# *Linear Regression*

Dr. L Jeganathan

School of Computing Sciences and Engineering

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

# Outline

# Description of the Problem

## Experience ( Data Set)

The input data is of the form

$$X = \{x^t, r^t\}_{t=1}^{N}, \quad \textit{where} \quad x = (x_1, x_2, ...x_d) \in R^d.$$

$x$ is a d-dimensional vector. $r \in R$.

- $x$ is the input variable. $x_i$'s are the input attributes.
- $r$ is the output variable.
- $x^t = ({x_1}^t, {x_2}^t, ...x_d^t)$ is the $t^{th}$-item in the input data
- $r^t$ is the $t^{th}$ output in the data.

# A Typical data set

| S.no. | $x_1$ | $x_2$ | $x_3$ | . | . | r |
|---|---|---|---|---|---|---|
| 1 | ${x_1}^1$ | ${x_2}^1$ | ${x_3}^1$ | . | . | $r^1$ |
| 2 | ${x_1}^2$ | ${x_2}^2$ | ${x_3}^2$ | . | . | $r^2$ |
| 3 | ${x_1}^3$ | ${x_2}^3$ | ${x_3}^3$ | . | . | $r^3$ |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| N | ${x_1}^N$ | ${x_2}^N$ | ${x_3}^N$ | . | . | $r^N$ |

## Task

- To find a relationship that involves $x_1, x_2, ...x_d$ and $r$.
- To predict $r$, for a given input $x$ using the arrived relationship.

## Method of accomplishing the task

- Hypothesis : Propose a relationship (an equation) involving $x_1, x_2, ...x_d$ and with weights $w_0, w_1, ...w_d$
- Sample hypothesis
  - $g(x) = g[(x_1, x_2, ..., x_d)] = w_1x_1 + w_2x_2 + w_3x_3 + ... + w_dx_d + w_0$
  - $g(x) = w_1x_1^2 + w_2x_2^2 + w_3x_3^2 + ... + w_dx_d^2 + w_0$
  - and so on
- One can choose the hypothesis in many ways.

## Learning is

To compute the values of the weights $w_0, w_1, ..., w_d$ in the hypothesis

$$g(x) = w_1 x_1 + w_2 x_2 + w_3 x_3 + ... + w_d x_d + w_0$$

by training (using) the data set $X$.

Note that:

- $g(x^t)$ is the value predicted by the hypothesis for the input $x^t$.
- For an input $x^t$, there will be an output $r^t$ (called as the actual output), in the given data set $X$.
- For an input, the difference between the actual output and the predicted output, $g(x^t) - r^t$, is the error due to the input $x^t$.

Performance of a learning is calculated in terms of the error generated by the 'learning'

## Performance Measure of the Learning

Average of the square of the errors (difference between the actual output and the predicted output using the hypothesis) made in each of the instance of E.

$$P = \frac{\sum_{t=1}^{N} [g(x^t) - r^t]^2}{N}$$

$N$ is the total number of instances in $X$.

# How to learn?

## Ideal strategy

To learn the weights $w_0, w_1, w_2, ..., w_d$ such that the performance measure is minimum.

## Learning Model involves optimization

To calculate the weights $w_0, w_1, w_2, ..., w_n$ based on the data set $X$, such that

$$\frac{\sum_{t=1}^{N} [r^2 - g(x^t)]^2}{N}$$

is minimum.

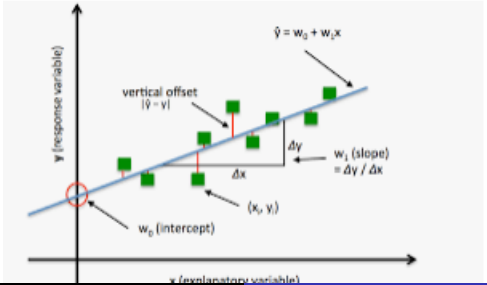
# Error function, $E(g|X)$

Here, we measure the performance in terms of the error involved, Error of the $g$ (hypothesis), given $X$, $E(g|X) = \frac{\sum_{t=1}^{N} [r^t - g(x^t)^2]}{N}$

- Here $r^t \in R$; $g(x^t) \in R$..
- The square of the difference $(r^t - g(x^t))^2$, can be viewed as the distance between the points $r^t$ and $g(x^t)$.
- We can also write the error function in terms of 'absolute value of the difference'. $E(g|X) = \frac{|r^t - g(x^t)|}{N}$

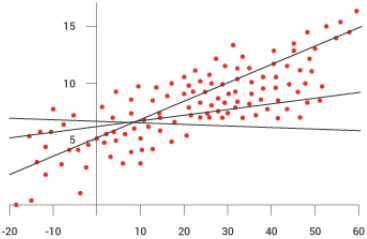
# Problem Description in a geometric sense

- Consider the data set, with d=1
- Data points can be plotted as points in x-r plane
- Hypothesis : $g(x) = w_0 + w_1 x$
- $w_0$ is the intercept. $w_1$ is the slope.

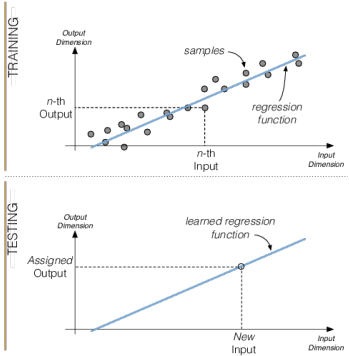


# Which line is best fit?

- Every line is a hypothesis.
- Can draw many lines that go through the data points.
- Which line is a ‘best fit’?- Line whose error is minimum.



# Predicting the output for a test data



# Types of Regression

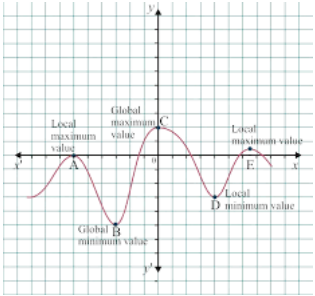
- Simple linear Regression
  - $g(x) = w_0 + w_1 x_1 + w_2 x_2 + ... + w_d x_d$, $x = (x_1), d = 1$.
  - i.e., $w_0 + w_1 x$
- Multiple linear Regression
  - $g(x) = w_0 + w_1 x_1 + w_2 x_2 + ... + w_d x_d$, $x = (x_1, x_2, ..., x_d), d \geq 2$
  - $g(x) = w_0 + w_1 (x_1)^2 + w_2 (x_2)^2 + ... + w_d (x_d)^2$, $x = (x_1, x_2, ..., x_d), d \geq 2$. This is linear in the w's
- Simple non-linear Regression:
  - $g(x) = w_0 e^{w_1 x}$, $x = (x_1), d = 1$
- Multiple non-linear Regression
  - $g(x) = \frac{w_1 x}{w_2 + x} + w_0, x = (x_1, x_2, ..., x_d), d \geq 2$

# Problem

- To minimise $E(g|X)$ and compute the weights $w_0, w_1, w_2, ..., w_d$
- $E$ is a function of d unknowns- parameters, written as $E(w_0, w_1, ...w_d|X)$.
- $(w_0, w_1, ..., w_d)$ is a point in the d-dimensional space.
- To compute the coordinates of a point such that the value of E at that point is minimum.

Given a function, say $f(x)$, to find the value of $x$ such that $f(x)$ is minimum at that point $x$.

# Mathematics of Minimum

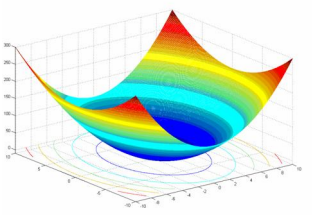



$f(x)$ attains a minimum at a point $x = a$ if $f'(a) = 0$ *and* $f''(a) > 0$.
$f(x, y)$ attains a local minimum at a point $x = a, y = b$, if , at the point $(a, b)$,

- $f_x = 0, f_y = 0$ at the point $(a, b)$ and
- $f_{xx}f_{yy} - f_{xy}^2 > 0$ at (a,b) & $f_{xx} > 0$ and $f_{yy} > 0$ at (a,b)

# Some insights

- The function which has to be minimised for our learning is the sum of the squares.
- Like $\sum_{i=1}^{N}(x_i)^2$
- If N=1, graph is a parabola



- If N=2, graph will look like:

## We infer

- Minimum point is the point where the function takes the minimum value.
- 'Sum square function' has only one minimum point. That point is local minimum as well as the global minimum.
- Slope of the tangent (the first derivative) at that minimum point is zero.
- We solve the equations obtained by equating the first derivatives to zero and obtain the stationary points (Points where the derivative becomes zero).

# Back to our learning process

- Our error function $E(g|X) = \frac{\sum_{t=1}^{N}(r^t - g(x^t))^2}{N}$, is the sum of the squares.
- There will be only one minimum point where the $E(w_0, w_1, w_2, ..., w_d|X)$ takes the global minimum.
- First derivative becomes zero at the minimum point.
- Hence, to find that minimum point (i.e., point $(w_0, w_1, w_2, ..., w_d)$), we have to identify the point where the derivative becomes zero.

# Pseudocode

0: Input : $X = \{x^t, r^t\}_{t=1}^N$, $x = (x_1, x_2, ..., x_d)$, $g(x) = w_0 + w_1 x_1 + ... + w_d x_d$
1: Compute $E = \frac{\sum_{t=1}^N [r^t - g(x^t)^2]}{N}$ { E is an equation in $w_0, w_1, ..., w_d$ }
2: **for** $i = 0$ to $d$ **do**
3: Compute $\frac{\partial E}{\partial w_i}$
4: Set $\frac{\partial E}{\partial w_i} = 0$ {We will get (d+1)-equations in (d+1)-unknowns}
5: **end for**
6: Compute $w_0, w_1, ..., w_d$ by solving the (d+1)- equations

# To compute $\frac{\partial E}{\partial w_i}$, $i = 0, 1$

Let $g(w_0, w_1|X) = w_1 x + w_0$, $x = (x_1, x_2, ..., x_d)$.

$$E(w_0, w_1|X) = E = \frac{\sum_{t=1}^{N} [r^t - w_1 x^t - w_0]^2}{N}. \quad (1)$$

## To compute $\frac{\partial E}{\partial w_0}$

$$\frac{\partial E}{\partial w_0} = \frac{\sum_{t=1}^{N} 2(r^t - w_1 x^t - w_0)(-1)}{N}$$

$\frac{\partial E}{\partial w_0} = \frac{\sum_{t=1}^{N} 2(r^t - w_1 x^t - w_0)(-1)}{N} = 0$

On simplification,

$\sum_{t=1}^{N} w_0 = \sum_{t=1}^{N} r^t - w_1 \sum_{t=1}^{N} x^t$

$$w_0 = \frac{\sum_{t=1}^{N} r^t}{N} - w_1 \frac{\sum_{t=1}^{N} x^t}{N} \tag{2}$$

Differentiating E w.r.t $w_1$ and equating to zero,

$$\frac{\partial E}{\partial w_1} = \frac{\sum_{t=1}^{N} 2(r^t - w_1 x^t - w_0)(-x^t)}{N} = 0$$

$$\sum_{t=1}^{N} x^t r^t - w_1 \sum_{t=1}^{N} (x^t)^2 - w_0 \sum_{t=1}^{N} x^t = 0$$

Substituting the value of $w_0$ from Equation (2), we have

$$\sum_{t=1}^{N} x^t r^t - w_1 \sum_{t=1}^{N} (x^t)^2 - (\frac{\sum_{t=1}^{N} r^t}{N} - w_1 \frac{\sum_{t=1}^{N} x^t}{N}) \sum_{t=1}^{N} x^t = 0$$

On simplification we get,

$$w_1 = \frac{(\sum_{t=1}^{N} x^t r^t) - \frac{(\sum_{t=1}^{N} x^t)(\sum_{t=1}^{N} r^t)}{N}}{\sum_{t=1}^{N} (x^t)^2 - \frac{(\sum_{t=1}^{N} x^t)^2}{N}}$$

Thus, we have computed the points $(w_0, w_1)$ such that $E$ is minimum:

$$w_0 = \frac{\sum_{t=1}^{N} r^t}{N} - w_1 \frac{\sum_{t=1}^{N} x^t}{N} \tag{3}$$

$$w_1 = \frac{(\sum_{t=1}^{N} x^t r^t) - \frac{(\sum_{t=1}^{N} x^t)(\sum_{t=1}^{N} r^t)}{N}}{\sum_{t=1}^{N} (x^t)^2 - \frac{(\sum_{t=1}^{N} x^t)^2}{N}} \tag{4}$$

All the terms in the R.H.S are known.

# Illustration with a sample data set

We illustrate the learning by Regression with a sample dataset. $X = \{x^t, r^t\}_{t=1}N$. Here, N=5,$x = (x_1)$. Input variable $x$ contains only one attribute.

| S.No | x | r |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 2 | 1 |
| 3 | 3 | 2 |
| 4 | 4 | 2 |
| 5 | 5 | 4 |

Calculation of $w_0$ and $w_1$, involves $(x^t)^2$, $(r^t)^2$ and $(x^t)^2(r^t)^2$, we make a table with those entries.

VIT
Vellore Institute of Technology

| S.No | x | r | $x^2$ | $r^2$ | $xr$ |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 2 | 1 | 4 | 1 | 2 |
| 3 | 3 | 2 | 9 | 4 | 6 |
| 4 | 4 | 2 | 16 | 4 | 8 |
| 5 | 5 | 4 | 25 | 16 | 20 |
| Sum | 15 | 10 | 55 | 26 | 37 |

$$w_1 = \frac{(\sum_{t=1}^{5} x^t r^t) - \frac{(\sum_{t=1}^{5} x^t)(\sum_{t=1}^{5} r^t)}{5}}{\sum_{t=1}^{5} (x^t)^2 - \frac{(\sum_{t=1}^{5} x^t)^2}{5}} = \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = 0.7$$

$$w_0 = \frac{\sum_{t=1}^{5} r^t}{5} - w_1 \frac{\sum_{t=1}^{5} x^t}{5} = \frac{10}{5} - (0.7)(\frac{15}{5}) = -0.1$$

Thus we have learnt $g(x) = (-0.1) + (0.7)x$.

# To predict the output for a given x

We have $g(x) = (-0.1) + (0.7)x$. Given $x = 4.2$, what is the value of $r$? $g(4.25) = (-0.1) + (0.7)(4.2) = 2.84$.

## Error computed with the training data

| S.No | x | r | $g(x)$ | $(g(x) - r)^2$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 0.6 | 0.16 |
| 2 | 2 | 1 | 1.3 | 0.09 |
| 3 | 3 | 2 | 2 | 0 |
| 4 | 4 | 2 | 2.7 | 0.49 |
| 5 | 5 | 4 | 3.4 | 0.36 |
| Sum | | | | 0.22 |

Empirical Error= $\frac{0.22}{5} = 0.044$

# Procedure for observing the Performance (Error) incurred in 'Learning'

Performance of a model is calculated using the data points, not used in training the model.

## Performance of the model

- Divide the data set $X$ in two parts: Training data set ($X'$), Validation data set $X''$.
- Usually, 75% of the total data points of X form the 'Training data set' $X'$.
- 25% form the 'Validation Data set'.
- Learn the hypothesis $g(X)$ by training the data set $X'$ such that $E(g|X')$ is minimum.
- Calculate Performance of $g(x)$ over $X''$.

Test data is the one which you have not seen earlier.

$Performance = \frac{E[g|X''|}{|X''|}$

where $|X''|$ is the total number of data points in the validation data $X''$

- The training set need not be chosen sequentially from X.
- If there are N data points in X, You can choose $K$ data points ($K < N$) in $\binom{N}{K} = \frac{N!}{K!(N-K)!}$ ways.
- For every chosen $X'$, one can compute the performance measure based on the respective $X''$.
- For each pair $(X_1', X_1'')$, let performance measure be $p_1$.

Average of all such $p_i$'s, is the performance measure of the learning model $g(x)$.

| Training | Validation | Performance |
|---|---|---|
| $X_1'$ | $X_1$'' | $p_1$ |
| $X_2'$ | $X_2$'' | $p_2$ |
| . | . | . |
| . | . | . |
| $X_k'$ | $X_k$'' | $p_k$ |

Performance Measure of $g(x) = \frac{\sum_{i=1}^{k} p_i}{k}$

# To choose the model that fits data set best

- For a data set, we can propose infinite models (hypothesis)
- For every hypothesis $g_i(x)$, we can compute the performance measure $P_i$
- Best model that best suits the given data set $X$ is the model which has the least $P_i$.

The pseudocode discussed so far, to learn the parameters , will work for any type of regression based learning models like:

- Simple/Multiple Linear Regression
- Simple/Multiple Non-linear Regression

Any Regression based Learning Model involves

- Model Proposal (Hypothesis) with parameters (weights)
- Compute the Error function - Objective function, Cost function
- Optimize the Error function and learn the appropriate values for the parameters.

# Exercise

| S.No | x | r |
|---|---|---|
| 1 | 2 | 1 |
| 2 | 2.1 | 1 |
| 3 | 1.8 | 2 |
| 4 | 4 | 2 |
| 5 | 3 | 4 |

- Propose a simple linear Regression Learning model and learn the parameters of the model by training the above data set.