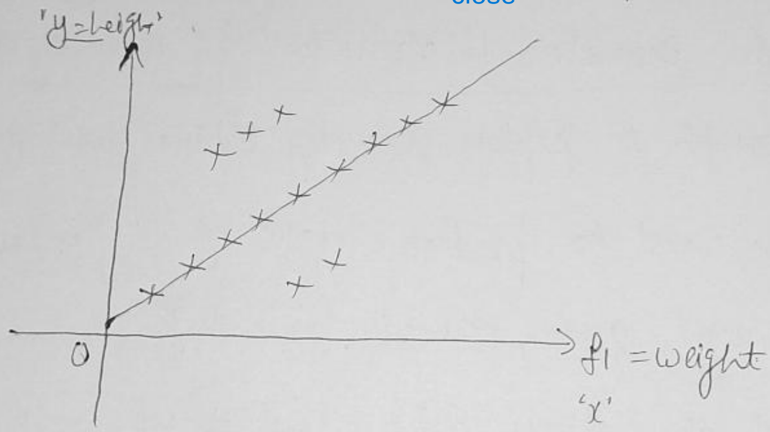
Let's say we have data as shown below:-

P.T.O



There will be some extreme points also, but in general one thing that we notice here is as weight increases height also increases.

In linear regression, we are effectively trying to find out a line that fits these points as ~~well~~ close as possible.



We want to find a line, that fits or that passes through the given data as ~~well~~ close as possible.

Here remember, we donot have any class labels, we are having a response label "y" which is basically equivalent to a class-label.

In '2D', if we are given just 'weight' & we want to predict 'height', we could say ~~that~~ ok, height of a person is some w_1 times weight plus some constant w_0

height = (w1 * weight) + w0 ⟶ linear form

So, our objective here in linear regression is to find 'w0' & 'w1'.

& w0 & w1 will represent our line that fits the points.

This is like.

$$y = mx + c$$

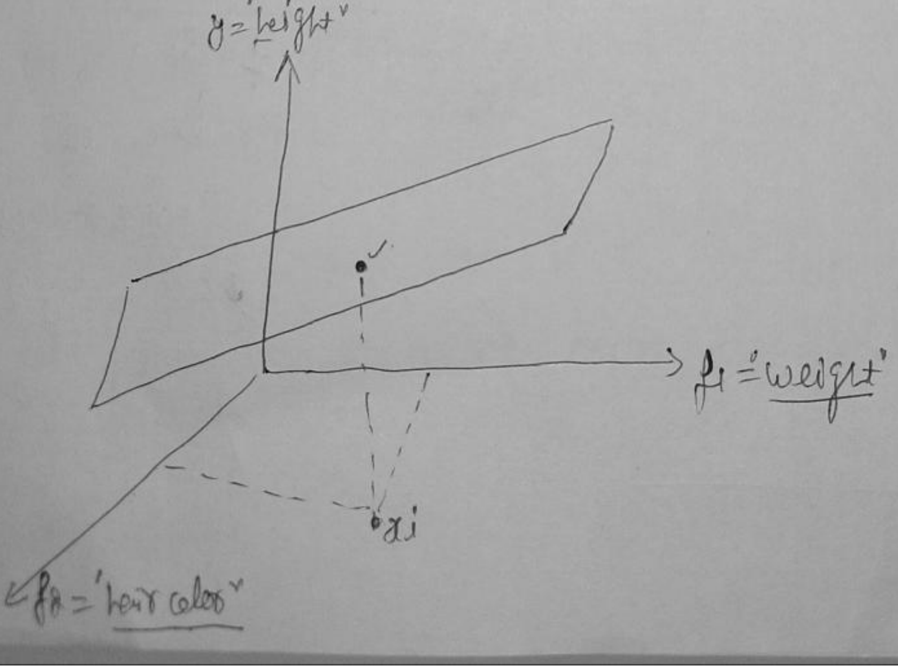
'height' ← y; m → 'slope'; x → 'weight'; c → 'y-intercept'

This is what we are trying to predict.

In 2d space, this is exactly like the equation of a line.

Our objective in 'Linear Regression' is to find the line that fits these points as ~~well~~ close as possible.

Let's now understand what happens in some higher dimensional space. Let's consider '3d space', let's assume we have two features 'weight' & 'hair color'. Now if we want to predict the height, we have to find a plane kind of structure.



Each point in the 3d space is represented using 2 features

$x_i = \langle x_{i1}, x_{i2} \rangle$

x_{i1} → 'weight'; x_{i2} → 'Hair color'

So, in a 3d case, given two features & a height to be predicted we have a plane to predict it.

So, in the case the way we predict it.

$$height = w_1 * f_1 + w_2 * f_2 + w_0$$

$$\boxed{y_i = w_1 * x_{i1} + w_2 * x_{i2} + w_0}$$

& Remember this is nothing but the equation of the plane.

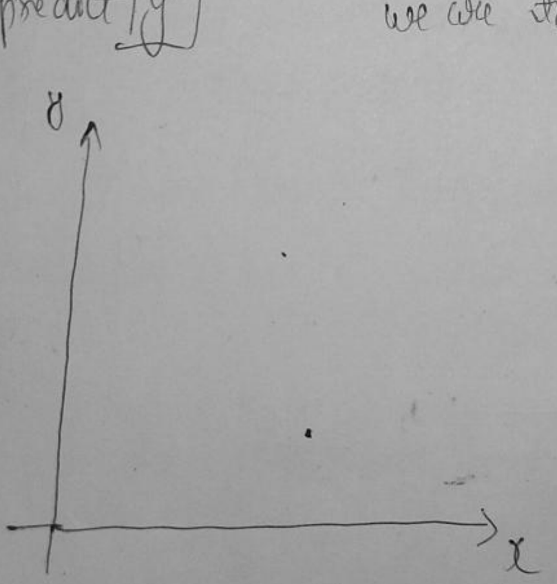
$\boxed{y_i = w^T x_i + w_0}$ → linear plane

We are trying to find a hyperplane.

So, our objective is to find a line or a plane or a hyperplane that best fits the datapoints.

Now let's see what does best fits means, it is very important to understand.

Let's assume we have two features 'x' & 'y', we are given $\boxed{\text{'x'}}$ & we need to predict $\boxed{\text{'y'}}$ we are trying to say $\boxed{y = f(x)}$



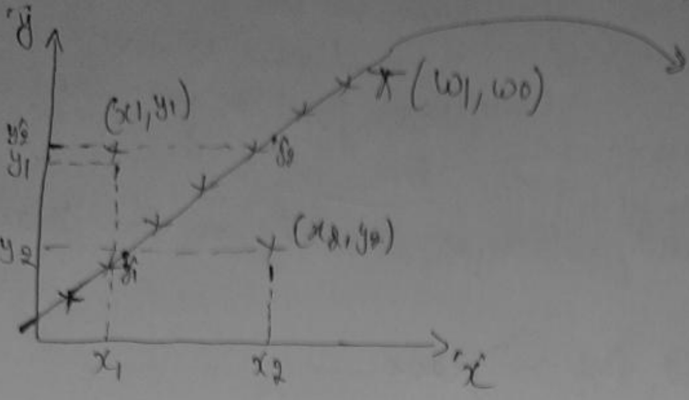
& $f(x)$ will be in the form of

$[w_1 x + w_0]$

$f(x) = w_1 x + w_0$

becoz I want to fit a line, it is a linear regression task.

Given this let's assume we have a bunch of points here. ⑤



This is the line that best fits the points, but remember, the extreme points are not lying on the line

Let take the first extreme point $x(x_1, y_1)$.

Now suppose, if given this I tried everything I came up with the line as shown above that best fits the points.

Now given a point like $x(x_1, y_1)$ which is not exactly on the line

Now given this point $x(x_1, y_1)$ let's try to predict the value using the function we formed

$f(x_1) = \hat{y_1}$ → This is the value we got.

& for this point $\hat{y_1} \neq y_1$

$\hat{y_1}$ is not exactly equal to y_1

Similarly let's take another extreme point $x(x_2, y_2)$

& find $\hat{y_2} = f(x_2)$

$[\hat{y_2} \neq y_2]$

So, if we decide that this is the line or plane that best fits the data then there is some error associated with these extreme points, which are not exactly on the plane.

& error there is. error $= y_1 - \hat{y_1}$ for point x_1

$$\begin{bmatrix} \text{error for } x_1 \text{ is } & y_1 - \hat{y_1} \\ \text{\& error for } x_2 \text{ is } & y_2 - \hat{y_2} \end{bmatrix}$$

becoz we are predicting $\hat{y_2}$ but actual value is something different becoz point is not lying on the line.

But if we take a point, which is exactly on the plane.

$x(x_3, y_3)$ ⟶ point lies on plane

then for this point $[\hat{y_3} == y_3]$

Which means error is zero.

So, we understood what is an error given this plane or line & we want to best fit the plane.. Best fit intuitively means, minimize the errors for each point..

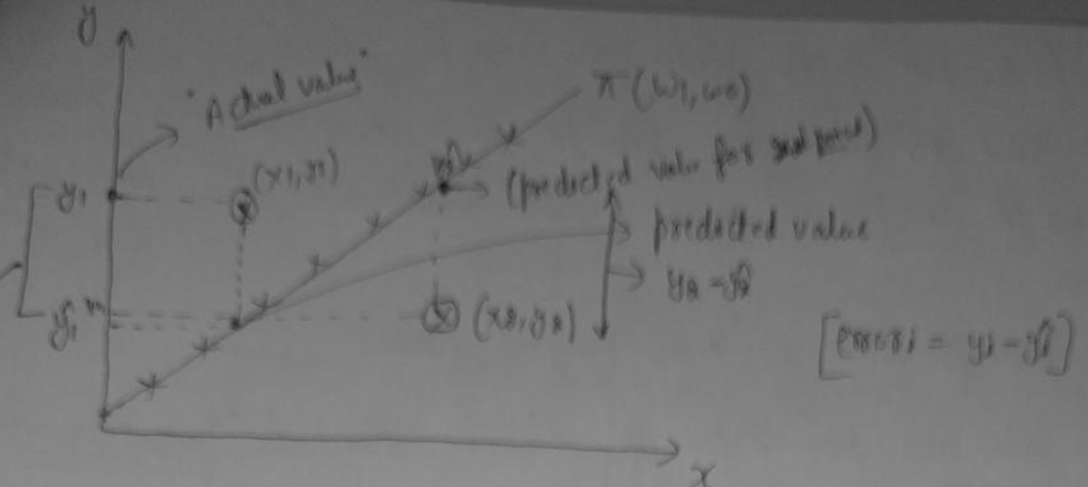
So, for each point there is an error, & the error is basically what is the model saying & what is the true value for the training data. So, we want to minimize the sum of errors across our training data.

Now let's see how to convert this english statement into "mathematical objective function" (optimization problem).

Now let's see what is the mathematical formulation of what we saw.

We said if this is "x" & this is "y"

P.T.O



& if we have a bunch of points, & this is the line or the plane that we have decided that best fits the points & there is a point above & below this decided line. Then we said given these type of extreme points, the error for the point 'i' is nothing but $y_i - \hat{y}_i$

$$error_i = y_i - \hat{y}_i$$

So if I subtract y_1 from $\hat{y}_1$ we get a positive value.

if we do $y_1 - \hat{y}_1$ we get a positive value.

Let's see what happens for the 2nd point. Since it is lying below the line.

& when we do $y_2 - \hat{y}_2$ we get a negative value.

for first point $x(x_1, y_1)$. we getting a positive value as an error.

& for the 2nd point $x(x_2, y_2)$ we got a negative value for the error.

$$\begin{bmatrix} error_1 = y_1 - \hat{y}_1 = +ive \\ error_2 = y_2 - \hat{y}_2 = -ive \end{bmatrix}$$

P.T.O

& for the rest of the points the error is <u>zero</u>

we need to find the best fit line or a best-fit plane.
A best fit line / plane means, which minimizes the sum of errors.

best fit-line = min. sum of errors

but there are two types of errors there are ["positive errors"] & there are ["negative errors"] (so take the sq. of errors)

So, the mathematical formulation for linear regression is very simple. we want to find a "w" & "w_0" (optimal w & w_0) such that minimizes the errors.

We know equation of a plane that doesnot passes through origin is

$$\pi : \boxed{w^T x + w_0 = 0}$$

Vector

scalar

it is y-intercept

$$(w^*, w_0^*) = \underset{w, w_0}{\operatorname{arg\,min}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\hat{y_i} = f(x_i) = w^T x_i + x_0$$

This is the problem that we are trying to solve.

= ~~arg min~~ trying to minimize both vector w & scalar "w_0" Summation $i = 1$ to n $y_i - (w^T x_i + x_0)$!

9

$$(w^*, w_0^*) = \underset{w, w_0}{\arg\min} \sum_{i=1}^{n} \underbrace{\{y_i - (w^T x_i + x_0)\}^2}$$

→ This is an optimization problem.

"Linear regression" is often known as "Ordinary least squares" or "Linear least squares"

This last term is called as "square loss" in optimization.

→ This is the optimization problem of "linear regression"