UNIT 11 REGRESSION

11.0 INTRODUCTION

In 1908 British biologist Francis Galton investigated the relationships between two variables to study the hereditary growth of children. In his research he categorised parents into two categories on the height: 1st category of the parents belongs to the family length smaller than average length of than parents' length and 2nd category of parents belong to the parents having lengththan the average length. This "regression toward mediocrity" gave these statistical methods thereprimarily the term regression describes the relationship between variables.

Simple regression $y=m*x + C$ describes the relationship between one independent and one dependent variable Where theueuse variable y varies with the value of x and thus a dependent variable,the value of variable x affected any variable hence is a independent variable and m is having some constant value.

Consider the following the parent-children's set

Parent	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5
Children	65.8	66.7	67.2	67.6	68.2	68.9	69.5	69.9	72.2

The mean height of the children is 68.44 whereas the mean height for the parents is 68.5.

The linear equation for the parents and children is

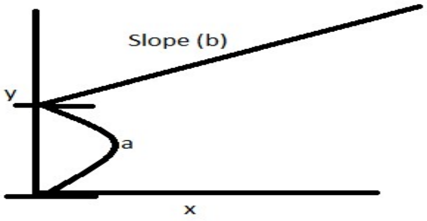
height_child = 21.52 + 0.69 * height_parent

Mathematically simple linear regression can be defined as $y=bx+c+\epsilon$.Where b is the slope of the regression lin , x is the variable which can change the value of y but can't be affected by another variable. Whereas y is a variable which varies with a change in the n value of x. And the known as error value majors between the actual value and predicted value. Variable y is described ass dependent variable or response variable, and variable x is defined as an explanatory or predictor variable.

Regression is a supervised machine learning model which describes the relationships between the response variable and predictor variables.So, regression model is used when it is required to determine the value of one variable using another variable.

If the variable to be predicted is a single variable, then the regression equation will be $y=a+bx$

To determine the value of dependent variable y we need to determine the slope b and constant value and thus by substituting the different set of values for variable x we can get the different value of variabley.



when x=0 then y=a, which means when there is no independent variable then the predicted variable constant value.Suppose we are having multiple independent variables $x_1,x_2,x_3,x_4 \ldots ,x_n$.Then the regressionequation will be $y=a+b_1x_1+b_2x_2+b_3x_3+\ldots\ldots box_.$

The regression line is also calasthma best fit line because the regression line aims to fit all the points or will be minimum. Regression is a linear regression when there is one predictor variable, and we can apply a linear regression model. Themultiple linear regression model came into existence when the number of predictorsvaries are than then one in number. When the relationships between variable y and x are not linear,we can apply-linear regression model.

Following are the ways used by regression analysis to determine the relationships between the response variable and predictor variables:

- **Find the relationship**: It is required to determine the relationships between the predictor variable and response variables. If any change in the independent variable will result inin the age of dependent variable there is an exitstance of relationship.
- **Strength of relationships**: By changing the value of one variable how much another variable changedetermine the strength of relationships.
- **Formation of relationships**: If a change in the value of the dependent variable will result in a change in independent variable, then formulate a mathematical equation to represent the relationships between both variables.
- **Prediction**: After formulation of the mathematical equation find the predicted value.
- **Another independent variable:**Another independent variable which is having impact on dependent variable. If there exist, then formulate the mathematical equation using these variables also.

Uses of Regression

- In a business scenario when it is required to determine the impact of different-different independent variables to find the target value regression can be used.
- When we want to represent in a mathematical expressionform, or we want to model a problem to determine the impact of different variables.

- It is very easy to explain about the business logic with the help of regression. Business logics can be explained very easily to the person.
- When the target variable is normally distributed having some characteristics, regression is very effective.

Examples of Regression

Relationship between uploading a picture on Facebook page and number of likes by the friends.

Relationship between the height of the child and their parents' heights.

Relationship between the average food intake and weight gain.

Relationship between the numbers of hours studied and marks scored by the students.

Relationship between the product consumption by increasing the product price.

Terminologies used in Regression Analysis

- **Dependent Variable:** A variable used to predict the output. It is also called as the target variable.
- **Independent Variable:** The variable which is having an impact on dependent variable is called independent variable. There may be one or more independent variable. This variable is also called as predictor variable. For example, salary of an employee depends on age,qualification,experience. Here salary is a dependent variable and age,qualification and experience are independent variables.
- **Outlier:**Outlier is a value which effect out output, very high value or very low value will affect the result. In case of regression first we have to remove the outlier first.
- **Multicollinearity:**If two values in our dataset are corelated to each other than other variables, such a condition is called multicollinearity. Example: age and date of birth of are correlated to each other. So, we have to avoid one of them.
- **Underfitting and Overfitting:**Overfitting results when our machine learning model work well with the training data set but it does not work well with test data set. Underfitting results when our dataset does not perform well even with our training data set.

11.1 OBJECTIVES

After completing this unit you will be able to:

- Understand the Regression Algorithm
- Understand and apply Linear Regresssion
- Understand and apply Polynomial Regression
- Understand and apply Support Vector Regression

11.2 REGRESSION ALGORITHM

Following are various types of regression algorithms.

Linear regression: Linear regression algorithm comes into existence when there is only one dependent variable and independent variables can be be one or more in numbers. If there is a single independent variable, then it is called as simple linear regression. In linear regression the relationships between the dependent and independent variables are linear i.e. of type y_i=a+b*x_i ; where y_i is a dependent variable and x_i is a independent variable. Variable b is the slope of the line and a is intercept with the axis. Example child height=a+b*(parent height)

Multiple Linear Regression: When there is only one dependent variable and more than one independent variables, then it results in multiple linear regression i.e. $y=a+bx_1+cx_2+dx_{3.}$; example weight = a+b * (daily meal)+ c* (daily exercise)

Logistic regression: In logistic regression algorithm dependent variable is binary in nature (False/True). This algorithm is generally used under cases like testing of the medicines, to detect the bank fraud etc.We had already discussed the concept of logistic regression in unit no. 10 of this course.

Polynomial regression: Polynomial regression is described with the help of polynomial equation where the occurrence of independent variable is more than one. There is no linear relationships between the dependent and independent variables. It results in a curved line instead of a straight line i.e. $y=c+ a*x+b*x^2$

Ecologic regression: Ecological regression algorithm is used when group data belongs to a group. Thus, data is divided into different groups and regression is performed on different groups. Ecologocal regression is mostly used in political research eg.party_votes %=.2+.5*(below_poverty_people_votes)

Ridge regression: It is a type of regularization. When data variables are highly correlated ridge regression is used. Using some constraints on regression coefficients, it is used to reduce the error and lower the bias. Mostly used in feature selection.

Lasso regression: Least absolute shrinkage and selection operator regression algorithm a penalty is assign the coefficients. Lasso regression uses shrinkage technique where data values are shrunk towards a mean.

Logic regression: In logic regression predictor variable and response variable both are binary in nature and applicable to both classification and regression problem.

Bayesian regression: Bayesian regression algorithm is based on Bayesian statistics. Random variables are used as a parameter to estimates. In this algorithm if the data is absent then some prior data is taken as an input.

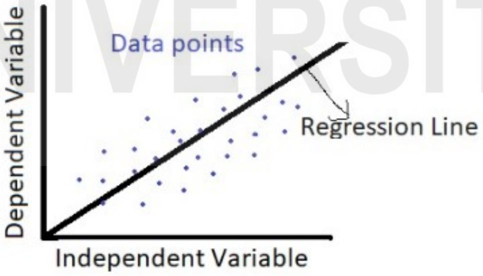
Quantile regression: This is used when the boundary of the quantile is of interest. Whenoverweight and underweight is considered for the health analysis it is consider as an quantile regression.

Cox regression: Cox regression algorithm is used when output of a variable depends on set of independent variables example patient_survival_after_surgery(Survived,Died)=(age,condition,BMI)

11.3 LINEAR REGRESSION

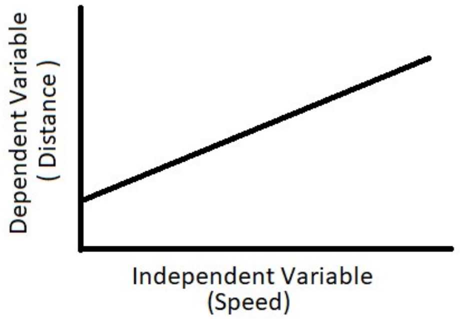
Linear regression is a mathematical method implemented where we want to find the response variable and predictor variables. When the relationships are linear then it is called as linear regression or otherwise it is called as a nonlinear regression. Linear regression makes prediction for continuous/real or numerical variables like age,salary,price etc.

As shown in figure x-axis represent independent variable and y-axis represent dependent variable. A Line with some slope is called linear regression line which shows the relationship between the independent and dependents variable and dots represents the point of the data sets, where some points lie on the line and some other points lie above and below of the line.

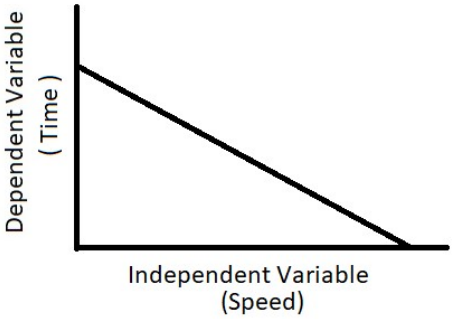


If DV is the dependent variable and IV is the independent variable, then the Positive Linear relationship results with the increases in dependent variable (DV)on the y-axis with respect to increase in value of independent variable (IV) on x-axis. For example, the distance traversed by the car increases when the speed of the car increases. Thus, the distance traversed by the car depends on the speed of the car.

And, the Negative Linear relationships result with the decrease in dependent variable (DV) on the y-axis with respect to the increase in independent variable (IV) on x-axis. For example, time taken by the car decreases with the increase in speed of car.



Positive Linear Relationship



Negative Linear Relationship

Now consider the following data points:

(x_i,y_i)={(45,75),(48,80),(51,100),(37,70)} for i =1,2,3 and 4.

(x_1,y_1)=(20,41)

(x_2,y_2)=(30,83)

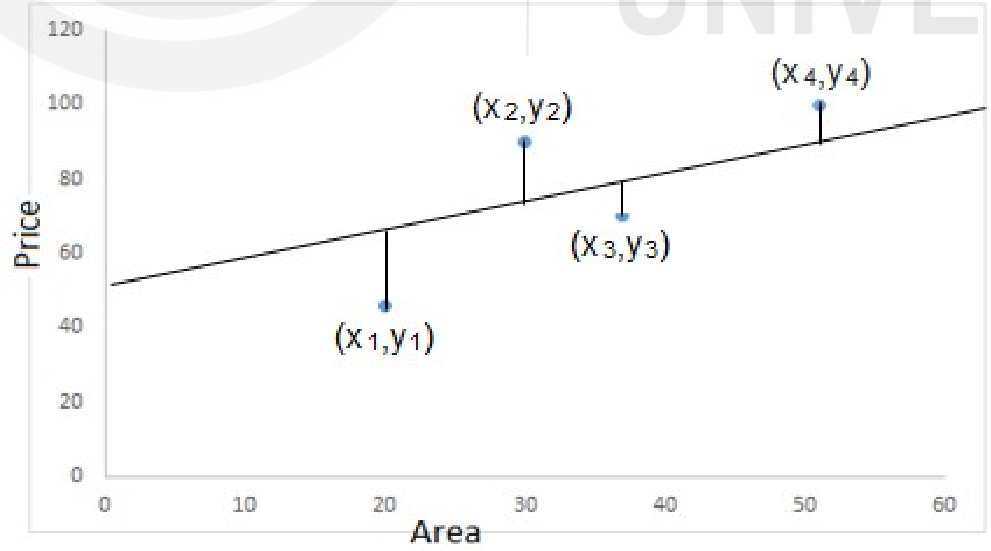
(x_3,y_3)=(38,62)

(x_4,y_4)=(52,100)

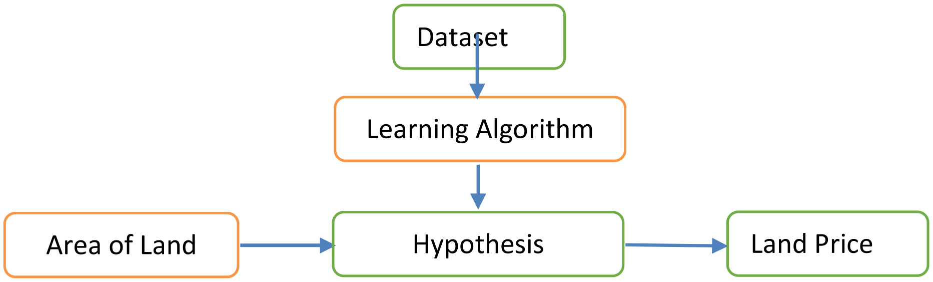
Where x=area given in meter square

And y= price of the land

Following graph draw the linear relationship between x and y



As shown in below diagram set of data is given as a input and learning algorithm will generate a output function conventionally known as a hypothesis (h). The role of the hypothesis function is to estimate the price by taking area as a input to the function. Mapping function h will map from area of land to price of land.



$h(y)=\Theta_0+\Theta_i x$, where h is a hypothesis of mapping from x to y.

we assume every point is described by our line on xy plane.

Total error=$\sum\left|\hat{y}^{(i)}-y^{(i)}\right|$

Where $\hat{y}^{(i)}$ is assumed data point and $y^{(i)}$ is the actual data point.

Average error=$\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}^{(i)}-y^{(i)}\right|$

But as we no error function is not differentiable for $-\infty<x<\infty$

So loss function will be

$J(\Theta)=\frac{1}{n}\sum_{i}^{n}\left|\hat{y}^{(i)}-y^{(i)}\right|^2$

$\hat{y}^{(i)}=h_\Theta(x^{(i)})=\Theta_1 x^{(i)}+\Theta_{0.}$

Now it is required to minimize our loss function(J(Θ)). A Gradient Descent approach will be used to minimize the loss function

Linear regression using least square method

Mathematical function is used to find the sum of squares (square of the distance of the points and the regression line) of all the data points. Least square method is a statistical method given by Carl Friedrich Gauss used to determine the best fit line or the regression line by minimizing the sum of squares. Least square method is used to find the line having minimum value of the sum of squares and this line is the best-fit regression line.

Regression line is y=m*x+c where

y= Estimated or predicted value (Dependent Variable)

x= Value of x for observation (Independent variable)

c= Intercept with the y-axis.

m= Slope of the line

Example :1

Consider the following set of data points (x,y), find the regression line for the given data points.

X	1	2	3	4	5
Y	3	4	2	4	5

Solution:

X	y	$(x - □)$	$(y - □)$	$(x - □)^2$	$(x - □)(y - □)$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
3	3.6	0	0	10	4

where m=$\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$

m= 4/10= 0.4

$\bar{x}$=mean of x=3

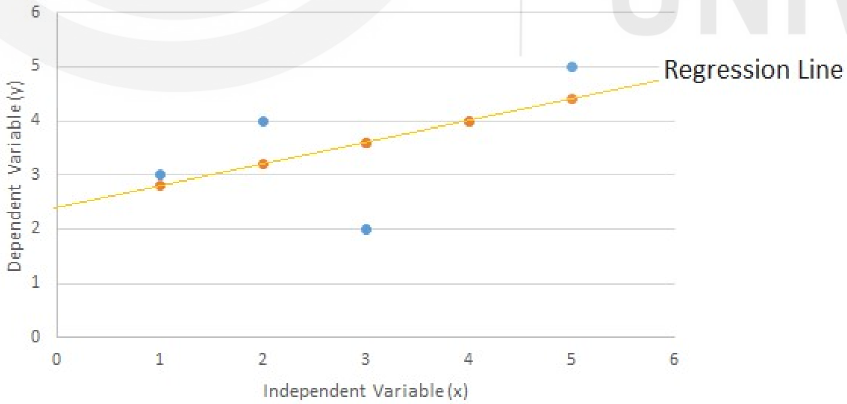
$\bar{y}$=mean of y=3.6

y=mx+c

m= 0.4

c=2.4

y=.4x+2.4



In the above figure blue points are the actual points and yellow points are the predicted points using least square method. Some points represented by blue color lie above the line while some other blue color points lie below the line. However some points represented by the yellow color lie on the line. All other points not lying on the line are the far away from the line with some distance. Thus, actual blue data points and the predicted yellow data points contain some distance between them. This distance or difference between the data points represent an error.

Cost function is used to find the distance between the actual data point value lying other than the regression line and the predicted value of data points lying on the regression line. Cost function optimizes the regression coefficient or weights. It measures how a linear regression model is performing.
Difference between the actual value y on y-axis and predicted value $\hat{y}$ is (y-$\hat{y}$), and cost=(y-$\hat{y}$)^2
if there are n number of data points then the cost function will be

$$\text{cost} = \frac{1}{2n}\sum_{i=1}^{n}(y - \hat{y})$$

or

$$\text{cost} = \frac{1}{n}\sum|(y - \hat{y})|$$

Since cost function provide the error between the actual value and predicted value so minimizing the value of cost function will improve the prediction value. Higher the cost function value will degrade the performance.

Mean Squared Error (MSE): The average of squared of the distance measured between the actual data points lying other than the line and predicted data points lying on the line is called as a mean squared error. It is written as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{n}\left(y_i - (a_1\, x_i + a_0)\right)^2$$

Where N = total number of data points
y_i = Actual value
$(a_1\, x_i + a_0)$ = Predicted value with slope a_1 and intercept a_0

Mean Absolute Error (MAE) is used to determine by calculation sum of all errors divided by the total number of errors in a group of predictions. While considering a group of data points their directions are not important. In other words, it is a mean of absolute differences among actual value and response value results where all individual deviations have even importance.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{m}\left|\hat{y}^{(i)} - y^{(i)}\right|$$

Check your progress 1

1. What is regression? Define linear regression.

2. Describe about overfitting and underfitting.

-

3. State True or False

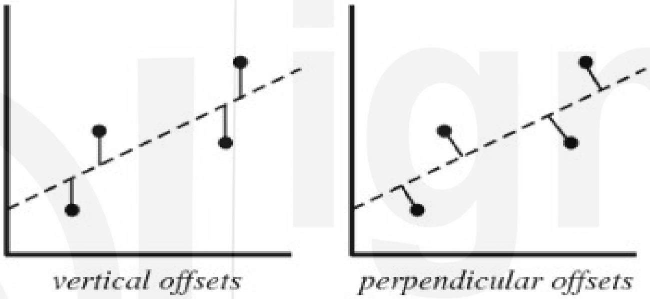
T	F

(a)To determine relationship between numeric variables Linear Regression is used.

(b) With the help of logistic regression, 0/1value attributes are predicted.

(c) In Linear Regression, Least Square Error is used to find the line fitted best.

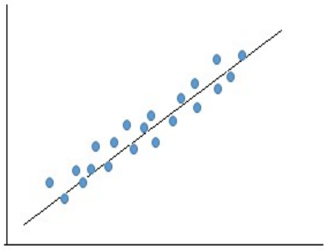
(d) If x-axis represent independent variable and y-axis represent dependent variable. Then vertical offset diagram shown below is used for least square line fit .

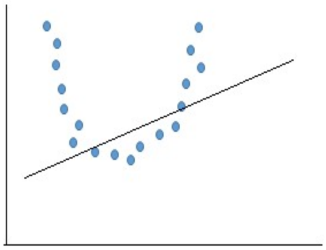


11.4 POLYNOMIAL REGRESSION ALGORITHM

Linear model can apply to data set having linear in nature, however if we have data set of nonlinear in nature then nonlinear model is to be applied.

As shown in figure all the data points are linear in nature. All points are close to the line, linear model regression model can be applied to the data sets. In figure 2 all the data points are nonlinear in nature so linear model cannot fit all the data points, only 2 or3 data points can be fitted to the linear model and all other points are far away from the line. Loss value for this graph will be very high and accuracy will be reduced.





y_1=a_0+bx is equation of linear regression, with slope b where a_0 is the intercept with the x axis.

$$y_2 = a_0 + b_1x_1 + b_2x_2 + b_3x_3 \text{............} b_nx_n = a_0 + \sum_{i=1}^{n}(b_ix_i)$$

where y_2 is a multiple regression equation with n independent variables.

Above two equations y_1 and y_2 are polynomial equations with degree 1.

Consider stock price S_p as a polynomial function of time.

$$S_p = a_0 + a_1T^1 + a_2T^2 + a_3T^3 + \varepsilon$$

Where S_p is a polynomial function and ε is an error and we need to find different values of $a_0, a_1, a_2 \; and \; a_3$ such that the difference between the value obtained from the above equation and from the model will be minimum.

Now we have data points $(T_{1,}S_{p_1})$, $(T_{2,}S_{p_2})$, $(T_{3,}S_{p_3})$……. $(T_{n,}S_{p_n})$.

Thus, the equation become

$$y_i = a_0 + a_1u_{i+}a_2v_{i+}a_3w_{i+}\varepsilon_i$$

Where $y_i = S_{p_i}$, $u_{i=}T_i$, $v_{i=}T_i^2$, $w_{i=}T_i^3$

$$Y= \begin{bmatrix} S_{p_1} \\ S_{p_2} \\ . \\ . \\ . \\ S_{p_n} \end{bmatrix} \qquad X= \begin{bmatrix} 1 & T_1 & T_1^2 & T_1^3 \\ 1 & T_2 & T_2^2 & T_2^3 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & T_N & T_N^2 & T_N^3 \end{bmatrix}$$

Extending to the mth order polynomial it becomes

$$X= \begin{bmatrix} 1 & T_1 & T_1^2 & T_1^m \\ 1 & T_2 & T_2^2 & T_2^m \\ . & \vdots & . & . \\ . & \vdots & . & \vdots \\ . & \vdots & . & \ddots \vdots \\ 1 & T_N & T_N^2 & T_N^m \end{bmatrix}$$

For all m<<N , and

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = (X^T X)^{-1} X^T Y \text{ known as left inverse of matrix X}$$

Squared Error (SE) is the error occurred between the predicted values and actual values used for polynomial regression line. It is written as:

$$S_r = \sum_{i=1}^{n} \left(y_i - \left(a_0 + a_1\, x_i + a_2\, x_i^2\right)\right)^2$$

Where n = total number of data points

y_i = Actual value

$\left(a_0 + a_1\, x_i + a_2\, x_i^2\right)$ = Predicted value

$$S_r = \sum\nolimits_{i=1}^{n} \left(y_i - a_0 - a_1\, x_i - a_2\, x_i^2\right)^2 \qquad \dots \text{(i)}$$

To minimize the error $\frac{\delta S_r}{\delta a_0} = 0, \frac{\delta S_r}{\delta a_1} = 0\ and\ \frac{\delta S_r}{\delta a_2} = 0$ …(ii)

On solving equation (i) we will get

$$\frac{\delta S_r}{\delta a_0} = -2\sum_{i=1}^{n} y_i - a_0 - a_1\, x_i - a_2\, x_i^2$$

Since $\frac{\delta S_r}{\delta a_0} = 0$

$$\dot{\Rightarrow} -2\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} 2a_0 + \sum_{i=1}^{n} 2a_1\ x_i + \sum_{i=1}^{n} 2a_2\ x_i^2 = 0$$

$$na_0 + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \qquad \dots.(iii)$$

Find $\frac{\delta S_r}{\delta a_1}$, on solving equation (i)

$$\frac{\delta S_r}{\delta a_1} = -2x_i \sum_{i=1}^{n} y_i - a_0 - a_1\, x_i - a_2\, x_i^2$$

Since $\frac{\delta S_r}{\delta a_1} = 0$

$$\dot{\Rightarrow} -2\sum_{i=1}^{n} x_i y_i + 2a_0 \sum_{i=1}^{n} x_i + 2a_1 \sum_{i=1}^{n} x_i^2 + 2a_2 \sum_{i=1}^{n} x_i^3 = 0$$

$$\dot{\Rightarrow}\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i \qquad \dots.(iv)$$

Find $\frac{\delta S_r}{\delta a_2}$, on solving equation (i)

$$\frac{\delta S_r}{\delta a_2} = -2x_i^2 \sum_{i=1}^{n} y_i - a_0 - a_1\, x_i - a_2\, x_i^2$$

Since $\frac{\delta S_r}{\delta a_2} = 0$

$$\Rightarrow -2\sum_{i=1}^{n} x_i^2 y_i + 2a_0 \sum_{i=1}^{n} x_i^2 + 2a_1 \sum_{i=1}^{n} x_i^3 + 2a_2 \sum_{i=1}^{n} x_i^4 = 0$$

$$\Rightarrow a_0 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i \qquad \ldots.(v)$$

From equation (iii),(iv) and (v)

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^3 \end{bmatrix} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}$$

where value of i varies from 1 to n.

Example 2. Consider the following set of data points (x,y). Find the 2nd order polynomial y=$a_0 + a_1 x_i + a_2 x_i^2$, and using polynomial regression determine the value of y when x is 40.

X	40	10	-20	-88	-150	-170
Y	5.89	5.99	5.98	5.54	4.3	3.33

Solution. From the given data points (x,y):

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
40	5.89	1600	64000	2560000	235.6	9424
10	5.99	100	1000	10000	59.9	599
-20	5.98	400	-8000	160000	-119.6	2392
-88	5.54	7744	-681472	59969536	-487.52	42901.8
-150	4.3	22500	-3375000	506250000	-645	96750
-170	3.33	28900	-4913000	835210000	-566.1	96237
$\sum_{i=1}^{n} x_i$ = -378	$\sum_{i=1}^{n} y_i$ = 31.03	$\sum_{i=1}^{n} x_i^2$ = 61244	$\sum_{i=1}^{n} x_i^3$ = -8912472	$\sum_{i=1}^{n} x_i^4$ = 1404159536	$\sum_{i=1}^{n} x_i y_i$ = -1522.7	$\sum_{i=1}^{n} x_i^2 y_i$ = 248304

$$\begin{bmatrix} 6 & -378 & 61244 \\ -378 & 61244 & -8912472 \\ 61244 & -8912472 & 1404159536 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 31.03 \\ -1522.7 \\ 248304 \end{bmatrix}$$

By solving above matrix the value a_0,a_1 and a_2 will be

a_0=6.07647

a_1=-0.00253987

a_2=-0.000104319

y= $6.07647 - 0.00253987\, x - 0.000104319\, x^2$

y(50)= $6.07647 - 0.00253987$ X 50 - 0.000104319 X 2500

= 5.68

Check your progress 2

1. Define polynomial regression.
   --Write down the general equation for the polynomial curve fitting.
   ----

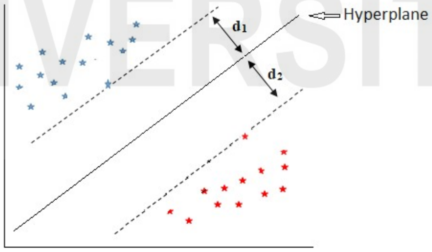
2. State True or False

   (a) A quadratic regression equation can be represented by $\hat{y} = b_0 + b_1x_1 + b_2x_2^2$, where $x_1$and x_2 are independent variables and y is one dependent variable.

   (b) Height of regression line is used to determine the intercept in multiple regression.

   (c) Multiple regression is used when dependent variable does not depend on more than one independent variable.

11.5 SUPPORT VECTOR REGRESSION



Support vector machine is used in solving both classification and regression problem. Consider a classification problem having two different categories as sown in figure.It is easy to separate these two categories by using a line between the two. There is a hyperplane between these two categories which will separate these two from each other. This hyperplane is used to divide the points into different categories lying opposite to the line. Other than hyperplane there are two marginal lines opposite to the hyperplane at a distance apart from the hyperplane. These two marginal lines are having a certain distance from the hyperplane so that all the points can be easily categorised.

Parallel to the hyperplane there are two parallel lines at a marginal distance from the hyperplane. Thus, we can say that there are three hyperplane ie., two line at a marginal distance are also hyperplane. These two marginal hyperplanes must pass through at least one of the closest datapoints. These data points are called **support vectors**. There can be more than one support vectors passes through the marginal hyperplane. These support vectors determine the marginal distance of these two lines from hyperplane (ie., d_1 and d_2). There can be more than one marginal hyperplane for the given data set.We have to choose the marginal hyperplane so that the distance d_1 and d_2 will be the maximum distance max(d_1+d_2).

Considering the data points of the given graph of Fig a. It is not possible to divide the points into two categories by using a linear hyperplane. Thus, we need to convert this graph into three-dimensional graph. The SVM kernel convert the two-dimensional data points in 3- or 4-dimensional data points as shown in Fig b, the hyperplane divide the data points of three dimension into two separate categories.

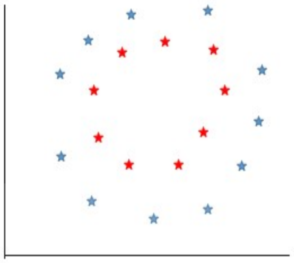


Fig a

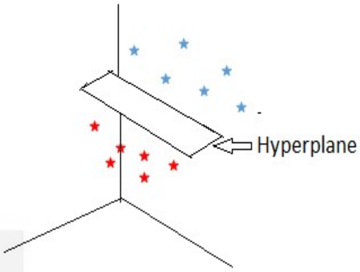


Fig b

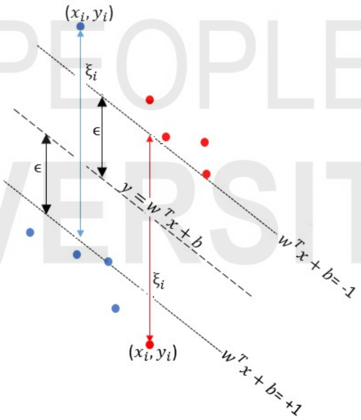
Consider the following graph

As shown in figure equation of hyperplane is $y=w^Tx+b$, where b is constant having value zero since line is passes through the coordinate (0,0) and the slope of line m is -1.

$y=w^Tx$ (since b=0)

Any point that lies below the hyperplane (w^Tx) contribute to the positive value of x, and so is an example of the blue data points, while any data point that lie above the hyperplane (w^Tx) contribute to the negative value of x, and so is an example of the red data points.

For a given margin value M we can say for any value of x where $w^Tx \geq M$ lies on blue points, and for any value of x where $w^Tx \leq -M$ lies on red points. Now consider a point x_i^+ that lies at the positive margin of the hyperplane then $w^Tx_i^+$=M. Here x_i^+ is a support vector. On travelling to the opposite direction perpendicular to the positive margine we will reach a point closest to the negative margin of the hyperplane called as x_i^-.



If x_1 and x_2 are two negative and positive regression vector lies on marginal lines $w^Tx+b=-1$ and $w^Tx+b=+1$, the distance between marginal lines can be determined by

$$w^T(x_2-x_1)=2$$

$$\frac{w^T(x_2-x_1)}{||w||}=\frac{2}{||w||}$$

We have to maximize $\frac{2}{||w||}$ subject to $w^T(x_2 - x_1)$≥ +1 and maximize$\frac{2}{||w||}$ subject to $w^T(x_2 - x_1)$≤ -1for all i=1, 2…...n. In generalized form, $y_i * w^T x_i + b_i \geq +1$, and we need to minimize $\frac{||w||}{2}$ (reciprocal of $\frac{2}{||w||}$).

If all the data points are classified by the marginal line, then it will overfit the machine. And this is not happening in real scenario. It is not always possible that all the data point lies on the right side of the classification. As shown in figure one of the red data points lie below positive margin and one of the blue data points lie above negative margin of the hyperplane. These two data points lie in opposite plane area. If ξis the distance of the data point from respective marginal line, we need to find out the error $\xi_i$for such points.

$$y_i - (w^T x_i + b_i) \leq \epsilon + \xi_i \text{ for each } \xi_i \geq 0$$

$$(w^T x_i + b_i) - y_i \leq \epsilon + \xi_i$$

Error computed = $C_i \sum_{i=1}^{n} \xi_i$ where C_i is the number of error and ξ_i is the error value

Thus it is required to minimize $(w^*, b^*) = \frac{2}{||w||} + C_i \sum_{i=1}^{n} \xi_i$

Where * represent the optimal value.

Example 3. For the given points of two classes red and blue:

Blue: { (1,1), (2,1), (1,-1), (2,-1)}

Red : { (4,0), (5,1), (5,-1), (6,0)}

Ploat a graph for the red and blue categories. Find the support vectors and optimal separating line.

Solution.

Now first support vector SV_1 with x-coordinate 2 and y-coordinate 1 is represented by

$SV_1=\begin{pmatrix} 2 \\ 1 \end{pmatrix}$

Similarly support vector $SV_2$with x-coordinate 2 and y-coordinate -1 and $SV_3$with x-coordinate 4 and y-coordinate 0 will be represented by

$SV_2=\begin{pmatrix} 2 \\ -1 \end{pmatrix}$ and $SV_3=\begin{pmatrix} 4 \\ 0 \end{pmatrix}$

Adding 1 as a input bias in support vector SV_1, SV_2 and SV_3

$$\overline{SV_1} = \begin{pmatrix}2\\1\\1\end{pmatrix}, \overline{SV_2} = \begin{pmatrix}2\\-1\\1\end{pmatrix}, \text{and } \overline{SV_3} = \begin{pmatrix}4\\0\\1\end{pmatrix}$$

To determine the value of α_1, $\alpha_2$and α_3 form the given linear equations we will assume that the support vector SV_1, SV_2 belong to the negative class and support vector SV_3belongs to the positive class.

$\alpha_1\overline{SV_1SV_1} + \alpha_2\overline{SV_1SV_2} + \alpha_3\overline{SV_1SV_3} = -1$ (-ve class)

$\alpha_1\overline{SV_1SV_2} + \alpha_2\overline{SV_2SV_2} + \alpha_3\overline{SV_2SV_3} = -1$ (-ve class)

$\alpha_1\overline{SV_1SV_3} + \alpha_2\overline{SV_2SV_3} + \alpha_3\overline{SV_3SV_3} = +1$ (+ve class)

$$\alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix}\begin{pmatrix}2\\1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\1\\1\end{pmatrix}\begin{pmatrix}2\\-1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}2\\1\\1\end{pmatrix}\begin{pmatrix}4\\0\\1\end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix}\begin{pmatrix}2\\-1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix}\begin{pmatrix}2\\-1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}2\\-1\\1\end{pmatrix}\begin{pmatrix}4\\0\\1\end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix}\begin{pmatrix}4\\0\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix}\begin{pmatrix}4\\0\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix}\begin{pmatrix}4\\0\\1\end{pmatrix} = +1$$

After simplification of above three equations, we get

$6\alpha_1 + 2\alpha_2 + 9\ \alpha_3 = -1$

$4\alpha_1 + 6\alpha_2 + 9\alpha_3 = -1$

$9\alpha_1 + 9\alpha_2 + 17\alpha_3 = 1$

After simplification of above three equations, we get

$\alpha_1 = \alpha_2 = -3.25$ and $\alpha_3 = 3.5$

The hyperplane that discriminates the positive class from the negative class is given by

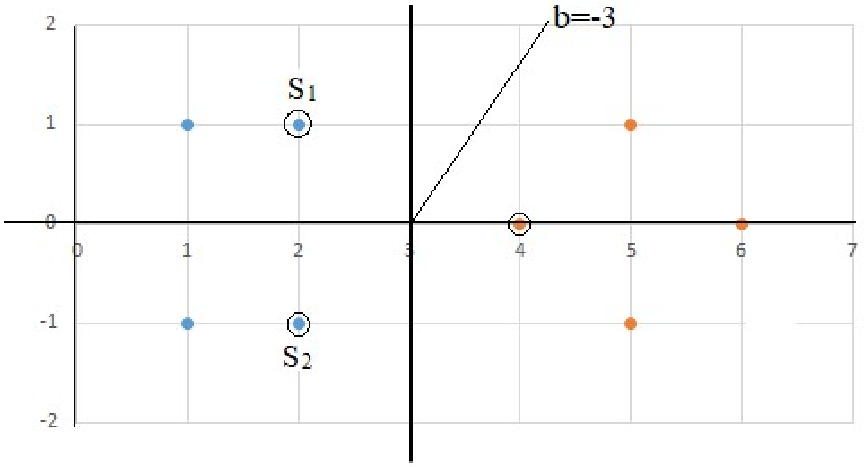
$$\overline{w} = \sum_i \alpha_i \overline{SV_i}$$

$$\overline{w} = \alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix}$$

$$\overline{w} = (-3.25) * \begin{pmatrix}2\\1\\1\end{pmatrix} + (-3.25) * \begin{pmatrix}2\\-1\\1\end{pmatrix} + (3.5) * \begin{pmatrix}4\\0\\1\end{pmatrix} = \begin{pmatrix}1\\0\\-3\end{pmatrix}$$

Hyperplane equation is y=wx+b

Where w= $\begin{pmatrix}2\\0\end{pmatrix}$ and b=-3 or b+3=0 is a line parallel to y-axis which separate both of the category red and blue.



Applications of Support Vector Regression

Used to solve supervised regression problems.

Can be used in both linear and non linear type of data.

Prediction of fire in forest during weather changes.

Prediction of electric power demand.

Check Your Progress 3

1. Define hyperplane.

----

2. Explain about support vector.

----

3. With the given set of points in two classes:

   Class A: $\binom{0}{1}\binom{0}{-1}\binom{1}{0}\binom{-1}{0}$

   Class B: $\binom{3}{1}\binom{6}{1}\binom{3}{-1}\binom{6}{-1}$

   Plot these two classes and find the line separating these two classes. Determine the margin and support vector of the two classes

11.6 SUMMARY

In this unit, we discussed about the concepts of regression – linear regression and nonlinear regression. We discussed about how to find relationship between response variable and predictor variable. Various terminologies used in regression are discussed with an example. Concept of dependent variable, independent variables and how to find the relationships are defined in this unit. Different types of regression also discussed in this unit.

This unit also focused on polynomial regression and how to plot a polynomial curve is also discussed. Concepts of overfitting and underfitting are also discussed in this unit.

In this unit support vector regression algorithm is discussed. Concept of hyperplane, marginal hyperplane and marginal distance are discussed with an example.

11.7 SOLUTIONS / ANSWERS

Check Your Progress 1

1. Regression is a supervised machine learning model which describes the relationships between response variable and predictor variables.So, regression model is used when it is required to determine the value of one variable using another variable.
   Mathematically simple linear regression can be defined as $y=bx+c+\epsilon$. Where b is the slope of the regression line, x is the variable which can change the value of y but can't be affected by another variable. Whereas y is a variable which varies with change in value of x. And ϵ is the known as an error value majored between the actual value and predicted value.
2. Overfitted results when it is unable to generalize well to new data. It results in high performance on trading data. Whereas underfitting results poor performance on training dataset
3. a. T
   b. T
   c. T
   d. T

Check Your Progress 2

1. Polynomial regression is a type of regression algorithm in which specifies the relationships between independent and dependent variable. But here the independent variables are of n^{th} degree polynomial.

2. General equation for the polynomial curve fitting

$$y = m_1x^1 \ + m_2x^2 \ + m_3x^3 \ +\dots\dots\dots..+ m_nx^n$$

$$y = \sum_{i=1}^{1=D} m_ix^i + C$$

3. a. T
   b. T
   c. F

Check Your Progress 3

1. Hyperplane is the line which categorise the data points into two categories.
2. Data points lying on two marginal hyperplanes are called as support vectors.
3. Now first support vector SV_1 with x-coordinate 1 and y-coordinate 0 is represented by

   $SV_1=\begin{pmatrix}1\\0\end{pmatrix}$

   Similarly support vector S_2 and S_3 will be represented by

   $SV_2=\begin{pmatrix}3\\1\end{pmatrix}$ and $SV_3=\begin{pmatrix}3\\-1\end{pmatrix}$

   Adding 1 as a input bias in support vector S_1, S_2 and S_3

   $\overline{SV_1} = \begin{pmatrix}1\\0\\1\end{pmatrix}, \overline{SV_2} = \begin{pmatrix}3\\1\\1\end{pmatrix}$, and $\overline{SV_3} = \begin{pmatrix}3\\-1\\1\end{pmatrix}$

   we need to find out 3 parameters α_1, $\alpha_2$and α_3 form the given linear equations by assuming that support vector S_1 ,S_2 belong to the negative class and support vector S_3belongs to the positive class.

   $\alpha_1\overline{SV_1SV_1} + \alpha_2\overline{SV_1SV_2} + \alpha_3\overline{SV_1SV_3} = -1$ (Negative class)

   $\alpha_1\overline{SV_1SV_2} + \alpha_2\overline{SV_2SV_2} + \alpha_3\overline{SV_2SV_3} = +1$ (Positive class)

   $\alpha_1\overline{SV_1SV_3} + \alpha_2\overline{SV_2SV_3} + \alpha_3\overline{SV_3SV_3} = +1$ (Positive class)

   $$\alpha_1\begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} + \alpha_2\begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} + \alpha_3\begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} = -1$$

   $$\alpha_1\begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} + \alpha_2\begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}3\\1\\1\end{pmatrix} + \alpha_3\begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} = +1$$

   $$\alpha_1\begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} + \alpha_2\begin{pmatrix}3\\1\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} + \alpha_3\begin{pmatrix}3\\-1\\1\end{pmatrix}\begin{pmatrix}3\\-1\\1\end{pmatrix} = +1$$

   After simplification of above three equations, we get

   $2\ \alpha_1 + 4\ \alpha_2 + 4\ \alpha_3 = -1$

$4\ \alpha_1 + 11\ \alpha_2 + 9\alpha_3 = 1$

$4\ \alpha_1 + 9\ \alpha_2 + 11\ \alpha_3 = 1$

After simplification of above three equations, we get

$\alpha_1 = -3.5,\ \ \alpha_2 = \alpha_3 = .75$

The hyperplane that discriminates the positive class from the negative class is given by

$$\bar{w} = \sum_i \alpha_i \bar{S}_i$$

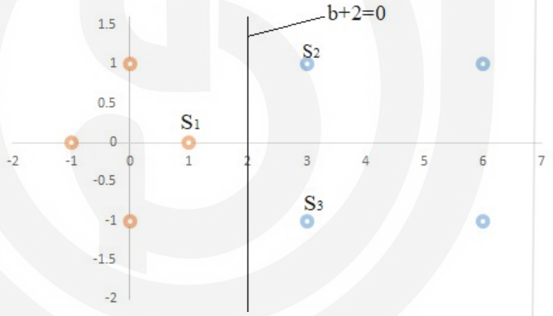
$$\bar{w} = \alpha_1 * \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \alpha_2 * \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + \alpha_3 * \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$$

$$\bar{w} = (-3.5) * \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + (7.5) * \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + (7.5) * \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}$$

Hyperplane equation is y=wx+b

Where w= $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and

b=-2 or b+2=0 is a line parallel to y-axis which separate both classes.



11.8 FURTHER READINGS

1. Machine learning an algorithm perspective, Stephen Marshland, 2nd Edition, CRC Press, 2015.
2. Machine Learning, Tom Mitchell, 1st Edition, McGraw- Hill, 1997.
3. Machine Learning: The Art and Science of Algorithms that Make Sense of Data, Peter Flach, 1st Edition, Cambridge University Press, 2012.