Week 1 : Session 1

1. Linear Regression

What is Regression ?

Consider the given data.

Year	Sales (Million Euro)	Advertising (Million Euro)
1	651	23
2	762	26
3	856	30
4	1063	34
5	1190	43
6	1298	48
7	1421	52
8	1440	57
9	1518	58

What to predict is "RESPONSE VARIABLE"

Prediction based on "REGRESSOR / EXPLANATORY VARIABLE"

Time k

Training Data

How to predict the **Sales**, as a function of Advertising ? (ii) How to predict the sales given the amount spent on advertising ?

REGRESSION is an Machine Learning (ML) technique, which precisely addresses this problem.

REGRESSION is an Algorithm to predict a RESPONSE variable based on a set of REGRESSOR or EXPLANATORY variable.

In this example, we are trying to predict the Sales using simply one regressor (ie) Advertising, but in general, there can be multiple regressors. For instance, consider the below example.

Year	Sales	Advertising		
		TV (x_1)	Radio (x_2)	Newspaper (x_3)
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

RESPONSE variable

Multiple REGRESSORS/ EXPLANATORY VARIABLES

There can be multiple Regressors, and the Regression vector $\bar{x}$ for the above example is given by

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \text{ where } x_1 \rightarrow \text{Cost of TV advertising}$$
$$x_2 \rightarrow \text{Cost of Radio advertising}$$
$$x_3 \rightarrow \text{Cost of Newspaper advertising}$$

In general, the Regressor $\bar{x}$ can be an n-dimensional vector. Thus,

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Vector of Regressors

These multiple regressors can be used to predict the RESPONSE variable.

Other Examples

① Let $y(k)$ = Price of Facebook stock at time k
$x_1(k)$ = Price of Microsoft stock at time k
$x_2(k)$ = Price of Google stock at time k
$x_3(k)$ = Price of Amazon stock at time k.

Problem (i) Predict the Price of Facebook stock at time k using the prices of related stocks (Microsoft, Google, Amazon) at the same time k.

Problem (ii) Predict the Prize of Facebook stock at time k using the past value of Prizes of the same stock (i) $y(k-1), y(k-2), \ldots$

This is called "AUTO REGRESSION".

(2) Let $y(k)$ = Sales of (SUV (Creta)) at time k ← RESPONSE variable

REGRESSORS:
- $x_1(k)$ = Sale of Bikes
- $x_2(k)$ = Sale of Cars
- $x_3(k)$ = Average income of people

Problem: Predict the sales of SUV (Creta) at time k using the related sales and unrelated data.

What is Linear Regression ?

The output $y(k)$ can be predicted, using a linear combination of REGRESSORS (or) EXPLANATORY variables $x_i(k)$.

$$y(k) = h_0 + h_1 x_1(k) + h_2 x_2(k) + \ldots + h_n x_n(k) + \epsilon(k)$$

(h_0: Bias / constant offset; $y(k)$: Response variable; $\epsilon(k)$: Prediction Error)

$$= \underbrace{\begin{bmatrix} 1 & x_1(k) & x_2(k) & \cdots & x_n(k) \end{bmatrix}}_{\bar{x}^T(k)} \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_n \end{bmatrix} + \epsilon(k)$$

($\bar{h}$: Regression coefficient vector; $\epsilon(k)$: Noise / Model Error)

Response at time k

$$\boxed{y(k) = \bar{x}^T(k)\,\bar{h} + \epsilon(k)}$$

This is the Regression Model for particular time k.

where,

- $\bar{x}^T(k) = [1 \; x_1(k) \ldots x_n(k)]$ is an $n+1$ dimensional vector. Though the number of Regressors is n, we have $n+1$ dimensional vector $\bar{x}^T(k)$, coz the first element 1 corresponds to the Bias, which gets multiplied by h_0. (Bias)
- $\bar{h} = \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_n \end{bmatrix}$ is an $n+1$ dimensional vector of Regression coefficients.
- $\epsilon(k)$ is the Model error.

This is termed Linear Regression and $h_0, h_1, h_2, \ldots, h_n$ are termed as Regression coefficients.

Essentially, in Machine Learning, we want to have a machine to learn the functional relationship via Learning these Regression coefficients.

By Learning the Regression coefficients, we are essentially able to learn the relationship between the EXPLANATORY Variables and the RESPONSE. So, this is essentially is the Machine Learning aspect.

This rule helps us predict the RESPONSE as a function of these REGRESSORS.

How do we learn the Regression coefficients ?

When we try to learn something, we learn by looking at other similar examples around that. For instance, We've seen the Weather/Rainfall as a function

of the Weather for several years. Using those previous instances, we've trained our mind to predict how the rainfall situation going to be for the next day.

So, essentially, one has to use the training examples (i.e) past occurrences of the Explanatory variables and Responses, to learn / infer the relationship between the Explanatory variables and Responses.

Thus, the Regression coefficients can be computed as follows.

Consider the availability of Training Pairs

$[y(k), \bar{x}(k)]$, for $k = 1, 2, \ldots, M$.

↳ No. of Training pairs.

$y(1)$ $\bar{x}(1)$
$y(2)$ $\bar{x}(2)$
$\vdots$ $\vdots$

Typically, Very large Training Set. Eg. Stock Market

Using this training data, how are we going to build our Linear Regression Model?

Note:
Of course, there are other Machine Learning algorithms, which do not have the Luxury of Training data. Those are known as Unsupervised Learning.

Week 1 : Session 2

We are trying to learn the Regression coefficient. For that we need the training data / training pairs. corresponding to each regression vector $\bar{x}(k)$ and its response $y(k)$. observed. So, we can write

$$y(k) = \bar{x}^T(k)\,\bar{h} + \epsilon(k)$$

$\epsilon(k)$ → error

$\bar{h}$ → Regression coefficients

As there are M such training pairs, we can write

$$y(1) = \bar{x}^T(1)\,\bar{h} + \epsilon(1)$$
$$y(2) = \bar{x}^T(2)\,\bar{h} + \epsilon(2)$$
$$\vdots$$
$$y(M) = \bar{x}^T(M)\,\bar{h} + \epsilon(M)$$

Each $\bar{x}^T(i)$ is a $1 \times (n+1)$ vector. And there are M rows of $\bar{x}^T(i)$. Hence the size $M \times (n+1)$

In vector form,

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(M) \end{bmatrix}_{M\times 1} = \begin{bmatrix} \bar{x}^T(1) \\ \bar{x}^T(2) \\ \vdots \\ \bar{x}^T(M) \end{bmatrix}_{M\times(n+1)} \bar{h} + \begin{bmatrix} \epsilon(1) \\ \epsilon(2) \\ \vdots \\ \epsilon(M) \end{bmatrix}_{M\times 1}$$

$\bar{y}$ = $X\,\bar{h}$ + $\bar{\epsilon}$

Therefore, the training data model is $\boxed{\bar{y} = X\bar{h} + \bar{\epsilon}}$

Typically $M \gg (n+1)$, where M → No. of Training Sets
n → No. of regressors

(ii) In Matrix X, # rows >> # columns.
Therefore X is a Tall Matrix, so inverse does not exists.

Now, how to determine the Regression coefficient vector $\bar{h}$? Typically, if X is a square matrix, we can determine $\bar{h}$ by simply multiplying by its inverse (X^{-1}). But here, X^{-1} does not exist, because X is a Tall matrix. So, we formulate this problem as follows.

We have; $\bar{y} = X\bar{h} + \bar{\epsilon}$

$$\Rightarrow \bar{\epsilon} = \bar{y} - X\bar{h}$$

In order to minimize the error, we have to minimize the Norm square of the error vector.

(a) Norm square of a vector is nothing but the square of the euclidean length.

$$\text{Minimize Error} = \min \|\bar{\epsilon}\|^2$$

$$= \min \epsilon^2(1) + \epsilon^2(2) + \cdots + \epsilon^2(M)$$

$$= \min \|\bar{y} - X\bar{h}\|^2$$

This is essentially, we are trying to determine $\bar{h}$ such that it gives us the best approximation to the observed responses $\bar{y}$.

So, to determine Regression coefficients $\bar{h}$, solve the problem

$$\min \underbrace{\|\bar{y} - X\bar{h}\|^2}_{\text{Minimize square of error}}$$

$\bar{\epsilon}$

Therefore, this problem is termed as Least Squares (LS) problem.

So, essentially, determining the Regression coefficient vector $\bar{h}$ from the training data $\bar{y}$, boils down to the LS problem where we are minimizing $\|\bar{y} - X\bar{h}\|^2$. This is known as the LS problem, where we want

$\bar{h}$ to be found as a vector, which minimizes the Least Squares Error $\bar{E}$.

And, the solution to the LS problem is given by the Regression coefficients $\bar{h}$ as

$$\bar{h} = (X^T X)^{-1} X^T \bar{y}$$

← Formula for Regression coefficients

This is how we determine the Regression coefficients $\bar{h}$ from the Matrix of Regressors X and the vector of responses $\bar{y}$.

Note: The Matrix $(X^T X)^{-1} X^T$ is termed as the pseudo-inverse of X, since $(X^T X)^{-1} X^T X = I$.

Recall, here X is a Tall Matrix (Not Square Matrix), and # Rows >> # columns. Hence X does not have inverse. Although X is not invertible, $(X^T X)^{-1} X^T$ acts like an inverse of X, since $[(X^T X)^{-1} X^T] X = I$.

($(X^T X)^{-1} X^T$ is termed as Pseudo-inverse of X.) ($[(X^T X)^{-1} X^T] = X^{-1}$)

Example:

Solve the LS problem.

$$\min \left\| \begin{bmatrix} -1 \\ 2 \\ -3 \\ 1 \end{bmatrix}_{4\times 1} - \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}_{4\times 2} \begin{bmatrix} h_0 \\ h_1 \end{bmatrix} \right\|^2$$

$\bar{y}$: M = 4 Training samples

X : # Rows > # columns, 4 > 2, Tall Matrix

$\bar{h}$: 2 Regression coefficients

We have, $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$, $\bar{y} = \begin{bmatrix} -1 \\ 2 \\ -3 \\ 1 \end{bmatrix}$, $\bar{h} = \begin{bmatrix} h_0 \\ h_1 \end{bmatrix}$

$$(X^T X) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

$$= \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}$$

$$A^{-1} = \frac{1}{|A|} \text{Adj}(A)$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1}_{2\times 2} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}_{2\times 2}$$

$$\bar{h} = (X^T X)^{-1} X^T \bar{y}$$

$$= \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \\ -3 \\ 1 \end{bmatrix}$$

$$\boxed{\bar{h} = \begin{bmatrix} -1/2 \\ 1/10 \end{bmatrix}}$$ This is the Regression coefficient vector.

(ii) For a given set of training data, and given a set of inputs and outputs, this is how we learn the Regression coefficients.