Correlation

→ Correlation means its the measure of strength and direction of a relationship between two variables.

→ It is a measure of the extent to which two variables are related.

Three possible results of correlation

1. positive correlation
2. Negative "
3. Zero "

1. Positive correlation — is a relationship betw two variables in which both variables move in the same direction ie, when one variable increases as the other variable increases, or one variable decreases while the other decreases.

eg: hours studied and exam score of a student.

2. Negative correlation — is the relationship betw two variables in which an increase in one variable is associated with a decrease in the other.

eg: Relationship betw speed of a car and time it takes to reach a destination.

In this case, one variable (speed) goes up, other variable (time) goes down.

3. Zero correlation - It exist when there is no relationship betw two variables.

eg:i) Relationship betw a person's shoe size and their exam mark

2) Relationship betw a person's height and their favorite type of music.

So, changes in one variable (height), have no predictable effect on the other (music performance)

Correlation equation

$$\text{Correlation } r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

where

r = correlation coeffient

x_i = individual data points in variable X (i^{th} value of data set)

y_i = individual data points in variable Y (i^{th} value of dataset)

$\bar{x}$ = Mean of X

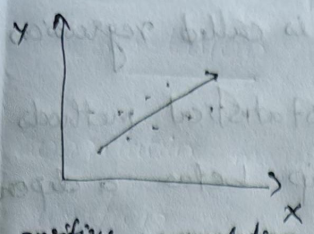
$\bar{y}$ = Mean of Y

$\sum$ = denotes the sum across all data points

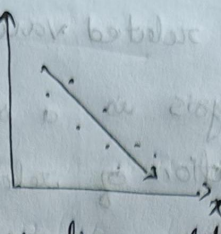
If $r=1$, Perfect positive correlation / association

$r=-1$, Perfect negative correlation / association

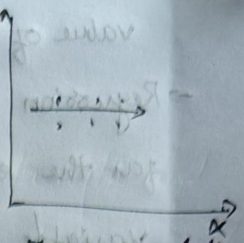
$r=0$, no correlation association



positive correlation association



negative correlation association



zero correlation association

-> The correlation coefficient (r) indicates the extend to which the pairs of numbers for these two variables lie on a straight line.

-> r lie is betw 1 and -1.

-> Values over zero indicates a positive correlation, while values unds zero indicates a negative correlation.

-> when r is,

* +1 : Perfectly positive
* -1 ; Perfectly negative
* 0 - 0.2 ; No or very weak association
* 0.2 - 04 ; weak association
* 04 - 0.6 ; Moderate association
* 0.6 - 0.8 ; Strong association
* 0.8 - 1 ; Very Strong to perfect association

Regression

-> correlation measures the Stength of the relationship betw two variables.

-> In Regression analysis, we can estimate the value of one variable with value of the other variable which is known.

-> The regression is a statistical method used to the unknown value of one variable from

value of the related variable is called regression.

→ Regression analysis is a set of statistical methods used for the estimation of relationship between a dependent variable and one or more independent variables. It can be utilized to asses the strength of the relationship between variables and for modeling the future relationship between them.

→ Independent variable or explanatory variable

The factors which affect the dependent variables or which are used to predict the values of the dependent variable are called independent variable, also called as predictor. (variable on x axis)

→ Dependent variable or response variable

The main factor in regression analysis which we want to predict or understand is called dependent variable, also called target variable. (y)

→ Independent variable cause an effect on the dependent variable.

→ Regression line

The line that best fits the data points on a scatter plot, with the independent variable on x axis and dependent variable on the y axis.

→ Regression equation

The equation that represents the slop of the regression line, the relationship betw the variable

and an estimate of the errors.

Some examples of regression can be:

1) Prediction of rain using temperature and other factors.
2) Determining market trends
3) Prediction of road accidents due to rash driving.

eg: Now, the company want to do the advertisement of $200 in the year 2019 and want to know the prediction about sales for this year.

advertisement	sales
$ 90	$ 1000
$ 120	$ 1300
$ 150	$ 1800
$ 100	$ 1200
$ 130	$ 1380
$ 200	??

-> To solve such type of prediction problems in machine learning, we need regression analysis.

-> Here advertisement is the independent variable and sales is the dependent variable.

Types of Regression

1. Linear Regression
2. Polynomial Regression
3. Logistic Regression
4. Ridge regression
5. Lasso regression
6. Decision tree regression.

1. Linear Regression