complete.

2 - Regression.

Q.1. What is regression?

→ ① Regression is a technique used to model and analyse the relationships between variables and how they contribute and are related to producing a particular outcome together.

② Regression analysis is a conceptually a simple method for investigating functional relationships among variables.

③ Regression predict a real and continuous value y for a given set of input x ~~(x=x1,2,~~ $(x = x_1, x_2, \dots)$

④ Regression is a supervised learning technique.

Q.2. Types of regression -

i] linear regression -

- Relation between independent & dependent variables is linear and usually expressed by straight line equation.

- $y = h(x) = w_0 + w_1 x.$
- $y = h(x) = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_n x_n$

① simple linear regression

- There exists only one dependent variable and related to only one independent variable.

- Ex - $y = h(x)$

  $= w_0 + w_1 x$

- We select simple linear regression when there is single input variable and single output variable with linear relationship.

② multiple linear regression -

The regression that has one output variable and more than one input/ independent variables with linear relationship between input & output is called multiple linear regression.

Ex - $y = h(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$.

2] Non-linear regression -

- Non-linear regression has a non linear relationship betⁿ independent variables and dependent variables.
- Number of independent variables can be one or more than one.
- Non-linear regression is expressed by a polynomial eqⁿ, hence also called as ~~non-linear~~ polynomial regression.
- Ex - $y = h(x)$

$$h(x) = w_0 + w_1 x_1^2$$

$$h(x) = w_0 + w_1 x_1 + w_2 x_2^2$$

$$h(x) = w_0 + w_1 \ln(x)$$

$$h(x) = w_0 + w_1 \sin(x)$$

$$h(x) = w_0 + w_1 e^x.$$

- We select non-linear regression when we have single/ multiple input variable and one output variable with non-linear relationship.

Q.3. Assumption for linear regression.

→ 1. Variables used, should be measured at the continuous level (variables need to be continuous variable).

2. There needs to be linear relationship between independent and dependent variables.

3. little or no multi-collinearity.

Multi-collinearity → one independent variable is co-related with other independent variable.

4. There should be not ~~outlinears~~. outliers.

An outlier is an observed data point that has a dependent variable value that is very different to the value predicted by regression eq^n.

Q.4. Hypothesis function for multiple linear regression:

→ Response or target variable $\hat{y}$ is defined as

$\hat{y} = h(x) = w_0 + w_1 x + w_2 x_2 + \dots + w_n x_n$.

where,

$x_1, x_2, x_3 \dots x_n$ are input/independent / ~~pere~~ predictor variables.

$\hat{y}$ is output variable.

$w_0, w_1, w_2, \dots, w_n$ are parameters or coefficients of regression.

Since there is possibility of difference bet^n actual output value & predicted value, we can write actual output as -

$y = \hat{y} + e = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + e$

$e = y - w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n .$

$= y - \hat{y}$

if e is negative, $e = \hat{y} - y$

Q.5. parameter estimation for multiple linear regression.

→ - Gradient descent algorithm is used to estimate parameters in multiple linear regression.

- The cost function is:

$$J(w) = \frac{1}{2n} \sum_{i=1}^{n} (h(x^i) - y^i)^2$$

where x^i = i^{th} input in dataset

y^i = i^{th} output in dataset.

<u>Basic gradient descent algorithm</u>

Repeat until converge

{

$$w_{new} = w_{old} - \eta \frac{\partial J(w)}{\partial w}$$

}

Repeat until converge

{

$$w_0\ new = w_0 - \eta \frac{\partial J(w_0, w_1, \ldots, w_k)}{\partial w_0}$$

$$w_k new = w_1 - \eta \frac{\partial J(w_0, w_1, \ldots, w_k)}{\partial w_k}$$

where $k = 1, 2, \ldots n$

}

$$\frac{\partial J(w_0, w_1, \dots, w_k)}{\partial w_0} = \frac{\partial \frac{1}{2n}\sum_{i=1}^{n}(h(x^i)-y^i)^2}{\partial w_0}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(h(x^i)-y^i)\,\frac{\partial(w_0 + w_1 x_1^i + w_2 x_2^i + \dots + w_n x_n^i)}{\partial w_0}$$

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

$$= \frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x^i - y^i)$$

$$\frac{\partial J(w_0, w_1, \dots + w_k)}{\partial w_1} = \frac{\partial \frac{1}{2n}\sum_{i=1}^{n}(h(x^i)-y^i)^2}{\partial w_1}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(h(x^i)-y^i)\,\frac{\partial(w_0 + w_1 x_1^i + w_2 x_2^i + \dots + w_n x_n^i)}{\partial w_1}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(h(x^i)-y^i)\,x_1^i$$

$$= \frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x^i - y^i)\,x_1^i$$

$$\frac{\partial J(w_0, w_1, \dots w_k)}{\partial w_k} = \frac{\partial \frac{1}{2n}\sum_{i=1}^{n}(h(x^i)-y^i)^2}{\partial w_k}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(h(x^i)-y^i)\,\frac{\partial(w_0 + w_1 x_1^i + w_2 x_2^i + \dots + w_n x_n^i)}{\partial w_k}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(h(x^i)-y^i)\,x_k^i$$

$$= \frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x^i - y^i)\,x_k^i$$

where $k = 1, 2, 3, \dots n$

Repeat until converge

{

$$w_0 \text{ new} = w_0 - \eta\left(\frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x_1^i + w_2 x_2^i + \dots + w_n x_n^i - y^i)\right)$$

$$w_1 \text{ new} = w_1 - \eta\left(\frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x_1^i + w_2 x_2^i + \dots + w_n x_n^i - y^i)\, x_1^i\right)$$

$$\vdots$$

$$w_K \text{ new} = w_K - \eta\left(\frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x_1^i + w_2 x_2^i + \dots + w_n x_n^i - y^i)\, x_K^i\right)$$

}

where $K = 1, 2, 3, \dots n$

Q. 6. Regression parameters

→ Regression model establish relation between response/dependent variable y with the independent/predictor variable x.

We can write the relationship using a hypothesis function as

$$y = h(x) \qquad x \rightarrow \boxed{h} \rightarrow y$$

where h is called hypothesis function

Hypothesis function describes the relationship between x & y variables.

If a relationship is linear then the ~~relationship between~~ regression is called linear regression.

If a relationship is non-linear then the regression is called non-linear regression.

Sometimes $h(x)$ is also written as $f(x)$.

$h(x)$ can be expressed in different way as:

$$h(x) = w_0 + w_1 x \quad - (1)$$

$$h(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots \quad - (2)$$

$$h(x) = w_0 + w_1 x^2 \quad - (3)$$

$$h(x) = w_0 + w_1 x_1 + w_2 x_2^2 \quad - (4)$$

Here, w_1, w_2 are called as coefficients of regression or model parameters.

x, x_1, x_2 are independent / predictor variables.

Q.7. Explain significance of cost vs. parameter curve in linear regression.

→

- Measure of how best the line fits to data or how best the hypothesis function predicts the output is specified by cost function.
- Different values of the weights (w_0, w_1) gives us different lines and our task is to find weights for which we get best fit.

Sometimes $h(x)$ is also written as $f(x)$.

$h(x)$ can be expressed in different way as:

$$h(x) = w_0 + w_1 x \quad - (1)$$

$$h(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots \quad - (2)$$

$$h(x) = w_0 + w_1 x^2 \quad - (3)$$

$$h(x) = w_0 + w_1 x_1 + w_2 x_2^2 \quad - (4)$$

Here, w_1, w_2 are called as coefficients of regression or model parameters.

x, x_1, x_2 are independent / predictor variables.

Q.7. Explain significance of cost vs. parameter curve in linear regression.

→

- Measure of how best the line fits to data or how best the hypothesis function predicts the output is specified by cost function.
- Different values of the weights (w_0, w_1) gives us different lines and our task is to find weights for which we get best fit.
-